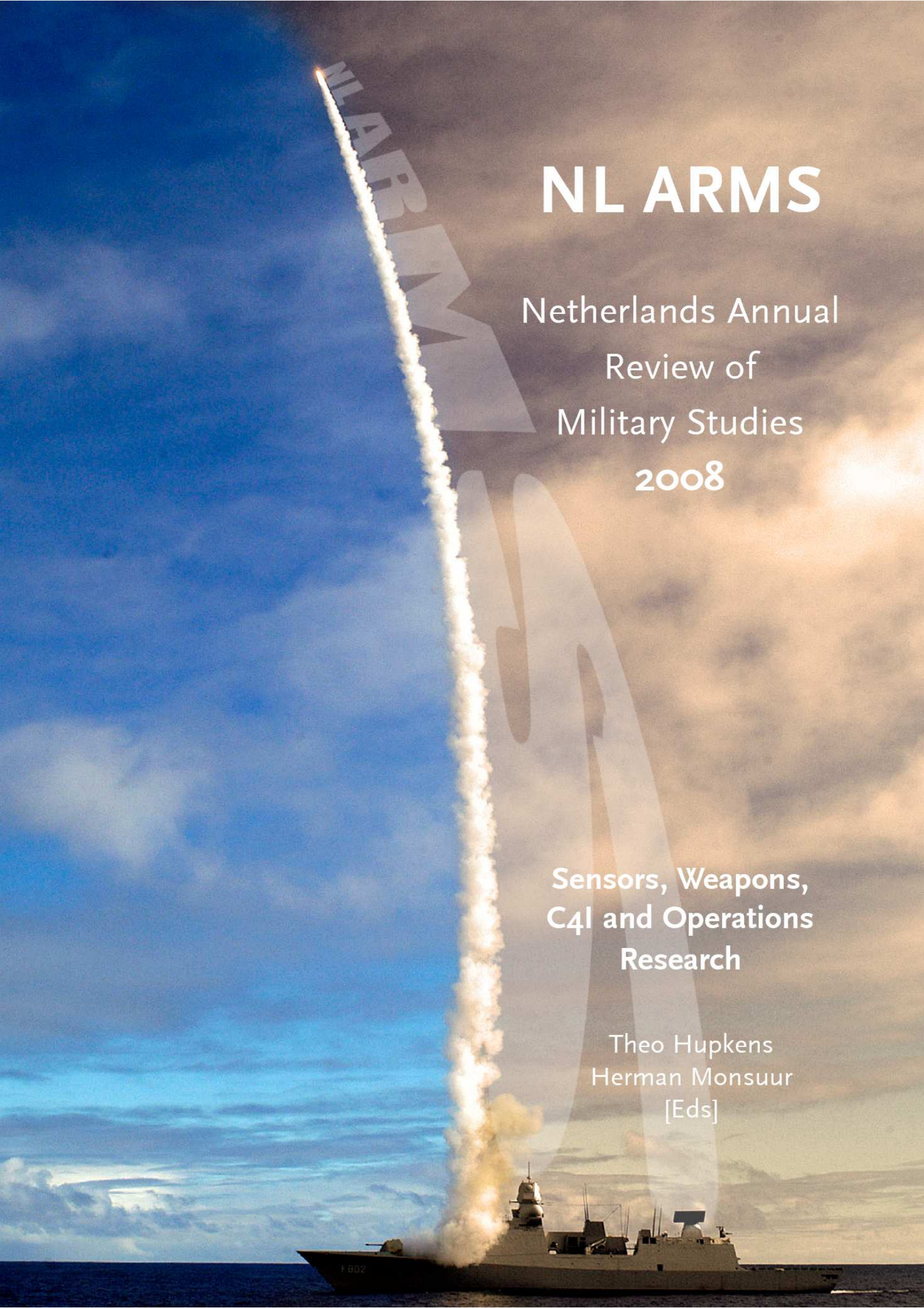


NL ARMS

Netherlands Annual Review of Military Studies 2008

Sensors, Weapons, C4I and Operations Research

Theo Hupkens
Herman Monsuur
[Eds]

NL-ARMS

Netherlands Annual Review of
Military Studies 2008

Sensors, Weapons, C4I and Operations Research

Theo Hupkens
Herman Monsuur
[Eds]

NL-ARMS is published under the auspices of the Dean of the Faculty of Military Sciences of the Netherlands Defence Academy (NLDA).
For more information about NL-ARMS and/or additional copies contact the editors, or the Faculty Research Office of the Faculty of Military Sciences of the NLDA, at the address below:
Faculty of Military Sciences of the NLDA
Faculty Research Office
P.O. Box 90.002
4800 PA Breda
phone: +31 76 527 33 17
fax: +31 76 527 33 22
email: RJ.Smits.01@NLDA.nl

NL-ARMS

1997	The Bosnian Experience *JLM Soeters, JH Rovers [eds]*
1998	The Commander's Responsibility in Difficult Circumstances *ALW Vogelaar, KF Muusse, JH Rovers [eds]*
1999	Information Operations *JMJ Bosch, HAM Luiijf, AR Mollema [eds]*
2000	Information in Context *HPM Jägers, HFM Kirkels, MV Metselaar, GCA Steenbakkers [eds]*
2001	Issued together with Volume 2000
2002	Civil-Military Cooperation: A Marriage of Reason *MTI Bollen, RV Janssens, HFM Kirkels, JLM Soeters [eds]*
2003	Officer Education – The road to Athens *HFM Kirkels, W Klinkert, R Moelker [eds]*
2004	Defense Logistics – Winning Supply Chain Networks *HFM Kirkels, W Ploos van Amstel [eds]*
2005/'06	Terrorist and counterterrorist operations *MGD Rothman, RJM Beeres, HFM Kirkels, JLM Soeters [eds]*
2007	Defense Accounting Control & Economics. Evidence from the Netherlands *RJM Beeres, EJ de Bakker, HFM Kirkels [eds]*
2008	Sensors, Weapons, C4I and Operations Research *ThM Hupkens, H Monsuur [eds]*

Cover design and portrait photography: Peter J. de Vries, Multimedia NLDA/KIM
On the front cover: The HNLMS De Zeven Provinciën launches a Standard Missile-2
(original photograph: CAVDKM)

Printed and bound by: Giethoorn ten Brink B.V., NL

ISSN: 0166-9982

Contents

Editorial Preface

Military forces have to deal with all kinds of operations. In peace keeping operations some units may have to win the *hearts and minds* of local communities that reside in (post-) conflict areas. At the same time they may have to learn how terrorist networks operate and how these networks recruit their members. In rescue operations, liaison officers have to assure seamless communication between their and other units. In other operations, some units may need to use robots, while other units fuse information from networks of sensors. These examples illustrate that military forces must have state-of-the-art equipment and up-to-date knowledge of performance and usability of sensor and weapon systems, C4I and Operations Research methods.

Some research that is necessary to keep ahead in military systems and technology is done within the Combat Systems Department (CSD) of the Netherlands Defence Academy (NLDA). We therefore are glad that we have been offered the opportunity to compile an issue of NL-ARMS. In this issue, the members of the Combat Systems Department inform the defence community and (military) research institutes about their current and future research projects. The contributions cover a large part of the field of technological academic research that military organisations require.

The Combat Systems Department
The CSD consists of the following sections:

- Sensor systems;
- Weapon systems;
- Command, Control, Communication, Computer and Information (C4I) systems;
- Operations Research.

In any military operation the capabilities and performance of sensors play a vital role in the outcome of such an operation. Therefore it is of crucial importance to evaluate the characteristics of the sensors deployed in the battle field, the range of detectable features with a certain sensor suite, the vulnerability of such sensors and their ability to share and fuse information from different sources. The research program of the **Sensor systems** section is directed towards exploitation of the sensor physics, connectivity of sensors in grids or networks, signal preparation and analysis.

The threats for operational military forces depend on the weapon systems of the opponent. In-depth knowledge about their systems makes it possible to estimate their performance and facilitates the development of counter measures. These measures have to eliminate the threat or to reduce its effect. Research of the **Weapon systems** section is relevant to defence in both conventional and (complex) asymmetric military conflicts, although the systems that have to be considered are different.

The research programme of the **C4I systems** section is intended to bridge the gap between Command and Control (C2) operations and technology, and between information technology and communications technology. For example, new operational C2 needs, such as Network Enabled Capabilities (NEC), Effects-Based Operations (EBO),

and information operations in a joint and combined force, make demands on the technological implementation of C2 systems. Also the convergence of information and communications technology makes it possible to apply ideas from the communications field to make C2 systems more mobile and to apply ideas from the computing field to make communications smarter.

The programme of the **Operations Research** section focuses on Defence Modelling, Simulation and Analysis. Subjects are, for example, search and detection, combat modelling, inspection strategies for counter-drugs operations, game theoretic models for conflict and network theory. Results of this research can provide analytic, conceptual and practical support for implementation and evaluation of methods and tools for NEC and Homeland Security.

The rules of play for authors
All contributions were reviewed by two colleagues. It was the editors' intent that all authors would focus on the principles of their research results and the military relevance of their work, rather than to give an absolutely sound scientific explanation with all the mathematical details and derivations. However, due to the nature of some research projects, some mathematics may still be unavoidable in the main body of a few contributions. Of course, all papers contain references to the scientific literature, including the author's own work, to assist the interested reader.

Outline of this volume
This volume starts with a review of non-lethal arms, by **Bart Koene** and two of his Belgian colleagues. There has been a growing interest in non-lethal weapons for both military and civilian purposes. Much research is needed to minimize the amount of damage caused by these non-lethal weapons. **Arthur Vermeulen** and **Eric Trottemant**, in their article on missile interceptor guidance laws, discuss trends in research on interceptor guidance laws for fast-flying and agile air targets. In particular they consider the robustness of the end-game guidance in military operations, using simulation results with different guidance laws. **Jan Martin Jansen** and two colleagues from the Radboud University in Nijmegen, discuss the so-called iTasks. These iTasks may change the current static workflow systems for the Web and also may prove to be useful in military planning and control. **Dick Ooms** discusses the need for a C4I architecture for the Netherlands armed forces and lists the progress that has been made up to now. Several challenges are identified, such as unique real-time requirements, interoperability and agility and unique security requirements. **René Janssen** and **Herman Monsuur** show how situation awareness is affected by the (stochastic) network topology that connects various military entities. Their research can contribute to a better understanding of interactions between networks, people and information. **Rien van de Ven**, in his article on search theory, shows that it is possible to increase the probability of detection of targets, such as drug boats, by slightly modifying the well-known linear and cross-over barriers when searching in a lane. In the article on operational sensor management, **Fok Bolderheij** describes an automatic system that is able to use several sensors in an optimal way, depending on the mission at hand. In their article on human-like visual perception for intelligent multi-modal information fusion, **Coen Stevens, Theo Hupkens** and **Léon Rothkrantz** describe a

first basal implementation of an automatic recognition system based on Gestalt principles. **Frans Absil** and **Jean-Pierre Hermand** discuss the principle of geoacoustic inversion techniques. Geoacoustic parameters, characterising the water column and the bottom, will obviously affect the performance of military sensor systems in, for example, Amphibious Operations. The RNLNC has been involved in two recent sea trials that are described in brief. In his article on acoustic sensing in uncertain environments, **Vincent van Leijen** investigates how acoustic information about the seabed can be obtained from bottom-reflected shipping noise. An innovative idea is the use of "sound sources of opportunity". **Raymundo Hordijk** and **Theo Hupkens** report on a new research project on cooperative robots, which is particularly suited to student participation. In their article on surface and air picture compilation with multiple naval radar systems, **Umesh Ramdaras** and **Frans Absil** investigate performance of sensor selection schemes in a target tracking scenario. **Ariën van der Wal**, in his article on sensor synergetics, illustrates how partial and soft decisions are aggregated via non-linear schemes. Fuzzy logic is used to formally describe sensor fusion. In another article, he then discusses how helicopters may fly autonomously using fuzzy logic, without crashing too often. Finally, **Theo Hupkens** describes a method to find barely visible moving objects in video recordings.

Acknowledgements
Looking back at the process of writing, reviewing and rewriting the various contributions, we would like to thank our colleagues for their cooperative and also critical attitude. We are glad that most members of the Combat Systems Department have been involved in writing or reviewing. The peer reviews often resulted in improvements of the original papers, but also generated new ideas for future and collaborative research.

We would like to thank LT CDR Michael Griffiths for the grammar corrections of the contributions. Thanks are due to Peter de Vries for the design of the front cover. We would also like to thank KLTZ Henk Munnik and Umesh Ramdaras for additional help.

It is the editors' wish that after reading some or all contributions, the reader will have a clear overview of the research of the Combat Systems Department. We invite the audience to give comments, get in touch with the authors for further discussion, initiate further cooperation with respect to the presented research, or at least enjoy reading this annual review.

Den Helder, June 2008

The editors,

Theo Hupkens
Herman Monsuur

Kinetic Non-Lethal Weapons

Bart Koene, Fatiha Id-Boufker[*] & Alexandre Papy[*]

Introduction

When Koene told a Belgian colleague that he had been asked to contribute to an issue of *NL-ARMS*, he understood the subject of this series was 'non-lethal' arms. It was explained that this was not the case. However, this is how the idea of this contribution was born.

NATO defines so-called 'non-lethal' weapons as follows (NATO mission statement, October 13, 1999):

> *Non-lethal weapons are weapons which are explicitly designed and developed to incapacitate or repel personnel,* ***with a low probability of fatality or permanent injury****, or to disable equipment, with minimal undesired damage or impact on the environment (our emphasis).*

According to the United Nations Institute for Disarmament Research (UNIDIR) non-lethal weapons (NLWs) can be defined as:

> *Non-lethal weapons are specifically designed to incapacitate people or disable equipment, with minimal collateral damage to buildings and the environment; they should be discriminate and not cause unnecessary suffering; their effect should be temporary and reversible; and they should provide alternatives to, or raise the threshold for, use of lethal force.*

Usually, in a definition of a weapon it is explained what a weapon should do, while the definitions of non-lethal weapons explain what a weapon should *not* do. In addition, these definitions have subjective and vague aspects: 'low probability' and 'unnecessary suffering' are not specified. As is apparent from these descriptions, the predicate non-lethal does not imply that these weapons cannot cause death, most important is the *intent* that they are non-lethal in case they are used and that death caused by the weapon employed is *as unlikely as possible*. Of course, there are weapons, which are 100% lethal, but thus far unfortunately there are no weapons, which are 100% non-lethal. Therefore, some people prefer using 'less-lethal' instead of 'non-lethal'. Non-lethality is dependent on the inherent nature of the weapon, the way a weapon is used and the vulnerability of the opponent or equipment. Several types of non-lethal weapons exist, for example: biological, electrical, chemical, electromagnetic and acoustic weapons.

In the last decade, the interest in non-lethal weapons has increased considerably. This is a consequence from both progress in non-lethal technology and growing interest from military forces and civil police for more sophisticated and proportional responses to violence. Many disciplines are involved in the study of non-lethal weapons. These include medicine, i.e. effectiveness, risks and treatment; legal, i.e. compatibility of the law and technology, legal protection for the professional user; social sciences and philosophy, i.e. dynamics of groups, perception of limited lethality and ethics; and ballistics, i.e. 'classical'

[*] Royal Military Academy, ABAL, Brussels, Belgium

ballistics, evaluation, mechanisms and projectile-target interaction and innovative developments.

While non-lethal weapons are a vast field, the topic of this paper is limited to so-called *kinetic non-lethal weapons* that can be applied to stop people performing harmful actions. While sometimes details are mentioned on the weapon used, the focus is on terminal ballistics, i.e. the projectile-target interaction. Well-known examples of non-penetrating kinetic NLWs projectiles are: baton rounds, beanbags, fin stabilised rubber projectiles, multi-ball rounds, rubber ball rounds, and sponge grenades. Munitions are fired at the target with relatively low velocity. As a consequence the maximum damage to people is limited; at the utmost they may be wounded.

Figure 1. The FN303 projectile (left) and the Bliniz projectile with its cartridge (right)

This is an orienting paper about kinetic non-lethal weapons. One of its purposes is to help the non-specialist reader better understand scientific and technological aspects of these new weapon systems. First, physical parameters and experimental methods are reviewed. Next, performance tests of two existing non-lethal weapon systems are shown, i.e. the FN303 and the Cougar with Bliniz projectile (Fig. 1). Results are compared with other projectiles: four different balls and a beanbag. They sketch the state-of-the-art and challenges of this developing field. Finally, a conclusion and summary is given.

Physical parameters and experimental methods

Impact on a human body is characterised by energy transfer to the body by an impacting projectile. This may cause injury. In evaluating the effect of projectiles, three domains are relevant, viz. physics, biomechanics and medicine. In order to understand these evaluations several physical parameters and experimental methods have to be understood. Next these parameters and methods are discussed.

Kinetic energy, momentum, impact area and energy density
Studies and evaluations of (possible) injury are often based on four physical parameters:

1. The kinetic energy of the projectile is equal to:

$$E_k = \frac{1}{2} m v^2 .$$

2. Its momentum given by:

   $p = mv,$

   with m the mass of the projectile and v its velocity.
3. The (effective) cross-sectional impact area:

   $$A = \pi\left(\frac{d}{2}\right)^2.$$

   (Note that the diameter d of the cross-sectional area may be different from the calibre of the projectile, for instance: in case of deformation it is larger.), and
4. The energy density e, i.e. the ratio of the kinetic energy E_k and the effective impact area A:

   $$e = \frac{E_k}{A} = \frac{2mv^2}{\pi d^2}.$$

Usually, experimental results are plotted or tabulated as functions of both kinetic energy E_k and energy density e.

Lyon et al. [1] mention that in 1975 a reasonable fit was accomplished by Clare et al. using an empirical four-parameter model, which included the projectile mass (m), the velocity (v), the projectile diameter (d) and the target mass (M). The parameters are plotted using the natural log of the projectile kinetic energy (E_k) vs. the log of the product Md. This model was extrapolated from the mass of the target animals to that of a typical adult male (70 kg). Fig. 2 includes two discriminant lines dividing the graph into three regions, viz. (1) a zone of low lethality, (2) a zone of mixed results, and (3) a zone of high lethality. This illustrates an important and difficult question in non-lethal weapon research: how can we minimise lethality of effective weapons?

The four-parameter model (Fig. 2) was later expanded into another empirical model, viz. a five-parameter model, the so-called blunt criterion (BC) [2]:

$$BC = \ln\left[\frac{\frac{1}{2}mv^2}{M^{1/3}Td}\right],$$

with T the thickness of the body wall of target. As the equation shows, the numerator represents the kinetic energy of the ammunition. The denominator contains those characteristics of the target that have been found to be related to its ability to tolerate the energy of impact.

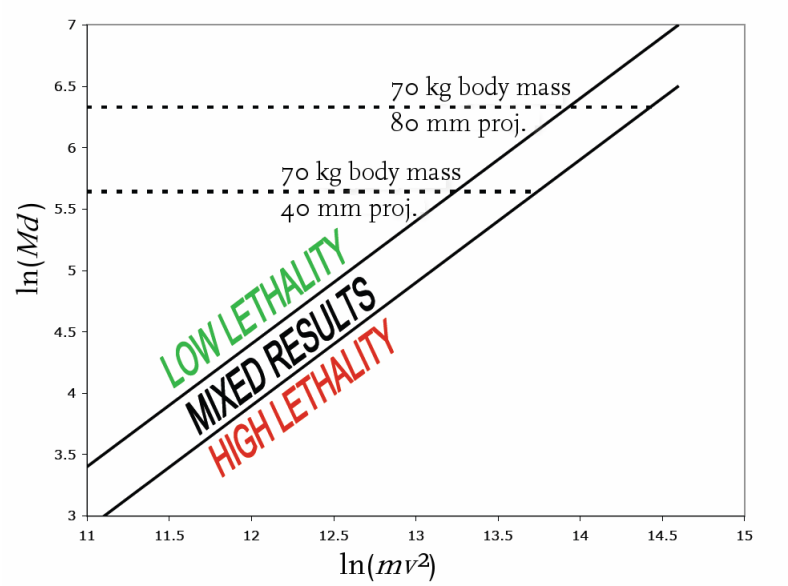


Figure 2. The empirical Four-Parameter Generalized Model for blunt impact, adapted from Lyon et al. [1]. On the vertical axis the log of the product of target mass M (kg) and projectile diameter d (cm) is shown. On the horizontal axis the log of the product of projectile mass m (g) and velocity v (m/s) squared is shown (see text).

Tissue and skin simulants

Animal tests are somewhat old-fashioned, for nowadays it is thought that tissue is inferior to a good simulant (MacPherson [3]), because a bullet impact on tissue has no special effects. Tissue is inhomogeneous and often gives irreproducible results. A good simulant is homogeneous, because it allows experiments to be repeatable and reproducible. Also, it should be practical, i.e. relatively easy to work with and available at acceptable cost. The most difficult requirement though is that of dynamic equivalence of simulant to tissue. Forces are produced on a bullet when it hits tissue and a good simulant should produce very similar forces on the bullet in the same conditions.

According to Viano and King [4] the biomechanical response of the body has three components: (a) inertial resistance by acceleration of body masses, (b) elastic resistance by compression of stiff structures and tissues, and (c) viscous resistance by rate-dependent properties of tissue. Three classical experimental evaluation methods exist, i.e. (1) clay back face signature tests, (2) ballistic gelatine tests, and (3) biomechanical surrogates (crash dummies) tests. In the following sections they are discussed.

1) Clay back face signature

This method examines the cavity created in standard materials, like plasticine, due to impact of the kinetic projectile. The ultimate objective is to determine the potential level of injury of a human being. In case the cavity depth is higher than a certain critical value, the result is a failure because the (possible) injury is considered too severe. The method is used to evaluate body armour.

2) Ballistic gelatine

Another method for the evaluation of the level of injury uses ballistic gelatine. Usually 10% or 20% (weight) is used. This method has been extensively used to model

penetrating impacts. For penetrating impacts 10% gelatine is considered better [3]. This material can be used to determine the rate of energy deposition and the total energy within a target by a penetrating projectile. High-speed cameras can illustrate the degree of temporary deformation and reveal other relevant impact phenomena.

However, this generally accepted procedure for penetrating projectiles has to be adapted and validated for the study of the effect of **non-penetrating** projectiles. Moreover, body tissues have a variable sensitivity for injury and resistance to impact. For her research of ballistic impact of the thorax Bir [5] concluded that 20% ballistic gelatine was better than 10% for both deflection and force data were closer to the human response. Moreover, in two of the impact conditions the 10% ballistic gelatine was penetrated. This indicates that ballistic gelatine and clay have possibilities as simulants in testing non-penetrating ballistic impacts, but with some limitations.

In testing kinetic non-lethal weapon systems it is necessary to include skin in the model. The FN303 projectile, for example, penetrates into gelatine but does not penetrate human skin. In Table 1 possible skin and tissue simulants for testing less lethal projectiles are shown. Moreover, one would also like to account for the effect of different types of clothing. We come back to skin and skin simulants later.

Table 1. Skin and tissue simulants. Ballistic gelatine is most commonly used as a soft tissue simulant.

Skin simulants	***Tissue simulants***
Pig skin	*Plasticine (clay)*
Rubber	*Ballistic gelatine, 10% or 20%*
Polyurethane [6]	*Soap*
Chamois leather [7,8]	

3) Biomechanical surrogates

In biomechanics surrogate tests have been developed as tools for injury evaluation for example for automotive industry (crash dummies), sports and aviation. A well-known method of analysis to determine the injury level to the thorax is the so-called viscous criterion (VC). This response is intended to predict the severity of soft tissue injury and cardio respiratory dysfunction caused by blunt impact. The method utilises measurements taken from the surrogate undergoing an impact event (see Fig. 3). The time-dependent product of the velocity of chest deformation (V in m/s) and the amount of (chest) compression (C in %) forms the figure VC [1,2,5]. The probability and level of injury can be assessed using the maximum VC, $(VC)_{max}$. So, according to this criterion not only the *amount* of compression is relevant, but also the *rate* of compression.

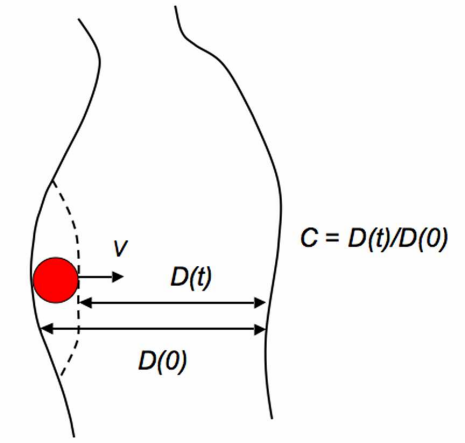


Figure 3. Sketch of chest deformation (V) and chest compression (C). Note that here V is the deformation velocity of the chest, not the velocity of the projectile.

It should be emphasised that impact phenomena associated with non-lethal projectiles are 'low-mass' and 'high-velocity', while 'high-mass' and 'low-velocity' impacts are typical for automotive collisions (Fig. 4). One of the challenges in this field is to validate tests for different non-lethal projectiles using test dummies. A factor that may complicate modelling these biomechanical experiments for blunt ballistic impacts is the dissipation of energy of the rounds upon impact. This method seems promising, also because one can measure forces on the test dummy during impact processes.

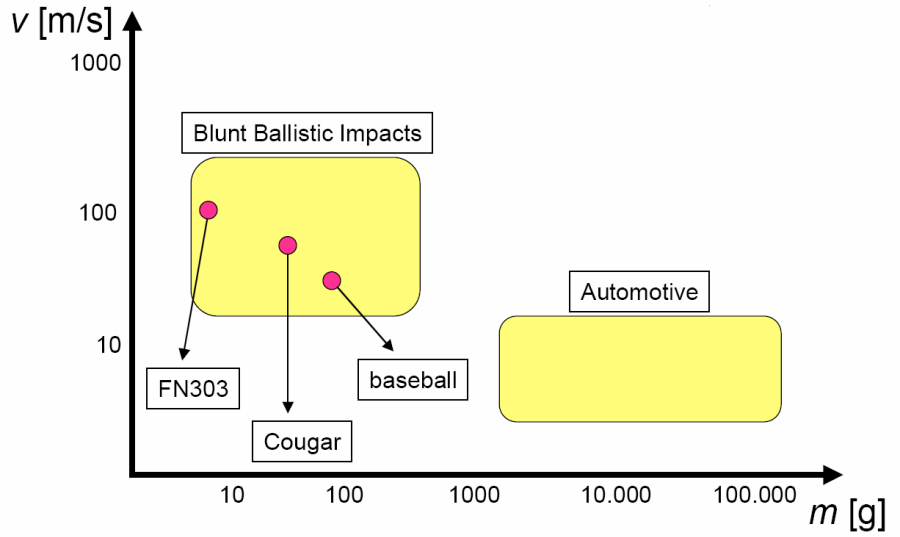


Figure 4. Logarithmic figure of projectile velocity v versus mass m adapted from Bir [5], with own data added. Note the differences in mass and velocity for automotive collisions and 'non-lethal' ballistic impacts.

Young's modulus and non-elastic behaviour of tissue

When a load is applied to a material, it undergoes deformation because the atomic bonds bend, stretch or compress. Tissue is no exception. Because bonds have been deformed they try to restore themselves to the original position. This generates a stress in the material. The applied force (F) causes a deformation (strain) and a restoring stress in the deformed bonds. Stress is a measure of a material's ability to resist the applied force. It is defined as $\sigma = F/a$, with F the force on the material and a the area of a cross-sectional

plane for the type of stress. The strain (ε) is defined as $\varepsilon = \Delta r/r$ with Δr the change of distance in a specific dimension with r the original value. The strain is often expressed as a percentage and the stress has the unit of pressure, Pa in SI units. Initially, many materials deform elastically. In this region the stress is proportional to the strain ε:

$$\sigma = E\varepsilon ,$$

where E is the modulus of elasticity or the *Young's modulus* of the material. Table 2 gives the Young's moduli of selected human tissues.

Table 2. Young's modulus of selected tissues and simulants.
The yield strength is the stress at which the material begins to deform plastically.

Tissue	*Young's modulus E kPa*	*Reference*
Human skin	*20-100*	*10*
	420-850	*11*
	200-300	*12*
Polyurethane (skin simulant)	*182 (yield strength: 2.583 MPa)*	*6*
Breast (fibroglandular)	*1.8*	*13*
Muscle	*0.675 ·10^6*	*14*
Soap	*21.45 ·10^3 (yield strength: 1.63 MPa)*	*15*
Gelatine 20%	*96 (0.001 s^{-1} strain rate)*	*16*
	124 (0.01-1 s^{-1} strain rate)	

Stress (σ) of a material should remains below a critical stress limit, the tensile strength; this is the point where a material breaks. Biological materials usually show different mechanical behaviour than more traditional materials. For instance (unlike homogeneous materials) skin has no unique, single, Young's modulus. This property varies depending on the strain applied. Typical behaviour for biological materials is a stress-strain diagram showing an *elastic part*, which may be linear or non-linear, and a history dependent *inelastic part*.

Variation of Young's modulus with strain is clearly illustrated in the research of Snedeker et al. [9] on kidney tissue. They used four descriptive material parameters to describe the behaviour of porcine and human kidney capsules, viz. ε_{max}, σ_{max} and E_1 and E_2. The ultimate strain and stress were denoted by ε_{max} and σ_{max}, respectively. E_1 was defined as the low-strain apparent modulus and was calculated from the slope of the best linear fit between 0 and 5% of the ultimate strain. The high-strain apparent modulus E_2 was defined similarly between 60% and 80% of the ultimate strain. 25 Tests were performed. For human kidney capsules E_1 and E_2 were 6.7 ± 1.9 MPa and 41.5 ± 11.1 MPa, respectively. ε_{max} and σ_{max} were equal to 33.4 ± 6.5% and 9.0 ± 2.9 MPa, respectively.

Ballistic gelatine has a Young's modulus of approximately 100-150 kPa. The dependence of the modulus on strain rate suggests that ballistic gelatine has a significant visco-elastic component to its mechanical behaviour. This is confirmed by dynamic measurements

from rheometer experiments [16]. Ballistic gelatine has similar properties as the soft tissues in Table 2, which is of course the reason why it is often used as a simulant.

Skin penetration

An aspect often ignored or discounted in wound ballistics is that of the initial penetration of human skin before the main wound is formed in the under-lying tissue. The skin should be taken into account in studying the effects of kinetic non-lethal weapons. In Table 3 data adapted from Warlow [17], threshold velocities and energy densities are given for human skin perforation.

Table 3. Threshold velocity and energy density for human skin perforation (adult upper thigh skin). Data were adapted from Warlow [17].

Composition	*Mass* g	*Sectional density* g/mm²	*Threshold velocity* m/s	*Energy density* J/cm²
4 mm – *0.157* in *Spheres*				
Glass	*0.08*	*0.0064*	*198 ± 23*	*9.7*
Steel	*0.26*	*0.021*	*126 ± 14*	*13.1*
Brass	*0.31*	*0.25*	*121 ± 13*	*14.5*
4.5 mm – *.177* in *Lead Pellets*				
Sphere	*0.54*	*0.034*	*110 ± 12*	*20.7*
Spire point	*0.56*	*0.035*	*109 ± 12*	*20.9*
Flat-nosed	*0.49*	*0.031*	*136 ± 17*	*28.3*
Hollow point	*0.44*	*0.028*	*133 ± 18*	*24.5*

For experimental ballistic research one is searching for good simulants for human skin. Fresh abdominal pigskin of 3-4 mm thickness has been shown to give the most comparable results. However, it is difficult to control its thickness, to store it and in obtaining convenient supplies. Haag [18] found that car inner-rubber tube of 1.3-2.0 mm thickness gave the next best results. This material is easy to obtain and offers no difficulties in storage. One can also think about latex rubbers of well-defined thicknesses like the type used for surgical gloves. Tests by Haag [18] and by Salziger [19] using such thin barriers in front of the gelatine test blocks only showed significant results in the case of 9 mm and .38 in handgun bullets at velocities below 100 m/s.

In a recent paper by Jussila et al. [21] the target values for ballistic skin simulant (30 year old man, chest) were a tensile strength (the point where material fails) of 180 ± 20 kPa, a threshold velocity of 94 ± 4 m/s and break at an elongation of 65 ± 5%. They found that the best skin simulant evaluated was semi-finished chrome tanned upholstery "crust" cowhide of 0.9-1.1 mm nominal thickness. Its threshold velocity was 90.7 m/s, tensile strength 20.89 ± 4.11 MPa and elongation at break 61 ± 9%. These values are close to the average for human skin. Of the synthetic materials, the authors considered 1 mm of natural rubber as a good possible skin simulant. However, its reported theoretical threshold velocity was only 82.9 m/s [20].

Di Maio et al. [21] performed a series of tests to determine the velocity necessary for lead air gun pellets (calibres .177 and .22) and calibres .38 bullets to perforate skin on human lower extremities. For calibre .177 air gun pellets of 8.25 grains (gr) a minimum velocity

of 101 m/s was required. The energy per area was 18.2 J/cm^2. Calibre .22 air gun pellets weighing 16.5 gr perforated at 75 m/s. The energy per area was 12.8 J/cm^2. A round nose calibre .38 lead bullet weighting 113 gr perforated skin at only 58 m/s. The energy per area was 18.9 J/cm^2. These values are in the same range as the values in Table 3.

One should understand that an important reason why velocity (kinetic energy) and penetration are not always correlated is that damage is not due to energy absorption, but to too much stress (σ_{max}). In the case that the strain is above a certain critical limit (ε_{max}) tissue is damaged. Moreover, for the threshold velocity the area (calibre), projectile material and shape should be taken into account.

Statistical injury risk assessment

In injury risk assessment the injury probability p is related to a biomechanical response x. With a special case of the logistic function, i.e. the standard *logistic function*, the injury probability is related to the response by

$$p(x) = \frac{1}{1 + \exp(\alpha - \beta x)} \ ,$$

where α and β are parameters derived from statistical analysis of biomedical data. This function gives a sigmoidal relationship with three distinct regions: for low biomechanical response levels there is a low probability of injury and, similarly, for very high levels the risk asymptotes to 100%. In the transition region between these two extremes there is risk proportional to the biomedical response. A sigmoidal function is typical of human tolerance because it can describe the distribution in weak through strong subjects in a population exposed to impact.

An example of use of a logistic function to assess injury is the following. The example is about lung injury due to non-penetrating impacts with stiff PVC cylindrical masses with 37 mm diameter but varied in mass from 0.069 to 3.0 kg. Determination of the lung weight and the volume of contusion measured injuries quantitatively. The ratio of lung weight and the expected healthy lung weight determined a parameter Q_i. Impacts resulting in a Q_i greater than 1.5 were considered as severe and impacts with a Q_i of less than 1.5 were considered minor to moderate. Experiments were performed and reported by Cooper and Maynard [22] and further analysed by Bir et al. [5,8] using logistic regression to relate these data to a single injury criterion. Deflection data was used to explore the viscous criterion VC as a means for predicting lung injury. The logistic function calculated is shown in Fig. 5. The R-value shows how strong the correlation is between the biomechanical response parameter $(VC)_{max}$ and the prediction of Q_i. The maximum value of R is 1 and represents a high correlation.

Based on this analysis one can see a $(VC)_{max}$ of 3.5 m/s will result in a 50% chance of sustaining a severe lung injury. A 25% risk of severe lung injury is predicted with a $(VC)_{max}$ of 2.8 m/s.

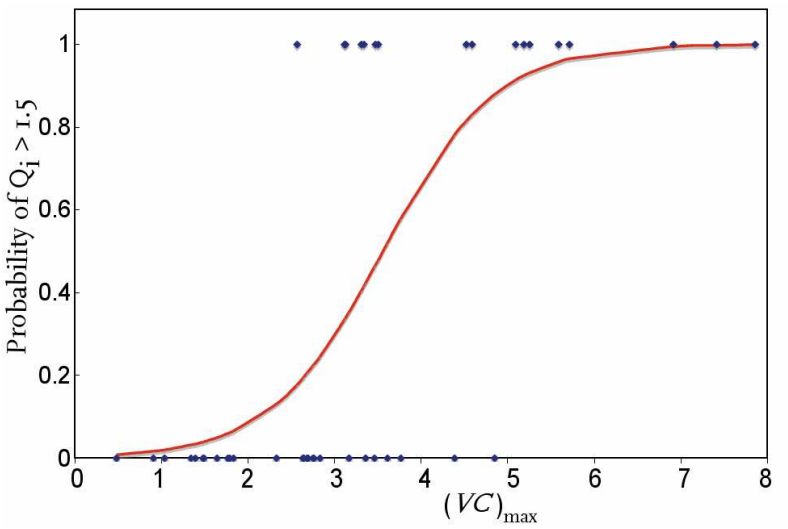


Figure 5. Logistic regression curve probability of $Q_i > 1.5$ versus $(VC)_{max}$ and experimental data points. The logistic function p calculated had $\alpha = 5.48$ and $\beta = 1.54$ (Quality fit: $X^2 = 25.085$, $p = 0.0000$ and $R = 0.67$). Data points with $Q_i > 1.5$ are 1 and points with $Q_i \leq 1.5$ are 0. Source data were provided by Cooper and Maynard [22] and Bir [5].

Human vulnerability models

At impact, a projectile will transfer energy and momentum to the human tissue. Depending on energy, momentum and impact location this will have an injurious or non-injurious effect. Numerical human vulnerability models can simulate the effect at impact on a human body. These models contain a detailed description of human anatomy. Through shot line analysis the damage to tissue involved along the penetration channel can be calculated as a function of the energy transferred to the tissue as a result of impact. For non-penetrating events, such as in the case of non-lethal projectiles, these models describe the stress waves propagated through the tissue. The stress waves are a result of the impacted projectile and can cause tissue damage. Computer codes such as FRAG/MAN IV and ComputerMan may thus provide *estimates* of the level of incapacitation corresponding to the probability that a targeted person will abort his intended actions (Griffioen-Young et al. [23]). However, these human vulnerability models are based on lethal data. Whether they are still valid enough for NLWs is an open question.

Performance tests of two existing NLW systems

Next, performance tests of two modern non-lethal weapon systems will be discussed, viz. (1) the FN303 weapon system and (2) the Cougar launcher with Bliniz projectile.

Experimental set up

The weapon that launches the 'non-lethal' missile is placed on a platform. The target is at a distance of 10 m. This is a typical distance for use of kinetic non-lethal weapons. As simulants for the human body plasticine is used. If test materials are penetrated, both diameter and depth of cavities are determined. Radar is used to measure both muzzle velocity and impact velocity [24].

Projectiles

The FN303 is developed by *FN Herstal* and operates with compressed air. Projectiles are stored in a rotating magazine with a capacity of 15 cartridges. The FN303 is a semi-automatic weapon and has a manual safety. The FN303 launches a fin-stabilised projectile that is made of brittle plastic. It contains bismuth granules and an amount of propellant, dependent of the type of projectile. All FN303 projectiles are non-toxic and environmentally friendly. The projectiles are designed to break at impact and thus avoid the risk of penetration wounds. The calibre is 17.3 mm, the mass is 8.5 g and the effective mass is 0.78 g. According to the manufacturer, the maximum effective range is 50 m because of the fin-stabilised design. The primary effect of the projectile is trauma; the shock immediately neutralises a person. Secondary effects can be caused by a chemical charge in the projectile, for instance *Oleoresin Capsicum* (OC) better known as pepper spray.

Figure 6. Cougar (left) and FN303 (right) launchers

The *Cougar* is a grenade launcher of the 'break-open'-type, developed for the firing of gamma grenades developed by Alsetex for preservation of order. The weapon can be used for both direct and indirect firing. Here direct firing is studied. The projectile is a flexible *Bliniz* projectile consisting of an amount of inert powder (flour) surrounded by a latex shell. The projectile is separated from the propellant by a plastic wad serving as a piston when the round is discharged; 4 sabots guide the projectile. When the muzzle is reached the elements serving for propulsion and guidance are removed from the projectile. Because of their relatively small mass and their relatively large surface they lose speed very quickly. Its calibre is 56 mm and the mass 82 g. The objective of the projectiles is trauma, the shock neutralises a person. The soft projectile transfers (a part of) its energy as it flattens upon impact; it folds around the shape of the struck area. According to the manufacturer, the effective range for direct fire with a Cougar launcher is between 5 and 25 m.

Other projectiles investigated are a 6.2 g so-called '1' rubber ball and a 0.60 g '15' flexible rubber ball fired with a shotgun mounted on the platform. The muzzle velocity is between 190 and 195 m/s. A beanbag is a bag made of cotton with 40 g lead in cartridges of 20 mm diameter. Its diameter of impact is variable up to 26 cm^2. Also, experiments are performed with tennis and squash balls. Other than with other experiments, targets are positioned at 5 m to improve the performance. Impact velocities of about 55.5 m/s are reached [24].

Results and Discussion

In Figs. 7 - 9 results for the FN303 and Bliniz (Cougar) projectiles are shown compared to a few other characteristic 'non-lethal' projectiles.

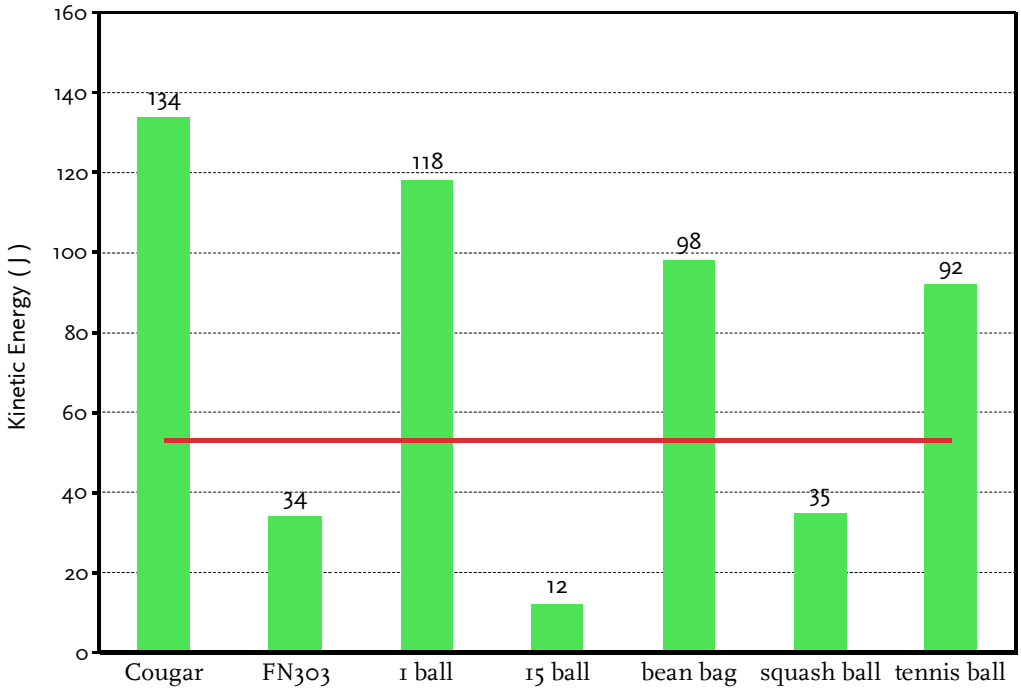


Figure 7. Kinetic energies of different projectiles. The red line represents a threshold value. Here 53 J is adopted (see text). The Bliniz projectile launched by the Cougar is over the threshold value.

Fig. 7 shows the kinetic energy of different projectiles. The kinetic energy of a 'standard' baseball 53 J [25] is taken as a threshold value. For the Cougar the kinetic energy is over the threshold value. From these data it seems evident that one should not aim at somebody's face. For this reason ballistic consistency and accuracy of fire are important.

From Fig. 8 it is apparent that for the Cougar the result for the kinetic energy density is better and below the threshold value of 6 J/cm^2, while for the FN303 the energy density is above this value. The threshold value has been suggested by Sellier and Kneubuehl [26]. Skin can perforate if the energy density is larger than 10 J/cm^2. The cornea can be perforated at a value of 6 J/cm^2 [26], but unfortunately also lower values can cause permanent injuries to the eye.

Fig. 9 shows the depth of cavities in plasticine for different projectiles. As threshold value 44 mm is adopted, which is taken from body armour tests [27]. While the Bliniz projectile launched by a Cougar weapon gives a rather high kinetic energy (over the threshold), its results for both energy density and depth of cavity in clay are well below threshold values. For FN303 it is the other way round, the kinetic energy is below the threshold value, and for both energy density and depth of cavity in plasticine, results are over the threshold values.

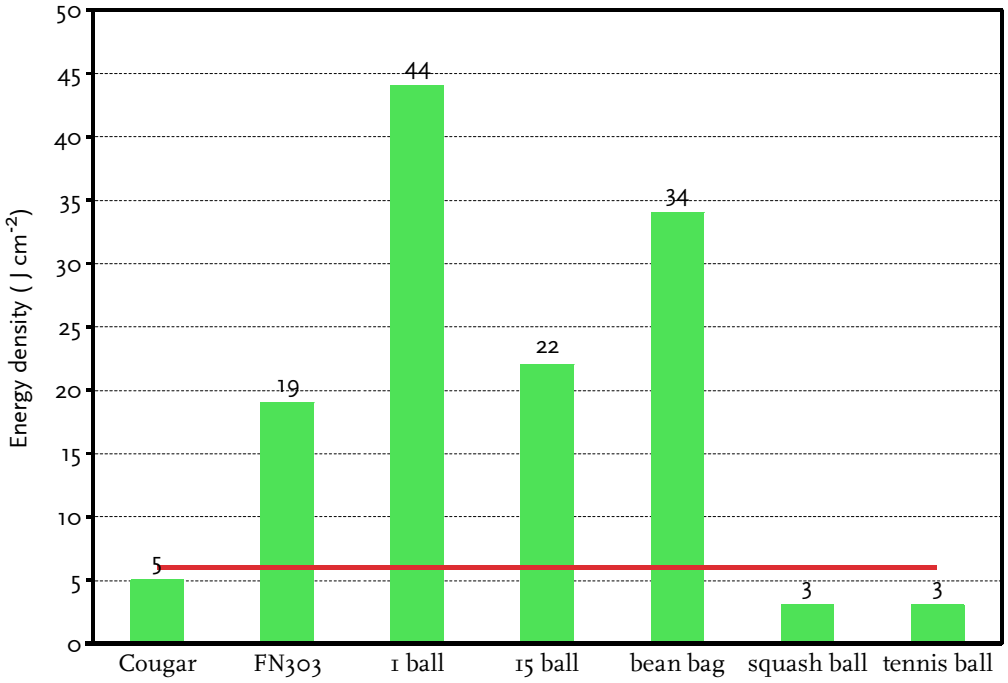


Figure 8. Kinetic energy densities of different projectiles. The red line represents the threshold value (6 J/cm², see text). Evidently, the kinetic energy density of the FN303 projectile is over the threshold value.

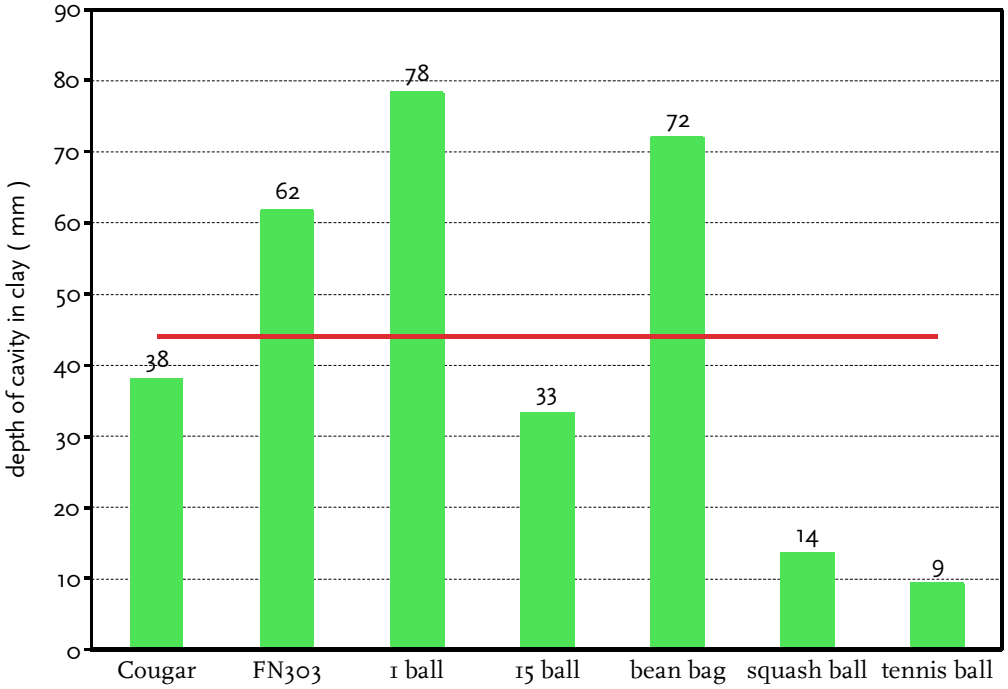


Figure 9. Depth of cavities in plasticine for different projectiles. The red line represents the threshold value (44 mm, see text). Evidently, the cavity due to the FN303 projectile is over the threshold value.

However, we do not know how much energy is transferred to the target and which stresses occur caused by impact. An indication of energy transfer to the human body is obtained by measuring the energy transfer to a ballistic pendulum. From the experimental results it is evident that the Bliniz (Cougar) projectile transfers 4-8% of its kinetic energy to the pendulum, while the FN303 transfers only 0.1-0.5%. This implies

that for the Bliniz projectile less energy is lost to deformation, rebound, heat and other possible loss factors than for the FN303. This may be attributed to the different nature of the projectiles. The FN303 projectile breaks upon impact and the Bliniz projectile only deforms. Also, the difference in mass of the projectiles can be expected to be important [24].

Discussion and summary

In the last decade, the interest in non-lethal weapons has increased considerably, see e.g. [28-32]. This is a consequence from both progress in non-lethal technology and growing interest from both military forces and civil police for more sophisticated and proportional responses to violence. At the moment there are *no* weapons that are 100% non-lethal. However, different classes of non-lethal or less lethal weapons are all intended to inflict as little physical damage as possible while still reliably subduing or incapacitating a person. The same is true for kinetic non-lethal weapons. These goals stand in contrast to traditional weapons development, which focuses on increasing the lethality of weapons. Fortunately, even now lethal effects and permanent injury due to these weapons is *much less likely* than with their conventional lethal counterparts.

New technologies for 'non-lethal' weapons require training of expert users, also because it further minimises harm. The same is true for knowledge of and development for medical treatments for people hit by a kinetic projectile. Also worth mentioning here are a few other important factors. One is the *shooting distance* in relation to safety of both the target and user of the weapon. Also important is the *accuracy* of the weapon. For some weapons a hit on a 'wrong' area (for instance the face) will result in permanent injury or lethality. Moreover, safety of personnel may require use of a lethal weapon. Also, when introducing non-lethal weapons one should be aware that criticism exists about the (ab)use of non-lethal weapons in the recent past. It is argued that these weapons can augment rather than replace lethal technology, see e.g. [33,34]. Like with conventional weapons, one would not like them to be available for the wrong people. So, governments and international communities should take measures to prevent proliferation.

In this orienting study experimental methods and physical parameters of so-called kinetic non-lethal weapons are investigated and reviewed. Some experiments were discussed to show the challenges and possibilities of this developing field. Two commercially available weapon systems, FN303 and Bliniz/Cougar, have been investigated and their impacts have been compared to tennis, squash, rubber balls and beanbags. Both systems demonstrate interesting kinetic weapon concepts. Though the kinetic energy and the kinetic energy density of projectiles are important parameters, we showed that they are not sufficient to classify the lethality of a weapon system. Apparently, there are other factors that co-determine the lethality. The reason why (kinetic) energy and damage are not always correlated is that damage is not due to energy transfer, but to too much stress in tissue. We think computer simulations are useful to augment our understanding of this impact process. However, this requires adequate physical and biomechanical input parameters. In case of kinetic non-lethal weapons the role of the skin tissue cannot be neglected. In this paper a number of useful figures have been collected which will serve

as a starting point for future work in this field. For this research we intend to use the finite element program AUTODYN.

Non-lethal weapons are a relatively new topic in the field of combat systems research. For non-lethal weapon systems different promising evaluation methods have been investigated. However, these methods require improvements and also new methods may be available in the future. At present, there are few consensuses in this field. This multi-disciplinary subject requires independent research by non-commercial institutes.

References

[1] D.H. Lyon, C.A. Bir and B.J. Patton, *Injury Evaluation Techniques for Non-Lethal Kinetic Energy Munitions*, Army Research Laboratory, Aberdeen Proving Ground, Report # ART-TR-1868, January 1999.

[2] C.A. Bir and D.C. Viano, 'Design and Injury Assessment Criteria for Blunt Ballistic Impacts', *J. Trauma*, **57** (2004) 1218-1224.

[3] D. MacPherson, *Bullet Penetration: Modeling the Dynamics and Incapacitation Resulting from Wound Trauma*, 2nd Printing, Ballistic Publications, El Segundo CA, 1994, 2005.

[4] D.C. Viano and A.I. King, 'Biomechanics of Chest and Abdomen Impact', in: D.R. Peterson and J.D. Bronzino (Eds.), *Biomechanics: Principles and Applications*, CRC Press, Boca Raton, 2008.

[5] C.A. Bir, *The Evaluation of Blunt Ballistic Impacts of the Thorax*, PhD Thesis, Wayne State University, Detroit, Michigan, 2000.

[6] D.J. Jones, *Skin and Tissue Simulant for Munitions Testing*, US Patent, #7,222,525 B1, May 29, 2007.

[7] B. Finley, J.H. Evans, J.F. North, T. Gibson, R.M. Kenedi, 'Dynamic and Structural Characteristic of Human Skin: A Comparative Study with Chamois Leather', in: *Environmental Progress in Science and Education*, 18th Annual Technical Meeting, New York, May 1972, *Proceedings*, Institute of Environmental Sciences, Mount Prospect, 1973.

[8] C.A. Bir, S.J. Stewart, M. Wilhelm, 'Skin penetration assessment of less lethal kinetic energy munitions', *J. Forensic Sci.*, **50**(6) (2005) 1426-1429.

[9] J.G. Snedeker, P. Niederer, F.R. Schmidlin, M. Farshad, C.K. Demetropoulos, J.B. Lee, K.H. Yang, 'Strain-rate dependent material properties of the porcine and human kidney capsule', *J. Biomech.*, **38** (2005) 1011-1021.

[10] R. Sanders, 'Torsional elasticity of human skin in vivo', *Pflügers Archiv European Journal of Physiology*, **342**(3) (1973) 255-260.

[11] P.G. Agache, C. Monneur, J.L. Leveque, and J. De Rigal, 'Mechanical Properties and Young's Modulus of Human Skin in Vivo', *Arch. Dermatol. Res.*, **269** (1980) 221-232.

[12] J. Serup and G.B.E. Gemec (Eds.), *Handbook of Non-Invasive Methods and the Skin*, CRC Press, Boca Raton, FL, 1995.

[13] A. Samani, J. Bishop, C. Luginbuhl, D.B. Plewes, 'Measuring the elastic modulus of ex vivo small tissue samples', *Phys. Med. Biol.*, **48** (2003) 2183-2198.

[14] Q. Grimal, S. Naïli and A. Watzy, 'A high-frequency lung injury mechanism in blunt thoracic impact', *J. Biomech.*, **38** (2005) 1247-1254.

[15] N. Ndompetelo, P. Viot, G. Dyckmans and A. Chabotier, 'Numerical and experimental study of the impact of smallcaliber projectiles on ballistic soap', *J. Phys. IV France*, **134** (2006) 385-390.
[16] T.F. Juilano A.M. Forster, P.L. Drzal, T. Weerasooriya, P. Moy, and M.R. VanLandingham, 'Multiscale mechanical characterisation of biomimetic physically associating gels', *J. Mater. Res.*, **21** (2006) 2084-2092.
[17] T. Warlow, *Firearms, the Law and Forensic Ballistics*, CRC Press, Boca Raton, 2005.
[18] M.G. Haag, 'Skin perforation and skin simulants', *A.F.T.E. Journal*, **34** (2002) 3
[19] B. Salziger and M. Strobele, 'Eindringtiefe von 9 mm Luger Geschossen in Gelatine', *Der Auswerfer*, March 1999, pp 29-31, BKA Wiesbaden, Germany.
[20] J. Jussila, A. Leppäniemi, M. Paronen, and E. Kulomäki, 'Ballistic skin simulant', *Forens. Sci. Int.*, **150**(1) (2005) 63-71.
[21] V.J.M. Di Maio, A.R. Copeland, P.E. Besant-Matthews, L.A. Fletcher, and A. Jones, 'Minimal Velocities Necessary for Perforation of Skin by Air Gun Pellets and Bullets', *J. Forensic Sci.*, JFSCA, **27** (1982) 894-898.
[22] G.J. Cooper and R.L. Maynard (1986) 'An experimental investigation of the biokinetic principles governing non-penetrating impact to the chest and the influence of the rate of body wall distortion upon the severity of lung injury.' In *Proceedings of IRCOBI European Impact Biomechanics Conference*, Zurich.
[23] H.J. Griffioen-Young, J.W.A.M. Alferdinck, T. Houtgast and J.L.M.J. van Bree, *The Human Response to Sound, Light, and Rubber Pellets: Theory Behind Non-Lethal Hand Grenades*, TNO report, 22 april 1999, TM-99-A031.
[24] C. Comhair and N. Lemmens, *Non-lethal weapons* (in Dutch), MSc Thesis, Royal Military Academy, Brussels, 2003.
[25] David C. Viano, J.D. McCleary, D.V. Andrzejak, and D.H. Janda, 'Analysis and Comparison of Head Impacts Using Baseballs of Various Hardness and a Hybrid III Dummy', *Clinical Journal of Sports Medicine*, **3** (1993) 217-28.
[26] K.G. Sellier and B.G. Kneubuehl, *Wound Ballistics and the Scientific Background*, Elsevier, Amsterdam, 1994.
[27] Office of Law Enforcement Standards, *Ballistic Resistance of Personal Body Armor*, NIJ Standard-0101.04, National Institute of Justice, June 2001.
[28] A.C. Kernkamp, A.W.G. van Oosterhout, I.M. Paarlberg, J.G.M. Rademaker and M.J.M. Voskuilen, *Technologie Verkenningen Niet-Letale Wapens* (in Dutch), TNO report, FEL-03-A126, Den Haag, 2003.
[29] N. Lewer (Ed.), *The Future of Non-Lethal Weapons*: Technologies, Ethics and Law, Frank Cass, London, 2002.
[30] N. Lewer and S. Schofield, *Non-Lethal Weapons*: A Fatal Attraction?, Zed Books, London, 1997.
[31] R. Mandel, *Security, Strategy, and the Quest for Bloodless War*, Lynne Rienner Publishers, Boulder/London, 2004.
[32] National Research Council, *An Assessment of Non-Lethal Weapons Science and Technology*, National Academics Press, Washington DC, 2003.
[33] S. Wright, 'The Role of Sub-lethal Weapons in Human Rights Abuse', *Medicine, Conflict and Survival*, **17** (2001) 221-233.
[34] B. Rappert, 'A Framework for the Assessment of Non-Lethal Weapons', *Medicine, Conflict and Survival*, **20** (2004) 35-54.

Trends in Missile Interceptor Guidance Laws

Arthur Vermeulen & Eric Trottemant

Introduction

How missiles manage to hit their target is a fascinating topic. Since the introduction of the first guided missiles such as the V-2 ([10], [7]) in World War II, their performance has been ever increasing. This increase has been possible because of more sophisticated seeker technologies (e.g., focal plane array), more accurate inertial platforms (e.g., ring-laser gyroscopes and solid-state accelerometers), better propulsion systems, stronger airframes and the development of more advanced guidance laws.

The time between the launch of the missile and the intercept of the target is usually considered in two guidance phases: (1) the (optional) midcourse phase in which the missile flies towards the target without observing it through its own sensors (for example it can use GPS or inertial navigation to fly to a predicted target position), and (2) the terminal phase, also called the end-game or homing phase. In this phase the missile observes the target motion through its own sensors and steers towards a minimal final miss distance according to a guidance law. These end-game guidance laws are the subject of this paper.

From a guidance point of view, the most challenging area of research is the intercept of fast-flying and agile (aggressive manoeuvring) air targets in the end-game. Traditional interceptor guidance laws perform satisfactorily against stationary or slow-flying targets although air targets have developed tactics for defeating them. The concept of these laws, such as proportional navigation and velocity pursuit, has been known since the early 50's [23] and relatively little research has been done since then to replace or improve them. These traditional laws can be successfully employed against most air targets because a small final miss can be compensated for by the lethal radius of the explosive warhead which consists of blast, fragments or an expanding rod. However traditional laws, which are believed to be currently in use, can be defeated by a fast-flying and agile air target. An example is a manoeuvring Tactical Ballistic Missile (TBM) which is a realistic threat in the (near) future. This new type of target is much less vulnerable than a manned aircraft and requires a very small miss distance or a direct kinetic hit-to-kill. Furthermore, it will not follow the predictable ballistic trajectory of a non-manoeuvring TBM such as the Scud ([21], [17]).

The objective of this paper is to summarise and to discuss trends in research on interceptor guidance laws for fast-flying and/or agile air targets. We will present an analysis of the implications of these trends on the missile performance and in particular the robustness of the end-game guidance in military operations. This analysis will be based on simulation results with different guidance laws but with the standard end-game geometry and the guidance system model of Zarchan [31].

It is relevant for the Dutch defence organization (Ministry of Defence and defence research institutes like TNO and NLR) to follow and understand research on new guidance laws because manufacturers do not in general share this information with their customers. Knowledge of state-of-the-art techniques facilitates the understanding of the trajectories of ones own missiles and the derivation – or smart guess – of guidance laws employed. This understanding enables the calculation of operationally relevant parameters such as the engagement range and coverage areas. Furthermore, knowledge of the theoretical possibilities are relevant when acquiring a new missile system and for estimating the potential – worst case – performance of the missile systems of the adversary.

The outline of the paper is as follows. First, the need for new interceptor guidance laws is illustrated with a simple example – a weaving target – in which traditional laws fail to achieve a successful intercept. In this section the simulation model that is used in the remainder of the paper, is presented. Second, modern techniques presented in the open literature to improve or replace the traditional interceptor guidance laws are summarised. Our focus will be on laws based on the theory of optimal control, differential games, and robust control. Examples of the performance of several laws will be shown. Finally, a summary with discussion is given and further directions of research are mentioned. Readers with little interest in the details of the methods are recommended to continue reading this final section.

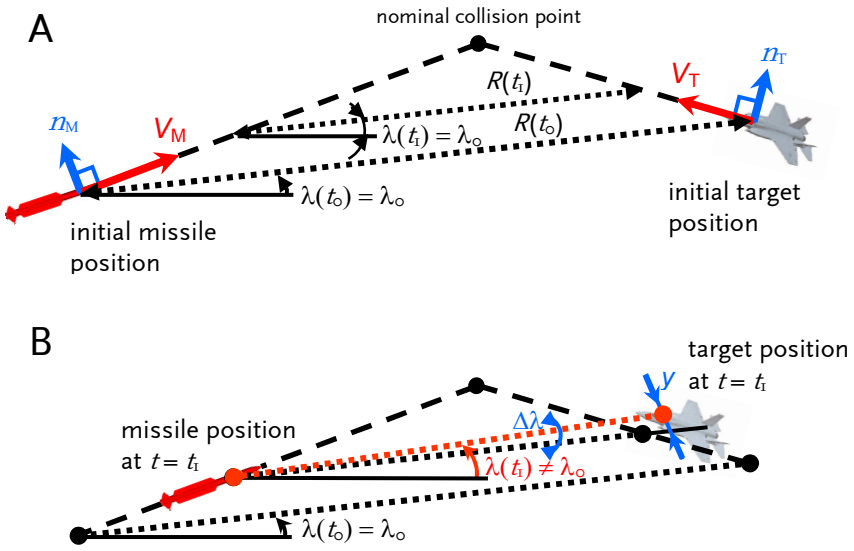


Figure 1. Two-dimensional missile-target engagement geometry in the endgame.

(A) The collision course in which the line-of-sight $\lambda(t) = \lambda_0$ does not change between successive points in time $t = t_0, t_1, \ldots$ and the missile will hit the target without manoeuvring at the nominal collision point at time $t = t_F$.

(B) A target manoeuvre results in a deviation of the collision course and a change in the line-of-sight angle $\lambda(t) = \lambda_0 + \Delta\lambda(t)$. The missile has to manoeuvre ($n_M \neq 0$) to hit the target.

When do traditional guidance laws fail?

Problem formulation

To better understand how guidance laws work, let us consider the planar (two-dimensional) engagement geometry as depicted in Fig. 1A. The target, denoted with the subscript "$_T$", and the intercepting missile, called missile and denoted with the subscript "$_M$", are at a range R of each other and they are flying with a constant velocity (V_T and V_M). The target is not manoeuvring so its lateral acceleration n_T is assumed to be zero and its heading does not change. Under these conditions a constant heading of the missile can be calculated so that the intercept will occur at the nominal collision point. In this case the trajectories yield the so-called (nominal) collision course or collision triangle.

In case the missile has the wrong heading or the target performs a manoeuvre ($n_T \neq 0$), the missile deviates from the collision course and it has to perform a lateral acceleration n_M in order to correct for the separation. This is shown in Fig. 1B for a target manoeuvre. The distance y is the missile-target relative separation of its position according to the collision course, so that the value $y(t_F)$ corresponds with the miss-distance at the time of intercept $t = t_F$.

In this paper it is assumed that the missile and target are in a head-on engagement and that their trajectories can be linearized around the collision course. The geometry of Fig. 1 is then described by the following equations ([31] or [6]):

$$\ddot{y} = n_T - n_M, \tag{1}$$

$$\dot{\lambda} = \frac{d}{dt}\left(\lambda_0 + \Delta\lambda(t)\right) = \Delta\dot{\lambda}, \tag{2}$$

$$V_c = -\dot{R}, \tag{3}$$

where $\dot{\lambda}$ is the line-of-sight rate ("$\dot{}$" represents derivative with respect to time) and V_c is the closing velocity. The total time of the intercept is given by $t_F = R(t_0)V_c^{-1}$ for a constant closing velocity so that the line-of-sight rate can also be written as

$$\dot{\lambda} = \frac{y + \dot{y}t_{go}}{V_c t_{go}^2}, \tag{4}$$

where $t_{go} = t_F - t$ is the time-to-go. This time is a very important variable in the design of modern guidance laws.

The geometry of the engagement is part of the model of the guidance system. This model consists furthermore of the guidance law, the seeker dynamics, an estimator in order to derive the necessary information for the guidance law, and the missile flight control system (including the missile dynamics and body sensors), see Fig. 2. This figure shows also that the output of the guidance law is the commanded lateral acceleration n_c and not the performed lateral acceleration n_M because of the dynamics of the missile flight control system, in particular its time lag.

For the design of the guidance laws a simplistic model of the seeker (i.e., only derivative-operation) and either a perfect system (zero-lag) or a first-order system for the flight control system is considered. The combination of the first-order flight control system and a limiter for the commanded acceleration n_c, will be called the **nominal model** in this paper. The saturation value for the limiter is set at $70g$. However, for a realistic engagement simulation a very sophisticated nonlinear model is necessary, see for example [33].

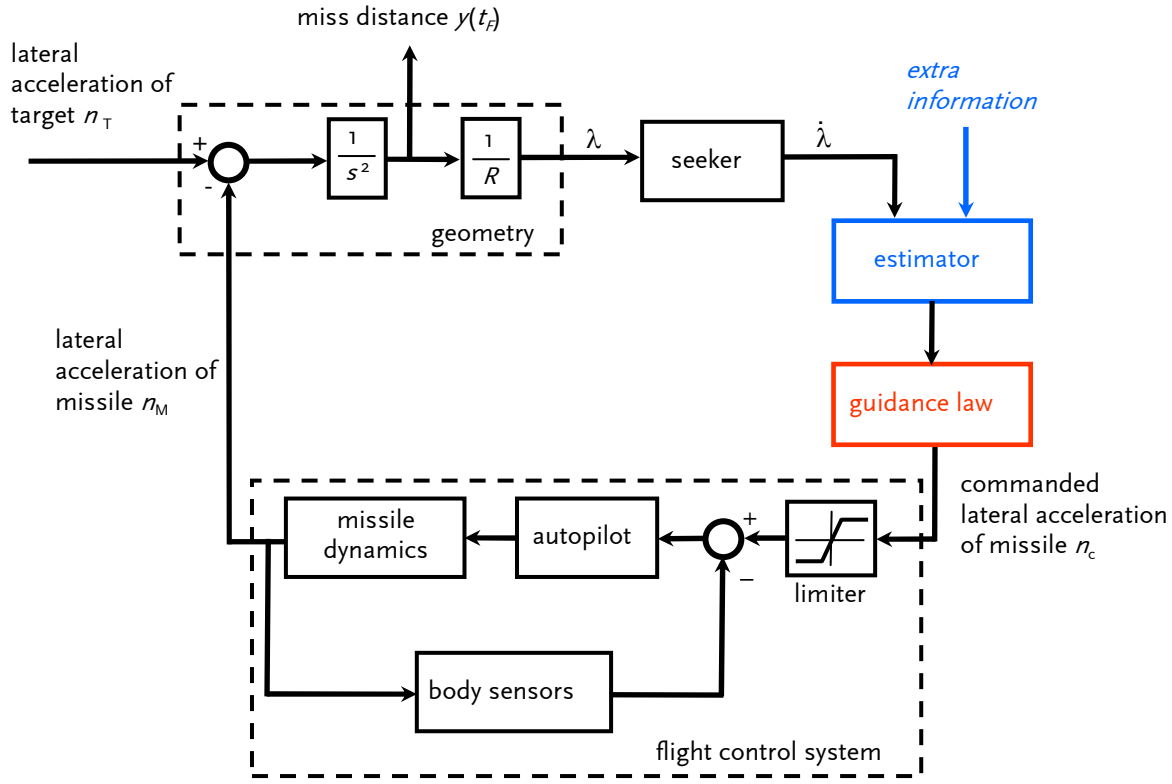


Figure 2. Model of the guidance system in the end-game. In the main loop, called 'homing loop', the guidance law (in red) has to drive the miss distance at intercept towards zero $y(t_F) = 0$ for all target manoeuvres n_T and model uncertainties. Extra information necessary for modern guidance laws can be a target acceleration model, the performed acceleration n_M, or the time-to-go until intercept, etc.

In the present paper no attempt is made to get to this detailed level and all simulations are made with the model of Zarchan [31] entitled "Fifth-order binomial model of guidance system with radome effects", with an additional limiter of $70g$ for the commanded acceleration n_c. It is this model that is called the **realistic model** in the present paper. It includes five first-order systems with time constant τ to model the guidance system, and two 'parasitic effects': (*i*) an unwanted feedback path created by the missile radome because the radome causes a bending of the incoming radar wave and thus a false target location [18], this path is modelled by the parameter $r \neq 0$, and (*ii*) the turning rate time constant T_α which can be defined as the amount of time it takes to turn the missile flight-path angle through a given angle of attack. The dynamics of the estimator are not taken into account (i.e., 'perfect estimator').

Example: a weaving target and proportional navigation

There are several different guidance laws which can be classified as traditional, see [31], [24] or [23] for an overview. As outlined in the introduction the focus in this paper will be on air targets so it is appropriate to consider a traditional law which is commonly used in both air-to-air and surface-to-air missiles; the well-known proportional navigation (PN) law. For this law the commanded lateral acceleration of the missile n_c is proportional to the line-of-sight rate,

$$n_c = NV_c\dot{\lambda}, \tag{5}$$

where N is a unitless designer chosen gain commonly known as the navigation constant. This guidance law is so popular because of its simplicity and the fact that it can easily be implemented with a gimballed seeker which measures the line-of-sight rate $\dot{\lambda}$ [23].

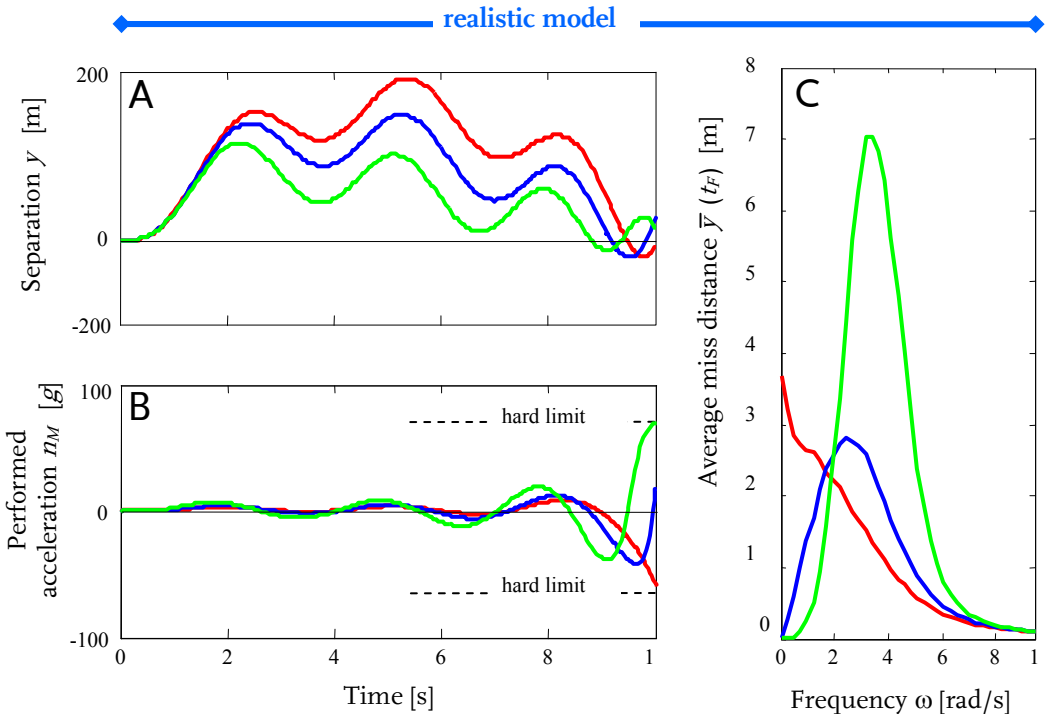


Figure 3. Failed intercept by using proportional navigation against a weaving target $n_T = A\sin(\omega t)$ *for different values of the navigation constant* $N = 2$ *(red),* $N = 3$ *(blue),* $N = 5$ *(green). (A,B) Example of missile-target relative separation* y *and missile lateral acceleration* n_M *as a function of time for a given frequency* $\omega = 2$ *rad/s and time to intercept* t_F*. The final miss distance* $y(t_F)$ *depends on the initial phase so an average miss distance* $\bar{y}(t_F)$ *is calculated. (C) It turns out that this distance depends strongly on the frequency* ω *and the target can easily escape by performing a weave manoeuvre with the appropriate frequency. Parameters:* $A = 12g$, $V_C = 1500$ m/s, $V_M = 1000$ m/s, $\tau = 0.4$ s, $T_\alpha = 2$ s, $r = -0.01$, $t_F = 10$ s.

The target can induce large miss distances if it initiates a maximum acceleration at the proper time before intercept. This maximum acceleration is limited by the pilot maximum g-load for manned air vehicles; a practical limit, when carrying a g-suit is $nT < 12g$ see [25]. Hence, higher values are possible for unmanned air vehicles. The type of manoeuvre called the barrel roll, or in a planar engagement the weave manoeuvre

given by $n_T = A\sin(\omega t)$ with amplitude A and frequency ω, is interesting because it is known to be difficult to counteract by an interceptor. Such a manoeuvre will thus deliberately be performed by the target. Fig. 3 shows an example of the simulation results and the average miss distance at intercept $\bar{y}(t_F)$ that can be obtained for different values of the navigation constant. A high value for the navigation constant ($N > 8$) results in an unstable system (not shown), but as can be noticed in the figure a small value (e.g., $N = 2$ or smaller) gives an enormous miss for a frequency $\omega = 0$ rad/s (i.e., a constant target acceleration) so it is also useless in practice. Useful values are thus in the range $2 < N < 8$.

The maximum value of the lateral acceleration n_c (and consequently n_M, see Fig. 3B) is limited to the impressive $70g$ as reported by Kopp [16] for air-to-air missiles. A decrease of this value will result in a further increase of the miss distance which is on average already above 10 m (this is roughly the lethal radius of the RIM-7 Sea Sparrow Missile as found on the Internet [1]) in the frequency range of $0.5-5$ rad/s for all values of the navigation constant. Proportional navigation is thus not suitable as a guidance law against this manoeuvre.

To conclude, it is interesting to note that the weave manoeuvre is also of great importance for tactical missile interceptors because a TBM naturally weaves into resonance (with frequencies $0.1-2$ rad/s and amplitudes $1-17g$ [13]) as it re-enters the atmosphere, see [20] or [9]. Recent simulations indicate that the oscillatory motion is even more complicated [13]. This point will be important in the section "Laws related to robust control theory" because it is related to the robustness of the guidance law for noise and model uncertainties.

Laws related to optimal control

Overview and theory

The well-known original proportional navigation and augmented proportional navigation laws are in fact related to the theory of optimal control, and in particular to linear quadratic optimisation. In this theory we assume a certain target manoeuvring strategy so the future target acceleration $n_T(t)$ is known to the missile, and we seek a guidance law so that the commanded missile acceleration n_c is a function of the system states (also called state-feedback),

$$n_c = K\vec{x}, \tag{6}$$

where K is a vector with different gains for each element of the state vector $\vec{x}$. A state vector specifies the state of the system. For example $\vec{x} = [y\ \dot{y}]^T$ (where "^{T}" indicates that the transpose is taken from this vector) for the simple geometry (Fig. 1) with a zero-lag flight control model. Higher order, and therefore more realistic, models will have a state vector with more elements.

The guidance law has to achieve a hit: so it should yield a zero miss distance at the time of intercept $y(t_F) = 0$. An additional requirement for the PN-law, and in fact most optimal

guidance laws, is that the missile uses a minimal effort (i.e., amount of energy used for lateral manoeuvres). The rationale for this later requirement is different for an endo-atmospheric and an exo-atmospheric interceptor. If an endo-atmospheric interceptor manoeuvres aggressively, it consequently will induce more drag. Because of this, its velocity decreases dramatically which is disadvantageous for the interceptor. For an exo-atmospheric interceptor drag does not exist but the interceptor relies on a limited amount of fuel (so energy) for its thrusters to perform lateral accelerations. The interceptor must thus have enough energy to last till the end of the engagement. The energy requirement is often formulated by using a quadratic performance index, also called cost function, for the acceleration n_c,

$$\text{minimize} \int_0^{t_F} n_c^2(t)dt \quad \text{subject to } y(t_F) = 0. \tag{7}$$

A (linear) model for the geometry and the dynamics of the flight control system is assumed and represented in the so-called state space form

$$\dot{\vec{x}} = \boldsymbol{F}\vec{x} + \boldsymbol{G}\vec{u}, \tag{8}$$

where $\boldsymbol{F}$ is the dynamics matrix, $\boldsymbol{G}$ is a matrix, and $\vec{u}$ are the inputs. This model is the nominal model for which the (derived) guidance law will be optimal.

The problem formulated by Eqs. (6)-(8) can be solved with the Schwartz inequality for relatively simple (low order) models [31].

Consider PN as an example. In order to derive this guidance law, Eqs. (1)-(3) describe the geometry and no flight control dynamics are considered ('perfect autopilot'; $n_M = n_c$). Furthermore it is assumed that the target does not manoeuvre ($n_T = 0$; so it continues flying in a straight line). Under these assumptions, the state space model is given by [6]

$$\begin{bmatrix} \dot{y} \\ \ddot{y} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y \\ \dot{y} \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} n_c. \tag{9}$$

The rows of this equation in matrix notation are thus only related to the geometry of the problem. The acceleration command, Eq. (6) can now be calculated as

$$n_c = \begin{bmatrix} k_1 & k_2 \end{bmatrix} \begin{bmatrix} y \\ \dot{y} \end{bmatrix} = k_1 y + k_2 \dot{y}, \tag{10}$$

with $k_1 = 3t_{go}^{-2}$ and $k_2 = 3t_{go}^{-1}$. Using the line-of-sight rate, Eq. (4), this can be expressed in the well-known form of Eq. (5) with a navigation constant $N = 3$. Hence, this commonly used value of the navigation constant corresponds with the optimal solution (i.e., zero miss distance and minimal effort) for a non-manoeuvring target with the most basic guidance model.

Different – more advanced and modern – guidance laws can be derived in a similar way by changing the model. First, we consider different manoeuvres. Augmented proportional navigation assumes a constant target acceleration (i.e., a turn to one side with a constant lateral acceleration) so a target acceleration model has to be added to the state space model which consists of the description of the target jerk $\dot{n}_T = 0$. Laws against other – more complicated – manoeuvres can be obtained in a similar way. Second, the assumption of a 'perfect autopilot' (zero-lag model) is modified into a first-order model for the flight control system. In this model a time constant reflects the time lag of the autopilot. Guidance laws which take this addition into account and (try to) compensate for the time lag, are commonly called "optimal guidance" [31] or "minimal energy laws" (MEL) [6]. As an example, we take into account the dynamics of the missile autopilot. The resulting law uses the same performance index Eq. (7) as PN but the state space model is changed to include the first-order model with time constant τ between the commanded n_c and performed n_M lateral acceleration. The commanded acceleration n_c becomes

$$n_c = k_1 y + k_2 \dot{y} + k_3 n_T + k_4 n_M, \tag{11}$$

with $k_1 = N' t_{go}^{-2}$, $k_2 = N' t_{go}^{-1}$, $k_3 = 0.5N'$ and $k_4 = N' K_M$. The new navigation constant N' and the gain K_M vary as a function of the time-to-go t_{go},

$$N' = \frac{6h^2\left(e^{-h} + h - 1\right)}{2h^3 - 6h^2 + 6h + 3 - 12he^{-h} - 3e^{-2h}}, \quad K_M = -\frac{e^{-h} + h - 1}{h^2}, \tag{12}$$

with $h = t_{go}\tau^{-1}$. The navigation constant is very high (N' much larger than 10) for a small t_{go} and it decreases for larger t_{go} towards the familiar value $N' = 3$. Notice that the augmented proportional law corresponds with the choice $N' = 3$ and $K_M = 0$ (i.e., the time constant $\tau = 0$ s).

Zarchan [31] has also derived an optimal law against a weaving target $n_T = A\sin(\omega t)$. The difference with MEL is that the target acceleration model is changed to the weave manoeuvre. The commanded acceleration n_c is now given by

$$n_c = k_1 y + k_2 \dot{y} + k_3 n_T + k_4 \dot{n}_T + k_5 n_M, \tag{13}$$

with $k_1 = N' t_{go}^{-2}$, $k_2 = N' t_{go}^{-1}$,

$$k_3 = N'\left(\frac{1 - \cos(\omega t_{go})}{\omega^2 t_{go}^2}\right), \quad k_4 = N'\left(\frac{\omega}{t_{go}} - \frac{\sin(\omega t_{go})}{\omega^3 t_{go}^2}\right), \quad k_5 = N' K_M, \tag{14}$$

with N' and K_M as given by Eq. (12). We denote this law the optimal weave law (OWL). Note that both modern laws, Eqs. (11) and (13), are state feedback laws and an extension of Eq. (5). However, the state vectors have additional elements such as the target acceleration n_T, its jerk $\dot{n}_T$ and the missile acceleration n_M. These values must thus be measured (or estimated) for the implementation of the guidance law.

Third, the same theory can be employed to shape the missile trajectory near impact. This can be done by adding additional requirements about the velocity at intercept $\dot{y}(t_F)$ [31]. This is useful if the lethality of the missile can be improved by hitting the target at certain strike angles. The requirement $\dot{y}(t_F)=0$ (i.e., no velocity at intercept) results in the so-called rendezvous problem which is well-known in astronautics.

Finally, we remark that for the guidance laws mentioned so far an analytical expression for the feedback gain K was given. However, this is not a necessity because the feedback gains can also be calculated numerically. This numerical approach makes it possible to derive guidance laws based on even more complicated and therefore more accurate, models of the flight control system and the missile flight condition. Mathematically a (nonlinear) matrix Riccati differential equation has to be solved and the (hard) constraint $y(t_F)=0$ has to be replaced by a so-called soft constraint, see for example [8]. This corresponds with a slightly different formulation of the performance index. For example Eq. (7) can be replaced by

$$\text{minimize } qy^2(t_F)+\int_0^{t_F} n_c^2(t)dt, \tag{15}$$

where $q>0$ is a weighting factor. Equation (15) is a weighted combination of miss distance and control effort. Using larger values of q makes the miss $y(t_F)$ more important than the amount of energy, and vice versa. Zarchan [31] has illustrated this approach for a second-order flight control system but, as stated before, any flight control system can be considered.

Numerical examples

The performance, in the sense of a smaller miss $y(t_F)$, of the modern guidance laws discussed in the previous section is superior to the traditional PN-law. This is illustrated in Fig. 4. The simulations in Fig. 4B and 4C are performed with the guidance system with radome effects so, despite the requirement of a miss distance of zero ($y(t_F)=0$), there is always a certain miss at intercept. It is not surprising that the overall miss and control effort is smallest for the optimal law designed for a weaving target, Eq. (13). Also MEL which is based on a very different target manoeuvre (i.e., a constant acceleration), outperforms the PN-law for frequencies below 6 rad/s, but its control effort is very large. The small miss distance for low frequencies seems thus to be related to the fact that a certain time-lag of the flight control system is taken into account. However, for high frequencies the traditional PN-law outperforms the optimal guidance laws. This is a consequence of the simulation model used. For the nominal model (Fig. 4A) OWL and MEL have better results than the PN-law at all frequencies. We can thus conclude that the performance of the optimal laws is promising, but that their robustness for model uncertainties is poor.

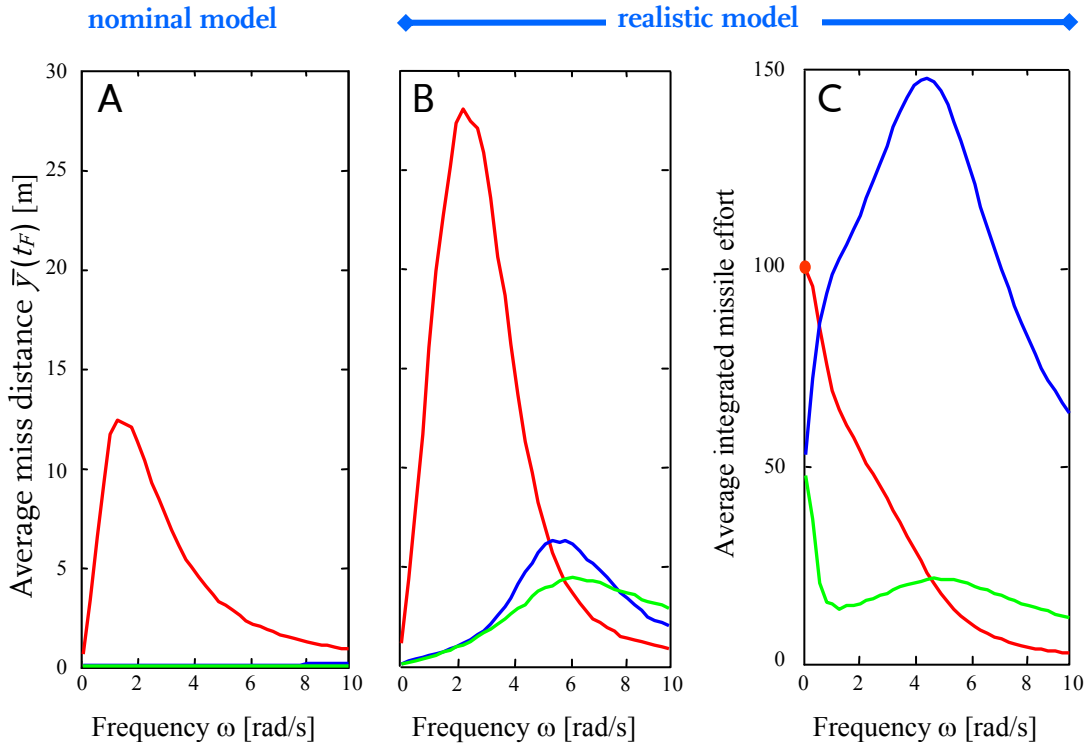


Figure 4. Comparison of three optimal guidance laws against a weaving target: (red) proportional navigation PN, (blue) the minimum energy law MEL which is designed for a constant target acceleration and (green) the minimum energy law OWL designed for a weaving target. (A,B) The average miss distance $\bar{y}(t_F)$ and (C) the missile effort shown as the average of the integrated lateral acceleration $\int n_M^2 dt$. The latter reflects the control effort which is expressed relatively to the effort obtained for PN at $\omega = 0.1$ rad/s (this value is taken as 100). For low frequencies OWL has the best performance but its miss distance will not be zero because of the difference between the nominal model and the model used for the simulation. Parameters as in Fig. 3, $N = 3$.

Laws related to differential games

Overview and theory

Guidance laws based on differential game theory [14] do not require assumptions on the future target manoeuvring strategy but only on the target manoeuvring capability. The word 'game' clearly illustrates what happens because the target manoeuvres to maximize the miss distance and the missile control effort while minimising its own effort, and at the same time the missile is manoeuvring to accomplish the opposite.

We restrict ourselves to linear quadratic (LQ) differential games, see [6] for a review. Mathematically such a game is a two-sided LQ optimisation, called a "zero-sum" game [3] in our context, and the cost function is written

$$\min_{n_c} \max_{n_T} \; q y^2(t_F) + \int_0^{t_F} \left(n_c^2(t) - \gamma^2 n_T^2(t) \right) dt, \tag{16}$$

where the missile (n_c) minimizes the performance index, and the target (n_T) maximizes the same index, and $q > 0$ and γ are two weighting factors. As in Eq. (15) a large value for q makes the miss distance more important. The relative manoeuvrability of target and

missile is indicated by γ. $\gamma > 1$ implies that the missile is more manoeuvrable than the target, which is normally the case. The guidance law will be of the state feedback form, Eq. (6).

The most basic guidance law based on the cost function Eq. (16) uses the state space description Eq. (9) and the assumption of perfect intercept $q \to \infty$. The law obtained is given in the form of the PN-law but with a navigation constant

$$N = \frac{3}{1-\gamma^{-2}}. \tag{17}$$

The navigation constant is thus adapted to the relative manoeuvring capability of missile and target. A more manoeuvrable target (smaller γ) increases the navigation constant so that the missile will react more aggressively. A hit is guaranteed for the nominal model with $\gamma > 1$ [6].

Different guidance laws can be obtained in a similar way as for LQ-laws. However, the design of LQ differential game guidance laws is more difficult because of so-called conjugate points for certain combinations of the design parameters (q, γ and t_F). Although these points might also be encountered by the design of one-sided LQ-laws, a conjugate point analysis always has to be performed for differential game guidance laws. The existence of a conjugate point can be recognised by the fact that one of the elements of the feedback gain K becomes infinite (unbounded) so that the missile has to perform an infinite – unrealistic – acceleration in order to hit the target. This law is thus not realisable because the target will escape in reality. An option is to redesign the law with different design parameters.

As an example we present the game-theoretic version of the MEL which we denote the differential game law (DGL). This law is mathematically equivalent to Eq. (11) but the navigation constant has changed to N^*,

$$N^* = \frac{6h^2\left(e^{-h}+h-1\right)}{6\tau^{-3}q^{-1}+2(1-\gamma^{-2})h^3-6h^2+6h+3-12he^{-h}-3e^{-2h}}. \tag{18}$$

$N^* = N'$ for a non manoeuvring target (γ much larger than 1) and the assumption of perfect intercept $q \to \infty$. $N^* > N'$ for a more manoeuvrable target (smaller γ) and $N^* < N'$ if a larger miss distance is accepted (smaller q). The design is very flexible and it depends strongly on the choice of the parameters q and γ but – as outlined before – a conjugate point analysis has always to be performed because there is not always a realizable solution, even if $\gamma > 1$.

To conclude this section we note that till now we have only considered the zero-sum game, or to be more precise zero-sum game Nash equilibrium solutions. There are also different equilibriums, such as the nonzero-sum Nash equilibrium. However, despite the

larger number of design parameters, the application of this latter equilibrium has not been beneficial to the guidance problem [27].

Numerical examples

Even with the simple model considered, there are several possible combinations for the weighting factors (or design parameters) q and γ, see Eq. (16). Our objective is not to perform a comparative study but only to demonstrate the influence of these parameters on the performance of the game-theoretic law DGL. Results are shown in Fig. 5.

A higher value of q decreases the miss distance. A smaller value of γ corresponds with a lower manoeuvre advantage so the missile has to act directly, more aggressively, on manoeuvres of the target. This can be observed in Fig. 5C by the increase of the control effort. The performance of the game-theoretic DGL approaches the performance of MEL (see Fig. 4) for high q, but in the present simulation setup it does not outperform MEL. Finally, we can repeat the remarks about the relatively poor robustness of the guidance law as done for the optimal guidance laws in the previous section.

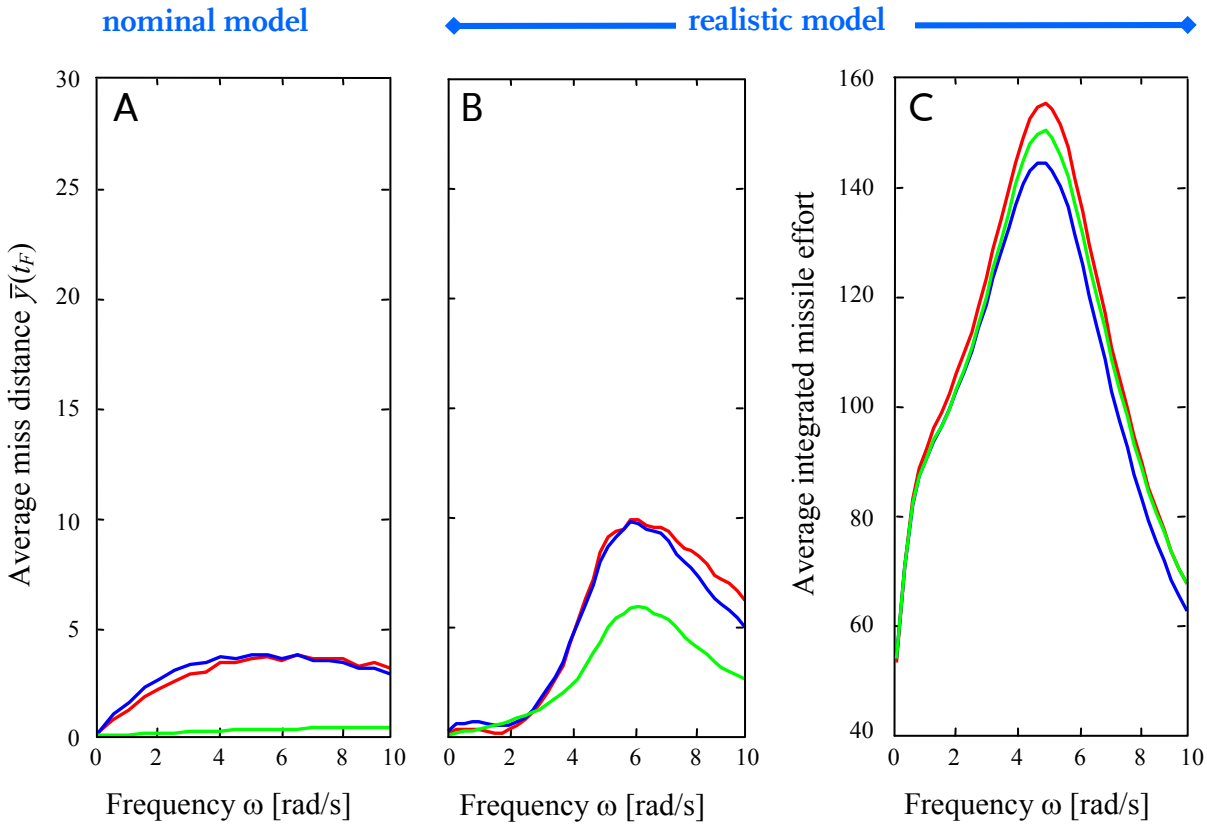


Figure 5. Performance of the game-theoretic guidance law (DGL). Outline as in Fig. 4. Design parameters: (red) $q = 10^4, \gamma = 10$, (blue) $q = 100, \gamma = 10$, and (green) $q = 100, \gamma = 3$. These figures show that a good nominal performance (red) can be obtained with proper tuning but that this performance deteriorates for the more realistic simulation model. Parameters as in Fig. 3.

Laws related to robust control theory

Robustness of a guidance law is a necessity for its implementation in a missile system. Besides the target manoeuvre, three types of error sources can be considered for which the law has to be robust: measurement noise, parametric uncertainties and dynamic uncertainties. The latter two are related to differences between the nominal model used for the design of the guidance law, and the 'real' model. A parametric uncertainty corresponds to an unknown value of a parameter, for example the time-to-go t_{go} or the time constant τ. A dynamic uncertainty means that the system dynamics of the real system and the nominal model are different. This is quite common in practice because (true) higher order dynamics of the system are often poorly known and thus neglected in the nominal model. In the simulations of the present paper a fifth order model with radome effects (see the subsection "Problem formulation") is used as the real model, while a zero-lag or first order model is the nominal model.

Research is performed on all three error sources. For example the noise is considered as a random 'manoeuvre' fitting in the framework of guidance laws based on optimal control theory or the noise is considered as a worst manoeuvre in order to fit in the framework of laws based on game theory. We will limit ourselves in this section to a short literature overview of research on parametric uncertainties, and in particular to estimation errors in two key parameters used in the design of the guidance laws: the time-to-go t_{go} and the time constant τ. Novice readers in the field of control engineering are recommended to skip the remainder of this section.

Ben-Asher and Yaesh [5] have proposed an optimal guidance law with reduced sensitivity to the estimation error in the time-to-go. This law, called reduced sensitivity law (RSL), consists of a differential game problem as discussed in the previous section, but with soft-constraints on all states at intercept. Mathematically this corresponds to replacing $qy^2(t_F)$ in Eq. (16) by $\vec{x}(t_F)Q\,\vec{x}(t_F)^T$. Intuitively this can be understood by realising that the velocity at intercept $\dot{y}(t_F)$ and also lateral accelerations of the missile are forced to be small so that an error Δt_{go} in the estimated time $t_{go}+\Delta t_{go}$ has limited consequences for the miss distance $y(t_F)$.

Robust design for an uncertain time lag in the dynamics of the flight control system of the intercepting missile can be accomplished by H_∞-control. This is a traditional technique for developing robust systems, see for example [32]. Here, it will not be further discussed but it is applied successfully to the guidance problem in [11] and [30], see also [6]. Finally, we note that the problem has also been tackled with the H_2-norm [29].

Currently, we are using robust programming to handle uncertainties and evader manoeuvres. In [26] parametric uncertainties are considered. A semidefinite program (SDP) with a corresponding linear matrix inequality (LMI) is derived for the guidance problem by using Lagrange relaxation techniques called the S-procedure [8]. This LMI is by definition convex (as all LQ problems) and can thus be solved easily with standard (optimisation) routines. The resulting guidance law is implemented in a receding horizon fashion which is common in model predictive control (MPC). Further discussion about these techniques is beyond the scope of this paper but it is important to realize that the

guidance laws obtained, as well as for H_∞- and H_2-control, *guarantee* a certain performance for all allowable uncertainties and disturbances.

Summary and discussion

Trends in missile interceptor guidance laws

In recent years, methods based on optimal control theory (linear quadratic optimisation), differential game theory and robust control theory have opened up new possibilities in the guidance field ([6], [17]). This paper has given a brief introduction to modern intercept guidance laws which are based on these theories.

Few comprehensive comparisons about modern guidance laws have been made in the literature. Anderson [2] has compared laws based on optimal control and differential game theory. As also shown in this paper, differential game guidance laws are less sensitive to errors in the estimate of target manoeuvres. Furthermore they can be made relatively robust against variations in the time-to-go estimate (not shown). Interestingly, Anderson also showed with a simple example that, depending on the scenario (i.e., a missile launched near the inner or outer launch boundary), a simple guidance law could outperform a (more) complicated one. It can thus be advantageous to switch between different guidance laws depending on the operational situation [28] and it is realistic to think that in the future the performance of a missile system will be increased by supplying information about the operational situation to the fire control system.

The major drawback of advanced and modern guidance laws is that more information about the missile and the target is needed for the implementation of the guidance law. On the one hand, this information consists of precise model knowledge such as time constants. This is in particular true for laws based on optimal control theory because these laws are very sensitive to model errors. Differential game guidance laws are less sensitive. On the other hand modern guidance laws require a larger number of measurements during the engagement. In this paper we have seen a few examples: target acceleration, the frequency of the target weaving manoeuvre, the time till intercept ('time-to-go'), missile-target relative separation and velocity, etc. All this information can be difficult to obtain and requires special processing. In particular, state estimation with Kalman filtering [31] is necessary to get an accurate estimate of the information. Furthermore, state estimation changes the overall guidance model. It introduces significant time delays, which might significantly deteriorate the performance as calculated for the nominal model [21]. The design of a fast and accurate estimator that guarantees robustness of the guidance system is an ongoing subject of research.

This paper has shown the natural evolution from guidance laws based on optimal control theory towards laws based on robust control theory. Most challenges and (new) developments are nowadays within this latter theory and its application to the guidance problem. For this reason the PhD-work of the second author is situated in this field.

Related research areas for missile guidance

The laws and methods discussed in this paper can also be used to address more operational questions in related areas. We will consider three examples and indicate our envisaged research for the coming years.

Weapon envelopes. The model of the guidance system used in the present paper is very simple. This facilitates the interpretation of the simulation results and thus the understanding of the influence of the different guidance laws studied. However, to predict the performance of a new guidance law in an operational scenario it is necessary to have realistic models of the missile and the target. In the open literature it is difficult to find these models, and in particular the values of the parameters within these models. One aspect of our future research is to fit the realistic models of Zipfel [33] within the framework of robust control and to derive a guaranteed miss distance. These distances can then be used to calculate different weapon envelopes which indicate the effectiveness of the weapon in a three-dimensional envelope that encloses volumes of air space around the weapon location [4].

Optimal evasive manoeuvres. Missile evasion can be considered an (one-sided) optimal control problem for the target if the guidance law employed by the intercepting missile is considered fixed and known, see [22] or [19]. Hence, the same methods as discussed in this paper can be used to calculate the best evasive manoeuvre according to a certain performance measure such as miss distance or control effort [15].

Use of countermeasures. In an operational scenario the target will use countermeasures, such as decoys and flares, to increase its survivability. Little (open literature) research has been performed to incorporate these measures into the simulation of the guidance system. Dionne et al. [12] use a multimodal probability density function to describe the target state (i.e., including the decoy) and they propose a new guidance law based on a predictive control approach. This law maximizes the probability that the position of the target lies within the reachable set of the missile. We want to initialise a theoretical study about maximising the survivability of the target by the optimal use of countermeasures in combination with evasive manoeuvres.

References

[1] http://www.fas.org/man/dod-101/sys/missile/rim-7.htm (visited on January 11, 2008).

[2] G.M. Anderson. Comparison of optimal control and differential game intercept missile guidance laws. *Journal of Guidance and Control*, 4(2):109–115, 1981.

[3] T. Başar and G.J. Olsder. *Dynamic noncooperative game theory, volume 23 of Classics in Applied Mathematics*. SIAM, Philadelphia, 2nd edition, 1999.

[4] R.E. Ball. *The Fundamentals of Aircraft Combat Survivability Analysis and Design*. AIAA Education Series. AIAA, Reston, Virginia, 2nd edition, 2003.

[5] J.Z. Ben-Asher and I. Yaesh. Optimal guidance with reduced sensitivity to time-to-go estimation errors. *Journal of Guidance, Control and Dynamics*, 20(1):158–163, 1997.

[6] J.Z. Ben-Asher and I. Yaesh. Advances in Missile Guidance Theory, volume 180 of *Progress in Astronautics and Aeronautics*. AIAA, Reston, Virginia, 1998.
[7] A. Bowdoin Van Riper. *Rockets and Missiles: the life story of a technology*. Greenwood Technographies. Greenwood Press, Westport, 2004.
[8] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.
[9] G.H. Canavan. *Miss distances in tactical missile intercepts*. Technical report, Los Alamos National Laboratory, 1993.
[10] C.K.S. Chun. *Thunder over the horizon: from V-2 rockets to ballistic missiles*. War, Technology and History. Praeger Security International, Westport, 2004.
[11] C.E. de Souza, U. Shaked and M. Fu. Robust h1 tracking: a game theory approach. *International Journal of Robust and Nonlinear Control*, 5(3):223–238, 1995.
[12] D. Dionne, H. Michalska and C.A. Rabbath. Predictive guidance for pursuit-evasion engagements involving multiple decoys. *Journal of Guidance, Control, and Dynamics*, 30(5):1277–1286, 2007.
[13] M. Guelman. Guidance games, 2007. (Slides of the presentation at the International Symposium on Guidance and Differential Games on November 8th, 2007).
[14] R. Isaacs. *Differential Games*. The SIAM Series in Applied Mathematics. John Wiley and Sons, New-York, 1965.
[15] J. Karelahti, K. Virtanen and T. Raivio. Near-optimal missile avoidance trajectories via receding horizon control. *Journal of Guidance, Control, and Dynamics*, 30(5):1287–1298, 2007.
[16] C. Kopp. Missiles in the Asia-Pacific. *Defence Today*, pp 62–67, 2005.
[17] B.-Z. Naveh and A. Lorber, editors. *Theater Ballistic Missile Defense*, volume 192 of *Progress in Astronautics and Aeronautics*. AIAA, Reston, Virginia, 2001.
[18] W.N. Nesline. Missile guidance for low-altitude air defense. *Journal of Guidance, Control, and Dynamics*, 2(4):283–289, 1979.
[19] S.Y. Ong and B.L. Pierson. Optimal planar evasive aircraft maneuvers against proportional navigation missiles. *Journal of Guidance, Control, and Dynamics*, 19(6):1210–1215, 1996.
[20] D.H. Platus. Ballistic re-entry vehicle flight dynamics. *Journal of Guidance, Control, and Dynamics*, 5:4–16, 1982.
[21] J. Shinar and T. Shima. Nonorthodox guidance law development approach for intercepting maneuvering targets. *Journal of Guidance, Control and Dynamics*, 25(4):658–666, 2002.
[22] J. Shinar and R. Tabak. New results in optimal missile avoidance analysis. *Journal of Guidance, Control and Dynamics*, 17(5):897–902, 1994.
[23] N.A. Shneydor. *Missile Guidance and Pursuit: Kinematics, Dynamics and Control*. Horwood Publishing, Chichester, 1998.
[24] G.M. Siouris. *Missile Guidance and Control Systems*. Springer Verlag, New York, 2004.
[25] A. Thornborough, editor. *Modern Fighter Aircraft Technolgy and Tactics*. Patrick Stephens Limited, Sparkford, 1st edition, 1995.

[26] E.J. Trottemant, C.W. Scherer, M. Weiss and A. Vermeulen. Robust minimax strategies for missile guidance design. *Proceedings AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2008. (Accepted).

[27] E.J. Trottemant, M. Weiss and A. Vermeulen. Zero-sum versus nonzero-sum differential game approach to missile guidance. *Proceedings of the 17th Symposium on Automatic Control in Aerospace*, 2007. Abstract 101.

[28] T.L. Vincent, R.G. Cottrell and R.W. Morgan. Minimizing maneuver advantage requirements for a hit-to-kill interceptor. *Proceedings AIAA Guidance, Navigation, and Control Conference and Exhibit*, 2001. AIAA 2001-4276.

[29] M. Weiss, M. Rol, W. Falkena and C. Scherer. Guidance performance analysis in the presence of model uncertainties. *Proceedings AIAA Guidance, Navigation, and Control Conference*, 2007. AIAA 2007-6786.

[30] I. Yaesh and J.Z. Ben-Asher. Optimum guidance with single uncertain time lag. *Journal of Guidance, Control, and Dynamics*, 18(5):981–988, 1995.

[31] P. Zarchan. Tactical and Stragetic Missile Guidance, volume 219 of *Progress in Astronautics and Aeronautics*. AIAA, Reston, Virginia, 5th edition, 2007.

[32] K. Zhou, J.C. Doyle and K. Glover. *Robust and Optimal Control*. Prentice-Hall Inc., Englewood Cliffs, 1995.

[33] P.H. Zipfel. *Modeling and Simulation of Aerospace Vehicle Dynamics. AIAA Education Series*. AIAA, Reston, Virginia, 1st edition, 2000.

Web Based Dynamic Workflow Systems and Applications in the Military Domain

Jan Martin Jansen, Pieter Koopman & Rinus Plasmeijer**

Introduction

A workflow system is a computer system that guides and monitors the tasks that have to be done by human workers in collaboration with computers. The total amount of work is a structured collection of tasks. The order of tasks and the assignment of tasks to workers is specified in the work flow specification. Typical application areas are activities where information has to be shared and enriched or checked by several people and/or systems from different functions, disciplines, departments, or even organisations. Examples are: insurance claim handling, purchase ordering, distributed planning, and distributed execution of standard operating procedures. In fact, every activity that involves the transformation of information from one person to another or between persons and automated systems can be modelled as a workflow system. In military applications, workflow systems have the potential to speed up operational Command and Control and supporting processes like planning, logistics and administration.

Commercially available workflow systems mostly use a special purpose (graphical) formalism to specify tasks and the flow of information between tasks. From this graphical representation normally an application is generated.

In this paper we describe the iTasks workflow system [7]. The iTasks system is a software library written in the declarative programming language Clean, which allows for the high-level specification of multi-user workflows. An important advantage of the iTasks system above commercially available workflow systems is that it is not a special purpose system, but embedded in a general purpose programming language which allows the user a much higher degree of freedom to specify complex tasks and actions within these tasks. For example, in the iTasks system it is possible to specify workflows where new tasks are dynamically dependent or created on the results of previous tasks. The system also allows for easy and flexible encryption of the information exchanged between partners.

Another important advantage of the iTasks system is that its user interface is entirely web-based. This makes it possible to use the system in collaboration with external partners – perhaps widely distributed geographically – without the installation of special software at the partner's side.

Finally, the iTasks system allows for client side execution of (sub)tasks using a special purpose interpreter running in the browser at the client side. This minimises the use of bandwidth for exchanging information between client and server and guarantees a quick response. It is even possible to handle subtasks without having a connection to the server.

* Institute for Computing and Information Sciences (ICIS),
Radboud University Nijmegen, the Netherlands.

This interpreter is implemented as a Java Applet and can run in all current available web-browsers, again without the installation of special software.

Workflow Systems

Commercially available workflow systems mostly use a graphical tool to specify the workflow application. In the graphical representation nodes connected with dependency arrows occur. Some nodes represent tasks that have to be fulfilled by the user, other nodes specify control. Typical control nodes are used to enable the parallel execution of tasks or to synchronise the results of several tasks. Ref. [2] contains an overview of the most important workflow patterns.

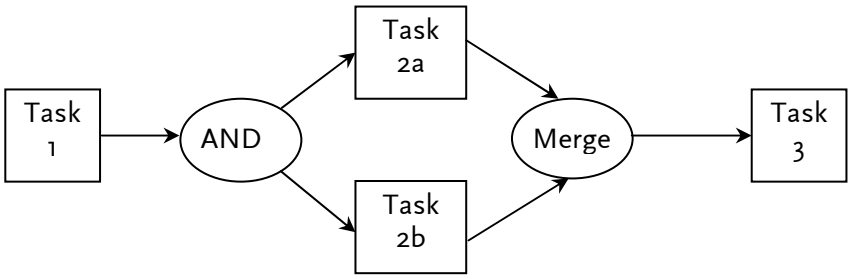


Figure 1. Example of a workflow specification

Fig. 1 shows an example of a workflow specification. Here the workflow consists of four tasks. Task 1 has to be executed first, e.g., the making of an initial plan. Tasks 2a and 2b can then be executed simultaneously, e.g., two subtasks that can be executed in parallel. After they both have completed, Task 3, e.g., the calculation of the total costs can be started. In this example the user tasks are represented by rectangular nodes and the control nodes by oval nodes. The specification only shows the flow of control and not the flow of information. The actual code for the workflow system is generated from the graphical representation by a code generator. The generated code mostly consists of a database program for storing information used in the workflow and several (web)applications for the users of the workflow. The applications interact with the database for retrieving information. In general, the database also contains the control information.

An important restriction of the use of a graphical tool is that the structure of the workflow is statically defined by the graphical tool and cannot dynamically change as a result of data produced by a preceding task.

The use of a code generator often leads to the creation of a large number of files, scripts, database tables and applications, where some of these must be further edited by the application programmer. This complicates the maintenance of the application and makes debugging in case of errors difficult.

Although workflow systems are useful for the support of all kind of activities in organizations, the actual use of these systems is limited. The main reason for this is that current commercial workflow systems do not accommodate the flexibility needed for practical use. They are only useful in situations where the actual flows can be formalised

beforehand in standard procedures. In practice more flexibility is needed because unexpected situations can occur that require ad hoc handling.

Dynamic Workflow systems, like iTasks, can be used to support processes with a dynamic nature. In this paper we look forward to the use of iTasks for planning and execution of military operations. These processes are characterised by a very dynamic nature. Planning processes need to be adapted because e.g., new intel information gives rise to changes. During operations e.g., feedback on the feasibility of plans or new sensor information may lead to adaptation of plans.

This paper gives an overview of the iTasks system, the motivation behind its implementation and looks forward to applications in the military domain. The structure of this paper is as follows. The paper starts with a justification of the use of web pages as interfaces for applications. Then the architecture of iTasks applications is discussed and an introduction to the iTasks library is given. The possible use of iTasks for the planning of complex military operations and the use for military operations themselves is discussed, together with a discussion of the use of iTasks in relation with Net-Centric Operations. Finally, the results are summarised and some conclusions are drawn.

Web interfaces

iTasks applications have a web-based user interface. In this section the advantages of web-based interfaces and the problems that arise from them are discussed. Using web pages as the interface for applications has gained much popularity during the last years. An important advantage of this approach is that no installation of special software is necessary on a computer to use the application. It is even possible to run the application on different platforms or operating systems (Microsoft Windows, Linux, Mac OS, Solaris, etc). Examples of applications with a web interface are: e-mail programs (web mail), online banking applications and web shops like Bol.com and Amazon.com. In fact, almost every (large) company nowadays uses web interfaces as the defacto standard to communicate with customers.

Despite this popularity and convenience for the user, for a developer of software it is still hard to write software for the web. The reason for this is that the web was originally developed to display information and links to other pieces of information. The current architecture of the web still reflects this original goal. An important problem that interactive web applications have to deal with is the fact that a user can move away from a web page at any moment and return to it later (e.g., by closing the browser or page, selecting another web page or by clicking the previous or next button). The web application must be able to deal with this. As a consequence of this transactions (e.g., the purchase of a book) are often not completed and the system should be capable of rolling back that part of the transaction that has already been completed.

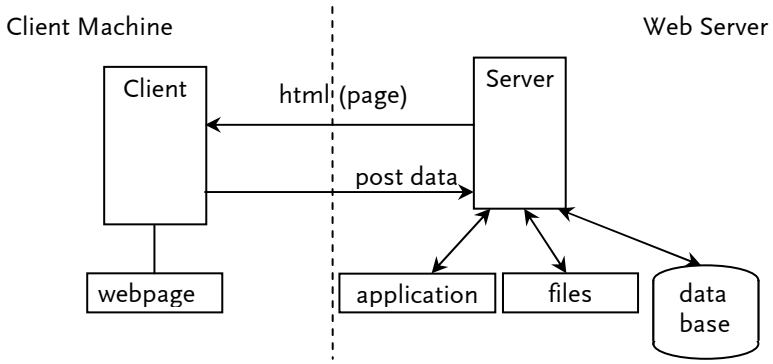


Figure 2. Architecture of a web application

Fig. 2 shows the typical architecture of a web application. The client browser (left from the dashed line) only displays the html generated by the (web) application running on the web server (right from the dashed line). This (web) application can read/write information from/to a database or files residing at the server side. The user can fill in web forms that are sent to the server and processed there. As a result the server produces a new web page that is displayed at the client side. The database and files are used to maintain information (e.g., the user login name or the purchase the user made). As already mentioned a difficulty that has to be dealt with is maintaining the state of a transaction. The server has to keep track of this state. The complication is that the user can move away from the web-page at any moment and come back to the web page at a later time.

In the classical setting the web server processes the web form filled in by the user and produces a complete new html page. A drawback of this approach is that the interaction can become rather slow. To overcome this, local processing at the client side can be done using JavaScript. JavaScript is a small programming language for which an interpreter is integrated in all modern web browsers. Web pages can be (partially) updated by JavaScript. In this way simple processing, that does not really need information available at the server, can be performed at the client side. To further enhance the performance of web applications it is even possible to make a request to the server from JavaScript where the results can be used to (partially) update the web page using so-called Ajax (Asynchronous JavaScript And XML) [3] technology. Fig. 3 shows the architecture of web applications using Ajax technology. Google extensively uses this technique in applications like Gmail and GoogleMaps to speed up their performance and make them more interactive. As a result of this, the developer of web applications has to write several programs: the server side program including access to the database and the generation of html pages; the JavaScript program for client side processing and Ajax interaction with the server. This makes software development of web applications a cumbersome and error prone activity. Again, like in the case of workflow software applications, there are tools that simplify this process. Most of these tools generate frameworks for web applications that must be filled in by the programmer. Again the developer has to deal with the problems of maintenance and debugging for these generated files and frameworks.

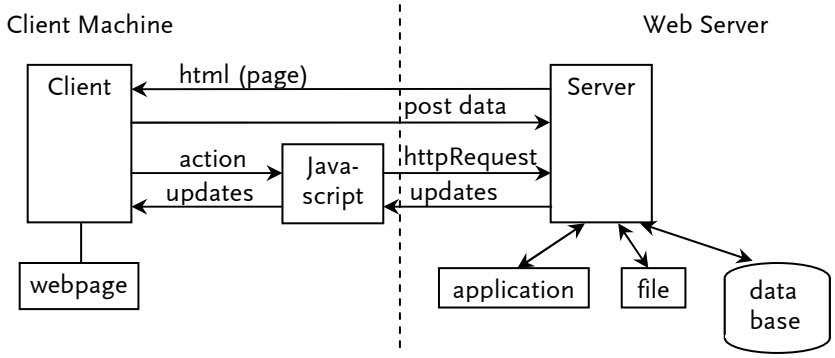


Figure 3. Architecture of a web application using Ajax technology

Architecture of an iTasks application

The iTasks library and iTasks applications are all programmed in the functional programming language Clean [8]. Clean is an example of a 'pure lazy' functional programming language. Another example of such a language is Haskell [1]. Important properties of pure lazy functional programming languages are:

- pure functional: a program consists of functions only and the result of each function is completely determined by the value it returns (there are no side effects);
- lazy: only the calculations necessary for obtaining the end result are done. This makes it possible to use infinite data structures like trees and lists without calculating them completely;
- memory (de)allocation is automatic;
- they have a powerful (strong) type system, which allows for the early detection of programming errors;
- they allow for higher order functions (functions that have other functions as arguments and result);
- they allow for generic functions (functions that can handle arbitrary data types).

The iTasks combinator library depends heavily on all these properties and can be seen as a major example of the application of generic programming techniques. For example, the generation of web forms from data structures, the (persistent) storage and retrieval of data in files or databases, the handling of user updated html forms, are all programmed using generic techniques. The consequence is that an application programmer gets this all for free and does not have to program any of these issues. The combinators of the iTasks library are all higher order functions. It is impossible to program a library like iTasks in a traditional programming language like C, C++ or Java in the same concise way. Parts (subtasks) of an iTasks application can be executed at the client side of the application by an interpreter. This is realised by giving the developer the possibility to annotate tasks in the program with the '`OnClient`' annotation.

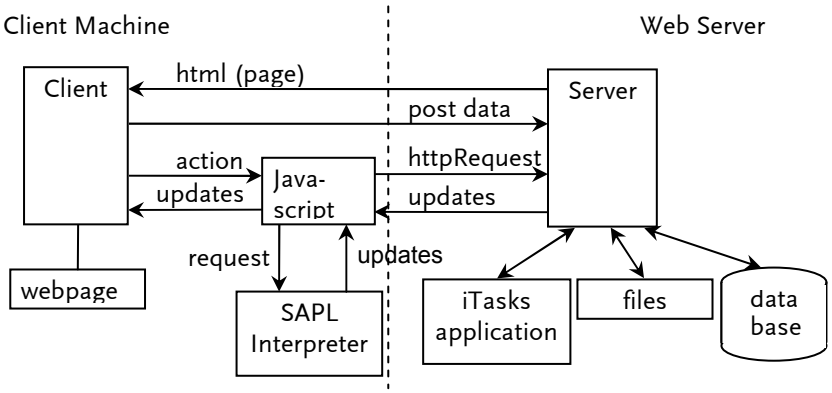


Figure 4. Architecture of an iTasks application

Fig. 4 shows the architecture of a typical iTasks application. This architecture greatly resembles the web architecture for Ajax applications, but there are important differences:

- At the client side the SAPL (Simple Application Programming Language) interpreter [4] is added. This interpreter is implemented as a Java Applet and is capable of executing Clean functions at the client side of the application;
- Both the server and client programs are generated from one single source programmed in Clean. From this source a server executable and a client SAPL program are generated by the Clean compiler. Both the executable and the SAPL source comprise the complete program. In theory it is even possible to run the complete application at the client, except for the storage and retrieval of information in files and data bases;
- All storage actions at the server side are automatically generated (at run time) from the data types in the program. No special user code is necessary, only the marking by the user of the data that have to be stored;
- The JavaScript at the client side is generic (the same for all iTasks programs). The JavaScript acts as an intermediary between client and server and client and SAPL interpreter. It takes care of updating the page with results from the server or from the interpreter and it transforms user actions in the forms into calls for the server or the interpreter;
- Web pages are generated within the application with the use of generic techniques, so the programmer does not have to program html pages;
- The developer only has to deal with the Clean source program. There is no need for editing generated program sources. This simplifies the maintenance of iTasks programs considerably.

As a consequence of this the programmer only has to deal with a single Clean program and not with JavaScript, html pages and data bases. This simplifies the creation of applications significantly. As will be shown, iTasks programs are very compact.

The Clean compiler produces both an executable and the input for the interpreter. The interpreter is a Java Applet, which is part of the initial html page and is loaded in the web browser when this page is loaded. After starting the interpreter the input file for the interpreter is loaded from the server by a JavaScript function.

Implementation aspects of iTasks

The iTasks library is built on top of the iData library [6]. iData is a library that supports the automatic generation of web pages. This library is based on the following two principles:

- Each user-defined data type can be turned into an (editable) web form using generic functions from a library included in iData. The web form is generated at the server side and transmitted to the browser as (part of) a web page;
- Each user edit action of such a web form is automatically transformed into an updated instance of a data type by a generic function. The user edit action is uploaded to the web server and further handled by the generic machinery. The data type of the value the user entered is checked automatically; if it is incorrect the update is not made.

The programmer may change the default generation of web forms by giving a specific instance of the above-mentioned generic functions for a data type. In this way custom-made web forms can be generated.

So, iData takes care of displaying (editable) information in a web form and updating the data structure in the application with changes made in these displayed web forms. A simple iData application consists of a single web page where a user can fill in forms. iTasks adds the following concepts to iData:

- Tasks are the basic units of an iTasks application. A task can consist of a single iData form (editable data type), but can also be a combination of simpler tasks (using combinators). The most important addition to an iData application is that the user can finish a task by clicking the 'Done' button that is added to a web page. At that moment the form corresponding to the data type of the task cannot be edited anymore and its content becomes available to other tasks;
- Task combinators enable the combination of tasks. A large number of basic combinators are available in the library, but the programmer can also define new combinators. Combinators are used to control the flow of processing and data from one task to another. Tasks can be performed sequentially, in parallel and distributed over several users.

An iTasks application starts as a single executable application at the server side. The iTasks application is re-executed for each client action and generates (part of) a new web page using as input the action and the current state. The state of an iTasks application can be encoded in the web page (in hidden fields), in server side files or in a server side database or a combination of these.

Not all tasks need to be executed by the server. Smaller tasks that do not need information stored in files or databases at the server side can be executed by the SAPL interpreter at the client side. The client side execution of tasks is completely transparent to the programmer of an iTasks application. Only the annotation of a task by the 'OnClient' keyword in the Clean source is necessary.

The client side execution of tasks increases the (execution) performance of iTasks application by reducing the internet traffic overhead (no client-server-client round trip necessary). The SAPL interpreter can also be used for the efficient implementation of (mouse) event processing for more interactive web elements like drawing canvases.

The iTasks combinator library

An iTasks application consists of a set of tasks (possibly for several users) that will be processed in the order specified in the workflow Clean program. iTasks allows for both sequential and parallel evaluation of tasks. The workflow evolves dynamically: the outcome of previous tasks determines the future behaviour of the workflow. Submission of a form by a worker will therefore generally influence the remaining work of this user as well as that of other workers: it may cause the creation of new tasks or the termination of existing tasks. Hence, a single click on a page in a browser may cause very complex state changes on the server and can affect the work of many workers.

The iTasks system is compositional: this means that complex tasks can be composed from other (smaller) tasks by using the iTasks combinators. The basic building blocks for tasks are editable data types. The iTasks system is capable of turning any data structure into an editor. The editor is displayed as a web form in a web page and can be edited by the user.

In iTasks it is also possible for the user to specify new combinators or parameterised workflows. Parameterised workflows can take other workflows as argument. Later on, an example of such a workflow will be given.

Below the most important iTasks combinators will be introduced by giving simple examples of their use. The examples serve to give the reader an idea of the potential and expressiveness of the iTasks system. With the iTasks toolkit it is possible to make much more complex workflows. Details can be found in [6].

Turning data types into tasks with editTask

The '`editTask`' function can turn an element of an arbitrary data type into a task. As a result the user can edit the data type element in a web form. The result of the edit action is fed back to the iTasks system as an element of the data type and can be further processed. If the type of the input does not fit the type of the data type the update is not made. '`editTask`' has two arguments: the name of the button that the user should press to end this task and the initial value of the editor. Here two examples of the use of this function are given: one for an integer argument and one for an element of type Person (together with the definition of Person).

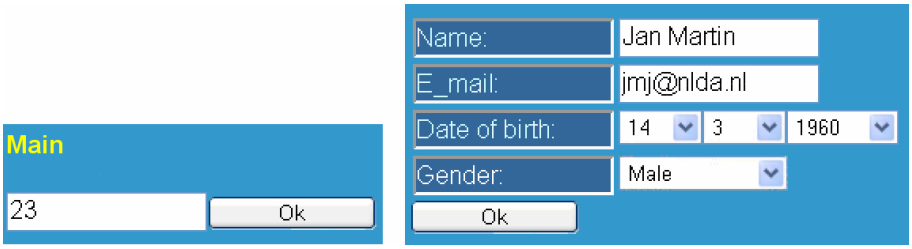


Figure 5. editTask for Int (left) and Person (right)

```
simpleInt :: Task Int
simpleInt = editTask "Ok" 0

:: Person =  { name        :: String
             , e_mail      :: String
             , dateOfBirth :: HtmlDate
             , gender      :: Gender
             }
:: Gender =  Female | Male

simplePerson :: Task Person
simplePerson = editTask "Ok" createDefault
```

Fig. 5 shows the resulting editors. Note that Person has 'createDefault' as initial value. The fields in the form will now get default values generated by the system.

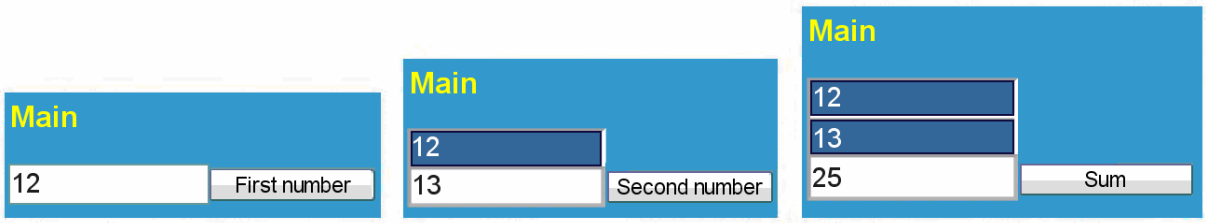


Figure 6. Sequence example (three consecutive windows from left to right)

Sequence

Workflows can be created by combining simple tasks where information of one task (e.g., the result of filling in a form by a user) flows to another task. The combinator in the iTasks system that takes care of this flow is the '=>>' operator. As an example consider the situation of a single user who has to provide two integer numbers in two consecutive forms and as a result gets the sum of these numbers.

```
simpleAdd :: Task Int
simpleAdd
=           editTask "First number" 0
  =>> \v -> editTask "Second number" 0
  =>> \w -> editTask "Sum" (v+w)
```

Fig. 6 shows the three consecutive appearing web-pages. '=>>' takes the result of the left side and gives it as an argument to the right side. '`\a -> body`' is a notation to introduce an inline anonymous function with argument '`a`' and body '`body`'. So '`=>> \v ->`' can be read as: make the result of the left hand side available to the right hand side under the name '`v`'.

Normally the final result of a task should not be editable. This can be achieved by replacing the last line by: '`=>> \w -> return_D (v+w)`'. Now the result is only displayed.

It as also possible to display a text next to the result:

```
=>> \w -> [Txt "Their sum is"] !>> return_D (v+w)
```

Sequence with multiple users

The previous example can easily be turned into an example with multiple users. Here user 0 first has to enter the first number, then user 1 should enter the second number and user 2 gets the result. The user number must be selected with a drop down box on the page, but can also be associated with a user login.

```
simpleAddMU :: Task Int
simpleAddMU
=           0  @:: editTask "First number" 0
  =>> \v -> 1  @:: editTask "Second number" 0
  =>> \w -> 2  @:: editTask "Sum" (v+w),
```

where '`k @:: t`' assigns task `t` to user `k`.

AND parallelism

The AND (`-&&-`) operator generates two tasks that both have to be finished before the results can be used.

```
simpleAndMU :: Task Int
simpleAndMU
=      (0  @:: editTask "Number entered" 0)
  -&&- (1  @:: editTask "Number entered" 0)
  =>> \(v,w) -> 2 @:: editTask "Sum" (v+w)
```

Now both user 0 and 1 may enter their numbers in parallel. The results are collected; their sum is calculated and displayed to user 2.

For AND also a multi-version '`andTasks`' exists, which handles a list of tasks. The task completes when all subtasks are completed.

OR parallelism

The OR (`-||-`) operator generates two tasks in parallel. As soon as one of them finishes the result of that task is available. The result of the other task is ignored.

```
simpleOrMU :: Task Int
simpleOrMU
=      (0   @:: editTask "A number" 0)
  -||- (1   @:: editTask "A number" 0)
  =>> \v -> 2 @:: editTask "First number" v
```

Both user 0 and 1 may enter a number, but only the first one that completes is received by user 2.

Also for OR a multi-version '`orTasks`' exists, which handles a list of tasks. The task completes as soon as one of the tasks completes.

With the help of sequence, AND and OR a great part of typical workflow applications can be programmed. The workflow literature often distinguishes between splitting and merging [2]. Splitting is the start of several tasks in parallel and merging is collecting the result(s) of one or more of the tasks that are split. In this way AND and OR both have the same splitting operation but have different merge operations. Because in iTasks the passing of control and results is integrated, splitting and merging are not distinguished.

Parallel tasks with a condition

In iTasks a special version of '`andTasks`' exists: '`andTasksCond`'. A number of tasks can be started in parallel. Each time one of the tasks is finished a condition is applied to all completed tasks. If the condition is met, '`andTasksCond`' is finished and the completed results are returned in a list.

```
simpleAndTaskCond :: Task Int
simpleAndTaskCond
=  andTasksCond pred [("User" +++ toString u,
                       u @:: editTask "Number entered" 0) \\ u <- [1..4]]
  =>> \xs -> [Txt "Their sum is"] !>> return_D (sum xs)
where pred xs = sum xs > 3
```

Here a parallel task for 4 users is started. They all have to enter a number. Here the condition checks if the sum of the already entered numbers is greater than 3. As soon as this is the case this task stops and the results are passed to another task where they are displayed.

This is a very powerful combinator because many other combinators can be expressed using it. For example the definitions of '`andTasks`' and '`orTasks`' can be given by:

```
andTasks xs = andTasksCond (\ys = length ys == length xs) xs

orTasks xs = andTasksCond (\ys = length ys == 1) xs
```

Tasks with check

Very often the data entered in a form must be checked before the user can proceed. For this a special version of '`editTask`', '`editTaskPred`' is added.

```
simpleCheck :: Task Int
simpleCheck
=             [Txt "Enter a positive number"] !>> editTaskPred 0 checknum
  =>> \num -> [Txt ("You entered: " +++ toString num)] !>> return_D num

checknum num = (num >= 0,[Txt "Number should be positive"])
```

The user should enter a positive number. If a negative number is entered, the user gets a warning and may enter a new number.

Cancelling a task

A frequently occurring action is that a user starts a task, but that he decides to cancel the task. For cancelling a 'Cancel' button should be added to the form and an appropriate action should be taken.

```
simpleCancel :: Task String
simpleCancel
=       (editTask "Number entered" 0 =>> \num -> return_V [num])
  -||-  (buttonTask "Cancel" (return_V []))
  =>>   handleResult
where handleResult []  = return_D "User cancelled"
      handleResult [n] = return_D ("User typed " +++ toString n)
```

Now together with the number field a 'Cancel' button is displayed as an OR task. The task that handles the result will determine the status of the result (empty list or list with one element). In case of an empty list the user has cancelled the task.

Parameterised tasks

It is also possible to specify tasks that have forms of any type as parameter. In this way abstract frameworks to be filled in later with a concrete type can be made. As an example consider the following task frame for double-checking a user form (note that '`a`' is the type of the parameter and '`val`' is its value).

```
doubleCheck  :: a (a -> (Bool, [BodyTag]))  -> (Task a) | iData a
doubleCheck  val pred
=                   [Txt "Please fill in the form:",Br,Br]
    ?>> editTaskPred val pred
    =>> \na ->    [Txt "Received information:", Br, Br,
                   toHtml na, Br, Txt "Is everything correct ?", Br]
    ?>> chooseTask      [("Yes", return_V True), ("No",  return_V False)]
    =>> \ok   ->  if ok (return_V na) (doubleCheck  na preda)
```

'`doubleCheck`' is an extension of '`editTaskPred`'. It first checks the value the user entered against the Boolean function '`pred`', and after this value is ok, it asks the user for a confirmation for this value.

Double check can for example be used to check a person form as follows:

```
doubleCheck createDefault personcheck
```

where '`personcheck`' is a Boolean valued function that checks a person type.

Shifting tasks

In the iTasks system it is possible to shift work from one user to another user. In the next example user 0 may start a task for user 1. This task consists of three steps. If user 1 completes the task the result is sent to user 0 and displayed. But if user 1 stops the task, the work is shifted to user 2 who has to complete the task.

```
shiftExample = simpleShift threeStepTask
```

```
threeStepTask :: Task Int
threeStepTask =   editTask "Done1" 0
      =>> \v1 ->  editTask "Done2" 0
      =>> \v2 ->  editTask "Done3" 0
      =>> \v3 ->  return_D (v1 + v2 + v3)

simpleShift :: (Task a) -> (Task a) | iData a
simpleShift task
=            button "Start the Work"
   #>> 1 @:: button "Stop" -!> task
   =>> \(stopped,TCl task) -> if (isJust stopped) (2 @:: task) (0 @:: task)
   =>> \result     -> return_D result
```

'`TCl task`' is a so-called task closure (partially evaluated task).

The possibility to shift partially evaluated tasks is very powerful and cannot be found in other workflow systems.

Other iTasks combinators

Up to now only a small number of simple examples are given. iTasks has other possibilities which will not be discussed in detail here.

- time limit for tasks, to put a deadline on a task;
- choose a number of tasks to complete from a list of parallel tasks;
- several kinds of repetitive tasks.

The task combinators from iTasks are not fixed, but can be easily extended with new combinators by an application programmer. This is possible because both the combinator library and the workflow program are written in the same language. In this way complex dependencies between data and tasks and even recursive tasks can be programmed.

Other uses of workflow formalisms

Workflow formalisms are not only useful to implement workflow systems, but can also be used to model procedures (or information flows) within a company or organisation. In this way a formal description of these flows or procedures can be made. The model can be used to check the completeness of the set of procedures. Are there steps in a procedure that do not have a follow up? Is the necessary input information available for all steps? The formal description can even be used for simulating the procedure and using this simulation to adapt the procedure.

Applications of iTasks in the military domain

The iTasks workflow library offers the possibility for the high level specification of complex workflows. This section will focus on the applications of iTasks in the military domain.

Workflows occur at various places within the military domain. First of all, standard workflow tooling for administrative processes like personal administration and travel expenses claim handling is used. Also for the material logistic processes, Enterprise

Resource Planning (ERP) tools are used. The Netherlands Ministry of Defence uses PeopleSoft, DIDO and SAP for these applications. In these areas the needs do not differ too much from the needs of other large companies and the commercially available tools are sufficient to deal with them.

Major operational processes are examples of more complex workflow problems. For these processes currently no workflow tools are used to support them, mostly because these processes are too complicated to fit in the formalisms of these tools. For operational processes one can distinguish between:

- the planning phase preceding a military operation;
- execution of the military operation.

Planning of military operations

The planning of operations comprises the following aspects:

- Logistics: Before people can be deployed, accommodation, power supply, water supply, food supply, etc. have to be arranged.
- Transport: Transportation is needed both for people and materiel (accommodation and supplies). A large part of the transportation has to be done beforehand (accommodation, fuel, infrastructure). Other transportation is needed during the entire deployment (food, ammunition, replacements).
- Intel: Prior to the deployment, but also during the operation, intelligence operations are needed. Examples of prior intel requirements are: What are the expected enemy forces, what is the available infrastructure (communication, resources (water, food etc))? What are safe routes for transportation? What are the local terrain conditions? How is the local climate? What kind of protection is needed for the initial transports? Examples of intel during the operations are: What is the enemy behaviour? What is the attitude of the local civilians?
- Communication: A communication infrastructure has to be built-up for the operation: radio, telephone (including GSM), satellite for communication with headquarters and allies including Non-Governmental Organisations, computer networks for the exchange of information, internet for home front communication.
- Budget: What will be the costs of the operation? Do we stay within the maximum allowed costs?

The planning process can be very dynamic, because, for example, intel information can influence already started other tasks. Planning of these complex operations often involves the commitment of large numbers of geographical distributed people over periods varying from several weeks to several months. Currently normal communication channels like telephone and e-mail are used for the exchange of information, while in general spreadsheet and database applications are used for maintaining information. This maintenance is in general on an individual or small departmental basis. This means that other people and departments do not have insight into this information and should make explicit requests (by telephone or e-mail) to obtain it.

It is clear that workflow tooling can be of great help during the planning phase of military operations. Here a summary of some issues the workflow system should support is given:

- access to information for the partners involved. We are dealing with a variable number of dislocated people causing a dynamic workflow topology;
- the automatic checking of deadlines and taking actions in case they are passed;
- initiation of actions (tasks) for several partners involved;
- monitoring the status of actions with the possibility to interrupt or reallocate tasks;
- the on-the-fly construction of new workflows by system end users;
- automatic checking of budget.

Use of workflows during operations

For the planning phase of military operations it is obvious that workflow tooling can be of great help. For the execution phase of military operations this is less clear. But many activities during the execution phase can be modelled as workflows. Military commands have to be distributed and refined. Feedback on the feasibility of these commands must be given. But also activities like weapon deployment or sensor allocation can be modelled as tasks to be allocated. The use of a flexible workflow system like iTasks gives new possibilities to study and better understand these processes and to implement them.

The applications that support military operations are often dedicated special systems like Combat (Battle) Management Systems. Workflow support with iTasks therefore requires interfacing of iTasks with these systems.

Concluding, the iTasks system is a candidate for use during operations, but the applications are less straightforward than for the planning phase. Pilot studies are needed to investigate the potential of iTasks.

Workflows and Net-Centric Operations

Net-Centric Operations aims to connect parties involved in operations by a communications network that can be used for the exchange of all kinds of information. Not only data and voice information can be exchanged, but also sensor and other operational information. In the ideal situation this leads to shared situational awareness, where all participants in the network have access to relevant information and can deploy appropriate weapon systems in the network. These operations pose great challenges. First of all, connecting all operational partners by a network in an operation is a complex task. Because of the mobility of participants, wireless communication means have to be used. Also the use of different (communication) standards by participants is a problem. But even if these technical problems are solved many challenges remain. Who should get which information at what moment? How is this information represented? What actions are the partners allowed to take? They have to respect (changing) Rules of Engagement. Again a flexible workflow system like iTasks offers a good starting point for the construction of applications for Network-Centric Operations. Other, more mathematical, properties of networks in Net-Centric Operations are discussed in [5].

One of the powers of a web-based toolkit like iTasks is that partners from other countries can participate as soon as they can make a network connection. No special installation of software is needed. This is a powerful feature because partnerships are likely to change on an ad hoc basis.

Future work

The iTasks formalism has a solid theoretical and implementation foundation. Although there is still interesting work to be done on these subjects, the focus for the near future will be on applications. It has already been indicated that iTasks could be useful for the planning phase of complex operations. The working out of a scenario for the planning phase of a realistic operation and the building of a demonstrator application for this scenario is planned. This demonstrator will be used to prove the usefulness of the iTasks approach and to obtain feedback from military experts. It is also expected that the implementation will lead to functionality demands for iTasks.

Current investigations already have lead to new demands for the iTasks system. One of these demands is the possibility of ad hoc creation of new tasks. This means that the user must be capable of creating new tasks (and dependencies between them) while using the workflow system. This on-the-fly task creation will not have the flexibility of the full iTasks system, but should be sufficient for simple sequential, AND, OR and repetitive tasks.

Conclusions

In this paper the iTasks combinator library for the construction of dynamic workflow systems has been described. First, a general description of workflow systems and a justification for applications with a web interface was given. No full description of the iTasks system was given, but instead the system was introduced by a number of simple examples that show the potential of the formalism.

The iTasks system has a number of important advantages in comparison with more traditional workflow formalisms:

- the system has a universal web interface. It can be used without the installation of special software. This allows for the on-the-fly join of new users;
- the system allows for easy security. Security can be based on both shared and public key encryption;
- iTasks has a compact and precise formalism. It allows for the formal reasoning about (dependencies between) activities;
- the formalism is extendable. iTasks has a number of predefined combinators, but it is possible for the application programmer to add new combinators;
- the iTasks formalism is compositional. New combinators can be made by combining existing combinators.

The possible usage of iTasks in the military domain was sketched. The most obvious candidate for its use is the planning phase of complex operations like the deployment of troops for longer periods. This planning involves a large number of people and is

characterised by a highly dynamic nature. Therefore, the traditional workflow tools are less useful. The building of a demonstrator application to investigate the usefulness of iTasks for these planning activities is planned.

It was also indicated that systems built with iTasks can be useful during complex operations and for Network Centric Operations.

References

[1] The Haskell Home Page. www.Haskell.org.

[2] W.M.P. van der Aalst, A.H.M. ter Hofstede, B. Kiepuszewski and A. P. Barros. Workflow Patterns. *Distrib. Parallel Databases*, 14(1):5–51, 2003.

[3] J.J. Garrett. Ajax. A New Approach to Web Applications. www.adaptivepath.com/ideas/essays/archives/000385.php, visited February 2008.

[4] J.M. Jansen, P.W.M. Koopman and M.J. Plasmeijer. Efficient Interpretation by Transforming Data Types and Patterns to Functions. In H. Nilsson, editor, *Proceedings Seventh Symposium on Trends in Functional Programming*, TFP 2006, Nottingham, UK, volume 7 of *Trends in Functional Programming*. Intellect Publisher, 2006.

[5] R.H.P. Janssen and H. Monsuur. Military Operations Research and Situation Awareness in Networks, NL-ARMS 2008 (this volume), p 73.

[6] M.J. Plasmeijer and P.M. Achten. The Implementation of iData. A Case Study in Generic Programming. In A. Butterfield, C. Grelck, and F. Huch, editors, *Implementation and Application of Functional Languages*, volume 4015 of Lecture Notes in Computer Science, pp 106–123. Springer, 2006.

[7] M.J. Plasmeijer, P.M. Achten and P.W.M. Koopman. iTasks: Executable Specifications of Interactive Work Flow Systems for the Web. In N. Ramsey, editor, Proceedings of the 2007 ACM SIGPLAN International Conference on Functional Programming, Freiburg, Germany, volume ICFP'07, pp 141–152.

[8] Software Technology Research Group, Radboud University Nijmegen. The Clean Home Page. www.cs.ru.nl/~clean.

Developing a C4I Architecture for the Netherlands Armed Forces

Dick Ooms & Tim Grant

Introduction

Motivation (1) – why an information architecture?

Why would someone want to develop an information architecture? Intuitively, we all know the purpose of an architecture when we think about it in the context of a building: it embodies the grand design, what it should look like when it is finished, how the different components contribute to the overall structure and form part of it, and how the components relate to each other. The architecture relates to the purpose of the building, the functionalities for its users, and expresses the vision of the architect about how these functionalities should be realised.

All of these attributes of the architecture of a building apply to an information (-systems, -services)[1] architecture as well. We can think about an information architecture as a composition of different components or building blocks, being information services, provided by information systems, supported by networks and communication systems, and supporting business processes. Unlike in the process of realising a building, these building blocks are usually not designed, developed and put into service in the same timeframe. On the contrary, they are developed, being used and ultimately being replaced in a continuous process. This is precisely why we need an information architecture: to improve coherence between new and existing building blocks, to provide guidance for new developments, and to ensure that the entire composition of building blocks supports the business processes by providing the information services required. To provide guidance for the development of new components, an information architecture usually depicts both the current situation (the *"ist"* situation) and the ideal situation in future (the *"soll"* situation), and provides guidance about the transition: how we should arrive from *ist* to *soll*.

Motivation (2) – why a C4I architecture and why is NLDA involved?

Development of information services, information systems and the ICT infrastructure for the Netherlands Armed Forces is guided by the Defence Information Architecture (*Defensie Informatie Voorzienings Architectuur,* DIVA). The Chief Director for Defence Information and Organisation (*Hoofddirecteur Informatie en Organisatie,* HDIO) is responsible for the development of DIVA, which is to be underpinned by a series of supporting architectures covering various architecture aspects[2] and defence policy areas[3].

---

[1] An information architecture defines organisational processes, the information flow required for these processes, services and systems which provide that information, and the technical means (ICT infrastructure: networks, communication systems, technical standards) required to support those systems. Such an architecture can be referred to as "information services architecture", "information systems architecture" or "ICT architecture", depending on which aspect prevails. In this chapter we will use the generic term "information architecture".

[2] DIVA Aspect Architectures cover aspects which are defence-wide and include information security and the ICT infrastructure (networks and communication systems).

[3] DIVA Sub Architectures cover policy areas such as operations (C4I), personnel, materiel, finance etc.

This is why a C4I architecture[1] is needed: it is one of the supporting architectures of DIVA. The business process it supports is the operational process. The C4I architecture defines the information flow required to support the operational process, information services that should be in place, and operational information systems which provide such services. The Commander in Chief of the Netherlands Armed Forces (*Commandant Der Strijdkrachten,* CDS) is responsible for operational policy and requirements, and for this reason also responsible for the development of the C4I architecture.

Why got NLDA involved? Since the creation of a new, amalgamated Defence Staff (*Defensie Staf,* DS) in 2005 as a follow-up of the separate staffs of the different services (navy, army, air force), various attempts have been made to create the C4I architecture, both by the DIO staff to assist CDS, and by the Netherlands Organisation for Applied Scientific Research (*Organisatie voor Technisch Natuurkundig Onderzoek*, TNO) as tasked by DIO. However, lack of capacity within DS halted further progress in this area. For this reason, CDS has requested the assistance of the NLDA to develop the first draft of the C4I architecture. It will be shown that this involvement will be beneficial for NLDA as well.

Theoretical context

The ISO-accepted *Recommended Practice for Architectural Description of Software-Intensive Systems* [ISO, 2007] defines a systems architecture as:

> *"the fundamental organisation of a (software-intensive) system, embodied in its components, their relationships to each other and the environment, and the principles governing its design and evolution".*

The intended purpose of developing a C4I architecture is essentially captured by *The Open Group Architecture Framework* [The Open Group, 2007]:

> *"an architecture description is a formal description of a system, organized in a way that supports reasoning about the structural properties of the system. It defines the (system) components or building blocks ... and provides a plan from which products can be procured, and systems developed, that will work together to implement the overall system. It thus enables you to manage ... investment in a way that meets (business) needs ..."*

This implies that for this research, the C4I facilities[2] of the Netherlands Armed Forces are collectively approached as one comprehensive system. This is a valid approach, since they collectively show the characteristics of a system as described in literature:

- they have a structure that is defined by its parts and processes;
- the Netherlands C4I system is a generalisation of reality;
- the various system parts have functional as well as structural relationships.

However, it should also be pointed out that, as laid down in the Netherlands Defence Doctrine (*Nederlandse Defensie Doctrine* (NDD), see [MOD NL, 2006]), deployed and

[1] Internationally, C4I has different meanings. Here we mean: Command & Control, Communications, Computers and Information / Intelligence.

[2] C4I facilities: these include operational information systems and mobile and deployable networks and communication systems.

mobile operational staffs and units of the Netherlands Armed Forces assigned on a mission will in principle always be operating as building blocks in an international force. This implies that their C4I facilities should also be building blocks of an international C4I structure consisting of national contributions from participating nations. This international C4I environment points at the necessary international dimension of the C4I architecture. Indeed, the international environment defines to some degree what the national C4I architecture should look like.

There is a great variety of architectural styles in the scientific literature, such as client-server architectures, component-based architectures, blackboard systems, model-view-controller, modular plug-in architectures, layered architectures and peer-to-peer architectures. In selecting an architecture style and framework, the aforementioned international dimension of the C4I architecture should be taken into account. The C4I architecture will comply with the principles of third-generation C2/C4I system architectures, as implemented in the NATO Architecture Framework (NAF), see [NATO, 2004], the US DoD Architecture Framework (DoDAF), see [US DoD, 2004], and especially DIVA. In 2007 TNO has performed a comparative study of these and other architectures [Riemens et al., 2008], the findings of which will be used in the development of the C4I architecture. Specific tools, model views and methods developed for these architectures could be applied for the Netherlands C4I architecture and could be proposed as additions to DIVA.

DIVA has mandated the *Service-Oriented Architecture* (SOA), in which software systems are built from software services. Services are relatively large units of functionality that are not *a-priori* associated with one another, i.e., they have no calls to one another embedded in them. Examples of services in a military context are: geographical and oceanographical data support, prediction of acoustic propagation, advice on Rules of Engagements in force and related legal implications; computation of fire control solutions; analysis of large amounts of sensor data (e.g., pattern recognition); analysis of electromagnetic intercepts; advice on weapon and target selection; etc. Instead of embedding calls to one another in their source code, services define protocols that describe how the services talk to one another. Based on these protocols, services can be linked and sequenced automatically in a process known as *service composition*. Research issues in SOA include protocol standards and service composition methods. Additional research issues specific to C4I include how to adapt services and SOAs to real-time requirements; bandwidth limitations; joint, combined and civil-military interoperability; agility and reconfiguration on-the-fly; and international regulatory constraints.

DIVA is a 3-level architecture (see Fig. 1), like NAF and DoDAF. The upper layer contains the business processes, the middle layer the information services which support the upper layer, and the bottom layer contains the technology required for the middle layer. For the C4I architecture, the business process is the operational process, for which the OODA Loop[1] will be adopted.

[1] As developed by Boyd. OODA: Observe, Orient, Decide, Act

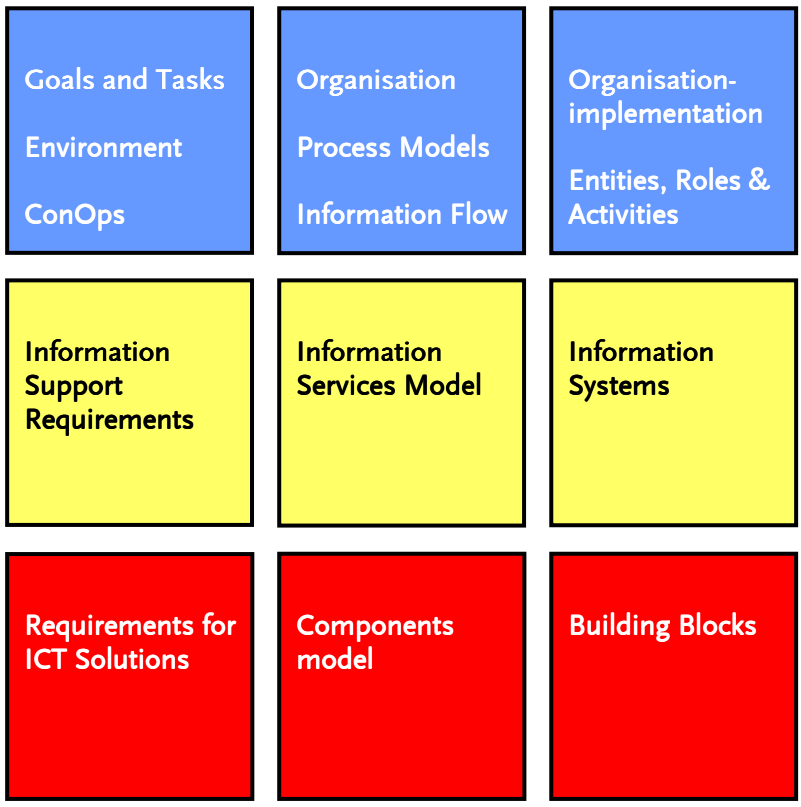


Figure 1. The DIVA 3 layer framework

Purpose, scope and structure of this chapter

This purpose of this chapter is to provide an overview of the progress made to date in developing a C4I architecture for the Netherlands Armed Forces. It starts with a discussion on the intended purpose and scope of the C4I architecture, because as prescribed by DoDAF, purpose and scope are the first subjects one has to deal with when developing an architecture, as they provide direction for all further activities. Once these have been defined, we take a quick tour around the C4I world, defining some (potential) challenges. These relate to some research issues mentioned above: bandwidth limitations and interoperability. Subsequently, it is shown how the C4I architecture could assist in coping with those challenges. Finally, we address the research into the actual development of the C4I architecture: an overview on the method of work adopted to arrive at the intended C4I architecture for CDS, the progress to date, and how this effort will be beneficial for the NLDA as well.

Purpose and scope

Purpose

Definition of the purpose of an information architecture could help to avoid a common pitfall in the world of information architectures: their size and level of detail, as developed by (over)enthusiastic information architects, tend to grow out of proportion, compared with the actual application of the end product, and thus the architecture seems to become

a goal in itself[1]. To avoid this trap, the practical purpose of the C4I architecture as viewed by the various stakeholders should be investigated from the outset[2]. The results of a first attempt are shown in Table 1. In addition to CDS and DIO, the following primary stakeholders have been identified: the Defence Materiel Organisation (*Defensie Materieels Organisatie,* DMO) which is responsible for the management and execution of C4I projects to realise C4I requirements as stated by CDS; the Centre for Automatisation of Mission Critical Systems (CAMS), which is responsible for the development of naval C2 systems, and its army-counterpart: the Command and Control Support Centre (C2SC), which is responsible for development of land-oriented C2 systems[3]. The major operational commands (maritime, land and air) are primary stakeholders as well, being the major users of C4I services and systems and as such involved in the identification of future C4I requirements. The required level of detail of the C4I architecture can thus be derived from its purpose, as viewed by its primary stakeholders.

Table 1. Primary stakeholders and purpose of C4I architecture as viewed by them

primary stakeholder	purpose of C4I architecture as viewed by stakeholder
CDS	supports the translation of C4I policy into C4I requirements, provides cohesion and priorities between C4I requirements
DIO	complements DIVA, provides specific requirements for the mobile and deployable ICT infrastructure[4] (DIO's responsibility)
DMO	provides guidance for C4I project architectures, specifies technical standards, provides coherence between C4I projects
CAMS & C2SC	provides priorities, guidance and coherence for development of new systems and services, specifies technical standards
major operational commands	provides a means to articulate information exchange requirements and insight in the realisation of these requirements

Although not considered primary stakeholders[5], NATO and operational partners could also be listed as stakeholders of the C4I architecture. They have an interest in the Netherlands C4I architecture as well, since it supports cohesion and interoperability in an international environment. Finally, even the C4I industry is to some extent a stakeholder, in view of the shift to more use of Commercial off-the-shelf (COTS) and Military off-the-shelf (MOTS) products, and the possibility of Public Private Partnerships.

---

1 Personal experience of the first author, confirmed in the first round of interviews with stakeholders.

2 This is in line with DoDAF, which mandates that as a first step in the development of the architecture, its intended use should be defined.

3 This would seem to leave out the development of air force C2 systems. A software development centre for air force C2 systems does not exist in The Netherlands for two reasons: firstly, the air force is using NATO C2 systems and proprietary C2 systems embedded in aircraft, which means less requirements for own C2 software development; secondly, some systems developed by C2SC are also in use by the air force, such as TITAAN (a deployable ICT infrastructure for deployed army and air force units).

4 The deployed and mobile ICT infrastructure is comprised of deployable **networks** to support deployed operational staffs and units, and deployed and mobile **communication systems** to create networks among mobile units and to link deployed and mobile networks into larger networks and into the static ICT infrastructure.

5 They are not listed as primary stakeholders because they do not define the required level of detail of the Netherlands C4I architecture.

Scope

The definition of the intended scope of the C4I architecture is closely related to discussion and even controversy about the responsibility for the armed forces deployable and mobile ICT infrastructure[1]. This is a sensitive issue in the operational world, because the deployable and mobile part of the ICT infrastructure is considered to be essential for deployed and mobile operational forces. This discussion can be traced back to the creation of DIO. At that time this caused discussion about the remaining responsibilities of the staffs of the various services (navy, army, air force). Since its inception, DIO has been responsible for the defence-wide information architecture, but the staffs of the services retained their responsibility to state, fund and realise requirements for their own mobile and deployable ICT infrastructure. When in 2005 the separate staffs of the services amalgamated into the new Defence Staff, the topic of discussion turned into the delineation of responsibilities between DIO and DS. A remaining responsibility for CDS was identified to state requirements for the deployable and mobile ICT infrastructure, while DIO retains the overall responsibility for the defence-wide ICT infrastructure: fixed, deployable and mobile.

Translated into architecture terms, this means that DIO is responsible for the DIVA aspect architecture of ICT Infrastructure, referred to as the Communications and Networks (aspect) architecture. CDS is responsible for the C4I architecture (a DIVA sub-architecture), which will articulate specific requirements, from an operational point of view, for the deployed and mobile ICT infrastructure. These requirements feed into DIO's Communications and Networks (aspect) architecture. Similarly, the C4I architecture will formulate specific requirements for information security systems and services, which feed into the Information Security (aspect-) architecture[2], and other requirements e.g., regarding operational logistics, which are catered for by other sub-architectures.

The discussion about scope is more than the reflection of old "territorial battles", which have by now been settled. It reflects a broader development: from "stovepipes", i.e. different specific ICT infrastructures for different services and different policy areas such as operations and logistics, into a common ICT infrastructure which supports all deployed and mobile staffs and units, and provides services for all policy areas.

C4I challenges

First we list some C4I challenges, both generic and specific for the Netherlands C4I situation. Subsequently we will show how a C4I architecture could help to cope with these challenges.

[1] The following information about internal discussion on scope and responsibilities is derived from personal author inside knowledge (from the first author), who served at the time as department head in the Naval staff and in the Defence staff. It provides useful contextual information and illustrates the shift from separate to common, from service-specific to joint systems and infrastructure.

[2] The Information Security architecture is a DIVA aspect architecture which is the responsibility of the Netherlands Defence Security Authority (*Beveiligings Authoriteit, BA*).

Common C4I challenges

In general, C4I systems have the following characteristics in common, which set them apart from "ordinary", i.e., non-operational information systems and which pose a challenge both for their design and for the supporting ICT infrastructure:

- **unique real-time requirements:** C4I systems supporting the C2 process are often real-time systems (e.g., supporting weapon engagements and providing the air picture) as opposed to most business-oriented processes[1], which generates specific requirements for processing speed and bandwidth;
- **bandwidth-limited environment:** C4I systems often have to operate in a bandwidth-limited environment[2] (mobile military communications and networks), which generates specific requirements for bandwidth-efficiency and -management;
- **interoperability and agility:** C4I systems and the supporting ICT infrastructure often operate in a dynamic environment with ad hoc arrangements, and a varying composition of partners: different forces, nationalities, non-governmental organizations, etc. The configuration of military units often changes on-the-fly during an operation, and the C4I system must itself change configuration accordingly. This generates requirements for interoperability (joint, combined, civil-military) and agility;
- **international architecture dialogue:** the international military C4I community is very much involved in the development of C2 concepts, C4I systems, the supporting ICT infrastructure, and in the choices to be made in the architectural development, which evolve in an ongoing international dialogue. For non-operational information systems used by the armed forces, one can and must conform to international standards that cannot be influenced, or even COTS;
- **unique security requirements[3]:** operating in an international coalition involves sharing of sensitive information and transport of information between national networks. At the same time, these networks carry highly classified national information that cannot be shared. Technical solutions should be accredited by all parties participating in the coalition.

C4I and NEC

C4I systems and the supporting deployable and mobile ICT infrastructure are an essential requirement for the realisation of the concept of Network Enabled Capabilities (NEC). The planned, phased realisation of this concept is laid down in the NEC Action

[1] Some non-military information systems have real-time requirements as well, such as Air Traffic Control and bank transactions. However, this is not true for the non-operational information systems used in the armed forces, which are non real-time. So, within the military context the distinction is valid. Moreover, non-military real-time systems do not require the same mobility and bandwidth as C4I systems. So, the combination of listed characteristics set C4I systems apart from military non-operational systems and from non-military systems.

[2] Some non-military information systems operate wireless as well, but mostly operate within commercial broadband coverage, which is not true for mobile operational units.

[3] Some non-military information systems also have special security requirements, but these are accommodated by commercially available products. C4I systems require specific non-commercial security solutions, which are to be certified at the national government level, and if necessary by NATO or partners.

Plan[1] which provides goals and milestones. C4I developments have to be synchronised with the planned realisation of NEC.

A specific requirement for Netherlands operational staffs and units is to become "*net ready*", i.e., to be able to make their weapon, sensor and C2 capabilities available to cooperating staffs and units, and vice versa: to be able to make use of such capabilities offered by cooperating staffs and units. While this requirement is recognised in general, translation into specific C4I requirements proves to be difficult.

C4I stovepipes
The term "stovepipe" refers to a C4I system which is dedicated to a specific service (navy, army, air force) or to a specific transmission channel, or a specific discipline or specialisation, or in another way shows a shortfall in the characteristics which are nowadays required in a network enabled battle space. These requirements are relatively new, which explains why many in-service C4I systems still show some stovepipe-characteristics. Until as recent as 2005, in the Netherlands Armed Forces each of the three main services developed its own C4I systems and supporting deployable and mobile ICT infrastructure, without much coordination with the other services. For many years, being interoperable with international partners was of far more importance than being interoperable within the Netherlands Armed Forces.

Of course not all stovepipes are bad. The different forces operate in different environments and this sometimes leads to other requirements and different choices. Examples are:

- restrictions in weight and space on board of military aircraft which leads to other choices for tactical datalink systems (i.e., only Link 16) than in the maritime environment, where coverage is the driver for continuation of use of HF datalink systems such as Link 11 and its successor Link 22, in addition to Link 16 for major units;
- interoperability at unit-level required for maritime operations, which has lead to extensive standardisation for communications equipment and operational information systems (i.e., MCCIS)[2], unlike in the land environment, where national internal interoperability prevailed in the past, and the approach now is to make use of national systems, linked by a common interface;
- the use of VLF radio specifically for submarine broadcasts, because these low frequencies can penetrate the water, allowing the submarine to stay submerged while copying the broadcast.

However, many current differences cannot be explained in this way and are simply caused in the past by a lack of coordination.

[1] NEC Action Plan: a yearly updated plan, developed by the Defence Staff, which governs the implementation of the NEC concept in the Netherlands Armed Forces.
[2] MCCIS: NATO's Maritime Command & Control Information System, initially intended for NATO command posts, now also widely implemented in national maritime headquarters and on board frigates and above.

Interoperability
This important aspect was already mentioned as one of the common C4I challenges. In an ideal situation, operational staffs and units in any mix of different services and nationalities should be able to interoperate seamlessly, and technical solutions to this aim should be transparent to the user. However, reality is still a far cry from this ideal end state. This means that from a national perspective, sometimes choices have to be made with which partners achieving interoperability has the highest priority, and whether national (joint) or international (combined) interoperability should prevail.

Solutions: the C4I architecture

Solutions for common C4I challenges
The fact that these challenges are common can be considered a blessing in disguise: it means that we can take a close look at NATO and partner nations to see how they cope with them. Having a Netherlands C4I architecture provides a means to implement possible solutions as embodied in e.g., NAF and DoDAF, by translating them into the Netherlands C4I architecture.

The C4I architecture could serve another purpose in relation to two of the listed challenges. Real-time requirements and bandwidth limitations could be considered a paradox: C4I systems pose high demands on bandwidth, while at the same time they have to operate in an environment that is characterised by its bandwidth limitations. This paradox will become even more apparent with the advent of many remote sensing systems, operated from satellites and UAVs. The C4I architecture could provide insight into the cumulative bandwidth requirements by various existing, planned and required C4I services and systems. This would reveal the total impact of these bandwidth requirements on the mobile and deployed ICT infrastructure. To put it the other way around, this could help in setting boundaries to unrestricted bandwidth claims. Rather than discussing bandwidth requirements ad hoc, each time when a new requirement pops up, the C4I architecture would allow a more structural approach.

Solutions with respect to NEC
The C4I architecture should describe both the current situation with respect to C4I services and systems *("ist")* and the situation required in future *("soll")*. The transition from *ist* to *soll* is to be specified in phases or *architecture stages*, which should be aligned with the different NEC maturity levels as specified in the NEC Action Plan. Admittedly, this could be a challenge, since the description of NEC maturity levels is non-specific as to C4I requirements. This would require that the NEC maturity levels are translated into specific C4I requirements, which then collectively can be depicted as C4I architecture stages. This translation should be performed in the context of the development of the C4I architecture.

With respect to the other challenge related to NEC: the C4I architecture could also be used to find a solution for the problem to define what it means to make units *net ready*. As mentioned in the previous paragraph, it could be used as a means to translate possible solutions by NATO and partners into the Netherlands C4I architecture.

Solutions for C4I stovepipes
Developing a common C4I architecture for the armed forces is probably a prerequisite to get rid of unnecessary stovepipes in a coherent and planned way. While investigating the *ist* situation, it should be questioned whether current differences are justified by differences in environment and deliberate choices. If they are not, they should probably no longer exist in the *soll* situation. The process of arriving at a shared view within the armed forces on what should be the *soll* situation, as part of the development of the C4I architecture, could prove to be very valuable in itself. Once the *soll* situation is agreed upon, a transition plan should be developed to arrive from *ist* to *soll*, and this coincides with the transition mentioned in the previous paragraph.

Solutions for interoperability
This aspect should be an essential feature of any C4I architecture. By investigating the information exchange requirements in different scenarios, the C4I architecture should support logical choices for setting interoperability priorities. At the systems and technical level, the applicability of different solutions should be investigated and principal choices should be made, such as the implementation of internationally agreed standards (e.g., NATO datalinks and waveforms) or implementing internationally agreed gateway solutions such as developed by the Multilateral Interoperability Programme (MIP). MIP developed a "common semantic core", which provides interoperability at the semantic level between nationally developed operational information systems (see [Chaum and Lee, 2008]).

Research into the development of the C4I architecture

Method of work
From a theoretical point of view, the research approach is *formulative* with *descriptive* and *evaluative* elements:

- it is formulative because the C4I architecture document formulates what the architecture should look like at a specific point in time to achieve the goals and milestones of the NEC Action Plan;
- it contains descriptive and evaluative elements because it describes the baseline, being the C4I components currently available, planned and being realised, and evaluates these components against the requirements defined in the C4I architecture.

Research methods include interviews, literature review, operational case studies and conceptual analysis of current C4I systems and projects.

The first phase of research consists of identifying stakeholders, defining purpose and scope, and ensuring leadership support. In the past years various C4I architecture efforts have been made as mentioned earlier in this chapter, the results of which are to be examined and used to the maximum extent possible, to avoid duplication of effort. It will also be investigated to what extent methods, tools and views from other architectures can be used for the development of the C4I architecture (see "theoretical context" above), and to what extent TNO will be involved.

The second phase of research consists of the collection of information to create the upper and middle layers of the C4I architecture. The upper layer describes the operational process and its information exchange requirements in various typical scenarios. The middle layer describes the information services and systems required to support the upper layer. To build the upper layer, interviews will be held with representatives from the operational commands, augmented with case studies and literature study. To build the middle layer, interviews will be held with representatives from the C2 development centres CAMS and C2SC and from the Defence Materiel Command (DMO)[1], augmented with conceptual analysis of current C4I systems and projects.

Building the upper layer should provide insight into the information exchange requirements in a number of standard operational scenarios. This should include whether these are currently being supported by available information services and systems, what is still missing and which deficiencies should be rectified first. Building the middle layer should result in the definition of a set of common operational information services which can be used both by CAMS and C2SC[2]. It should also provide an overview of information services currently being provided by C2 systems and being developed and planned. Comparing the information from the upper and middle layers could show discrepancies between what is required (upper layer) and what is being developed (middle layer), and could help setting priorities for further development of services.

The third phase of research will be aimed at providing the bottom layer, which completes the C4I architecture. This layer will define technical standards for C4I services and systems, and technical requirements for the supporting ICT infrastructure, e.g., the cumulated capacity requirements for communication links (see "solutions for common C4I challenges" above). This development effort is a logical follow-up of the building of the middle layer, and will use the same information sources mentioned above.

The C4I architecture covers a vast area. To keep the development efforts manageable, initially the scenarios to be studied will be kept as simple as possible, covering standard situations. As follow-on, more complex scenarios should be examined, up to the maximum level of ambition for deployment of the Netherlands Armed Forces[3], using the experience from the first architecture efforts.

Progress to date

Phase one has been largely completed. Working arrangements have been established with DS, in close coordination with DIO. This has resulted in a first definition of purpose and scope, an outline of the method of work, and an initial framework for the C4I architecture document ([Ooms, 2008]). This version has been discussed with DS and DIO. The report

[1] Although only national players in the C4I field will be interviewed, this should not imply a primarily national focus. As a rule, Netherlands C4I projects are embedded in international developments, which is strongly promoted by Netherlands C4I professionals.

[2] DIVA already contains operational information services, which will be used as a starting point. As a first impression, a finer granularity seems required.

[3] As politically agreed, this is for the army a deployed brigade, and for navy and air force the equivalent.

of the comparative study conducted by TNO ([Riemens et al., 2008], see "theoretical context" above) is being studied, and the possible use of methods, tools and views from other architectures will be discussed with TNO. Involvement of TNO in architecture work in 2008 has been agreed in principle with DS and TNO and will be formalised in the near future. Leadership support is being ensured at two-star level within DS and DIO.

Phase two has been initiated. For the middle layer, initial contact with C2SC has been made and information provided on architecture efforts and C4I projects is being studied. CAMS will be contacted at short notice. For the upper layer, the staff officer C4I of the Defence Staff Operations Center (DOPS/J6) has been interviewed and as follow-on his counterparts in the operational commands (maritime, land and air) will be approached at short notice.

Benefits for NLDA
The information derived from the involvement in the C4I architecture can directly be integrated into the study material for the Bachelor CICS course and various C2/C4I related subjects of the Bachelor MS&T course, such as computer networks, C2 architecture, military communications, and subjects within the C4I profile. Furthermore, the C4I architecture document could provide a starting point for various BSc thesis projects. From a wider perspective, the architecture research efforts will increase the visibility of NLDA defence-wide and will show how its scientific know-how can be applied for the armed forces.

References

ISO (2007) *Recommended Practice for Architectural Description of Software-Intensive Systems,* ANSI/IEEE standard 1471-2000, in 2007 accepted by the International Standards Organisation (ISO) JTC1 as ISO/IEC DIS 42010:2007. ANSI: American National Standards Institute, IEEE: (US) Institute of Electrical & Electronics Engineers.

The Open Group (2007) *The Open Group Architecture Framework* (TOGAF), developed by the Architecture Forum of The Open Group, continuously evolved since mid-90's. Current version is 8.1.1.

MOD NL (2006) *Netherlands Defence Doctrine (Nederlandse Defensie Doctrine)* (NDD), first edition, Netherlands Ministry of Defence (MOD NL), 2006

NATO (2004) *NATO C3 System Architecture Framework* (NAF) version 2, AC/322-D(2004)0041.

US DoD (2004) *DoD Architecture Framework* (DoDAF), version 1.0, Vol I-III, US Department of Defense, February 2004.

Riemens, J.M.J., Hekken, M.C. van, Lasschuyt, E. and Niet, M. de (2008) *Een architectuurmetamodel voor Defensie; meer samenhang en herkenbaarheid,* TNO-DV 2007 A117, 30 January 2008 (in Dutch).

Chaum, E. and Lee, R., (2008) *Command and Control Common Semantic Core Required to Enable Net-centric Operations,* Proceedings of AFCEA – GMU C4I Center symposium "Critical issues in C4I".

Ooms, D. (2008) *DIVA Deelarchitectuur operationele informatie voorziening,* (in Dutch), in preparation.

Military Operations Research and Situation Awareness in Networks

René Janssen & Herman Monsuur

Introduction

Operations Research

What is the best we can do when we only have limited time to decide? Who must act? When? Where? How? As we live in a world where change is both far reaching and fundamental, we are probably facing such questions. It is good then to know that Operations Research (OR) can help you to take the right decision. This scientific discipline emerged from a process of solving optimization problems with strategic, operational, technical and tactical aspects. With the analytical and computer, and also conceptual, based approach, OR provides people and organizations with a rational basis for decision-making.

OR plays an important role, both in the public and private sectors. Whether one thinks about transportation, inventory planning, communication network design, risk management, health care, reverse logistics (reuse or disposal of products and material) or e-business. OR is a powerful tool in the hands of everyone involved in management. When OR is applied to the military domain, we speak of Military Operations Research (MOR). This special branch of OR focuses on subjects like search and detection, combat modelling, multi-criteria analysis, planning and logistics of military operations, inspection strategies for counter-drugs operations, measures of effectiveness for new weapon systems, game theoretic models for conflict, network theories, homeland security, value and effect of battlefield information, deployment of UAV's, decision analysis, war on terror, detection and mitigation of threats, bio-attacks, terrorists' networks, etc. Like many fields of scientific research, Military Operations Research has become a highly multi-disciplinary endeavour.

The MOR section of the NLDA has its 'home base' at the Royal Netherlands Naval College in Den Helder. This section concentrates on search and detection, combat modelling, homeland security and Network Enabled Capabilities (NEC). The most used scientific methods are statistical analysis, simulations, network theories, game theory, decision analysis, etc.

(Military) Network Science

'Information dominance and superiority' and 'network centric warfare' are terms that have become part of the lexicon associated with the transformation of the military force in the 21st century. Also well-known are the four tenets of network centric warfare: 1. A robustly networked force improves information sharing and collaboration; 2. Such sharing and collaboration enhance the quality of information and shared situational awareness; 3. This enhancement, in turn, enables further self-synchronization and improves the sustainability and speed of command; 4. This combination dramatically increases mission effectiveness. The concept of NEC involves the use of complex systems, consisting of many components that are heterogeneous in functionality and capability with both non-local and non-linear interactions and effects. The network aspect of this all

captures the essence of transformation and is also a central element in improving combat effectiveness. Studying social, cognitive, physical and information networks is therefore of paramount interest. But, in spite of this dependence on networks, in military sciences, little is known about, for example, the relationship between architecture and functioning of a network: given some task, will a hierarchical network be outperformed by certain other network configurations that are more flexible? What about information propagation through a network? What are the mechanisms that explain why and how networks change over time? Or, with respect to security, which networks are prone to deliberate attacks on their nodes? What about the modelling of networks characterised by noisy and incomplete data? The answer to these questions contributes to the desired situation awareness at higher command levels and thus to the decision-making process.

In this contribution, we focus on the relation between the network's architecture and situation awareness, which forms the basis for decision-making.

Networks and network value

The concept of a network has become a basic and central notion in several scientific disciplines. In sociology and economics, many issues, like the interaction between individual entities, are formulated in terms of networks. In literature one may find several illustrations of this network approach; for comprehensive introductions see, for example, [Barabási, 2003], [Watts, 1999], [Dutta, 2003] or [Wasserman and Faust, 1994]. In military sciences, the concept of network centric operation has also attained considerable attention. The issue here is how operations are affected by the topology (the structure of links and nodes) of information networks, physical networks and social networks. See [Cares, 2004], [Darilek et al., 2001], [Monsuur, 2007b] or [Perry et al., 2002, 2004] for more information. Fig. 1 represents an example of a topology for a generic network.

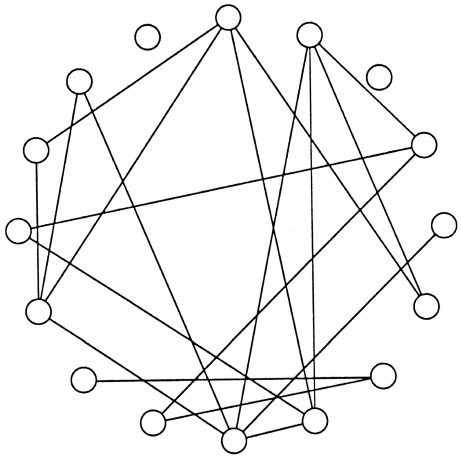


Figure 1. Illustration of a network

We assume that the network generates value for itself as a whole, but also for the individual nodes that are connected through links of the network. In a social network for example, the network value of a particular node may be something like status or prestige that a node derives from characteristics of its local network structure. For example, if a node is in a brokery position, meaning that the network becomes disconnected if it removes links, its status may be high. An important characteristic of a social network is

that if a node somehow succeeds in gaining extra status, this also adds to the status of neighbouring nodes. So, network value is transferable and the extent of transferability depends on the strength of the tie between the two nodes. For a military network, where nodes exchange information they have gathered and processed, the network value may be situation awareness. The level of situation awareness is a result of the functioning of the information exchange network. Transferability depends on usability of information: is the information that is relayed to a particular node relevant, timely, concise, and is it highly regarded in terms of source and content? For an arbitrary network, this network value may also depend on exogenous (network independent) or unique characteristics, such as abilities and resources. In the case of military networks, these exogenous characteristics of the individual units may be their decision making facilities. For an overview, see the following table taken from [Monsuur, 2008].

Table 1. Network value, transferability and exogenous value

Type of network	*Network value*	*Transferability depends on*	*Exogenous value*
social network	*status or prestige*	*strength of tie*	*abilities and resources*
alliance network	*monetary gains/ competitiveness*	*corresponding standards*	*internal organization*
military information network	*situation awareness*	*usability of information*	*information fusion capabilities*

Networks consist of a number of distinct entities that may be similar or dissimilar (components, people, military formations). These entities interact in such a way that new properties or behaviours emerge that are beyond the capabilities of any of the entities acting alone. In general, networks may be considered from three perspectives: *Network structure* (links, nodes, connection rules), *Network evolution* (behaviour of network: deterministic or stochastic and its adaptation to the environment) and *Network dynamics* (mechanisms for networked effects). We will elaborate on these perspectives showing that (abstract) network theories are useful for studying emergent behaviour and emergent properties (shared situation awareness, agility, robustness) in the NEC domain.

Network structure

Consider the ancient Chinese game of 'GO' in which players capture stones and occupy territory. The board on the left of Fig. 2 shows a traditional grid; the board on the right shows a grid designed for a complex network. There are large hubs, clusters and long distance connections. It is clear that in order to win this game, new strategies will have to be developed. For example, on the left, a traditional strategy creates advantage from a great number of adjacent stones, while new strategies have to take into consideration hubs and long distance connections between cleverly placed clusters. An appealing property of these 'battles of networks' (also from a military point of view) is that complex networks prevent competitors from guessing the specifics of their strategies.

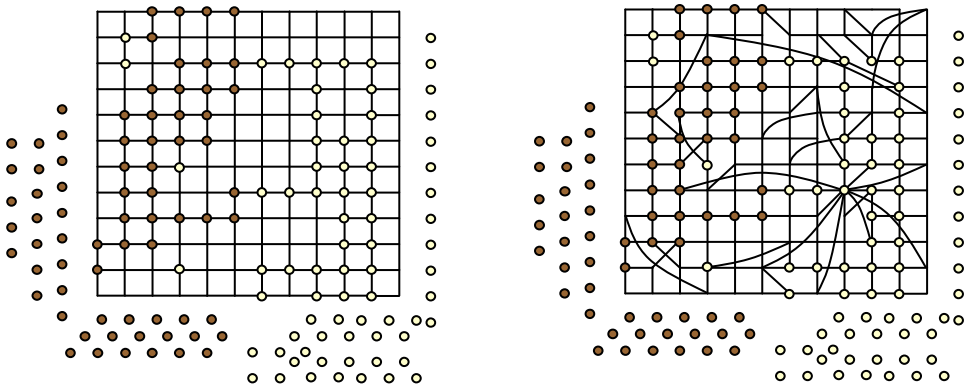


Figure 2. The game of GO on a regular and complex network structure (Modification of a figure from Harvard Business Review 2006: Battle of the Networks, [Cares, 2006])

It is clear that network structure matters, also for NEC. For another illustration we refer to [Grant, 2006], where the chaos that resulted from the 9-11 attack is analysed using a more flexible and agile network, instead of a fixed hierarchical network.

Network evolution

Suppose that a military commander has suggested the following network, in order to control a certain area of operation, see Fig. 3. Nodes represent decision facilities, information fusion centres, combat units, ground based air defence, etc. The nodes relay information they gather and process in order to increase situation awareness for the whole network (or just for the military commander).

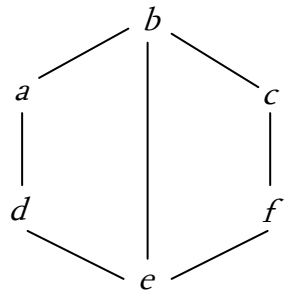


Figure 3. The planned information network

After a few hours of successful, autonomous functioning of the network, the commander considers his job done. The next day, he is informed that the network has rearranged itself. The nodes *a* and *c* now directly exchange information; moreover, nodes *b* and *e* were planning to finish their direct link. How can the actions of these individual, autonomous nodes be explained? A possible explanation might be that in the original network node *a* is covered by node *e*. This means that all of its connections that provide information (from *b* and *d*), also provide information to node *e* and, in addition, node *e* also receives information from node *f*. Something similar holds for node *c*. Note that, if nodes *a* and *c* decide to connect, they become uncovered. So, covered nodes, suffering a structurally visible position, will have an incentive to rearrange the network, either by adding or by severing links. As is proved in [Monsuur, 2007a], only a few network topologies can emerge from this process of actions of the individual nodes: complete networks, star networks and simple cycles.

In the literature one may find many models that try to explain how certain network topologies do emerge. A well-known example is *preferential attachment*. This mechanism assumes that a new node connects to nodes of the existing network with probability proportional to the number of connections these nodes already have. The result is that the 'rich-get-richer': a small number of very well-connected nodes (hubs), a medium number of moderately connected nodes and a large number of sparsely connected nodes. This resembles the structure of the internet. Fig. 4 is an artist impression of the internet (www.andreae.com/images/Pictures_and_Logos/Internet-map.gif), showing a few large hubs that are connected to an extreme number of other nodes.

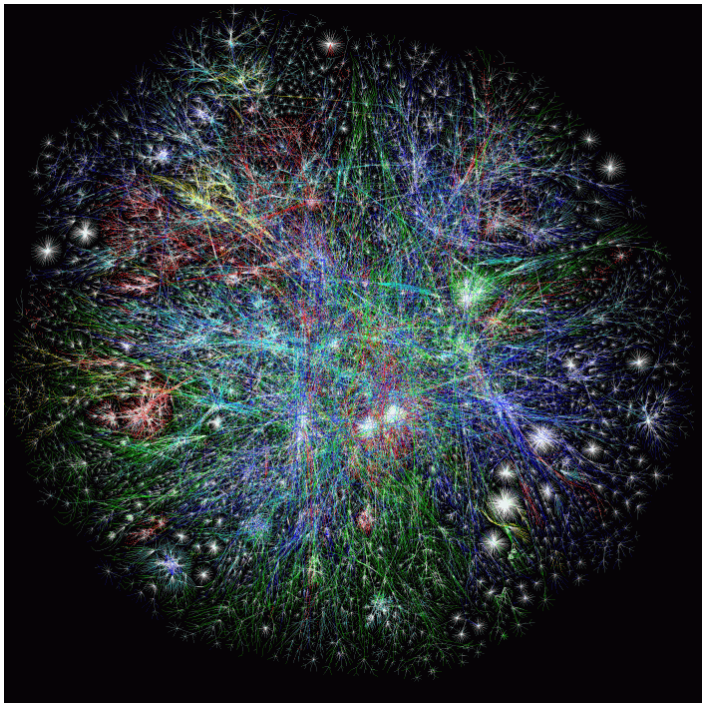


Figure 4. The internet as a scale-free network

Network dynamics

Because of the many ways in which forces could network efficiently, there is an interest in modelling and experimentation, and therefore there is a corresponding interest in metrics for network centricity. These metrics may shed light on the interplay between network topology and dynamics, where the role of the network's topology is to serve as a skeleton on which dynamic processes, such as the transfer of information, takes place. In the literature on NEC, one may find several metrics for the network effects, such as self-synchronisation and situational awareness, see Table 2, taken from [Fewell and Hazen, 2003].

For example, *speed of command* is the time required to complete one full cycle of the observe-orient-decide-act loop; *force agility and massing of effect* has to do with the ability to achieve a massed effect at some critical point in the battle space, and then to reorganise quickly to amass effects elsewhere as the situation develops; *self-synchronisation* means that the units' efforts are such that they are mutually supportive in the accomplishment of the overall goal, without the need for detailed centralised control.

Table 2. Characteristics of network-centric military systems

Top level – force level characteristics: • *speed of command* • *self-synchronisation* • *effect-based operations* • *information superiority* • *force agility and massing of effects* • *shared situational awareness* • *reachback* • *interoperability*
Second level – characteristics of decisions • *speed* • *soundness*
Third level – characteristics of information • *relevance, clarity* • *accuracy* • *comprehensibility* • *value* • *timeliness* • *consistency* • *secrecy* • *deg. of interoperability* • *age, currency* • *completeness* • *authenticity*
Fourth level – general characteristics of networks • *availability* • *reliability* • *concurrency* • *survivability* • *coverage* • *security*
Base level – physical properties • *bandwidth, network topology, server speed etc*

Many studies attempt to measure the effect of networks on (military) operations. Some studies compare the effects of centralization and decentralization [Dekker, 2002, 2005], other studies try to measure the effect of information for the army or navy [Perry et al., 2002, 2004]. Other studies investigate network statistics such as link to node ratio, connectivity and cluster coefficients. These statistics are investigated to find suitable metrics for networked effects described in the table above, for example see [Cares, 2004].

Situation awareness

Underlying the concepts of Network Enabled Capabilities or Network Centric Warfare is the belief that a decision-maker – for example a military unit – can take better decisions if more or better information is presented to him. To be more precise, better information yields better situation awareness of a decision-maker, which in turn enables him to take better decisions. Usually more or better information can be obtained by sharing information with others. The concept of situation awareness is generally understood as 'knowing what is going on', implying the possession of knowledge and understanding to achieve a certain goal. It is the *perception* of the elements in the environment, the *comprehension* of their meaning and the *projection* of their status in the near future [Endsley, 1995]. Nodes represent decision facilities, information fusion centers, combat units and so on. There are several factors that influence this awareness of the situation that is presented to a decision-maker, or more generally, a node in the network. Among these factors is, firstly, the quality of the information available at individual nodes. Quality of information has several aspects, for example: completeness, correctness and currency [Perry et al., 2004]. Combined, they add to the situation awareness of a node. Secondly, as networks provide an opportunity for cooperating entities to share information, situation

awareness of a particular node also depends on its positioning within the network and the network topology. Thirdly, it depends on characteristics of the individual decision-makers themselves, such as experience and training, quality of information fusion facilities, the rate at which information can be processed, the location within the area of operation, the psycho-social environment, organization, prior knowledge, etc.

In the process of transformation towards networked operations, military organisations face high demands regarding their flexibility and coherent integration of sensor, weapon and decision-making capacities. To be successful and to achieve the desired result, it is important to improve coordination of operations through sharing situational information. In this contribution, we focus on the second factor mentioned above, which is an important research area: *the situation awareness of networked elements* [National Research Council, 2005, 2007]. Our results can be used to gain insight into the role of the configuration of the network regarding the improvement and exchange of situation awareness. It also makes possible the comparison of alternative investment into C4I.

Our approach and results
In our model, we distinguish two independent aspects of situation awareness of a node in a network. First of all, we distinguish the exogenously (network independent) given attributes and characteristics of the individual nodes. Examples are its decision-making and information fusion facilities, its training and its positioning within the area of operation. Secondly, we study the importance of a node with respect to the distribution of information. This follows from its local network surroundings and the network topology. We combine these two influences to assess situation awareness.

In the base model, the vector of situation awareness v, is determined using the following mathematical (recurrent) relation:

$$v = \alpha A v + b,$$

where $b > 0$ is a vector of given characteristics regarding situation awareness. The matrix A represents the network structure and contains all the relevant features of the network, its nodes, links and transferability of information. In general, b and A can be determined using techniques from multiple-criteria decision analysis that aggregate various characteristics into a single scalar, see [Monsuur, 2007b]. The parameter α may be interpreted as the relative importance of Av with respect to b. So, the vector v representing situation awareness is the weighted sum of two components: the vector b, (the 'stand-alone' situation awareness) and secondly, the improvement that results form transferred situation awareness, Av, of this final vector v itself.

It is important to realize that the mathematical relation formulated above may also be interpreted as follows: The situation awareness of the set of nodes, as represented by the vector v, is 'confirmed' by the network structure (links and transferability of information) and exogenously given characteristics or private information of each node, as represented by the vector b. We will show that the process of updating situational information using links of the network is equivalent to solving our functional relation between v, A and b.

We also present stochastic variations on this model to include uncertain willingness or possibility of individual nodes to transfer information to adjacent nodes (as may be experienced in practice). We introduce a network performance metric that can be used to compare different network configurations and which takes into account stochastic behaviour of the nodes. It also can be used to compare investments in the (C4I) structure of the network. We illustrate our results by simulation. This shows that:

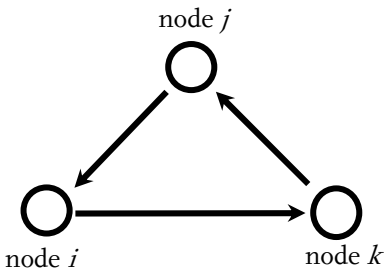
- Updating information using the network topology (information flowing along links) yields new quality characteristics of the nodes (new situation awareness);
- The network performance metric can be used to compare various alternative network configurations.

For other (social) network analysis methods used to analyze military architectures, we refer to [Dekker, 2002, 2005], [Ling et al., 2005], [Perry and Moffat, 2004] or [Cares, 2004]. For example, in [Dekker, 2002], a methodology is introduced that combines social network analysis techniques with military thinking about organizational structure. Metrics are calculated, such as number of links, degree of nodes and distances in networks, to conduct delay analysis, centrality analysis and intelligence analysis.

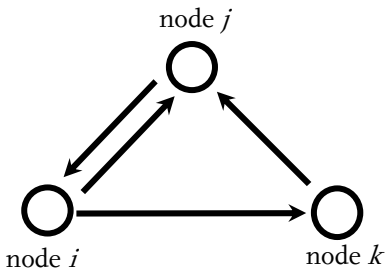
The calculation of situation awareness in a network of (partially) cooperating nodes

As stated previously, cooperating nodes in a network get better situation awareness by sharing information. This process of updating information within the network only depends on given deterministic characteristics of the nodes themselves and the joint network structure. So individual nodes are always in a position to receive information from adjacent nodes or hand over information to others if possible *and* they are always prepared to do so. This assumption states that the nodes behave fully deterministically. However, this will later on be relaxed by allowing uncertain behaviour of the nodes.

A network is a structure made of nodes (i.e. military units) that are tied by links. Each link is assumed to be a *one-way link*, i.e. the information flow along a link will always be in *one direction only*. From the viewpoint of an information receiving node *i*, node *j* is an adjacent node of this node *i* if there is a link from node *j* to node *i*. In return, node *i* is an adjacent node for node *j* if a link exists from node *i* to node *j*, see Fig. 5.



Node *j* is an adjacent node of node *i* but not the other way round. Node *j* can only receive information from node *i* via node *k*.



Node *i* and node *j* are adjacent nodes of each other.

Figure 5. Adjacent nodes

We assume that the nodes are labelled from 1 to n. With each node i we associate a real nonnegative number, b_i, which is called the 'stand-alone' situation awareness of this particular node. In general each b_i can be determined using techniques from multiple-criteria decision analysis which aggregate various characteristics of a node into a single scalar. The vector b contains all the individual values b_i.

If there is a link from node j to node i, we associate a real nonnegative number, a_{ij}, with this link which represents the usability of the information flowing from node j to node i from the point of view of the receiving node i, see Fig. 6. These numbers a_{ij} can also be determined using techniques from multiple-criteria decision analysis. If there is no link from node j to node i, we put $a_{ij} = 0$. The $n \times n$ matrix A with the entries a_{ij} is called the adjacency matrix.

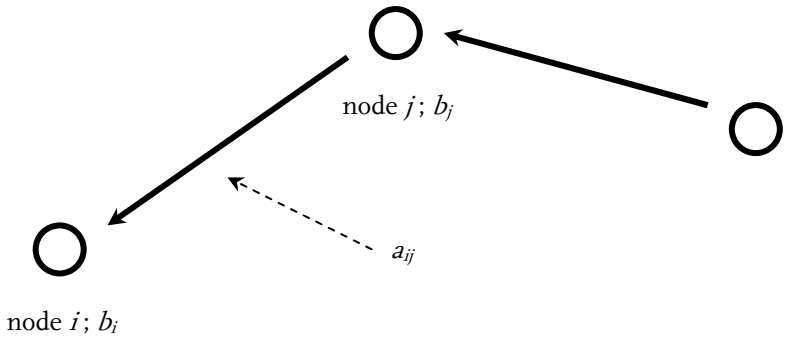


Figure 6. Usability of information

Next we introduce a discount factor α, $0 < \alpha < 1$, which brings in the fact that the usability of information which is flowing along links will decay over time, i.e. information will lose its usability if it is getting older. Before sharing information the situation awareness of the nodes is given by the vector $b = (b_1,...,b_n)^T$. After each node has received information *only* from its adjacent nodes, the new situation awareness of the nodes is given by $b + \alpha Ab$. By iteration information can be updated through the network, so that nodes also receive information from nodes which are not adjacent nodes, but are two, three, or more steps away. Updating information in M-steps yields the situation awareness, v_M, which for $M \geq 1$ is defined recursively as follows:

$$v_0 = b;\ v_1 = b + \alpha A v_0;\ \ \ldots\ldots;\ v_M = b + \alpha A v_{M-1}.$$

Taking the limit of M tending to infinity, we get the following result:

$$v = \lim_{M\to\infty} v_M = \lim_{M\to\infty} \sum_{k=0}^{M} \alpha^k A^k b = (I - \alpha A)^{-1} b.$$

We call v the (definite) situation awareness of the nodes after sharing information. This vector v satisfies the equation

$$v = \alpha A v + b.$$

In this sense we can say that v is 'confirmed' by the network structure and the 'stand-alone' situation awareness of the nodes, b.

Example: Consider the network of Fig. 7, with nodes Joint Strike Fighter, Ground Based Air Defence, etc. Assume that we take

$$b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix} = \begin{pmatrix} 1.00 \\ 0.50 \\ 0.85 \\ 1.00 \\ 0.30 \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 & 0.5 \\ 0.5 & 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0 & 0.5 & 0 \end{pmatrix}, \qquad \alpha = 0.25.$$

Solving the equation $v = \alpha A v + b$ yields the following situation awareness:

$$v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{pmatrix} = \begin{pmatrix} 1.09 \\ 0.67 \\ 1.01 \\ 1.26 \\ 0.59 \end{pmatrix}.$$

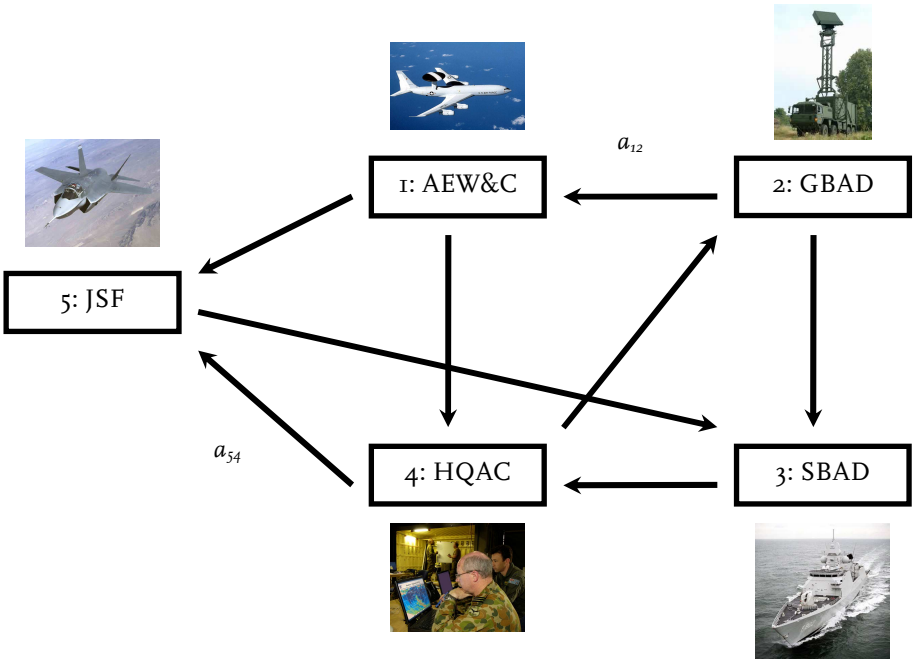


Figure 7. A communication network

So we can conclude that each node has more situation awareness. For example the situation awareness of node 5 was 0.30 and has become 0.59, so it has almost doubled.

Next we introduce a network performance metric which combines given characteristics of the nodes with the Network Topology in order to compare different network

configurations. As stated before, updating information in M-steps yields situation awareness, v_M. For $M \geq 1$ we define the network performance metric NTb_M by

$$NTb_M = \frac{e^T v_M}{e^T b} = \frac{e^T \sum_{k=0}^{M} \alpha^k A^k b}{e^T b} = \sum_{k=0}^{M} \alpha^k \frac{e^T A^k b}{e^T b},$$

where e is a vector of 1's. Taking the limit of M tending to infinity, we get the following result:

$$NTb = \lim_{M \to \infty} NTb_M = \frac{e^T v}{e^T b} = \frac{e^T (I - \alpha A)^{-1} b}{e^T b}.$$

Example (continued):

$$NTb = \frac{\sum_{i=1}^{5} v_i}{\sum_{i=1}^{5} b_i} = \frac{1.09 + 0.67 + 1.01 + 1.26 + 0.59}{1.00 + 0.50 + 0.85 + 1.00 + 0.30} = 1.27$$

We can calculate this number also for other network structures in order to compare the different networks. Notice that instead of the fixed value $\alpha = 0.25$, we can plot NTb as a function of the variable α. This yields Fig. 8.

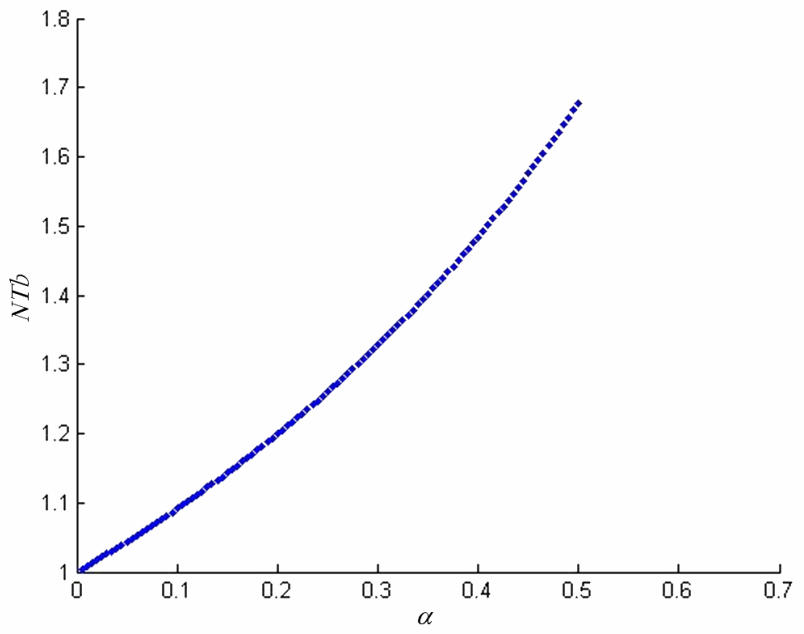


Figure 8. NTb as function of α

Up to now we assumed that individual nodes are always in a position to receive information from adjacent nodes or hand over information to other nodes if possible *and* they are always prepared to do so. However as may be experienced in practice, the process of updating information not only depends on given deterministic characteristics of the nodes itself and the joint network structure. It also depends on the *uncertain* willingness or possibility of individual nodes to receive and transfer information. We will now take into account this uncertainty.

At each stage k, $k \geq 1$, of the process of updating information within the network, the uncertain behaviour of the nodes is modelled by a collection of independent and identically distributed random variables $Y_{k,ij}: \Omega \to \{0,1\}$, $1 \leq i,j \leq n$, such that

$Y_{k,ij} = 1$ if an information flow between node j and node i is possible;
$Y_{k,ij} = 0$ if an information flow between node j and node i is not possible.

For a fixed outcome ω in the sample space Ω the process of updating information in M-steps yields the situation awareness, $V_M(\omega)$, which for $M \geq 1$ is defined as follows:

$$\begin{aligned}
V_0(\omega) &= b \\
V_1(\omega) &= b + \alpha A_1(\omega) V_0(\omega) = b + \alpha A_1(\omega) b \\
&\vdots \\
V_M(\omega) &= b + \alpha A_M(\omega) V_{M-1}(\omega) \\
&= \ldots = b + \alpha A_1(\omega) b + \alpha^2 A_2(\omega) A_1(\omega) b + \ldots + \alpha^M A_M(\omega) \cdots A_1(\omega) b \\
&= b + \sum_{k=1}^{M} \alpha^k \left(\prod_{s=0}^{k-1} A_{k-s}(\omega) \right) b
\end{aligned}$$

where the matrices $A_k(\omega)$ have the entries $a_{ij} Y_{k,ij}$. Note that we obtain the former, deterministic expression for the situation awareness v_M, if all the matrices $A_k(\omega)$ are equal to A. The network performance metric that combines the given characteristics of the nodes with the Network Topology is defined by

$$NTb_M = \frac{E\left(e^T V_M\right)}{e^T b} = 1 + \sum_{k=1}^{M} \alpha^k \frac{E\left(e^T \left(\prod_{s=0}^{k-1} A_{k-s}\right) b\right)}{e^T b},$$

where $E(.)$ denotes the expectation and e is a vector of 1's. Most of all we are interested in the case when M tends to infinity. So let $\{D_1,\ldots,D_N\}$ be the collection consisting of all outcomes of $A_1(\omega)$. Notice that this collection is always finite, because $N \leq 2^n$.

Suppose $\alpha \leq \left(\max_j \left(\sum_{i=1}^{n} a_{ij} \right) \right)^{-1}$. Then the network performance metric that combines the given characteristics of the nodes with the Network Topology is defined by

$$NTb = \frac{\int e^T x \, d\mu(x)}{e^T b}.$$

Here μ is the unique probability measure which satisfies the equation

$$\mu = \sum_{m=1}^{N} P(A_1 = D_m) \, \mu \circ f_m^{-1},$$

where each f_m is the affine mapping $f_m : x \mapsto b + \alpha D_m x$. The existence and uniqueness of this probability measure follows from the fact that $\{f_1,\ldots,f_N; P(A_1 = D_1),\ldots, P(A_1 = D_N)\}$ is an iterated function system with probabilities. The integral $\int e^T x \, d\mu(x)$ in the expression of NTb can be determined by applying Elton's theorem [Elton, 1987], i.e. fix a sequence of

matrices $\{A_k(\omega)\}_{k \geq 1}$ for some outcome ω in the sample space Ω. Let the orbit $\{x_n\}_{n=0}^{\infty}$ be defined by $x_0 = b$ and $x_{n+1} = b + \alpha A_{n+1}(\omega)\, x_n$. Then with probability one

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} e^T x_k = \int e^T x \, d\mu(x) \ .$$

So in order to determine *NTb*, we use the fact that *NTb* equals the expression

$$\lim_{n\to\infty} \frac{1}{n+1} \left(1 + \sum_{k=1}^{n} \frac{e^T x_k}{e^T b} \right) .$$

Example (continued):
Assume that p is the probability that an information flow between node j and node i is possible, independently of the choice of i and j. Then for $p = 1$ we get of course the former (deterministic) result: $NTb = 1.27$. For $p = 0$ the nodes don't share information, so we get: $NTb = 1$. For values of p between 0 and 1, the graph of the function NTb as a function of p is as in Fig. 9.

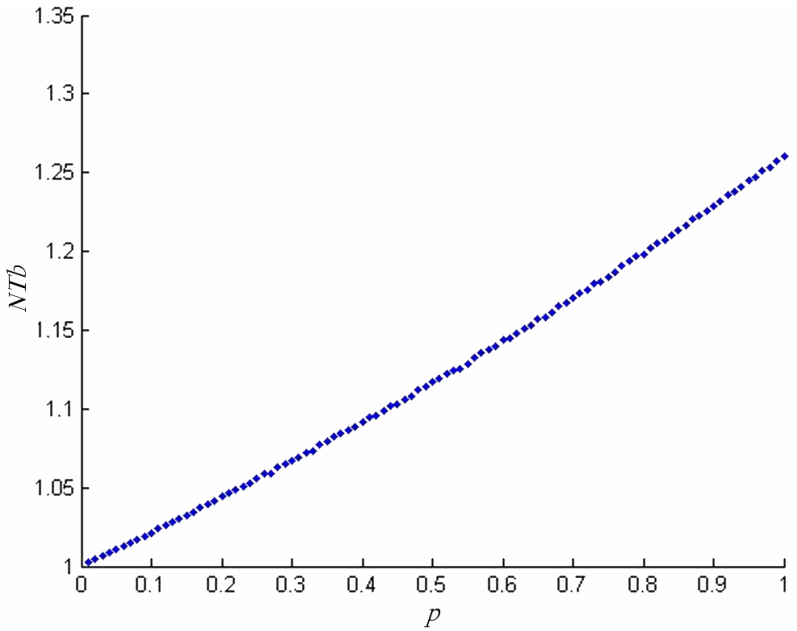


Figure 9. Graph of NTb as function of p

Conclusion

'The achievement of military effect will, in the future, be significantly enhanced through the networking of existing and future military capabilities, under the banner of Network Enabled Capabilities (NEC)' [Ministry of Defence UK, 2005]. NEC has three overlapping and mutually dependent dimensions: Networks, Information and People. At the heart of NEC is a network that is used to distribute information. It enables Defence forces to acquire, generate, distribute, manipulate and utilize information. Information is gathered from a variety of sources, enters the network and is then disseminated through the network to improve situation awareness. It can be exploited, leading to decisions to achieve a desired outcome. Decision-makers must identify what information is required and available to

make the right decision. Therefore, effective information management will grow in importance, especially in a networked environment. People will need to learn how to share and find information from a variety of sources and then use that information to optimise their decisions.

Our results and approach can contribute to a better understanding of the interaction between these three dimensions of NEC. We give two examples.

Network and information
We are able to calculate the impact of the specific configuration of the network on the improvement and exchange of situational information. This enables the comparison of alternative investments into C4I.

Network and people
Another spin-off of our approach and results is the possibility to investigate the interaction between the various types of networks and the various types of agents' behaviour, when they have the task to solve complex problems. For example, given a fixed network topology, and a complex problem that has to be solved, one may investigate the effects of changing the behaviour and decision-making qualities of agents or people. Or, consider a given set of agents, each having some fixed decision-making qualities. Then one may investigate the influence of changing the network topology (for example by deleting the hierarchical structure and moving towards a more flexible network structure) on how these agents (or people) take advantage of new technology. Results of the kind we presented can also be used if one wants to validate new technologies or concepts in a military environment.

Generally speaking, findings of this kind of research can be used to support the modelling of warfare, decisions on force structures, trade-offs among the platforms' weapons and C4ISR systems. Last but not least, it may significantly contribute to changes in doctrine and tactics, techniques and procedures.

References

Barabási A-L. (2003) *Linked*. A Plume Book.

Cares, J. (2004) *An information age combat model*. Alidade Incorporated. Produced for the Director, Net Assessment, Office of the Secretary of Defense (under contract TPD-01-C-0023).

Cares, J. (2006) Battle of the Networks. (The HBR List of Breakthrough Ideas for 2006). *Harvard Business Review*, 40-41.

Clark, T., Moon, T. (2002) Assessing the Military Worth of C4ISR Information. *Proceedings 7-th International Command and Control Research and Technology Symposium*.

Darilek, R., Perry, W., Bracken, J., Gordon, J. and Nichiporuk, B. (2001) *Measures of Effectiveness for the Information-Age Army*. RAND, Santa Monica.

Dekker, A.H. (2002) Applying Social Network Analysis Concepts to Military C4ISR Architectures. *Connections* 24(3): 93-103.

Dekker, A.H. (2005) Network Topology and Military Performance. In: Zerger A. and Argent R.M. (eds.) *Modsim* 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2005, pp 2174-2180.

Dutta, B. and Jackson, MO (Eds.) (2003) *Networks and Groups, Models of Strategic Formation,* Springer.

Elton, J.H. (1987) *An Ergodic Theorem for Iterated Maps,* Ergodic Theory and Dynamical Systems, volume 7, pp 481-488.

Endsley, M.R. (1995) Toward a Theory of Situation Awareness in Dynamic Systems. *Human Factors* 37(1): 32-64.

Fewell, M.P., Hazen, M.G. (2003) *Network-Centric Warfare; Its Nature and Modelling.* Report DSTO-RR-0262.

Forman, E.H., Gass, S.I. (2001) The Analytic Hierarchy Process – An Exposition. *Operations Research* 49: 469-486.

Grant, T.J. (2006) *Measuring the Potential Benefits of NCW: 9/11 as a case study.* Proceedings 11th ICCRTS Conference.

Ling, M.F. (2005) Nonlocality, Nonlinearity and Complexity: on the Mathematics of Modelling NCW and EBO. *Proceedings 22nd International Symposium on Military Operational Research.*

Ling, M.F., Moon, T., Kruzins, E. (2005) Proposed Network Centric Warfare Metrics: from Connectivity to the OODA Cycle. *Military Operations Research Society Journal* 10: 5-13.

Ministry of Defence UK (2005) *Network Enabled Capabilities.* JSP 777.

Monsuur, H. (2007a) Stable and emergent network topologies: a structural approach. *European Journal of Operational Research,* 183 (1), 432-441.

Monsuur, H. (2007b) Assessing Situation Awareness in Networks of Cooperating Entities: A Mathematical Approach, *Military Operations Research*, volume 12 number 3, pp 5-15.

Monsuur, H. (2008) Network induced powerbase: the appreciation of contributing to the value of other nodes. To appear in: *Social Networks, Development, Evaluation and Influence.*

National Research Council (2005) *Network science.* The National Academies Press.

National Research Council (2007) *Strategy for an Army Center for Network Science, Technology, and Experimentation.* The National Academies Press.

Perry, W., Button, R.W., Bracken, J., Sullivan, T., Mitchell, J. (2002) *Measures of effectiveness for the information-age navy. The effects of network-centric operations on combat outcomes.* National Defense Research Institute, RAND.

Perry, W., Moffat, J. (2004) *Information sharing among military headquarters: the effects on decisionmaking.* RAND Monograph 2004.

Perry, W., Signori, D., Boon, J. (2004) *Exploring information superiority: A methodology for measuring the quality of information and its impact on shared awareness.* National Defense Research Institute, RAND.

Watts, D.J. (1999) *Small Worlds, The dynamics of networks between order and randomness,* Princeton Studies in Complexity, Princeton University Press.

Wasserman, S., Faust, K. (1994) *Social network analysis: methods and applications.* Cambridge University Press, Cambridge.

A Further Optimization of Crossover and Linear Barriers in Search Theory

Rien van de Ven

Introduction

Context

Since the start of World War II (WWII), technological advances have dramatically changed tactical and strategic operations. The science of Operations Research (OR) grew out of the need to solve problems related to evaluation and optimal use of these new technological advantages. Solving problems which occur in countering enemy technologies and newly implemented tactics was also important. An initial specialisation in OR was Search Theory. The tactics that were developed to search for the enemy played a very important role in the Allied efforts against German U-boats during WWII, see Fig. 1.

Figure 1. U-744 forced to surface, March the 6th, 1944 by depth charging in North Atlantic. Source: Naval Museum of Québec (http://www.mnq-nmq.org), with permission.

Interesting discussions regarding the search for submarines in the Bay of Biscay may be found in [2] and [3]. For example, it was discovered that aircraft maximised sighting distance by approaching at 45 degrees to U-boat tracks. Most search patterns ran either NW-SE or NE-SW across some assigned coverage area.

Nowadays, applications of Search Theory can be found practically everywhere. For example, the Navy and Air Force search for hostile submarines, Special Forces search for terrorist groups, Unmanned Aerial Vehicles search for nuclear plants or launching facilities of opposing forces. Of course, there are also many non-military applications, such as the search and rescue of drowning persons and counter drug operations.

Using mathematical and probabilistic models, Search Theory has developed several interesting search patterns that optimise the probability of detecting a target. For an overview of classical search theory, we refer to [1].

Many papers nowadays focus on multi-agent systems and simulations. For applications, we refer to [2] and [3]. In this contribution we follow an analytical approach.

The reason for doing so, is that we are able to demonstrate analytically that further optimization is possible by slightly modifying the well-known *crossover barrier* (Fig. 2) and *linear barrier* (Fig. 3), where the search area is a *lane* (i.e. a Southwards going channel), see [4].

We assume that targets intend to traverse this lane Southwards. We also assume that target speed U is constant along its path. This assumption is not far from reality because, after reaching cruise level, the target usually maintains a steady speed. A further assumption is that we know the intent and capabilities of the target. More precisely, we assume that we know its speed. Its position, however, is unknown, so arbitrary. Examples are rescuing a person floating in the water, detecting fast drug boats, etc.

We assume the observer is protecting the lane while moving at speed V through the lane according to some fixed pattern. Any target that closes the observer to within his *sweep radius* R is detected. So the observer's detecting device is binary: the target is either detected or not detected. A further assumption is that there is enough time for the observer to detect targets.

The *crossover barrier* starts on the left of the lane and crosses to the opposite side (track OA) in such a way that its Southwards movement equals that of a hypothetical target which simultaneously moves from B to A. Next the observer moves Northwards (track AB), crosses the lane to the left (track BC), and finally moves Northwards to the starting-point (track CO). In this way a *butterfly* search pattern is created. After completing one basic movement the pattern is repeated several times. The *crossover* model is chosen, when the speed of the observer is greater than that of the target.

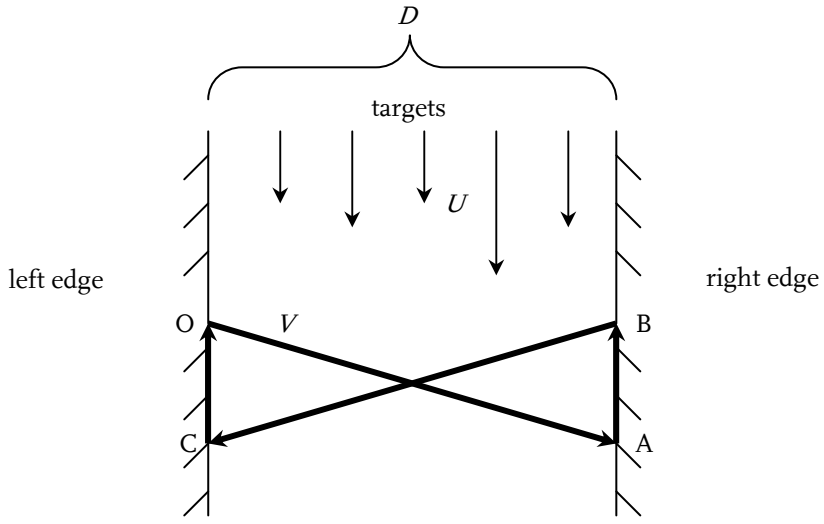


Figure 2. Crossover barrier (changing course at the edge)

The *linear barrier* moves in a straight line from West to East, i.e. its course is perpendicular to that of the targets. If the observer reaches the edge of the lane, it will reverse course. In this way a *linear* search pattern is created. After completing one basic movement the pattern is repeated several times. The *linear* model is chosen, when the speed of the observer is less than or equal to that of the target.

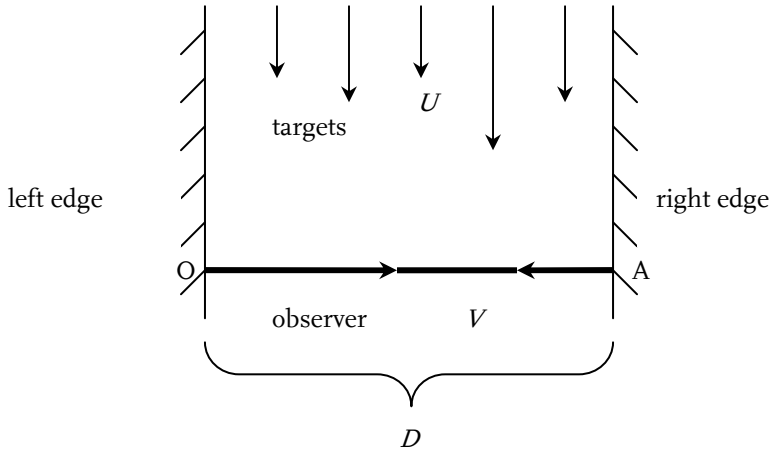


Figure 3. Linear barrier (reversing course at the edge)

Problem definition
It is well-known that instead of changing course exactly at the edge, changing course when the sweep radius reaches the edge (i.e. changing course at distance R from the edge) generally yields higher probabilities of detection.

In this contribution we shall investigate whether or not an even higher probability of detection may be obtained, by turning at distance xR from the edge. The *turning-factor* x is some real number between 0 and 1. So, if $x = 0$, the *barrier* changes course at the edge. If $x = 1$, the *barrier* changes course when the sweep radius reaches the edge.

We will discuss two questions:

1. If the barrier changes course at distance R from the edge, does this situation – compared with changing course at the edge – always lead to a higher probability of detection?
2. If the barrier changes course at the edge or at distance R from the edge, does one of these two situations lead to a maximum probability of detection?

The construction of this contribution will be as follows: first in two different sections, we present the *crossover barrier* as well as the *linear barrier*. In both models we will go into three scenarios:

- the barrier changes course at the edge;
- the barrier changes course when the sweep radius reaches the edge;
- the barrier changes course at distance xR from the edge.

In the section *Results and Discussion*, an overview of the results will be presented for both models. A discussion on the choice between the two models is also carried on. Finally in

the section *Conclusions*, the answers to the questions as formulated above will be summarized.

We will calculate all probabilities of detection choosing a *lane width* D with magnitude 24 nautical miles (NM), while *sweep radius* R equals 4 NM. Hence a comparison between the two models and the three scenarios can be made. We will vary the ratio of observer's speed V and target's speed U between just greater than one and four (*crossover* model), respectively half and four (*linear* model). This quotient V/U is called the *speed ratio* and is denoted by ρ. Hence $1 < \rho \leq 4$ (*crossover* model), respectively $0.5 \leq \rho \leq 4$ (*linear* model).

Crossover barrier patrol

Changing course at the edge

We assume $V > U$ (i.e. $\rho > 1$). If $V < U$ or $V \approx U$, the *crossover* model is not practicable, because the angle – in relation to the horizontal axis – chosen by the *crossover barrier* is not defined or close to $\frac{1}{2}\pi$. The latter is not desirable because it will cost the *barrier* too much time to reach the opposite side of the lane. The lane is D wide, while the sweep radius of the observer equals R. We assume $D > 2R$, because if $D \leq 2R$ the observer could restrict himself to a position in the middle of the lane.

Let t_1 be the time it takes for the observer to reach the opposite side of the lane. We can use Pythagoras' Theorem (applied in a right-angled triangle with hypotenuse Vt_1 and catheti D and Ut_1) to determine $t_1 = \dfrac{D}{\sqrt{V^2 - U^2}}$.

So, observer and target move according to $\begin{pmatrix} D \\ -Ut_1 \end{pmatrix}$, respectively $\begin{pmatrix} 0 \\ -Ut_1 \end{pmatrix}$ from their initial position.

To determine the probability of detection, we keep the position of the target fixed with only the observer and its detection circle moving *relatively* to the target. The relative movement of the first leg (track OA) is obtained by calculating the difference of the observer's absolute movement and that of the target:

$$\begin{pmatrix} D \\ -Ut_1 \end{pmatrix} - \begin{pmatrix} 0 \\ -Ut_1 \end{pmatrix} = \begin{pmatrix} D \\ 0 \end{pmatrix}.$$

Let t_2 be the time it takes for the observer to proceed up the channel. The length of this *upsweep* is equal to Ut_1. Hence:

$$Vt_2 = Ut_1 = \frac{D}{\sqrt{\left(\frac{V}{U}\right)^2 - 1}}.$$

So, it follows:

$$t_2 = \frac{D}{V\sqrt{\rho^2 - 1}}.$$

Hence the relative movement of the second leg (track AB) satisfies:

$$\begin{pmatrix} 0 \\ Vt_2 \end{pmatrix} - \begin{pmatrix} 0 \\ -Ut_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{D}{\sqrt{\rho^2-1}} + \frac{D}{\rho\sqrt{\rho^2-1}} \end{pmatrix} = \begin{pmatrix} 0 \\ \frac{D}{\rho}\sqrt{\frac{\rho+1}{\rho-1}} \end{pmatrix}.$$

The expression:

$$S = \frac{D}{\rho}\sqrt{\frac{\rho+1}{\rho-1}} \tag{1}$$

is called the *spacing* and is denoted by S.
Hence the relative movement of the observer in relation to the target is given by $\begin{pmatrix} D \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ S \end{pmatrix}, \begin{pmatrix} -D \\ 0 \end{pmatrix}$ and $\begin{pmatrix} 0 \\ S \end{pmatrix}$. In this way a *meander* search pattern is created, see Fig. 4, where a strip with width $2R$ around the relative track will be swept.

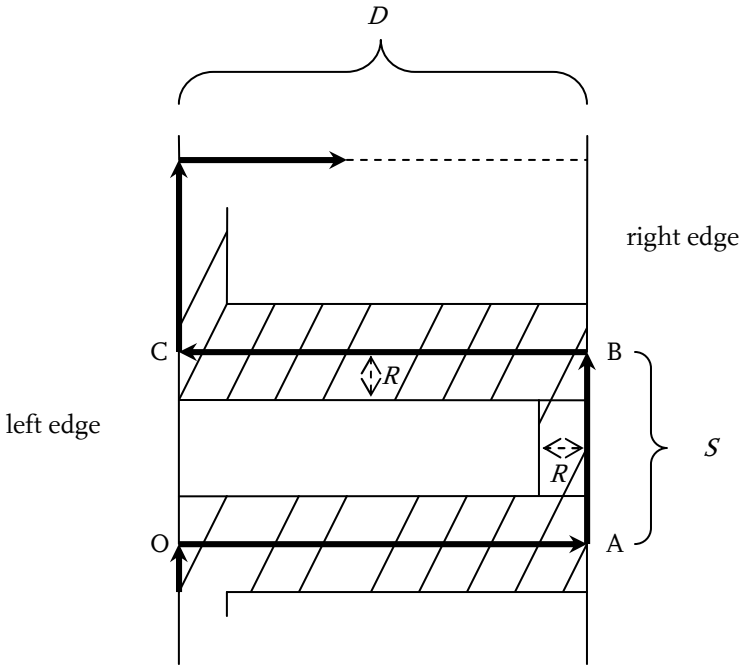


Figure 4. The relative area swept for the crossover barrier changing course at the edge ($R \leq \frac{1}{2}S$)

Since the target has a fixed and random position, the probability of detection is determined by taking the ratio of the shaded area and the total area. We can – in view of the regularity in the pattern of the movement – restrict ourselves to one distinctive part of the relative track, i.e. the track OA → AB → BC in the rectangle OABC. The dimensions of this rectangle are *lane width* D and *spacing* S.

If $R \geq \frac{1}{2}S$, the total area will be swept. So the probability of detection P_{det} will be equal to 1.
If $R \leq \frac{1}{2}S$, then:

$$P_{\text{det}} = \frac{2DR + (S - 2R)R}{DS}. \tag{2}$$

Changing course when the sweep radius reaches the edge

If the observer changes course when the sweep radius reaches the edge, the geometry will be more complex, but the general idea is the same. In a distinctive part of the relative movement the ratio of the swept area and the total area will be calculated. As Fig. 5 shows, the swept area consists of rectangles and sectors of a circle.

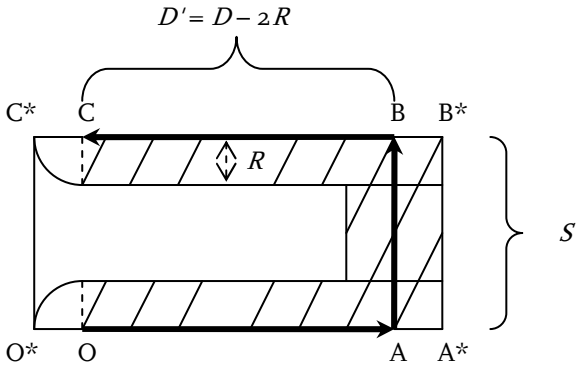


Figure 5. The relative area swept for the crossover barrier changing course when the sweep radius reaches the edge ($R \leq S/2$)

The area of the two sectors of a circle equals $\frac{1}{2}\pi R^2$. There are also two rectangles with dimensions $D - R$ and R and a rectangle with dimensions $S - 2R$ and $2R$.

Because OA equals $D - 2R$, *spacing* S is calculated on the basis of $D' = D - 2R$. Hence:

$$S = \frac{D'}{\rho}\sqrt{\frac{\rho + 1}{\rho - 1}}. \tag{3}$$

If $R \leq \frac{1}{2}S$, then:

$$P_{\text{det}} = \frac{2R(D-R) + 2R(S-2R) + \frac{1}{2}\pi R^2}{DS}. \tag{4}$$

If $R \geq \frac{1}{2}S$, the rectangles overlap, as do the sectors, see Fig. 6.

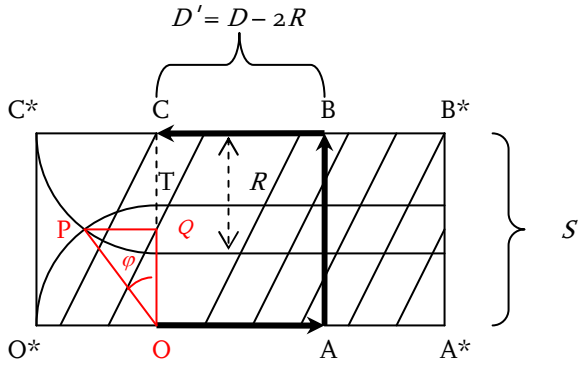


Figure 6. The relative area swept for the crossover barrier changing course when the sweep radius reaches the edge ($R \geq \frac{1}{2}S$)

The area of the overlap of the sectors of the circle is obtained by first calculating sector OPT, then by calculating triangle OPQ. The overlap is twice the difference of these two results, i.e. the difference of twice sector OPT and twice triangle OPQ.

Using $\varphi = \arccos(\frac{S}{2R})$ we obtain the following probability of detection:

$$P_{\text{det}} = \begin{cases} \dfrac{2R(D-R)+2R(S-2R)+\frac{1}{2}\pi R^2}{DS} & \text{, if } R \le \frac{S}{2} \\ \dfrac{(D-R)S+\frac{1}{2}\pi R^2 - R^2\arccos(\frac{S}{2R})+\frac{S}{2}\sqrt{R^2-\frac{S^2}{4}}}{DS} & \text{, if } R \ge \frac{S}{2} \end{cases} \tag{5}$$

Changing course at distance xR from the edge

If the observer changes course at distance xR ($0 \le x \le 1$) from the edge, geometry gets even more complex, but the general idea is the same. The ratio of the swept area and the total area will be calculated in a distinctive part of the relative movement.

Again we distinguish cases $R \le \frac{1}{2}S$ and $R \ge \frac{1}{2}S$.

If $R \le \frac{1}{2}S$ the swept part consists of rectangles and truncated sectors of a circle, see Fig. 7.

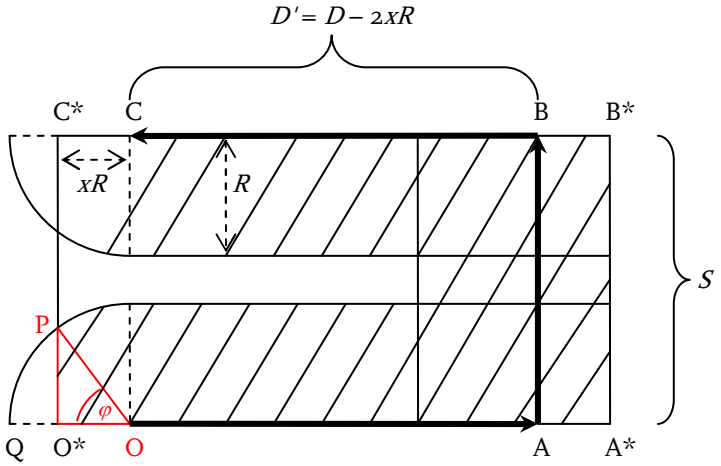


Figure 7. The relative area swept for the crossover barrier changing course at distance xR from the edge ($R \le \frac{1}{2}S$)

If $R \ge \frac{1}{2}S$, the rectangles overlap, as do the truncated sectors, see Fig. 8. If PQ is longer than xR, the overlap is so much that the swept area is equal to OABC. Hence the detection probability equals 1. This occurs when $xR < \text{PQ} = \sqrt{R^2 - \frac{S^2}{4}}$, i.e. $x < \sqrt{1-\left(\frac{S}{2R}\right)^2}$.

If $x \ge \sqrt{1-\left(\frac{S}{2R}\right)^2}$, we refer to Fig. 8.

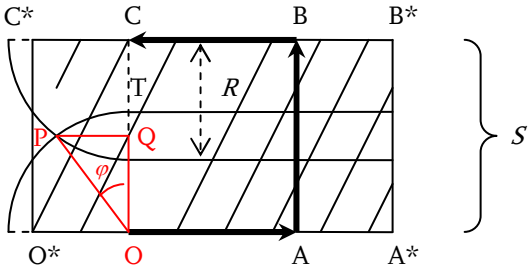


Figure 8. The relative area swept for the crossover barrier changing course at distance xR from the edge ($R \ge \frac{1}{2}S$)

Calculations similar to (5) lead to the following probability of detection:

$$P_{\text{det}} = \begin{cases} \dfrac{2(D-xR)R+(S-2R)(R+xR)+\frac{1}{2}\pi R^2 - R^2\arccos x + R^2 x\sqrt{1-x^2}}{DS} & \text{, if } R \le \frac{S}{2} \\ \dfrac{(D-xR)S + R^2(\frac{1}{2}\pi - \arccos x + x\sqrt{1-x^2} - \arccos(\frac{S}{2R})) + \frac{S}{2}\sqrt{R^2-\frac{S^2}{4}}}{DS} & \text{, if } x \ge \sqrt{1-\left(\frac{S}{2R}\right)^2} \end{cases} \quad (6)$$

Substituting $x = 0$, respectively $x = 1$ in (6) it is easy to check that the probability of detection satisfies (2), respectively (4) and (5).

Linear barrier patrol

Reversing course at the edge

As mentioned before: a *linear barrier* is preferred if $V < U$ or $V \approx U$. The *linear barrier* moves from West to East, respectively from East to West. Targets always move Southwards. If t is the time it takes for the observer to reach the opposite side of the lane, then the relative movement of the track satisfies:

$$\begin{pmatrix} Vt \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ -Ut \end{pmatrix} = \begin{pmatrix} Vt \\ Ut \end{pmatrix}, \text{ respectively } \begin{pmatrix} -Vt \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ -Ut \end{pmatrix} = \begin{pmatrix} -Vt \\ Ut \end{pmatrix}.$$

In this way a *ladder* search pattern is created as a relative track. A strip with width $2R$ around this track will be swept.

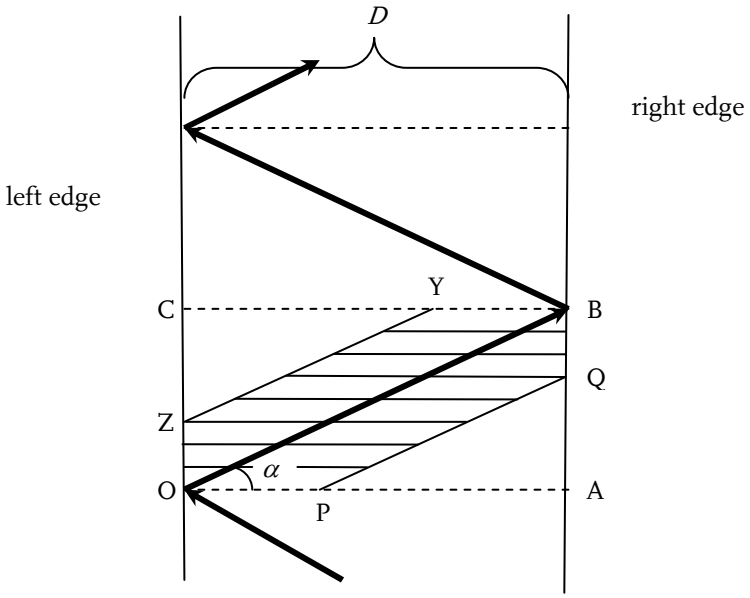


Figure 9. The relative swept area for the linear barrier changing course at the edge ($R/\sin\alpha \le D$)

Since the target has a fixed and random position, the probability of detection is determined by taking the ratio of the shaded area and the total area. We can – in view of the regularity in the pattern – restrict ourselves to one distinctive part of the relative track, i.e. the track OB in rectangle OABC. See Fig. 9[1].

[1] We assume P to be to the left of A, i.e. OA > OP. Hence $D > \frac{R}{\sin\alpha} = R\sqrt{1+\rho^2}$. If $D \le R\sqrt{1+\rho^2}$, the swept area coincides with OABC. So, the probability of detection equals 1. If $D = 24$ and $R = 4$, this will only happen if $\rho \ge 6.0$. Hence we can disregard this situation.

Using $\sin\alpha = \dfrac{U}{\sqrt{U^2+V^2}} = \dfrac{1}{\sqrt{1+\rho^2}}$ we obtain:

$$P_{\text{det}} = \frac{2R}{D}\sqrt{1+\rho^2} - \frac{R^2}{D^2}\left(1+\rho^2\right). \tag{7}$$

Reversing course when the sweep radius reaches the edge

If the observer reverses course when the sweep radius reaches the edge, the geometry will be more complex, but the general idea is the same. In a distinctive part of the relative movement the ratio of the swept area and the total area will be calculated. As Fig. 10 shows, the swept area consists of a hexagon and two sectors of a circle[1].

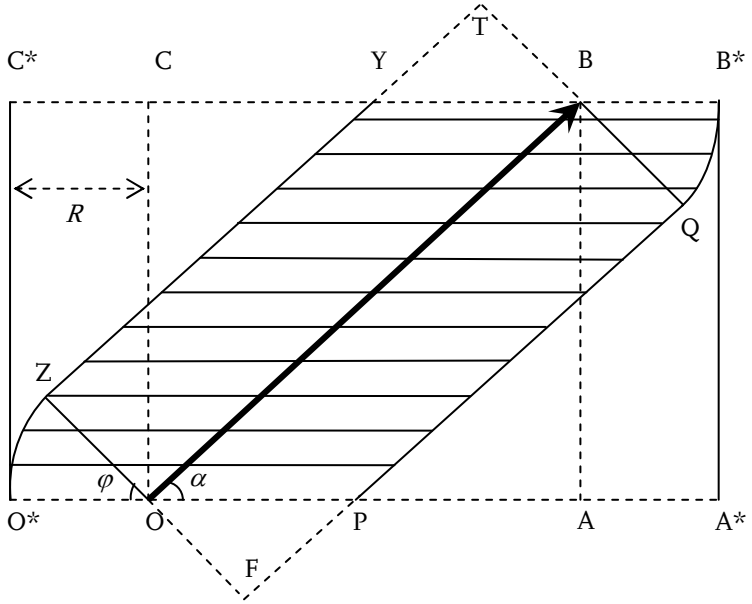


Figure 10. The relative area swept for the linear barrier reversing course when the sweep radius reaches the edge

Calculating the swept area (i.e. hexagon OPQBYZ and sectors O*OZ and B*BQ) and the whole area (i.e. rectangle O*A*B*C*) and using $\alpha = \operatorname{arccot}\rho$ gives:

$$P_{\text{det}} = \frac{2R}{D}\sqrt{1+\rho^2} + \frac{R^2}{D(D-2R)}\left\{\left(\tfrac{1}{2}\pi - \operatorname{arccot}\rho\right)\rho - \rho^2\right\}. \tag{8}$$

Reversing course at distance xR from the edge

If the observer reverses course at distance xR ($0 \le x \le 1$) from the edge, the geometry gets even more complex, but the general idea is the same.

Two cases arise:

Case A: Z inside the lane (see Fig. 11)

[1] We assume Q to be above AA*. If $\rho \le 4$, $D = 24$ and $R = 4$ (so AB ≥ 4), this assumption will be satisfied.

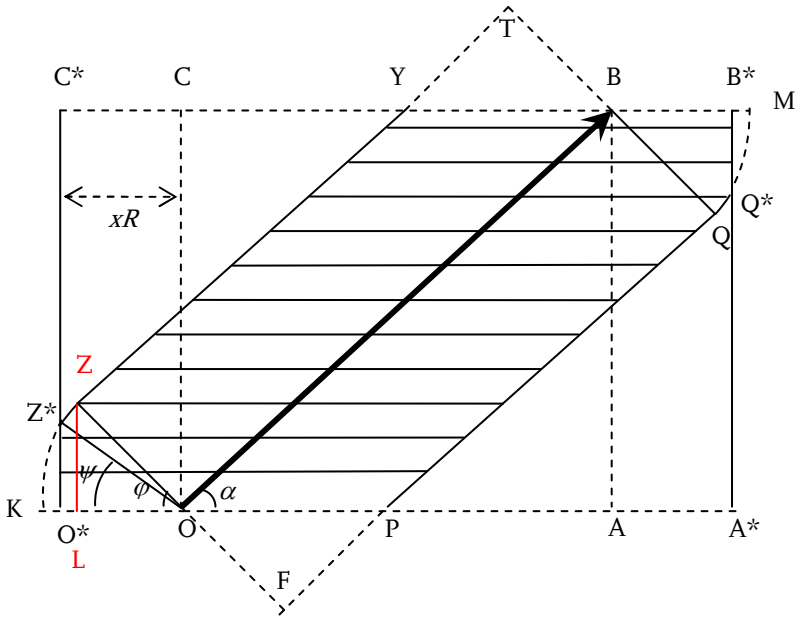


Figure 11. The swept part for the linear barrier reversing course at distance xR from the edge ($x \geq \frac{1}{\sqrt{1+\rho^2}}$)

This occurs when $O^*O = xR \geq LO = R\cos\varphi = R\cos(\frac{1}{2}\pi - \alpha) = R\sin\alpha = \frac{R}{\sqrt{1+\rho^2}}$, i.e. $x \geq \frac{1}{\sqrt{1+\rho^2}}$. In this case, the swept area consists of hexagon OPQBYZ and truncated sectors OO*Z*Z and BB*Q*Q.

Case B: Z outside the lane (see Fig. 12)

This occurs when $x < \frac{1}{\sqrt{1+\rho^2}}$.

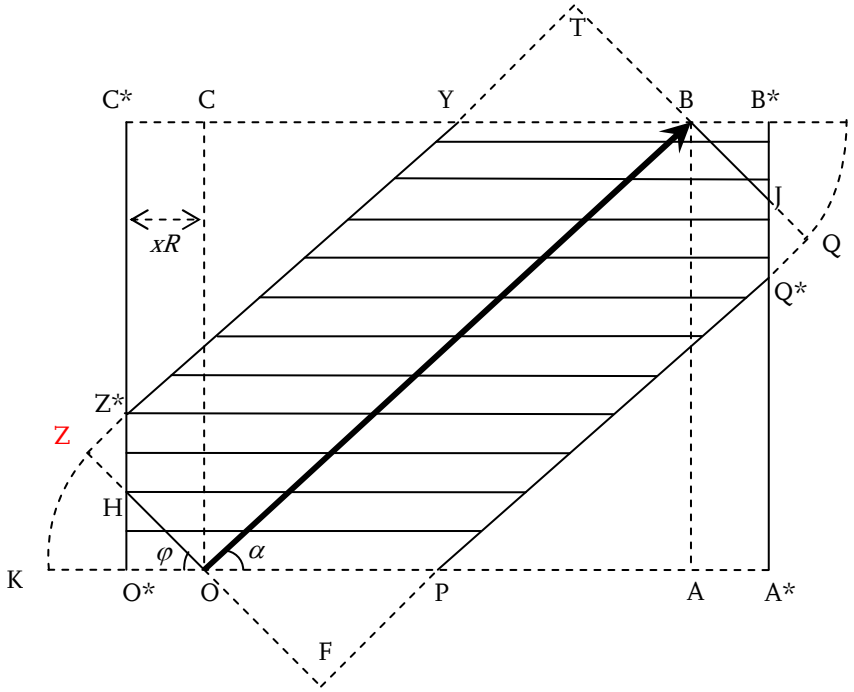


Figure 12. The swept part for the linear barrier reversing course at distance xR from the edge ($x < \frac{1}{\sqrt{1+\rho^2}}$)

In this case, the swept area is obtained by first calculating rectangle ZFQT, then subtracting triangles FOP, BTY, HZZ* and JQQ*, and then adding triangles HOO* and BB*J.

We can prove that the probability of detection satisfies:

$$P_{\text{det}} = \begin{cases} \dfrac{2R}{D}\sqrt{1+\rho^2} + R^2\left\{\dfrac{\left(\frac{1}{2}\pi - \operatorname{arccot}\rho - \arccos x + x\sqrt{1-x^2}\right)\rho - \rho^2}{D(D-2xR)}\right\}, & \text{if } x \geq \frac{1}{\sqrt{1+\rho^2}} \\ \dfrac{2R}{D}\sqrt{1+\rho^2} + \dfrac{2xR^2\sqrt{1+\rho^2} - R^2\rho^2 - (x^2+1)R^2}{D(D-2xR)}, & \text{if } x < \frac{1}{\sqrt{1+\rho^2}} \end{cases} \quad (9)$$

Results and discussion

We have chosen $D = 24$, $R = 4$ and $1 < \rho \leq 4$ (*crossover*), respectively $0.5 \leq \rho \leq 4$ (*linear*). The probability of detection is calculated for several values of the *speed ratio* ρ and the *turning-factor* x.

Crossover barrier patrol

In Table 1, we also mention the magnitude of the *spacing* S_o (corresponding to $x = 0$), respectively S_1 (corresponding to $x = 1$).

Table 1. Probability of detection as a function of speed ratio ρ and turning-factor x for the crossover barrier

	ρ	1.5	2.0	2.5	3.0	3.5	4.0
x	S_o	35.8	20.8	14.7	11.3	9.2	7.7
0.00		0.353	0.487	0.621	0.756	0.891	1.000
0.05		0.363	0.498	0.633	0.768	0.905	1.000
0.10		0.372	0.508	0.644	0.781	0.918	1.000
0.15		0.382	0.519	0.656	0.793	0.931	1.000
0.20		0.392	0.530	0.667	0.806	0.945	1.000
0.25		0.401	0.541	0.679	0.819	0.959	1.000
0.30		0.411	0.551	0.691	0.832	0.973	1.000
0.35		0.421	0.562	0.703	0.844	0.987	1.000
0.40		0.431	0.573	0.715	0.857	0.998	1.000
0.45		0.440	0.584	0.727	0.870	0.999	1.000
0.50		0.450	0.595	0.739	0.883	0.998	1.000
0.55		0.460	0.606	0.751	0.896	0.998	1.000
0.60		0.470	0.616	0.762	0.909	0.998	1.000
0.65		0.480	0.627	0.774	0.922	0.997	1.000
0.70		0.489	0.638	0.786	0.935	0.997	1.000
0.75		0.499	0.649	0.797	0.947	0.996	1.000
0.80		0.509	0.659	0.809	0.959	0.994	0.999
0.85		0.518	0.669	0.820	0.971	0.993	0.998
0.90		0.527	0.679	0.831	0.977	0.990	0.996
0.95		0.536	0.689	0.841	0.974	0.987	0.993
1.00		0.545	0.698	0.850	0.970	0.982	0.988
	S_1	23.9	13.9	9.8	7.5	6.1	5.2

As *speed ratio* ρ increases, *spacing* S_x decreases. The swept area of the rectangle will assume growing importance in relation to the total area, see Figs. 4 – 8. Hence the probability of detection will increase as *speed ratio* ρ increases. This result is true for all values of $x \in [0.00; 1.00]$.

Comparing changing course at the edge and changing course when the sweep radius reaches the edge leads to the conclusion that – at a given ρ – *spacing* S will be smaller. Changing course at distance R from the edge thus leads to a higher probability of

detection. $\rho = 4$ is an exception: when changing course at the edge *spacing* S will be less than $2R$. So the probability of detection equals 1. Hence the probability of detection – in the case of changing course at distance R from the edge – will turn out to be smaller[1]. If $\rho \leq 2.5$, the probability of detection is maximal, if the observer changes course at distance R from the edge. If $\rho = 3$ or $\rho = 3.5$, then changing course at distance R from the edge gives a higher probability of detection when compared with changing course at the edge, but the probability of detection is maximal at a turning-distance smaller than R. As ρ increases, the optimal turning-distance decreases. If $\rho = 4$, the probability of detection is maximal, if the observer changes course at the edge[2].

Linear barrier patrol

As *speed ratio* ρ increases, the probability of detection increases too (see Table 2). This is illustrated in Fig. 9: as ρ increases, gradient α decreases, as does AB. The swept area will assume growing importance in relation to the total area. This holds for every $x \in [0.00; 1.00]$.

Table 2. Probability of detection as a function of speed ratio ρ and turning-factor x for the linear barrier

x \ ρ	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
0.00	0.338	0.416	0.511	0.606	0.696	0.776	0.845	0.902
0.05	0.340	0.419	0.514	0.610	0.700	0.780	0.849	0.906
0.10	0.343	0.422	0.518	0.614	0.704	0.785	0.853	0.909
0.15	0.345	0.425	0.521	0.618	0.709	0.789	0.857	0.913
0.20	0.348	0.428	0.524	0.622	0.713	0.793	0.861	0.916
0.25	0.350	0.430	0.528	0.626	0.717	0.797	0.865	0.920
0.30	0.352	0.433	0.531	0.630	0.721	0.801	0.869	0.923
0.35	0.354	0.436	0.535	0.633	0.725	0.805	0.873	0.927
0.40	0.356	0.438	0.538	0.637	0.729	0.810	0.877	0.930
0.45	0.358	0.441	0.541	0.641	0.733	0.814	0.881	0.933
0.50	0.360	0.444	0.544	0.645	0.737	0.818	0.884	0.936
0.55	0.362	0.446	0.548	0.649	0.741	0.821	0.888	0.939
0.60	0.363	0.448	0.551	0.652	0.745	0.825	0.891	0.941
0.65	0.365	0.451	0.554	0.656	0.749	0.829	0.894	0.943
0.70	0.366	0.453	0.557	0.659	0.752	0.832	0.896	0.944
0.75	0.368	0.455	0.560	0.662	0.755	0.834	0.898	0.945
0.80	0.369	0.457	0.562	0.665	0.758	0.837	0.899	0.944
0.85	0.370	0.459	0.565	0.668	0.760	0.838	0.899	0.943
0.90	0.371	0.461	0.567	0.670	0.762	0.839	0.899	0.941
0.95	0.372	0.462	0.568	0.671	0.762	0.838	0.896	0.936
1.00	0.372	0.462	0.569	0.671	0.761	0.835	0.891	0.929

Comparing both situations – reversing course at the edge and reversing course at distance R from the edge – leads to the conclusion that – at a given ρ – the latter has a higher probability of detection. If $\rho \leq 1.50$, the probability of detection is maximal, if the observer

[1] If $\rho = 4$, then – if changing course at the edge – S equals 7.7, i.e. less than $2R$. So, the probability of detection equals 1. *Spacing* S is a function of D' (see Eq. (3)): as D' increases, S increases too. Hence a probability of detection with magnitude 1 will be obtained at a higher value of ρ. Therefore the mentioned exception is only valid when $D = 24$ and $R = 4$. If $D = 20$, the exception is true if ρ is greater than, or equal to 3.5.

[2] If $\rho = 4$ the maximum probability of detection equals 1. The mentioned value of x is only one possible solution. Every choice of $x \in [0.00; 0.65]$ leads to a maximum detection probability.

changes course at distance R from the edge. If $\rho \geq 2.0$, then changing course at distance R from the edge gives a higher probability of detection when compared with changing course at the edge, but the probability of detection is maximal at a turning-distance smaller than R. As ρ increases, the optimal turning-distance decreases.

Choosing between crossover and linear

As mentioned before: the choice between the *crossover* model and the *linear* model depends on the magnitude of the observer's speed in relation to that of the target. If $V < U$, a *linear* model has to be chosen. If $V \approx U$, a *linear* model is preferred, since – in case of a *crossover barrier* – the observer's absolute course would be almost Southwards: if $V = 10.5$ and $U = 10$, then observer's (absolute) course equals 162 degrees. The latter is not desirable, because it will cost the *barrier* too much time to reach the other side of the lane. If $V >> U$, a *crossover* model is preferred. Crossing will be almost West/Eastwards: if $V = 30$ and $U = 10$, then observer's (absolute) course equals 110 degrees.

The decision between *crossover* and *linear* is not only influenced by *speed ratio*, but also by $D' = D - 2xR$. In this contribution – we have chosen $D = 24$ and $R = 4$ – only the *turning-factor* x varies ($0 \leq x \leq 1$). Hence $16 \leq D' \leq 24$.

Linking both models in such a way that the model with maximum detection probability is chosen, we can give a clear idea of the dependence on *speed ratio* ρ and *turning-factor* x (see Table 3). Detection probabilities using the *linear* model are represented in green, those using the *crossover* model in blue. The maximum detection probability is represented in red.

Table 3. Probability of detection as a function of speed ratio ρ and turning-factor x for the linear barrier, respectively the crossover barrier

x \ ρ	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
0.00	0.338	0.416	0.511	0.606	0.696	0.776	0.891	1.000
0.05	0.340	0.419	0.514	0.610	0.700	0.780	0.905	1.000
0.10	0.343	0.422	0.518	0.614	0.704	0.785	0.918	1.000
0.15	0.345	0.425	0.521	0.618	0.709	0.793	0.931	1.000
0.20	0.348	0.428	0.524	0.622	0.713	0.806	0.945	1.000
0.25	0.350	0.430	0.528	0.626	0.717	0.819	0.959	1.000
0.30	0.352	0.433	0.531	0.630	0.721	0.832	0.973	1.000
0.35	0.354	0.436	0.535	0.633	0.725	0.844	0.987	1.000
0.40	0.356	0.438	0.538	0.637	0.729	0.857	0.998	1.000
0.45	0.358	0.441	0.541	0.641	0.733	0.870	0.999	1.000
0.50	0.360	0.444	0.544	0.645	0.739	0.883	0.998	1.000
0.55	0.362	0.446	0.548	0.649	0.751	0.896	0.998	1.000
0.60	0.363	0.448	0.551	0.652	0.762	0.909	0.998	1.000
0.65	0.365	0.451	0.554	0.656	0.774	0.922	0.997	1.000
0.70	0.366	0.453	0.557	0.659	0.786	0.935	0.997	1.000
0.75	0.368	0.455	0.560	0.662	0.797	0.947	0.996	1.000
0.80	0.369	0.457	0.562	0.665	0.809	0.959	0.994	0.999
0.85	0.370	0.459	0.565	0.669	0.820	0.971	0.993	0.998
0.90	0.371	0.461	0.567	0.679	0.831	0.977	0.990	0.996
0.95	0.372	0.462	0.568	0.689	0.841	0.974	0.987	0.993
1.00	0.372	0.462	0.569	0.698	0.850	0.970	0.982	0.988

If $\rho \leq 1.5$, the *linear* model is preferred, if $\rho \geq 3.5$ the *crossover* model is preferred. If $2.0 \leq \rho \leq 3.00$, the situation is mixed: if the *barrier* is changing course near the edge

($x \approx 0$), the *linear* model is preferred; if the *barrier* is changing course near distance R from the edge ($x \approx 1$), the *crossover* model is preferred.

Conclusions hold in this particular situation ($D = 24$, $R = 4$), but it is possible to demonstrate that similar conclusions can be drawn for other values of *lane width* and *sweep radius*.

Conclusions

In the *Introduction* to this contribution two questions were formulated:

1. If the *barrier* changes course at distance R from the edge, does this situation – compared with changing course at the edge – always lead to a higher probability of detection?
2. If the *barrier* changes course at the edge or at distance R from the edge, does one of these two situations lead to a maximum probability of detection?

We have investigated two models: the *crossover* model and the *linear* model. In both models we made a distinction between changing course at the edge, changing course when the sweep radius reaches the edge and changing course at an alternating distance.

If $D = 24$ and $R = 4$, the following conclusions can be drawn:

1. If $\rho \leq 3.5$, changing course when sweep radius reaches the edge will – compared with changing course at the edge – lead to a higher probability of detection. If $\rho \geq 4.0$, changing course at the edge will give a better result.
2. If $\rho \leq 2.5$, changing course when sweep radius reaches the edge will lead to a maximum probability of detection. If $\rho \geq 4.0$, changing course at the edge will give the best result. If $\rho = 3.0$ or $\rho = 3.5$, maximum probability of detection will be obtained when the *barrier* changes course at distance from the edge less than R. As *speed ratio* increases, turning-distance from the edge will decrease.

Following an analytical approach, we were able to demonstrate that – obtaining the probability of detection – further optimization is possible. More precisely, the probability of detection increases by a few percent under certain circumstances, if we choose the turning-distance of the *barrier* carefully.

References

[1] Benkoski, S.J., Monticino, M.G. and Weisinger, J.R. (1991) *A Survey of the Search Theory Literature*. Naval Research Logistics, 38: 469-494.

[2] Carl, R.G. (2007) *Search Theory and U-Boats in the Bay of Biscay*. Air Force Institute of Technology, Wright-Patterson Air Force Base.

[3] Hill, R.R., Carl, R.G. and Champagne, L.E. (2006) *Using Agent-Based Simulation to Empirically Examine Search Theory Using a Historical Case Study*. Journal of Simulation, 1: 29-38.

[4] Operations Analysis Study Group (OASG), United States Naval Academy (1977) *Naval Operations Analysis*. Naval Institute Press, Maryland.

Mission-Driven Sensor Management

Fok Bolderheij

Introduction

Managing the sensor systems onboard modern naval vessels demands an increasing amount of operator knowledge due to the fact that these vessels are equipped with state-of-the-art sensor systems that provide more functionality and more accurate information at the cost of more complex control mechanisms. Furthermore, the shift of operational areas to littoral waters with often dense civil traffic and rapidly changing geographical and meteorological conditions calls for a much more dynamic adaptation of these sensor controls in comparison with the more stable environment of traditional operational areas in the Atlantic Ocean. The lowering of defence budgets on the other hand creates a need for crew reduction and shorter education/training times, thus reducing the synergy created within teams of operators and the knowledge and experience of individual operators.

From the above it can be concluded that sensor management requires an increasing amount of operator knowledge, while in effect, the available amount of knowledge is decreasing. The consequences of incorrect sensor management may however be severe: if the sensor systems of a ship fail to detect threatening objects, the vessel may be incapacitated or even neutralised and consequently mission objectives will not be met. This observation justifies research into the ways in which operators deploy and optimise the available sensor systems to observe the environment and how these observations contribute to mission success. This will result in the support of, or ultimately, the automation of the deployment of complex sensor systems in a versatile maritime environment.

This paper describes research into generic sensor management principles that enable the development of a support system that is capable of bridging the growing gap between the available knowledge and the required sensor management related knowledge.

Sensor management issues

As already mentioned in the introduction, sensor management is currently executed by operators, who have to translate the goals of a mission into technical sensor settings while taking operational and political constraints like Emission Control (EMCON) plans and Rules Of Engagement (ROEs) into consideration. Because these technical controls are sensor specific, the operator must be familiar with the meaning of each setting and how changing this setting affects the performance of the sensor. Furthermore, the operator has to account for and, if possible, compensate for the prevailing environmental influences on the quality of the information (QoI) that is delivered by the sensor. Furthermore, the operators must be aware of the complementary properties of the different sensors and actually have to consider the deployment of the complete sensor suite as opposed to setting each individual sensor.

The observation that system-specific sensor management is a complex task that requires extensive operational and technical knowledge is also recognised in literature and various papers can be found that propose methods and algorithms to support this task. Strömberg et al. [Strömberg et al., 2002] have conducted a literature survey that presents an overview of relevant principles and methods concerning sensor management. Most of the methods reviewed by them provide a technical, sensor-oriented approach that strives for obtaining optimal sensor settings but leaves the translation of the operational requirements into technical sensor settings to the operator and therefore do not provide a solution to the identified problem. McIntyre and Hintz have compiled *a Comprehensive Approach to Sensor Management*, consisting of three papers [McIntyre and Hintz 1999-I, 1999-II and 1999-III], that describe a survey of modern sensor management systems, a new hierarchical model and goal lattices. In their first paper [McIntyre and Hintz 1999-I], they present the concept of the sensor management process and recognise sensor management as a process that contributes to the realisation of the mission goals; how this may be achieved is however not made clear.

Interviews with operational experts [Bolderheij and Absil, 2006] showed that a good picture of the operational environment, also referred to as the Operational Picture (OP), the Recognised Picture (RP), the Recognised Maritime Picture (RMP) or the Common Operational Picture (COP) can be regarded as a critical success factor for mission success. According to these experts, the OP consists of all objects in the neighbourhood of the platform and that two important conditions have to be met in order to consider it a good picture:

1. the OP must be complete;
2. the OP must be accurate.

Sensor management should therefore support the compilation of a good OP because the sensors are the resources that provide the required information about the environment. This means that the deployment of the sensor systems must be aimed at satisfying the identified conditions. Bolderheij et al. [Bolderheij et al., 2005] argue that these requirements can be met by constructing the OP from objects that represent the mission-relevant elements in the environment. They state that the OP can be considered *complete* if each relevant element in the environment is represented by at least one (preferably by only one) object in the OP and that the *accuracy* of the OP can be pursued by reducing the uncertainty in the information about the object. To maintain the completeness of the OP, sensor systems have to be deployed to *search* the environment for the presence of these elements while the accuracy can be increased by *tracking* and *classifying* them.

Integrating sensor management in the Command and Control process

The discrepancy between the available and the required amount of sensor management related knowledge described in the introduction gave rise to the question whether and, if so, how the sensor management process could be embedded in the Command and Control (C2) process because this process is currently executed by an operator who utilises mission related data to deploy the sensor systems.

The Allied Joint Doctrine [AJP – 01(B)] defines C2 as the process that plans, directs, coordinates, controls and supports an operation and therefore inherently has to direct,

coordinate, control and support the deployment of the available sensor systems. In the previous section we saw that the complete and accurate picture of the environment is a mission critical success factor and since this picture is compiled from sensor observations, it is clear that the sensor systems are important resources that are essential for mission success. The sensor management process can therefore be regarded as a key sub-process of the C2 process.

Command and Control

The definition of the C2 process that was presented above in itself explains why the RNLN considers the C2 process of vital importance and why a substantial amount of research has been and is still funded to analyse the nature and layout of this process. With this research the RNLN wants to increase its efficiency and support the automation of the process, thus enabling further crew reduction. A review of related literature yielded a *cognitive* C2 process model [van Delft and Schuffel, 1995] which forms the basis for subsequent research into the C2 process founded by the RNLN. This C2 process model distinguishes four main sub-processes:

1. The provision of Situational Awareness (SA): this process gathers data about events in the environment of the platform and compiles a picture of the environment.
2. The execution of Threat Assessment (TA): this process enhances the information compiled in the picture of the environment by reasoning about the imposed threat.
3. The support of Decision Making (DM): this process makes decisions about the deployment of the available systems based upon the threat in the environment.
4. The execution of Direction and Control (DC): this process executes the decisions with respect to the deployment of the ship's systems or resources, thus striving for mission completion.

The layout of the process model is shown in Fig. 1.

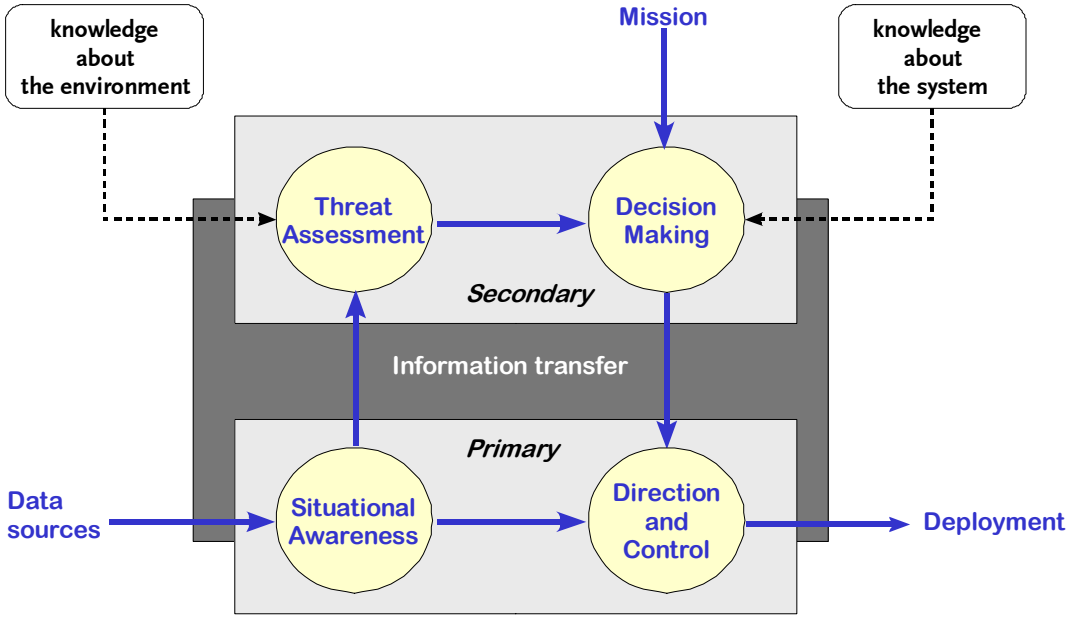


Figure 1. The cognitive C2 process model

Fig. 1 shows the three different types of input to the DM process:

1. Input from the TA process (information about threats in the environment).
2. Input from the mission (e.g. goals, requirements and constraints).
3. Knowledge about the system (e.g. available sensors and weapons).

From the description of the SA and the TA processes it can be concluded that these processes in effect execute the picture compilation process: the SA process gathers the information from the available sensors, associates, correlates and fuses this information and provides an *objective* view on the environment; the TA process now uses this information to infer the consequences of the elements in the environment in the (near) future. To accomplish this task, assumptions have to be made and therefore the picture becomes more *subjective* as a consequence of the reasoning process. At this point, the function of the direct connection between the SA process and the DC process became unclear. The report of Van Delft and Schuffel [Delft and Schuffel, 1995] states that predefined set points in the second level of information transfer enable the usage of this 'fast track'. Because the report takes a *Human Resources* point of view to the C2 process, this means that operators are required to control the combat systems by a nearly instantaneous appreciation of the situation based on intensive training. This *recognition-primed decision making process* as it is called by Klein and Grandall [Klein and Grandall, 1996] only functions if the situation at hand bears a resemblance to a training situation. This observation gave rise to the question whether this link can be maintained in the future, because due to the financial reasons mentioned in the introduction, training and education time will be limited. Because this research focuses primarily on the modelling and implementation of the sensor management related knowledge, research into the implementation of training would divert the attention too much from this objective and therefore this link was (temporarily) removed from the model.

The sensor control loop

After the removal of the link between the SA and the DC process, the C2 process showed a striking resemblance to the Observe, Orient, Decide and Act loop (OODA-loop) as proposed by Boyd [Boyd, 1987-1992]. This loop was initially intended to explain victory in air-to-air combat, but is nowadays also used within a wide variety of applications like in the design of business strategies. It describes how data is upgraded into information that in its turn leads to knowledge that can be used for actions that contribute to the realisation of mission goals. The consequences of these actions can now be observed as changes in the environment, observations that, after analysis may trigger more decisions and subsequent actions. From this description it is clear that the C2 processes from Fig. 1 can be directly mapped on the OODA processes. This provided an opportunity to redesign the C2 process as a loop, see Table 1.

Table 1. OODA and C2 processes

OODA Process	C2 Process
Observe	Provide Situational Awareness
Orient	Perform Threat Evaluation
Decide	Perform Decision Making
Act	Execute Direction and Control

To close this C2 loop, a feedback connection from the DC process to the SA process needs to be implemented. At first sight it may not be apparent what form this connection should take, but if one recalls the statement from the previous section that sensor management consists of at least two sub processes of which one sub process controls/tunes the sensor system, it can be seen then, that the loop can be closed by means of the sensor systems: the sensor settings that are produced within the DC process generate new sensor observations that are again inserted into the SA process. This results in a revised, OODA-loop based C2 concept. Fig. 2 depicts this new C2 process model in combination with the knowledge required for the sensor management process including some other resources that are controlled by the DC process.

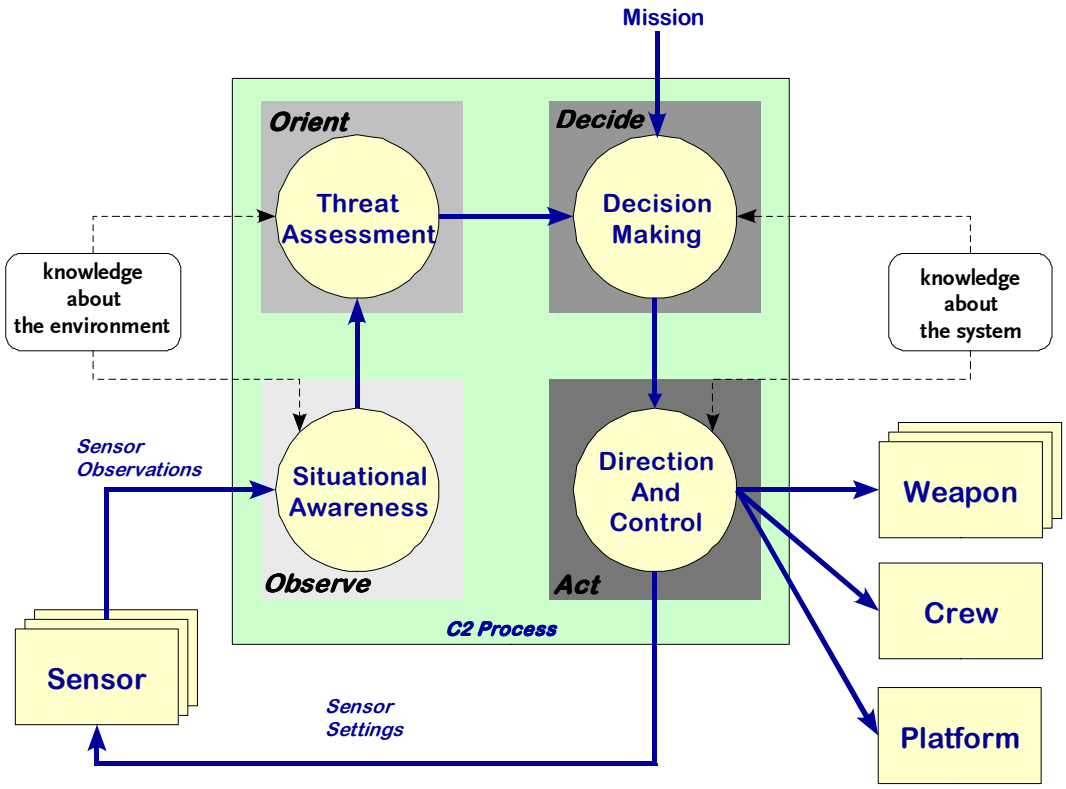


Figure 2. The revised C2 process model including the controlled resources

The adapted C2 process model now shows the outline of a sensor control cycle that induced further research.

The object-centred C2 model

The C2 process that was developed in the previous section is described in a *functional* way. Fig. 3 shows the processes (functions) that provide situational awareness and assess threats by processing and analysing the data received from the sensor systems to make

decisions about the direction and control of the available resources. Data/ information flows from one process to another following the arrows in the diagram. In the development process this data flow has to be described because it is not evident what data is exchanged between the processes. Interviews and literature reviews made it clear however, that all data exchange has some relation to the OP which acts as a central element within the C2 processes: all processes either add information to the OP, update the information that is contained by its objects, or use this information to make decisions about deploying the available resources. These results enabled the first step towards a new, OP-centred, or *object-centred* C2 concept.

The introduction of the OP as the central element in the C2 process is visualised in Fig. 3. This rearrangement of the C2 processes provides the option to utilise the OP as a virtual *blackboard*: all processes write their object related information onto this blackboard or use the information that is written on it. The OP may also be regarded as a virtual *marketplace* where *agents* representing objects in the environment negotiate their need for resources.

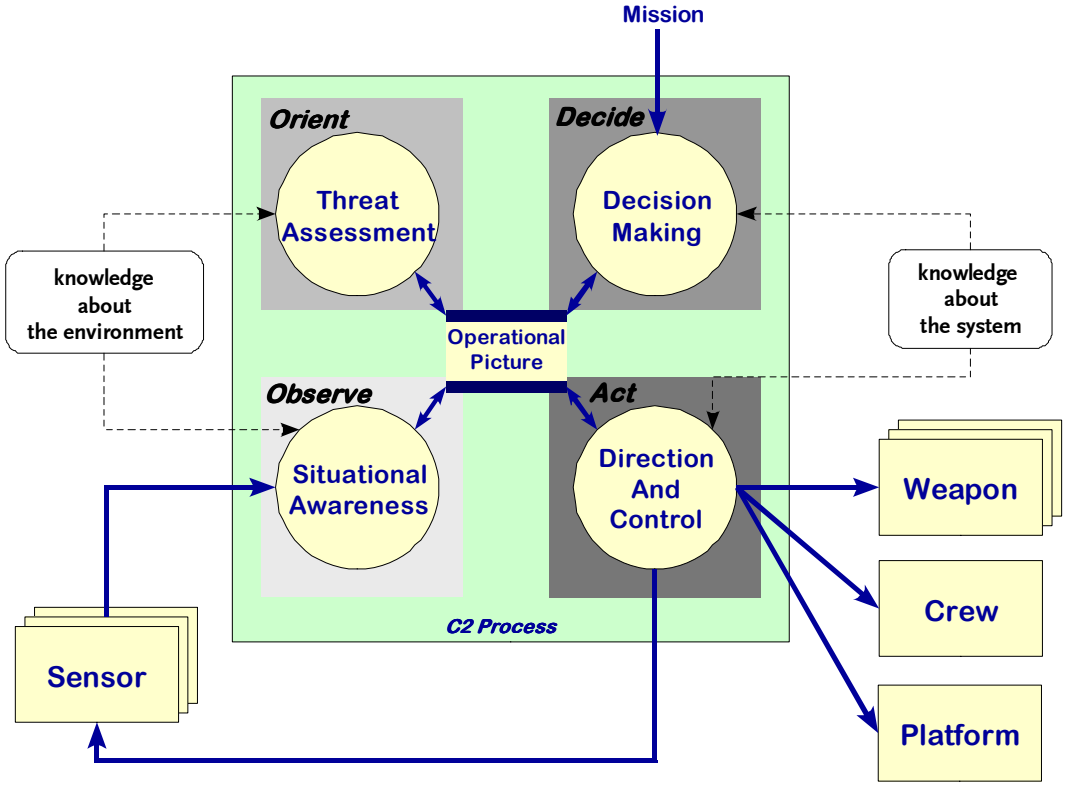


Figure 3. The object-centred C2 process

The reorganisation of the C2 processes that is introduced here also enables the implementation of fast, pre-programmed reactions triggered by specific observations and results in the reinstatement of the *recognition-primed decision-making* process [Klein and Grandall, 1996] that was briefly discussed previously.

A closer look at Fig. 3 shows that the initial OODA (SA-TA-DM-DC) loop based C2 concept of Fig. 2 is in fact abandoned, because all processes interact directly with the OP without a specific sequence. This is even more evident if the SA, TA, DM and DC processes are broken up in their constituting sub-processes: there is no apparent reason

to maintain a specific execution sequence among those sub-processes. These sub-processes may even be executed concurrently if this is facilitated by the infrastructure. It can be seen however, that the concept of a sensor control cycle introduced previously still holds, as observations are provided by the sensor, are stored in the OP and are used to select a sensor and determine the sensor settings. This sensor control cycle is depicted in Fig. 4.

The upper-half of the cycle is formed by the processes that use the sensor information to compile the OP and the lower-half is composed of the processes that use the information stored in the OP to select the most appropriate sensor(s) and to control each sensor.

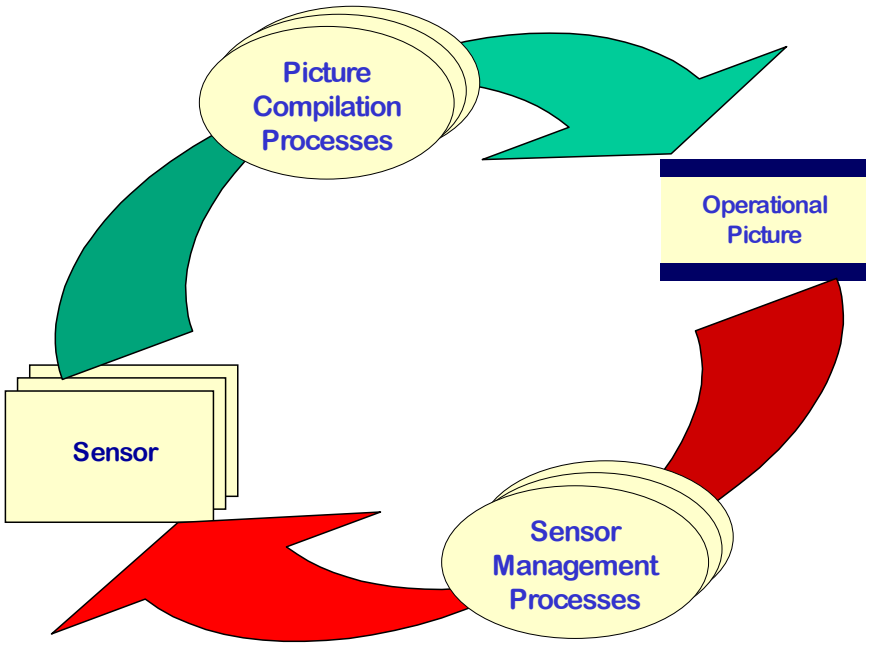


Figure 4. The sensor control cycle

The sensor management process

In their overview of sensor management methods and principles, Strömberg et al. [Strömberg et al., 2002] pose two relevant questions with respect to sensor management: "tasks want to know 'what sensor to select?' and sensors want to know 'what action to take?'" Based on these questions, two sensor management sub-processes can be identified:

1. A sensor tasking process that decides which sensor is the most appropriate for a specific task.
2. A sensor scheduling process that controls the sensor that is tasked by the sensor tasking process.

These sub-processes however are not identical to the sensor management functionalities proposed by Blackman and Popoli [Blackman and Popoli, 1999] who describe a template for a sensor management design consisting of two loops:

1. A loop controlled by a 'Macro Sensor Manager' assigning the tasks that need to be accomplished to satisfy the overall goals (here: the generation/maintenance of the Operational Picture);
2. A loop controlled by a 'Micro Sensor Manager' optimising the assigned tasks.

Both sets of functionalities can be reconciled by considering that the micro-level functionality, being the process that determines how a specific task is performed, is in fact a combination of the sensor tasking process and the sensor scheduling process. The combination of the sensor management tasking functionalities and the sensor management sub-processes yields a new sensor management concept that takes the shape of a three-stage sensor manager:

1. A task-composing stage that produces sensor tasks.
2. A sensor-selecting stage allocates (a) sensor(s) to these tasks.
3. A controlling stage that optimises the sensor settings with respect to the allocated task.

The first stage of this sensor manager analyses the information stored in the attributes of each object in the OP and uses information about the mission to determine if more *accurate* information is required and composes a sensor task based on these requirements. The Quality of Information (QoI) is determined by the accuracy of the sensor that provided the observation and the quality of the process that filters the observations (if present). Therefore this accuracy, or rather uncertainty, should be stored in combination with the information itself in the attributes of the object. This stage of the sensor manger was implemented as a rule-base. The second stage uses the available knowledge about the system from Fig. 1 to select the most appropriate sensor or combination of sensors. The third stage uses the information about the object, the knowledge of the system and the knowledge about the environment to determine the sensor settings. If too many tasks are assigned to this sensor, the task is returned to the second stage of the sensor manager and this stage now assigns the task to the next best sensor. These stages can therefore be combined in a single algorithm as described in [Van Norden et al., 2005].

From the description of the sensor management stages it can be seen that the first and second stage of the sensor manager can be situated within the DM process as they decide about tasks that need to be executed and the sensor that will be assigned to execute them. The third stage however has to be positioned within the DC process because the sensor is *controlled* to optimally execute these tasks.

Initiation of the object store

An analysis of the object-centred C2 process in combination with the three-stage sensor manager from the previous section reveals an initiation problem: at start-up time of the C2 process, no sensor observations are available and therefore no objects are present in the OP. Because the three-stage sensor manager uses the information stored in the object attributes to compose a sensor task, to select a sensor and to determine the corresponding sensor settings, no new measurements are generated. In order to enable the detection of objects in the environment of the platform, surveillance functions have to be initiated. These initial surveillance functions could be operator-controlled, but then the original

problem that prompted this particular research reappears and therefore the sensor control loop needs to be initiated in a different way.

A solution to this problem is provided by a further exploitation of the mission information showed in Fig. 1. Within the planning stage of a mission, among other activities, the resources of the opposing forces are reviewed. This review process yields a set of potential weapon carriers and related weapon systems likely to be deployed by the opposing forces. These objects can also be stored in the OP as *virtual objects* because they represent real world objects that might be present in the environment of the platform, but are not (yet) observed. The information stored in the attributes of the instantiated objects, can be utilised by the sensor manager to select the most appropriate surveillance sensor and to determine the optimal sensor settings for these surveillance tasks. There may also exist more or less *uncertainty* about some aspects of the opponent's resources: depending on *quality* of the available intelligence information, it may not be known for sure what type of weapon systems the carriers are fitted with and their location may not be precisely known. This uncertainty resembles the uncertainty resulting from the sensor accuracy and this uncertainty can also be exploited by the sensor manager to determine the most appropriate sensor for the task and the corresponding task parameters.

Budget allocation and prioritisation needs

The object-oriented approach that was described in the previous sections supports the scheduling of the available sensor systems in the case of sufficient resources. In traditional sensor suites, different types of tasks are assigned to different types of sensors: search tasks are assigned to (long range) surveillance radars; target acquisition is accomplished by radars that provide a higher range and cross range resolution than the standard surveillance radars and tracking and illumination is done by track radars. Because dedicated radars are available to perform different tasks, there is not much need for dividing the sensor budgets. The only experience available in task scheduling and budget allocation is related to the deployment of mechanical Single Target Trackers (STTs) for Weapon Assignment (WA) purposes: once the decision has been made to deploy a guided weapon system, a scheduler selects the missile in combination with a fire control radar. The characteristics of this type of scheduling mechanism fit the needs of a sensor manager for tracking and weapon direction; it does not however reserve sensor capacity for not yet detected but potentially more dangerous objects (the so-called virtual objects) and is therefore not suited for scheduling Multi Function Radars (MFRs, sensors that provide surveillance, tracking and sometimes weapon guidance and/or classification capability) or complete sensor suites. In modern active MFRs, allocated search budget is not available for tracking purposes, and the illumination of an object for weapon guiding purposes will seriously drain the available time/energy budget (TEB). This means that in a scenario with a lot of 'neutral' traffic, these tracking tasks would consume the entire TEB while surveillance tasks are omitted and any missile in the vicinity of the platform would remain undetected. Therefore priorities have to be assigned to these different tasks.

Various scheduling mechanisms dealing with this problem are proposed in literature. For instance Huizing and Bloemen [Huizing and Bloemen, 1996] suggest a scheduling algorithm for an MFR, based upon an operator assigned priority of the sensor function type. The question that remains to be answered here is on what basis these priorities have to be assigned and furthermore, two similar sensor function types may require different priorities. Komorniczak et al. [Komorniczak et al., 2002], describe a prioritising mechanism based upon the kinematical properties, its Identification Friend or Foe (IFF) identity and an operator assigned rank of a threat object once this object is detected; this mechanism could be used to assign the priorities required for the tracking functions [Huizing and Bloemen, 1996] but it needs to be expanded for assigning priorities to surveillance functions.

According to Huizing and Bloemen, the prioritising mechanism has to placed in an operational perspective; this requirement closely fits the demand for the maximisation of the probability of mission success mentioned earlier, because depending on the intentions of the operators of the objects in the environment of the platform, these objects may damage or even destroy our platform. In a novel approach to solve his problem, the probability of mission success is maximised by ranking these objects with respect to their capability to cause mission failure and assigning the available sensor budget in accordance with this ranking. This *threat ranking* process can be executed by estimating the risk composed of the lethality of the object and the probability of occurrence of the damage that can be inflicted by this object. The risk estimation process is described in detail in [Bolderheij and van Genderen, 2004].

The new C2 concept with embedded sensor management

From the descriptions in the previous sections, the C2 model shown in Fig. 3 and inherently the sensor control cycle from Fig. 4 was developed in more detail, by identifying the processes that contribute to the picture compilation process and combining them with the three-stage sensor manager. The results are shown in Fig. 5. This figure outlines how sensor observations are merged into the OP and how this information can be used to track and classify objects and to infer their threat. The information is then utilised in combination with the related uncertainty to update the deployment of the sensors in order to keep the OP complete and accurate.

Simulation and results

To demonstrate the validity of the newly developed C2 model and the sensor management principles and the sensor manager that was designed along the lines of these principles, a prototype was developed and subsequently tested in a simulation environment. Operational experts were asked to assist with the composition of a sufficient realistic maritime scenario. In this scenario the deployment of a MFR consisting of four active antenna arrays was simulated because it has been shown in practise that this type of sensor is hard to control as it incorporates different sensor functions that need to be deployed simultaneously. The results of the deployment are logged and analysed after the mission has been completed.

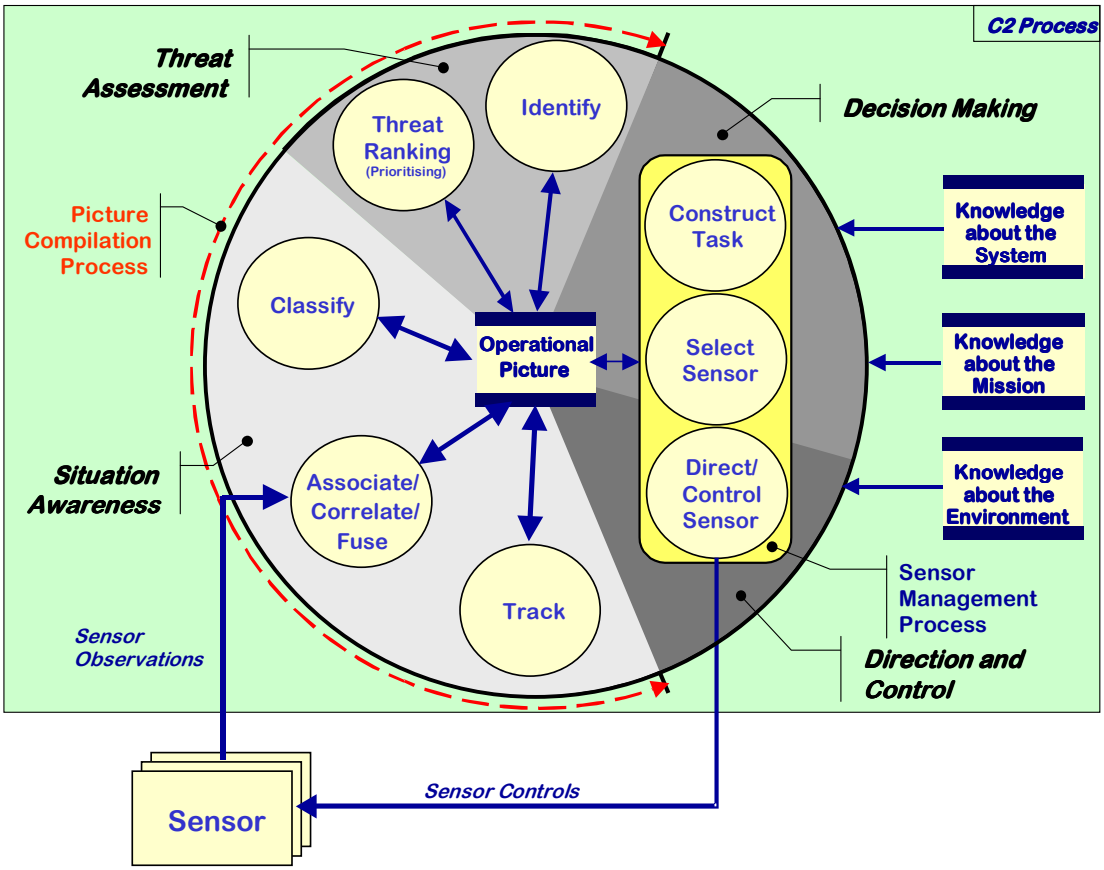


Figure 5. The more detailed C2 concept with embedded sensor manager

The scenario

In the constructed scenario that is depicted in Fig. 6, the important regional power Orange Country has just extended its territorial waters to include some important maritime oil fields, a claim that is heavily disputed by the international community.

To demonstrate the determination of the international community in this matter, an RNLN Air Defence and Command Frigate (ADCF) is tasked to sail along a navigation track (solid blue line) which is laid out just outside the original territorial waters (dashed green line) but well within the new territorial waters. The sensor suite of the ADCF consists of a Volume Search Radar (VSR), an MFR, several navigation radars, an Electronic Support Measures (ESM) system, an Infrared Search and Tracking system (IRST) and a Trainable Electro-Optical Observation System (TEOOS). Orange Country has deployed two land-based mobile missile launchers and a small aircraft carrier, which is positioned just within the border of the newly claimed territorial waters. Both launcher 1 (SSM site 1) and the aircraft carrier are fitted with subsonic Surface-to-Surface Missiles (SSMs) that are launched in the direction of a predicted hitting point and activate their internal radar after a pre-programmed time delay. Launcher 2 (SSM site 2) is loaded with an SSM type that first follows a set of predetermined waypoints before it activates its internal radar. Intelligence sources have made this information also available onboard the ADCF.

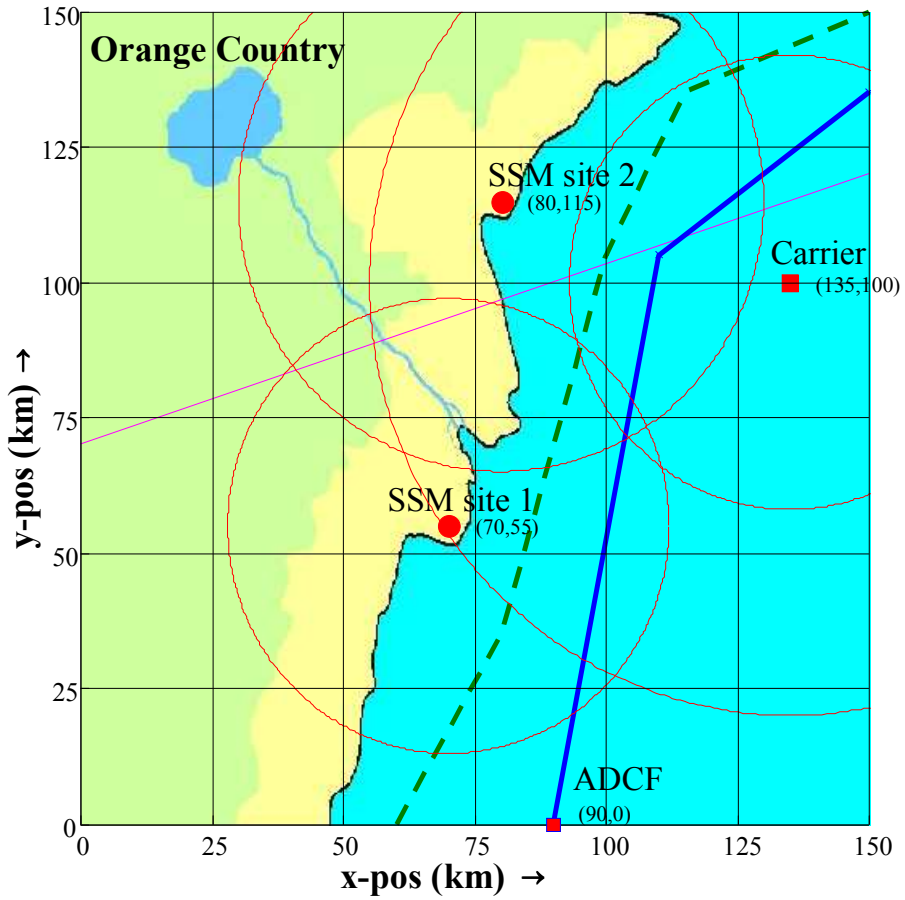


Figure 6. The maritime test scenario

At the starting time of the mission, the conflict between Orange Country and the international community has just escalated to a full-scale war. After the ADCF has entered the range of SSM site 1, it is discovered by a patrol aircraft originating from the carrier and subsequently four missiles are launched. Furthermore, two missiles are fired from SSM site 2 at ADCF when it comes within range. While the scenario is rolling, several civilian aircraft move through the area following the airway represented in Fig. 6 by the solid cyan-coloured line. Because intelligence sources are not sure whether all platforms with weapon carrying capabilities operated by Orange Country have been identified, a constant threat of Subsonic Sea-skimming missiles (SBSs) is assumed during the mission.

Results

The sensor control cycle was initiated by inserting the SSM launchers and the aircraft carrier in the system as virtual objects and the related prior information was stored in their attributes: this information initiated Limited Volume Search (LVS) functions when the ADCF entered the known (estimated) range of the particular missile range. Similarly, the generic SBS triggered a Horizon Search (HS) function because no specific direction could be assigned to this threat. Observations resulting from these surveillance functions caused the real objects within sensor range to be detected, tracked and eventually illuminated when this was operationally required. The allocation of the sensor budget during the execution of the mission is shown in Fig. 7. In this figure, the allocation of the time budget is shown as bar graphs in time steps of 10 seconds for each of the four

antenna faces of the MFR. It can be seen from this figure that during the first stage ($t = 0$ s until $t \approx 4{,}500$ s) of the mission, only HSs are executed, resulting from the presumed presence of a generic SBS. The scheduler only allocates time to this task because no other tasks need to be performed. The amount of time is based on the object properties and (in future versions) on environmental factors. As the ADCF enters the range of the SSMs from site 1, an LVS is assigned in the direction of its presumed location (Face 1 and 4). The beam of this function is determined by the uncertainty in the position of this site. The height of the search pattern is derived from the maximum flight level of the missile type that may be launched from this type of launcher. While the ADCF continues its navigation track, several airliners are transiting following the air lane situated at an altitude of approximately 10 km. Because of the curvature of the earth these airliners may be detected coincidentally by an LVS. The reconnaissance aircraft takes off from the aircraft carrier at $t \approx 6{,}000$ s, and is finally detected by the HS function; the aircraft then turns and when it is outbound, the priority is reduced. The four missiles that are fired from launcher 1 at $t \approx 6{,}700$ s after the detection of the ADCF, are detected by the LVS, are subsequently tracked and a weapon guidance function is initiated. This is the only moment within the simulation that the sensor manager had to drop a sensor function due to insufficient budget. The HS is temporary dropped in favour of the illumination and the LVS because this function has the lowest priority.

The scenario had to be reconstructed several times to create this overload situation. If in future versions of the sensor manager more sensors are incorporated, the HS function has to be assigned to the next best sensor.

After the ADCF enters the range of SSMs from site 2 ($t \approx 14{,}500$ s), two missiles are fired, but since these missiles first follow a predetermined track leading them away from ADCF and because the launcher is still behind the radar horizon, these missiles are only picked-up at a very late stage by the HS function. At that time the risk posed by these missiles has already risen substantially and therefore a weapon guidance function is scheduled immediately at $t \approx 15{,}500$ s. In the remainder of the scenario, the MFR performs LVS functions to observe the aircraft carrier.

Because presently no comparable sensor management systems exist, operational experts were consulted about this prototype. When presented with the loggings of the MFR deployment, they remarked that this way of controlling a sensor would likely put too much strain on human operators, as it requires a constant update of the parameters that *drive* the deployment of the different sensor functions. Especially the repeated re-evaluation of the involved risk posed by the OP-objects and the resulting changes in priorities could create overload situations in these types of scenarios. This is less likely to happen when traditional sensors are used because then the OP is compiled from VSR observations that may be augmented by IFF data or information from other sensors.

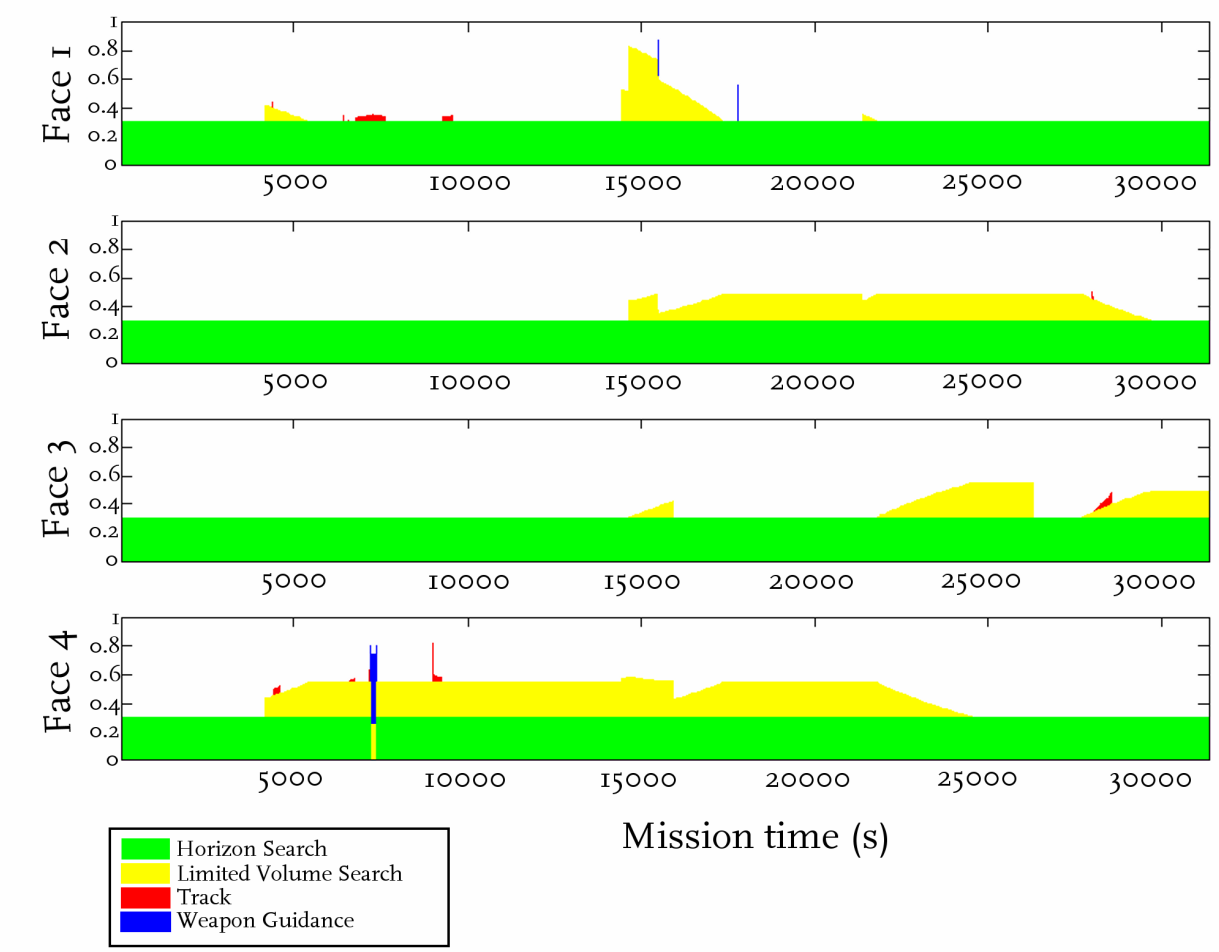


Figure 7. The allocation of the MFR budget for each of the four antenna faces

Threatening objects are automatically assigned to one of the two or three available STTs based on their classification and calculated Time on Top. In this case, the picture is compiled by means of several different sensors, where each sensor has its own controller or operator, provides only limited functionality and has its own TEB. Because an MFR is able to track and illuminate many more targets than two or three STTs while it is also capable of performing several different types of surveillance functions (semi) simultaneously, many more monitor and control actions are required in comparison with traditional sensors. As each function draws its TEB from the same source, the allocation of this budget among the different sensor functions must be controlled by a single operator/controller in order to avoid conflicts. This situation provides many more opportunities for the creation of overload situations causing sub-optimal sensor deployment. Because the MFR has only been recently introduced, no management guidelines exist and only limited MFR management experience is available and therefore it is difficult to compare the performance of the sensor manager with a human operator. Nevertheless, the general feeling of the operational experts was that the sensor manager was very well capable of outperforming a human operator, especially in terms of dynamic adjustment of the different control parameters and integral management of the complete sensor suite.

Other use
The object-oriented C2 concept was also tested in a prototype mission manager that controls the deployment of Unmanned Arial Vehicles (UAVs) and in an experimental combat management system that was developed to prevent firing at own forces (so-called 'blue on blue engagements') [De Jong et al. 2008]. In both systems, the applied concept resulted in an enhancement of the situational awareness and showed promising results with respect to the (re)deployment of the available resources (UAVs, soldiers) and the prevention of 'friendly fire'.

Conclusions

In this paper, sensor management was approached from an operational perspective and was described as the process that determines *what* sensor functions are required, *which* sensor(s) should be selected to execute these functions and *how* these sensors should be controlled to compile a *complete* and *accurate* OP.

An important new concept resulting from this research lies in representation of the OP as a set of objects that represent both the *observed* and the *expected* mission-relevant elements in the environment. The information about the properties of these elements in combination with the related uncertainty is stored in the attributes of these objects. This representation of the OP allows the picture compilation processes to be defined as the processes that seek to determine the properties of these objects and reason about the threat they pose. These processes now form half of a novel sensor control cycle. The other half of the cycle consists of a newly designed object-oriented, three-stage sensor manager. The *first stage* of this sensor manager analyses each of the objects in the OP to determine what tasks need to be executed to reduce the uncertainty related to the properties of the object in order to improve the *accuracy* of the OP. The *second stage* then selects the most appropriate sensor for this task based on the QoI delivered by each of the available sensors. Finally, the *third stage* controls the settings of the selected sensor or hands the task back to the second stage if the resources of the sensor are depleted. To initiate the sensor control cycle, *virtual objects* were introduced that enable the allocation of surveillance functions necessary to detect the expected objects, thus contributing to the required *completeness* of the OP. The objects that are detected by these surveillance functions are now tracked, classified, and identified to ensure the *accuracy* of the OP using sensor observations.

The prototype C2 system and sensor manager that were developed from this design were tested by managing an MFR model in a simulation environment. The execution of the simulation showed that autonomous deployment of complex sensors like the MFR by means of the information stored in the OP is feasible.

Because the sensor manager assumes the existence of an object-oriented OP, the integration with existing C2 systems either involves the reengineering of those C2 system components that execute the picture compilation process or the development of an interface between the existing C2 system and the sensor manager. These interfacing issues have to be resolved before the sensor manager can be successfully integrated into existing C2 systems.

References

Blackman, S. and Popoli, R. (1999) *Design and Analysis of Modern Tracking Systems*, Artech House, Norwood, MA, pp 1004-1018.

Bolderheij, F. and Van Genderen, P. (2004) Mission Driven Sensor Management.: *Proc. 7th Int. Conf. on Information Fusion*, Stockholm, pp 799-804

Bolderheij, F., Absil, F.G.J. and Van Genderen, P. (2005) Risk-Based Object-Oriented Sensor Management, *Proc. 8th Int. Conf. on Information Fusion*. Philadelphia.

Bolderheij, F. and Absil, F.G.J. (2006) Mission Oriented Sensor Management.: *Proc. Cognitive systems with Interactive Sensors,* Paris.

Boyd, J.R. (1987-1992) A Discourse on Winning and Losing. Unpublished briefing notes. Various editions. Available from: http://www.d-n-i.net/richards/boyds_ooda_loop.ppt

De Jong, J.L., Burghouts, G.J., Hiemstra, H., Te Marvelde, A., Van Norden, W.L., Schutte, K. and Spaans, M. (2008) Hold Your Fire!: Preventing Fratricide in the Dismounted Soldier Domain, *Proc. 13th Int. Command and Control Research and Tech. Symp.* Bellevue, WA.

Huizing, A.G. and Bloemen, A.A.F. (1996) An Efficient Scheduling Algorithm for a Multi Function Radar, *Proc. IEEE int. symp. on Phased Array Systems and Technology*, Boston, MA, pp 359-364.

Klein, G.A. and Crandall, B.W. (1996) *Recognition-Primed Decision Strategies*, ARI Research Note 96-36, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA.

Komorniczak, W. Kuezerski, T. and Pietrasinski, J.F. (2000) The Priority Assignment for Detected Targets in Multi-Function Radar, *Proc. 13th Int. Conf. On Microwaves, Radar and Wireless Communications*, Mikon, pp 244-247

McIntyre, G.A. and Hintz, K.J. (1999-I) A Comprehensive Approach to Sensor Management, Part I: A Survey of Modern Sensor Management Systems, *IEEE Transactions on SMC*.

McIntyre, G.A. and Hintz, K.J. (1999-II) A Comprehensive Approach to Sensor Management, Part II: A new hierarchical model, *IEEE Transactions on SMC*.

McIntyre, G.A. and Hintz, K.J. (1999-III) A Comprehensive Approach to Sensor Management, Part III: Goal Lattices, *IEEE Transactions on SMC*.

NATO. AJP - 01(B) (NATO/PfP unclassified): Allied Joint Doctrine.

Strömberg, D., Andersson, M. and Lantz, F. (2002) On Platform-Based Sensor Management, *Proc. 5th Int. Conf. on Information Fusion*, Annapolis, ML, pp 1374-1380.

Van Delft, J.H. and Schuffel, H. (1995) *Human factors onderzoek voor toekomstige commando centrales KM*, TNO-TM 1995 A-19 (in Dutch).

Van Norden, W.L., De Jong, J.L., Bolderheij, F. and Rothkrantz, L.J.M. (2005) Intelligent Task Scheduling in Sensor Networks, *Proc. 8th Int. Conf. on Information Fusion*, Philadelphia.

Modelling Human-like Visual Perception for Intelligent Multi-modal Information Fusion

Coen Stevens, Theo Hupkens & Léon Rothkrantz

Introduction

Military sensors are being used to create situational awareness. In a process of 'multi-sensor data fusion', a sensor grid containing a multitude of similar and different sensor types contributes to the overall awareness by gathering and combining input data. Such data fusion has been applied in numerous military applications including ocean surveillance, air-to-air defence, battlefield intelligence, surveillance and target acquisition, and strategic warning and defence [Hall, 2001]. Regarding the sensor grid there are several recent developments, i.e., sensor types become multimodal and mobile sensor deployment will be increasingly autonomous. Multimodal means using multiple modalities, which are different types of physical phenomenon that can be sensed, such as light and sound. In terms of military applications one can think of a combination of radar and electro-optical systems, or electro-optical combined with acoustic. Non-military examples of incorporating multimodal fusion include: enhancing automatic speech recognition with visual features, and person identity verification.

The aim of our research is to design and implement an autonomous and adaptive surveillance system based on a video and acoustic sensor. In order to achieve our goal, we need to implement a suitable data fusion framework and fitting fusion techniques. The fusion of the two modalities: vision and audio, has to solve the problems of ambiguity, redundancy and synchronicity in a seamless manner. The idea is that the surveillance system takes over the task of the human observer, which means interpreting the scene and spotting for aggressive or other illegal activities that will have to be reported back to surveillance personnel who can then take the appropriate action.

In order to achieve autonomous surveillance with a multimodal, intelligent sensor we believe that understanding and modelling human perception is at the crux of making intelligent context sensitive systems that try to make sense out of an overwhelming amount of data coming in through their sensors. In this article we will focus purely on the visual part of our fusion model and present our computational model for visual perception including the results we have so far.

Modelling human perception

Humans unconsciously utilize audiovisual information fusion continuously. For example, when listening to a speaker, we also tend to look at his or her lip movements (and other non-verbal signs, like gestures), which help us to improve speech recognition by utilizing the complementary information in vision and audition. Not only do we receive more information using multiple senses, but multimodal processing can help us to resolve ambiguous information within any single modality. This enhances our situational understanding and awareness.

Neurological evidence suggests that multimodal fusion is only done at a higher level following the perception of each of the separate modalities. However it appears that high-level perception, the level at which concepts and representations come into play, is not separable from low-level perception (basic processing of incoming data), being deeply intertwined [Chalmers et al., 1992]. Not only will low-level perception influence high-level concepts (bottom-up), also the conceptual influence keeps perception flexible given any context (top-down). For example when we have prior knowledge of a situation, say we are given a picture and are told in advance that there will be a man in the picture, we use this high-level concept of 'a man' to group the low-level input by looking for 'man-features'. Or another example: when we appear to see a face, we tend to interpret the features in the face as eyes, mouth and nose, even when the picture is so unclear that we would not have recognized a nose if the area of the nose would have been presented in isolation. This is the power of context, which is the interpretation of lower level features given the higher-level concepts (e.g., face). Now this is not something exclusively for visual or auditory perception, it also works between these two modalities. When you hear meowing you expect to see a cat! Most work to date in visual and auditory perception has been targeted at either bottom-up or top-down processing. The main challenge for future models of perception is the integration of such top-down influences with bottom-up processing [Riesenhuber and Poggio, 2000]. We believe that such an integration can be accomplished by modelling auditory, visual and audiovisual perception on an underlying theory of 'self-organization' and 'emergent behaviour', which will be explained in detail in the following sections.

Background

Emergent perception

Emergent behaviour can be found in nature among many species where their local actions and interactions result in a global behaviour of the entire group which is novel with respect to the behaviour of every single member of the group. For example ants leaving pheromone trails while gathering food, lead to an effective path-finding strategy of food sources for the entire group of ants that follow the strongest (reinforced) trails.

A lot of psychophysical and neurological evidence suggests that perception deploys emergent mechanisms resembling the above mentioned emergent behaviour. The emergent properties of perception were shown to exist by the Gestalt psychologists and their Gestalt theory, which started in 1921 with the Max Wertheimer founding paper [Wertheimer 1923]. Gestalt theory was a reaction to the established notion of structuralism introduced by W. Wundt in 1879. Structuralism stated that perception is built up from atomic elements, called sensations, which together by mere addition of all elements constitutes the overall perception. Gestalt theorists on the other hand believed that the whole can be different than the sum of its parts. They emphasized the interaction of the parts and the organizational process as a dynamic process. Gestalt theorists often describe perception as a self-organizing system that spontaneously takes on the 'best' or simplest arrangement in given conditions. During the process of self-organizing perception Gestalts (organized wholes) emerge from the data gathered by our sense organs. Gestalt psychologists provide a theoretical framework based on psychophysical

experiments with perceptual laws of organization with which they emphasize the interaction of the parts and the organizational process as a dynamic process.

The following are some of the Gestalt laws of organization (see Fig. 1 for an illustration of Gestalt principles):

- Proximity: when objects are close to one another we tend to perceive these as a group.
- Similarity: when objects look similar we tend to perceive one object rather than separate objects.
- Good continuation: when the eye tends to be led from one (series of) objects to another one, we tend to perceive these as a group.
- Closure: of several possible perceptual organizations, ones yielding "closed" figures are more likely than those yielding "open" ones.
- Orientation and symmetry: objects oriented with horizontal and vertical axes, or ones that are symmetric, are more often perceived as figures.

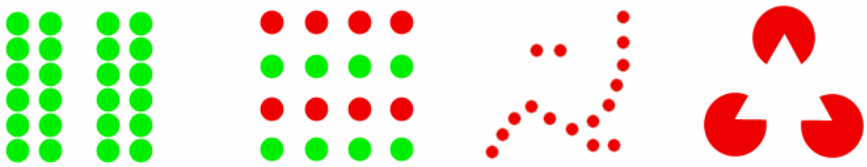


Figure 1. Gestalt principles from left to right, proximity (two groups), similarity (four groups), good continuation (three groups, where one group forms a continuous curved line) and closure (forms a closed triangle)

Camouflage is obviously related to colour and patterns, but also to Gestalts. For example, in Fig. 2 it is according to the law of similarity that the same colour without a border merges into its environment, or as the Gestaltist M. Wertheimer puts it: "If an object is to be hidden by blurring its boundaries, then it is important that besides the colouring, its texture and fine detail are matched to those of its environment."

Figure 2. Camouflaged soldier (from natural gear: http://www.naturalgear.com/backgrounds/natgear/1024/3a.jpg)

In Fig. 3 we can see another example of camouflage. Here the ship is painted with random areas of black and white.

Figure 3. HMS Furious, (British Aircraft Carrier, 1917-1948)
Source: http://www.bobolinkbooks.com/Camoupedia/DazzleShips.html

The texture pattern breaks up the visual outline of the ship when it is seen across the water, and makes it more difficult to tell which way the ship is heading, and to discriminate the different parts of the ship. This type of 'camouflage' does not hide the object from the viewer, but dazzles him.

To summarize, perception as in organizing the input into coherent subsets containing a single object or structure, is the result of competition and cooperation of laws of organization, rather than mathematical bottom-up segmentation. We like to see the organization laws as grouping pressures which try to push and group the input into a particular arrangement. Interaction between grouping pressures at the lowest level of perception give rise to emergent coherent structures (objects), which are novel with respect to the individual cues (e.g., pixels). So it is legitimate to say that the whole is more than the sum of its parts. The reason why interaction among grouping pressures is such a key ingredient arises from the necessity for dealing with the contradictory and incomplete set of cues present at any real-world input caused, among other things, by occlusions, distortions, and reflections. By letting these pressures actively push each other with no centralized interference, structures may emerge, which amount to a reconstruction of the shared fate of the constituent elements.

Computational models of Gestalt principles

Researchers have designed computational models for several Gestalt principles separately, e.g., good continuation, closure and organized contours [Desolneux et al., 2003]. However many have taken a strictly mathematical bottom-up approach, and computed absolute thresholds of meaningful groupings, where they neglected the crucial top-down (contextual) pressures and dynamic interaction among grouping pressures. Grouping pressures (like the Gestalt laws) on their own do not create strong coherent structures. Instead only those supported by an abundance of evidence by other grouping pressures constitute coherent groupings. Take for example the left dot-pattern in Fig. 4,

which contains a dotted line that can be easily recognized by humans. In the presence of lots of non-aligned dots it becomes much more difficult to observe the same line. Here seeing the alignment would not be the result of a single 'line-detection' grouping pressure, but is the result of a myriad of grouping pressures that interact and exploit redundant information. Different grouping pressures that propose similar groupings, provide more (redundant) evidence for a coherent strong structure. Examples of such grouping pressures are proximity, good-continuation, regular-orientation and regular-distance. Alternatively (bottom-up) mathematical line detection algorithms could quite easily find the same line in the left dot-pattern of Fig. 4. However they would also still find the same line when more random points are added to the same example, even when humans would no longer see the alignment (see the right example in Fig. 4).

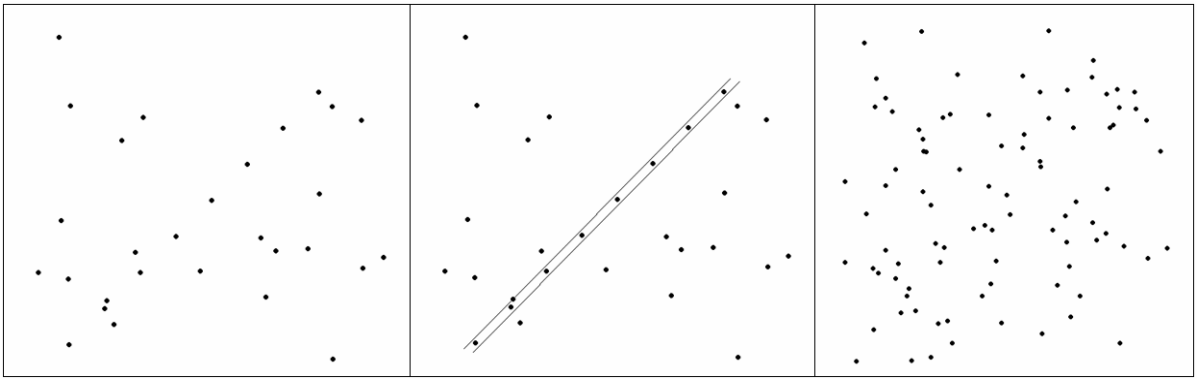


Figure 4. Left: 21 uniformly randomly distributed and eight aligned dots. Middle: this meaningful alignment is detected as a large deviation from the random pattern. Right: same alignment, but with 81 random dots. The alignment is no more meaningful (and it is not seen by the average human observer). In order to be meaningful, it would need at least 11 aligned points. (Examples from Desolneux et al. 2003).

We propose a computational model for visual perception based on the general underlying theory of implementing perception by self-organization [Stevens et al., 2008], which is founded on "The Ear's Mind", an architecture that supports emergent processes, self-organization, and context sensitivity, for the primitive perception of sound [Dor, 2005]. By implementing several visual grouping pressures that utilize the emergent architecture, we will demonstrate their importance and necessity for a computational model of visual perception. Our preliminary results agree with expected human visual grouping behaviour and support our ongoing work on audiovisual fusion.

The Ear's Mind

The Ear's Mind theory, offers a general architecture for simulating emergent sensory perception and specifically for the segregation of auditory scenes [Dor, 2005]. The model of The Ear's Mind was inspired by the 'Copycat' model [Mitchell, 1993; Hofstadter and FARG, 1995]. The Copycat computer program [Mitchell, 1993] models the mechanisms of analogy-making in a letter-string micro-domain. It was designed to be able to discover insightful analogies, and to do so in a psychologically realistic way. In the Ear's Mind, Copycat is abstracted from its original micro-domain and specific sort of analogy-making paradigm. The Ear's Mind is designed to model the unconscious, automatic auditory grouping pressures in humans. Such pressures, it seems, steer the perception of sound by cooperative and competitive interactions, resulting in the grouping of sound elements into context-sensible entities. These are the pressures we talked about in the previous

sections. A software prototype, simulating the most basic functionality of the proposed model has already been implemented and presented with sound excerpts of standard psychoacoustic experiments [Bregman, 1990]. Preliminary results agree with expected human auditory grouping behaviour [Dor and Rothkrantz, 2008].

A computational model for visual perception

Based on the same architecture as the Ear's Mind, our emergent system works as a non-supervised collection of independent local primitive agents which represent and act as local grouping pressures (e.g., proximity and regularity) that will try to force a specific grouping onto the input. These agents compete and cooperate to build or destroy bridges in the data-landscape they work on, resulting in the creation of high-level structures out of low-level input. Different grouping pressures that propose similar groupings, provide more evidence for a coherent strong structure. We take the visual Gestalt laws of organization as our initial starting point for modelling several different grouping pressures, but we do not take only the Gestalt laws as an exhaustive list of possible pressures. Before delving into more detail on the grouping pressures we first turn to the overall architecture of our visual perception model.

Architecture

The architecture, illustrated in Fig. 5, is based on four major building blocks: the Pre-processor, Workspace, Slipnet, and Coderack containing Agents.

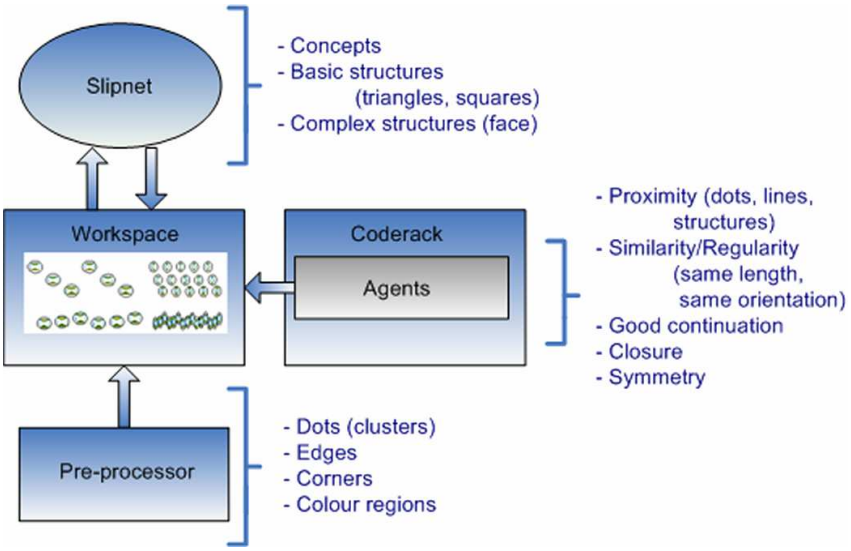


Figure 5. The architecture of our visual perception model

The *Pre-processor* analyses the input image and produces a list of salient cues, containing cue type, location and any other properties needed for a cue's definition. It is up to the pre-processor to fill the workspace with the most primitive cues and not with higher level interpretations or groupings of primitive cues. We do not propose an exhaustive list of primitive cues, but rather keep the option open to include more different cues as we go along. Currently we have only a single type of cue, namely dots, to study and model organizations of dot patterns. Later on, for more complex input images we certainly need to resort to other primitive cues, e.g., density, gradients, colour and edges. One thing we have to keep in mind is that these artificially constructed dot patterns are in a way more

difficult than real-life images, in the sense that in complex images we can find a lot more redundant information for building coherent structures.
The *workspace* is where the actual construction is taking place with the building of perceptual structures on top of the cues. When the workspace is filled with cues, we are ready to start launching local agents. We have implemented the following four different types of agents, which are described in more detail in the following sections: Proximity, Regular-orientation, Good-continuation and Regular-distance. Over time, through the actions of these agents, cues in the workspace gradually acquire various descriptions, and are linked (bonded) together by various perceptual structures. It is important to see that the strength of the architecture lies in the combination of multiple agents and their interaction, and not so much in any single agent. One imperfect agent that suggests a particular grouping that is in conflict with the grouping of a structure built by many other agents supporting each other, will by no means affect the overall good outcome of the system. Hence we do not try to build perfect agents, but we want to find and gather as much grouping evidence as possible. Initially we randomly launch the agents on the workspace, which means that the system works strictly in a bottom-up fashion. Later on the architecture provides in the necessary top-down influences to direct the launching of agents in an appropriate way and to focus on the most relevant cues and structures given the context of the input image. The agents will be placed in the so-called '*Coderack'*, which is a waiting room filled with agents that will investigate possible structuring in the workspace and making probabilistic decisions. Agents are stochastically selected from the Coderack (the name 'Coderack' is taken from Copycat, where agents were called 'Codelets'). For example if we would not incorporate top-down pressures and in a particular case start and continue with a high portion of the agents being regular distance agents, one is bound to find regularity in the end, even though we might not perceive this regularity due to stronger structures in the context, which are not found because we only focused on finding regularity in the first place. Therefore we need to regulate the agents to be launched. If like in the previous case regularity is hard to find, then less agents need to be launched to search for this type of grouping, especially when another structure based on non-regular evidence is being formed.

The *Slipnet* is responsible for the top-down influences, which is a network of interrelated concepts, where each concept is represented by a node and is surrounded by potential associations and slippages (changing from one concept into another). Conceptual relationships represented as links have a numerical length, which resembles the 'conceptual distance'. Conceptual links in the Slipnet adjust there lengths dynamically as the conceptual distances gradually change under the influence of the evolving structure in the workspace. In the Slipnet each of the concepts can become active when instances of them are noticed in the workspace. Also agents can provide feedback to the workspace by creating a top-down pressure to look for further instances of active concepts. Furthermore, concepts can spread activation to their neighbours.

Building bonds and relations
On the workspace we distinguish two types of bonds: 'cue-bonds' and 'relation-bonds'. The cue-bond is proposed between two cues, like in the middle example of Fig. 6, where we have the three dot cues, from the left example of Fig. 6, and a bond represented by an

arrow that starts in the most right dot-cue and points to the middle one. What the bond represents is a local view from the right cue stating that it groups together with the middle cue, based on the grouping pressure that proposed and built the bond. In our model 'Proximity' is an example of a grouping pressure that constructs such cue-bonds. Some other grouping pressures, like regular distances, are not cue-bonds, but bonds among the distance relation between two cues and the distance relation between two other cues. If they have equal distances, then we can speak of regular distances. To express these groupings between two sets of cues we use the relation-bond. For example in the right dots-pattern of Fig. 6 we displayed a bond between two relations, where the two dashed lines between the first and second dot, and the second and third dot represent two relations, which are bonded by the grey pointing arrow. Distance and orientation are the two relations we used in our model. For a relation-bond to be built we first need to build the relations on the workspace.

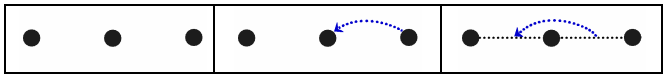


Figure 6. Left: three dot cues. Middle: Cue to Cue bond. Right: Relation to Relation bond.

The actual proposing and building of bonds is split into work for two different types of agents: *scouts* and *builders*. Scouts search the workspace for cues to bond, and the builder-agent builds the bond. Initially we launch 'Propose Bond'-scouts, which follow two rules:

1. Land on two or three cues.
2. IF fitting pressure description THEN Propose bond and put Bond Builder in the coderack ELSE terminate.

Next, when the Bond Builder is launched it acts as follows:

1. Check for resistance to bond. It is possible that existing bonds oppose building the proposed bond.
2. IF NO resistance THEN build bond ELSE fight: Resulting in either building or deleting the proposed bond.
3. IF proposed bond is built THEN post Bond Extender scout.

The Bond Extender scout on its turn does the following:

1. Lands on the bond to be extended.
2. Checks for extensions to propose new bonds.
3. IF proposing a new bond THEN put Bond Builder in the coderack ELSE terminate.

Now that we have explained the general bond building mechanism, we can move on to the specific grouping pressures.

Grouping pressures

We have implemented the following four grouping pressures, which are modelled after the Gestalt laws of visual perception: *Proximity, Regular-orientation, Good-continuation and Regular-distance*. It is important to remember that the strength of our model lies in the

combination of multiple grouping pressures and their mutual supporting evidence, and not so much in any of them in isolation. Different grouping pressures that propose similar groupings, provide more (redundant) evidence for a coherent strong structure. Next we will describe each implemented grouping pressure in more detail including their grouping results on our alignment example from Fig. 4.

Proximity

The Proximity scout proposes and builds bonds between cues based on the distance between cues. The purpose of the Proximity agent is to bond each cue from a local perspective to other cues which are the closest. When we land with our proximity scout on a dot (cue) we take this cue as the centre point of a circular search zone for which we make a list of all the cues within this zone. For each cue we find in our search zone we calculate the Euclidean distance to the centre cue, and use these distances to set up a probability for being a candidate for a proximity bond. The shorter the Euclidean distance the higher the chance the cue will be chosen to build a proximity bond. Strong proximity relationships are between those cues that both have proximity bonds that point to each other (two-way proximity), which shows that from the local perspectives of each of the two cues the other cue is proximate.

The results on alignment examples are displayed in Fig. 7 after 100 scouts were launched onto the workspace. In the left result we can see many two-way proximity bonds including bonds between the 8 dots (see Fig. 9) that form the visible line among 21 uniformly randomly distributed dots. Only the two bottom-left dots are not bonded together due to another (random) dot that lies really close to the alignment. This suggest that there is proximity evidence for grouping some parts of the line together, but solely on proximity one would not perceive the line. It is interesting to see (although quite messy) that based on proximity the alignment in the right result of Fig. 7 has no support whatsoever and is totally disturbed by interfering close by random dots, which is just as one would expect to see based on proximity.

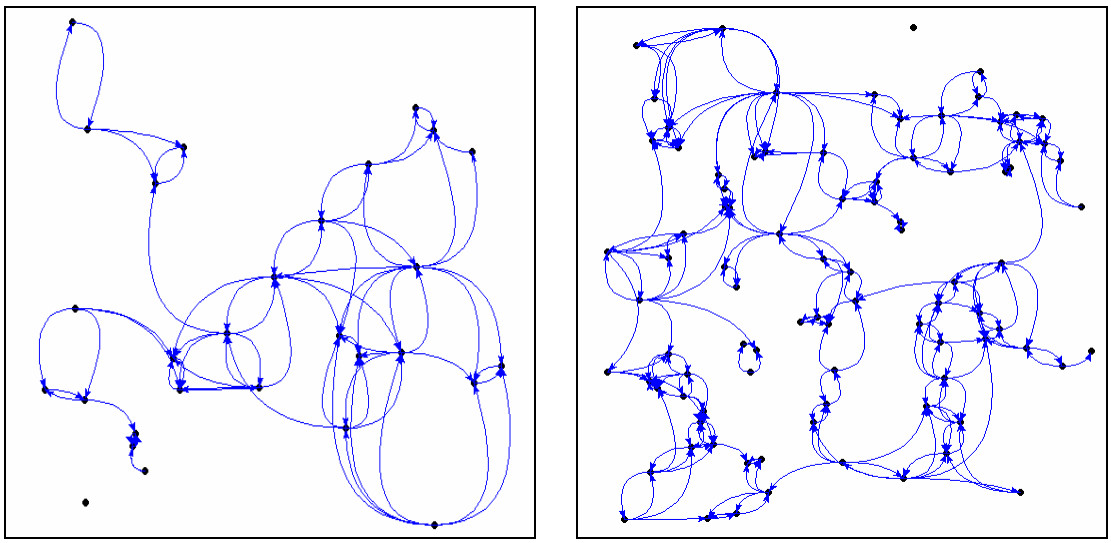


Figure 7. Left: Proximity bonds on 21 uniformly randomly distributed and eight aligned dots. Right: Proximity bonds on 81 uniformly randomly distributed and the same eight aligned dots.

Regular-orientation

The Regular-orientation scout proposes and builds bonds between orientation-relations that have the same orientation and share one cue, which essentially means a straight line through three dots. We explain how the Regular-orientation scout operates by the use of the example given in Fig. 8, where we initially start with three dots (leftmost dot-pattern).

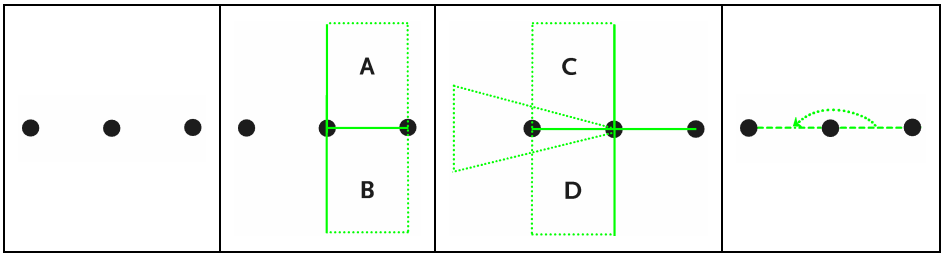


Figure 8. Leftmost: three initial dots. Middle left: free zone check. Middle right: free zone check and triangle search zone. Rightmost: Regular-orientation bond.

First the agent lands on a random dot cue, which is in this example the dot (cue) to the far right and we take this cue as the centre point of a circular search zone for which we make a list of all the cues within this zone. The diameter of the search-zone should be sufficiently large, not to exclude the finding of lines over large distances. We find two cues and for both cues we calculate the distance to the rightmost cue, and use these distances to set up a probability for being a candidate for the second dot on the line. The shorter the distance the higher the chance the cue will be chosen (just like we did with the proximity scout). Say we choose the middle point as the second dot. Now we check for free zones that need to be free of interfering dots, illustrated in the middle left example by two rectangular zones (A and B). The zone's width is proportional to its length, which is the distance between the first and second dot. If both A and B would contain any dots, then the scout will terminate. On the other hand if only one of them includes a dot or if they are both free of them, then we continue the search for a third dot. The reason why we introduce the concept of free zones, is that it helps to home in on 'clear' lines by avoiding dense clusters of dots. In search for the third dot the scout constructs a triangle search zone in the direction from the first to the second dot, starting from the second dot (see middle right example). The length and width of the triangle are proportional to the distance between the first and second dot. From all the dots found in the triangle zone we calculate the distances to the middle cue and use these distances to set up a probability for being a candidate for the final third dot on the line. In our example we find only one dot, and also here we check for interfering dots between the second and third dot, just like we did between the first and second dot with rectangular zones (C and D). We have three conditions under which we abort proposing a regular orientation bond, because under these conditions both sides of the alignment would have interfering dots:

- If there are cues in rectangle A and D.
- If there are cues in rectangle B and C.
- If there are cues in rectangle C and D.

If none of these conditions apply then the scout proposes to bond the orientation relation between the first and second cue, and the same relation between the second and third cue as shown in the rightmost example of Fig. 6.

The results of the regular orientation scout on alignment examples are displayed in Fig. 9 after the actions of 100 scouts on the workspace. In the left result we can see that the scout for this grouping pressure easily finds and groups the alignment of dots together. Additionally it finds even more dots that form other alignments. These alignments are correct. However, when we look at the total dot pattern, we are not drawn to these other alignments and will not mark them as something interesting. This is just one opinion of one type of agent, which unless it is supported by any other grouping pressure remains a weak structure. In the right result of Fig. 9 we see that the found alignments are all over the place and none seem to fit the 'hidden' 8-dot alignment, which matches expected human visual grouping in this particular example.

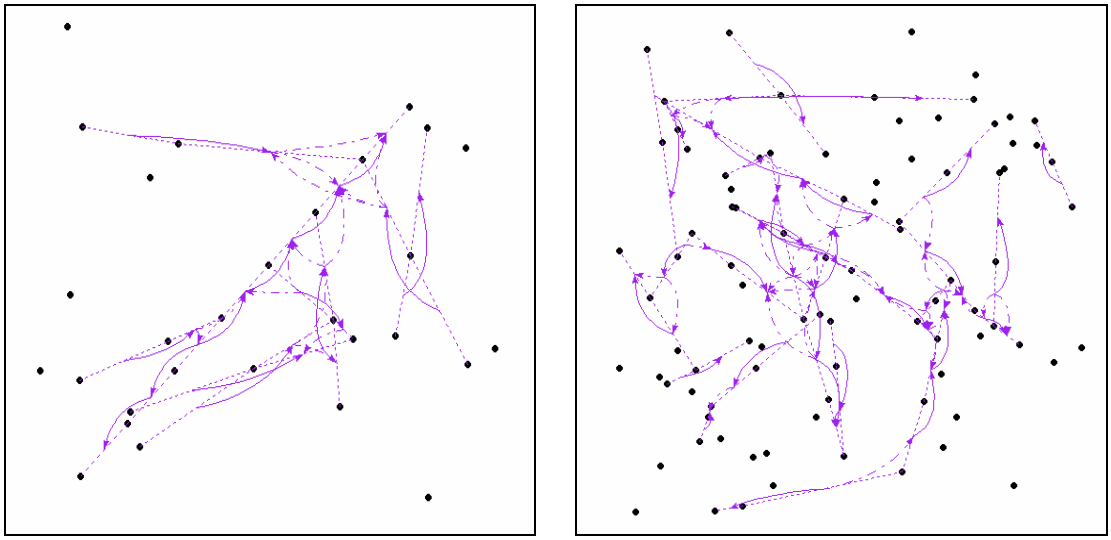


Figure 9. Left: Regular-orientation bonds on 21 uniformly randomly distributed and eight aligned dots. Right: Regular-orientation bonds on 81 uniformly randomly distributed and the same eight aligned dots.

Good continuation

The Good-continuation scout works in an almost identical way as the Regular-orientation scout, working also with orientation relations. Only where the Regular-orientation scout spots straight lines, the Good-continuation scout finds the best continuation of a line, which could be slightly curved. For this behaviour the scout follows the same steps as the Regular-orientation scout, only allows the triangle search zone for the third dot to be much wider and has a different selection criterion for the best candidate dot in the triangle zone. The selection criteria is no longer based on being nearer to the second dot (the point where the triangle cone begins), but based on best fitting of the orientation between the first and second dot. The results of the good-continuation scout are presented in Fig. 10 and resemble the results of the regular-orientation scout, with only minor differences.

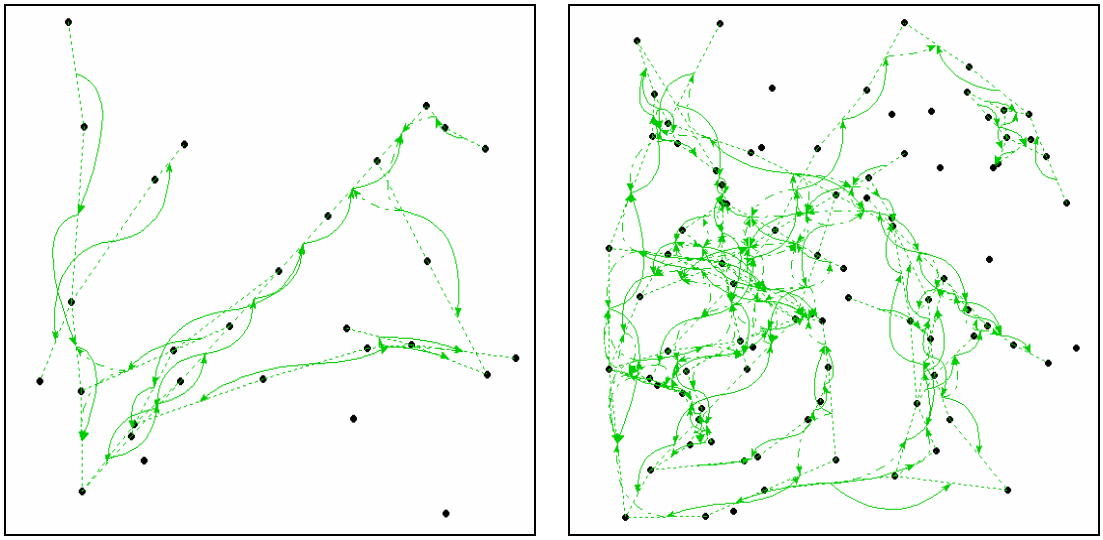


Figure 10. Left: Good-continuation bonds on 21 uniformly randomly distributed and eight aligned dots. Right: Good-continuation bonds on 81 uniformly randomly distributed and the same eight aligned dots.

Regular distance

Finally the fourth scout we have implemented, the Regular-distance scout tries to bond cues together that have the same inter distance. With the help of the example in Fig. 11 we will demonstrate how this agent operates.

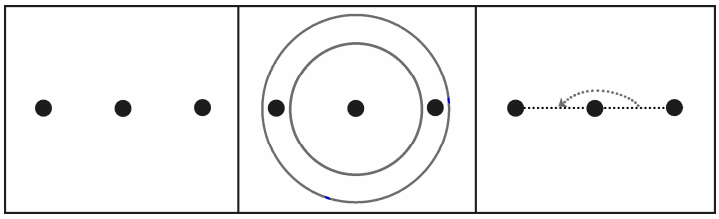


Figure 11. Left: three dot cues. Middle: margin space. Right: Regular-distance bond.

First we land on a random cue in the workspace, which in this example is filled with three dot cues (leftmost dot-pattern). The agent lands on the middle cue and we use this dot as the centre point of a circular search zone for which we make a list of all the cues within this zone, and find two other cues (the leftmost and the rightmost).

For both found cues we calculate the distance to the middle cue, and use these distances to set up a probability for being selected for the second step. The shorter the distance the higher the chance the cue will be chosen. Say we would have chosen the far right cue to perform the second step of the agent, which is finding other cues that have the same distance to the middle cue. We make a list of all the cues with the same distance, given a small error margin, illustrated in the middle figure by two circles. In our example we find the leftmost dot within the margins. The next step the scout proposes to bond the distance relation between the middle and rightmost cue, and the same relation between the middle and leftmost cue as shown in the rightmost example of Fig. 11.

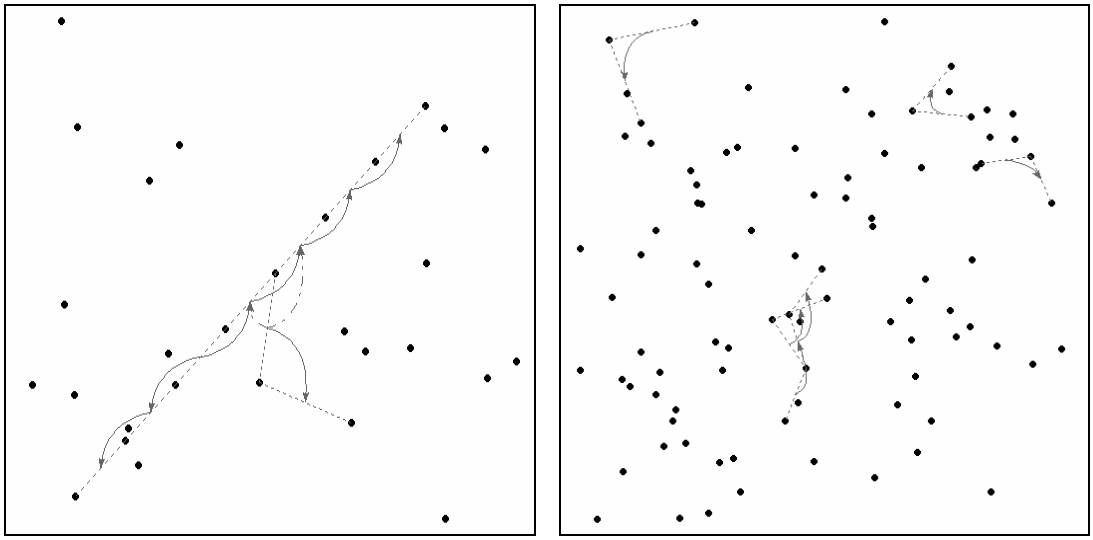


Figure 12. Left: Regular-distance bonds on 21 uniformly randomly distributed and eight aligned dots. Right: Regular-distance bonds on 81 uniformly randomly distributed and the same eight aligned dots.

The results of the Regular-distance scout are presented in Fig. 12, and just like with Regular-orientation and Good-continuation, this scout flawlessly discovers the alignment and this time it is almost the only thing it finds apart from one other bond. Furthermore, as expected the scout finds none of the 'hidden' 8-dots in the Bottom example.

Joint Grouping pressures

Fig. 13 depicts the combined results of the Regular-orientation, Good-continuation and Regular-distance grouping pressures. The proximity grouping pressure was left out for clarity. The alignment of mutually supportive grouping pressures can be clearly seen in the left result of Fig. 13. From this example, the advantage of using mutually supportive grouping pressures is contrasted with the interpretation power of each grouping pressure on its own. Consequently, only those bonds supported by multiple grouping pressures may lead to the formation of higher level structures.

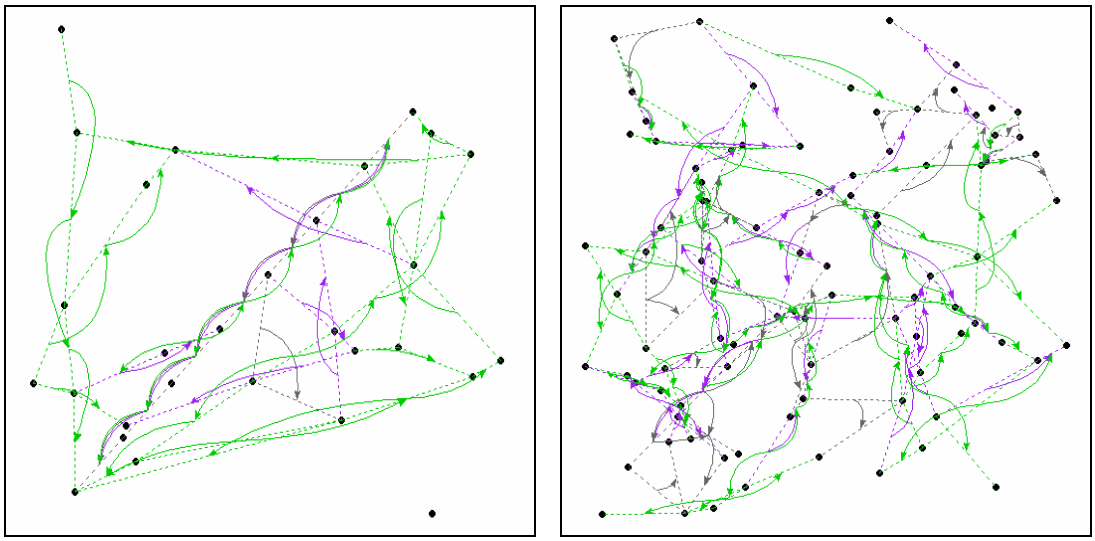


Figure 13. Combined results where purple-bonds = regular orientation, green bonds = good continuation, grey-bonds = regular distance. Left: Combined results on 21 uniformly randomly distributed and eight aligned dots. Right: Combined results on 81 uniformly randomly distributed and the same 8 aligned dots.

Future work

A plan for a second implementation phase has been devised for taking the computer program closer to the proposed theoretical model both by extending the cue and agent repertoire and by implementing higher-level capabilities. In addition, the visual perception model together with the Ear's Mind is used as a template for implementing an audiovisual fusion model for multimodal perception (see [Stevens et al., 2007]). Such an audiovisual model is expected to enhance the capabilities of real-world scene segmentation in comparison with single modality models. A working model for audiovisual perception can later be augmented and combined with other input data, which is foreign to human perception, like for instance infrared and echolocation. Our model will finally be used to construct an autonomous and adaptive surveillance system based on a video and acoustic sensor. Applications of such a system are harbour protection and battlefield surveillance (detection and recognition of friend or foe).

Conclusions

The aim of our research is to design and implement an autonomous and adaptive (context sensitive) surveillance system based on a video and acoustic sensor. In our approach we try to implement a working model of human audiovisual perception, because we believe that understanding and modelling human perception is at the crux of making intelligent context sensitive systems that try to make sense out of an overwhelming amount of data (e.g., smart surveillance systems). Our initial goal, which we described in full detail, was to focus on visual perception and to implement a working model of primitive visual perception, integrating top-down influences with bottom-up processing, and mimicking perceptual grouping behaviour of human subjects. We proposed the visual perception model as a novel emergent, self-organizing model, supported by neurological and psychological evidence. The model consists of an open architecture allowing the addition of new features, pressures and interaction methods, making it possible to define more agents and extend the model's capabilities. Following the design phase, the model was implemented as a software prototype, and was used for testing Proximity, Good-continuation, Regular-orientation and Regular-distance grouping pressures. Results so far show that the implemented model forms a promising foundation for further research and expansion for dealing with more complex images.

References

Bregman, A.S. (1990). *Auditory Scene Analysis: The Perceptual Organisation of Sound,* 2nd paperback ed. 1999, MIT Press, Cambridge.

Chalmers, D.J., French, R. M., and Hofstadter, D. R. (1992) High-Level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology. *Journal of Experimental and Theoretical Artificial Intelligence* 4, 185–211.

Desolneux, A., Moisan, L. and Morel, J.-M. (2003) Computational Gestalts and Perception Thresholds. *Journal of Physiology-Paris* 97, 311–324.

Dor, R. (2005) *The Ear's Mind: A Computer Model of the Fundamental Mechanisms of the Perception of Sound,* Technical report 05-16, Delft University of Technology.

Dor, R. and Rothkrantz, L.J.M. (2008) The Ear's Mind: An Emergent Self-Organizing Model of Auditory Perception, Submitted to the *Journal of Experimental and Theoretical Artificial intelligence*. In press.

Hofstadter, D.R. and FARG (1995) *Fluid Concepts and Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought*. Basic Books, New York.

Mitchell, M. (1993) *Analogy-Making as Perception: A Computer Model*. MIT Press, Cambridge.

Riesenhuber, M. and Poggio, T. (2000) Models of Object Recognition. *Nature Neuroscience* 3, 1199–1204.

Stevens, J.C., Dor, R. and Rothkrantz, L.J.M. (2007) Boiling Down Emergent Self-Organizing Soups to Solid Multimodal Perception, *Proceedings of Euromedia 2007*.

Stevens, J.C., Dor, R., Hupkens, Th.M. and Rothkrantz, L.J.M. (2008) Modelling Grouping Pressures for Emergent and Self-Organizing Visual Perception, *Proceedings of Euromedia 2008*.

Wertheimer, M. (1923). Laws of Organization in Perceptual Forms, *Psychologische Forschung* 4, 301-350.

Rapid Environmental Assessment System: Concept, Geoacoustic Inversion and At-Sea Experiments

Frans Absil & Jean-Pierre Hermand

Introduction

Maritime operations lead to an increasing focus on environmental effects in littoral areas and coastal zones. Shallow waters with water depths below 200 m, where amphibious operations are likely to take place and military systems will be deployed, have special characteristics. Typically, the water depth is rapidly changing and the physical parameters of the environment, including sea surface, water column, seafloor and sub-seafloor, exhibit a strong temporal and spatial variability. Seasonal, morning and afternoon effects in the water column may affect the sound velocity profile. Local variations in the bottom properties, such as the thickness, density, compression/shear sound speeds and attenuation of unconsolidated sediment layers and the sub-bottom are common. The environmental properties and therefore the frequency-dependent properties of the acoustic propagation medium are likely to vary from one area to the next, just a few tens, hundreds, or thousands of meters away. In these areas there is often intense human activity, with many man-made objects and noise sources in the water, or on the bottom.

Environmental parameters characterising the water column and the bottom, referred to as the set of *geoacoustic* parameters, will obviously affect the performance of military sensor systems in Anti-Submarine Warfare or Amphibious Operations. Predicting sonar performance during deployment is important, and adaptive sonar signal processing may yield significant improvements in situational awareness. This is not straightforward since the propagation characteristics are complex in a (very) shallow water acoustic waveguide where the interaction with the boundaries controls the propagation. Knowledge of the geoacoustic parameters is crucial to sensor deployment as part of a military operation. Obviously, in well-known operational areas there will be regular surveys and monitoring of the environment. A database, charting the temporal and spatial variability, will be made available; this is a standard task for hydrographic services around the world.

However, in the framework of the European Defence and Security Policy (EDSP) there is an increasing demand for rapid characterization of the environment in less-known and unknown shallow water areas. The word “rapid” implies that weeks to days ahead of an actual military operation one would like to do a brief survey of the relevant area. That should yield the environmental parameters that are passed on to the operational commanders who will then deploy their sensors and know what to expect from the sonar systems during the operation. The process of quick medium characterization is called *Rapid Environmental Assessment* (REA), as part of the process of *Battlefield Preparation* (BP). Preferably, this should be done as a covert operation, with a minimum set of equipment, without the need for having big ships or other military platforms in the area, and with a near real-time processing cycle for the data.

A number of REA system concepts have been considered. Many configurations are possible but the illustration in Fig. 1 shows one such approach. It shows a number of

system components in the littoral zone; any ship, helicopter or aircraft might deploy the single sound source and the set of drifting acoustic-oceanographic buoys. These will acoustically monitor the environment for a given period and transmit their data via telemetry to a monitoring station (possibly relayed via aircraft or satellite). At the monitoring site the acoustic data is used and processed to yield the geoacoustic parameter set. In scientific terms this is called *geoacoustic inversion*; this technique will be discussed in more detail below.

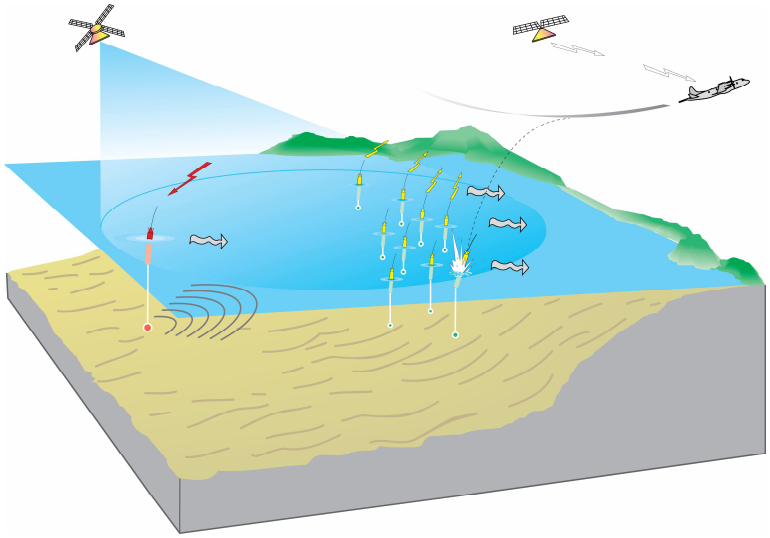


Figure 1. Shallow water Rapid Environmental Assessment (REA) system concept based on a single sound source (left), a set of drifting acoustic-oceanographic buoys (centre) and telemetry of data to a remote monitoring station such as a satellite or aircraft (top).

The system components shown in Fig. 1 are representative of specific system functions; there is a sound transmitter plus a number of receivers. Obviously, one may think of alternative realisations of these functions. The sound source might be a passing surface ship or even ambient noise in the littoral zone. The receiver hydrophones might be part of a horizontal towed array behind a ship, or of a moored vertical array. Instead of the drifting source and receivers one might deploy an Autonomous Underwater Vehicle (AUV) with sensor, navigation and communication sub-systems; in a distributed systems concept there might be a coordinated group of such vehicles. However, there is a definite advantage in limiting the set of system components. Therefore, the concept of a *sparse* set of receivers is highly relevant, and from a systems performance point of view the feasibility of REA with a sparse set of hydrophones will have to be investigated.

In 2002 the Royal Netherlands Naval College (RNLNC, now part of the Netherlands Defence Academy) started a research program on the topic of REA with a sparse set of acoustic-oceanographic sensors. The research focused on:

- the development of an effective and efficient geoacoustic inversion scheme with a novel approach;
- acquisition of experimental data at sea for validation of the REA geoacoustic inversion approach; and
- optimisation of the search strategy for geoacoustic parameter determination.

A first approach to the segmentation problem was presented in [13], using modern time-frequency methods, such as Gabor atoms to segment the downrange domain for a set of drifting receiver buoys. This limited investigation did not yield a convincing outcome; segmentation was most critical to algorithm tuning and tentative interpretation.

This paper discusses the principle of geoacoustic inversion. That will demonstrate the need for validation with experimental data. The RNLNC has been involved in two recent sea trials that will be described in brief. Results for the geoacoustic parameter characterization for typical examples will be shown and discussed. Finally, the paper presents conclusions, future recommendations and references for further reading. This NLARMS volume contains another paper about the REA research project (see contribution by Van Leijen).

Geoacoustic inversion and backpropagation techniques

The principle of geoacoustic inversion is shown in the diagram in Fig. 2 [1, 2]; in clockwise orientation the essential steps are shown. Remember that the goal is to obtain an estimate for the set of geoacoustic parameters shown in the lower part of the upper left plot: the sound velocity profile $c_1(z)$ in the water column (where z is the depth coordinate), the density d_2, sound speed c_2 and gradient g_2 and sound attenuation α_2 in the sediment layer, and the density d_3, sound speed c_3 and attenuation α_3 in the sub-bottom.

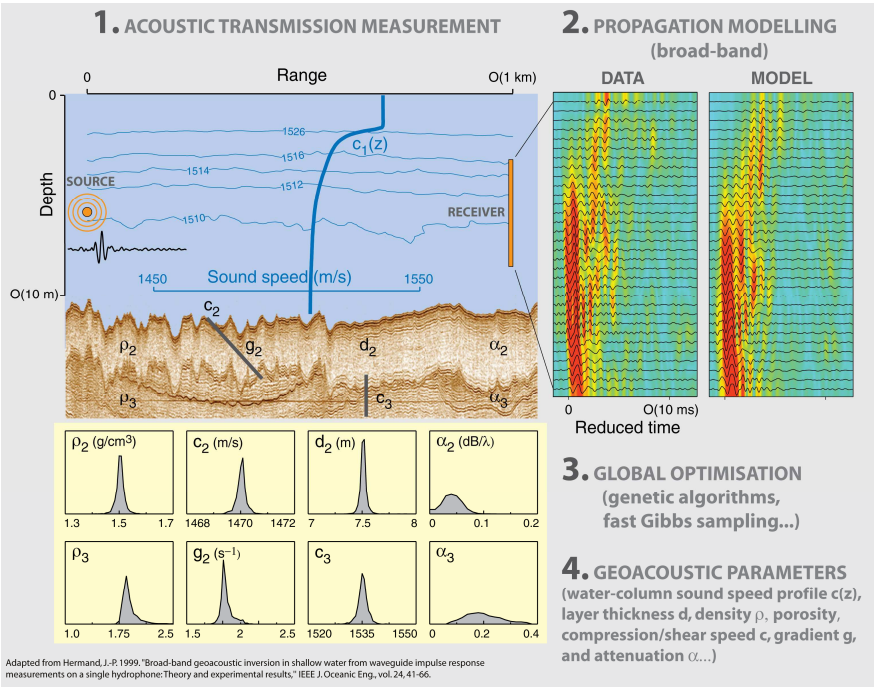


Figure 2. Principles of geoacoustic inversion using a broadband matched-filter approach. The essential steps that yield the set of geoacoustic parameters from the acoustic measurements are shown in clockwise sequence.

Step 1 involves the acoustic experiment: the sound propagation from the source to the receiver (the figure shows a vertical hydrophone array) in the shallow water acoustic waveguide is measured. In **Step 2** the measured data is compared with a receiver data replica field obtained with an acoustic propagation model. This propagation model

requires an initial guess for the geoacoustic parameter set that represents the environment. The acoustic wave equation, either in the exact or some approximate form, will be solved at multiple frequencies. The *modelling over a broad range of frequencies* is an essential aspect of the current geoacoustic inversion approach, in combination with a sparse set of hydrophones. In **Step 3** the difference between the model and real data is used as input to a *backpropagation* scheme that will yield update information for the geoacoustic parameters. Scientifically, the search for the set of environmental parameters that will minimize the mismatch between the measured and the acoustic propagation model data corresponds to an *optimisation* problem.

Backpropagation, as shown in diagram in Fig. 3 [12], does the search by iteratively updating the set until the final solution is reached in **Step 4** (in this case a set of parameter value distributions in the lower left of Fig. 2). In the diagram the initial guess of the set of geoacoustic parameters is represented by the vector γ. With that initial estimate the *forward model* determines the acoustic propagation from the source ($r = 0$, where r is the range) to the receiver ($r = R$) for a given geometry; the resulting sound pressure field at the receiver hydrophone set is represented by $u(\gamma; R, z)$ (note the multiple frequency, broad-band approach).

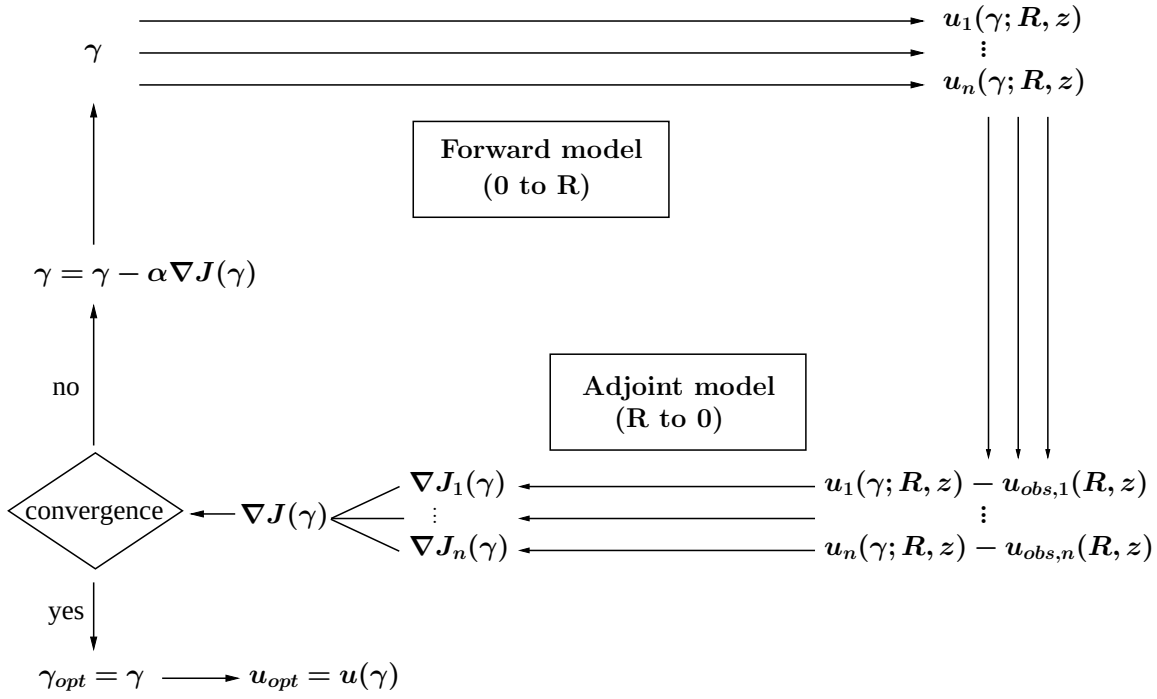


Figure 3. The backpropagation scheme for geoacoustic inversion. The gradient of the cost function J controls the search for the set of geoacoustic parameters γ that will minimize the difference between the model and measured data $u–u_{obs}$ at the receiver array. Shown in the figure is the broad-band, multiple frequency approach to both the forward and the backpropagating adjoint model [12].

The mismatch between the model and measured field at the receiver $u_i - u_{obs,i}$ is then backpropagated from the receiver to the source using the *adjoint model* (see the next section), yielding a gradient of the cost function $J(\gamma)$. The cost function weighs the quadratic difference between the model and measured field on the receiver array over the multiple frequencies. It may contain additional *regularisation* terms to control the

character of the solution. The cost function gradient is used to update the set of environmental parameters (note the term $\alpha\nabla J(\gamma)$ in the figure, with α a tuning parameter) and an *iterative optimisation scheme* is obtained, as shown by the clockwise loop in the figure, until after a number of iterations the final estimate γ_{opt} is reached.

The adjoint method of optimal control for REA

This section discusses the elements and techniques in the geoacoustic inversion scheme, and summarizes the salient features of the PhD thesis work of Meyer [12], as part of the RNLNC REA research project. Consider the separate blocks from the diagram shown in Fig. 3:

- The forward model is a parabolic-type approximation to the acoustic wave equation. The Wide-Angle Parabolic Equation (WAPE) generates the replica model of the receiver data for the range-independent shallow water acoustic waveguide. It is demonstrated to give a sufficiently accurate propagation pattern, when compared with a more complex and computationally demanding full-field coupled normal model, such as Kraken-C [14]. The multiple frequency approach involves 7 different source frequencies between 200 Hz and 800 Hz.
- The sound speed profile in the water column is represented in varying detail, depending on the type of inversion. In ocean acoustic tomography the sound speed profile is determined, and one type of representation used in this research is through Empirical Orthogonal Functions. In geoacoustic inversion the range-independent interface between the water column and the bottom is represented by a non-local boundary condition [11], which means that at a certain position the solution is affected by the boundary parameter values further upstream and downstream.
- The gradient of the cost function is determined by backpropagating the residual sound field, i.e., the mismatch between the measurements and the replica, from the receiver to the source. Backpropagation is based on an adjoint model approach, an optimal control method [7, 10] used in various fields such as meteorology, geophysics, fluid dynamics, but also in missile guidance performance evaluation. Application of the adjoint method to underwater acoustics is fairly recent. In this research a semi-automatic adjoint generation [5] via a modular graph approach [15] for the WAPE has been used.
- The gradient approach for updating the estimates of the set of geoacoustic parameters has been compared with other search techniques. Iterated Local Search and population-based meta-heuristic techniques, such as Genetic Algorithms and Ant Colony Optimisation [8] were evaluated for the same shallow water scenarios.

A meta-heuristic can be seen as a general algorithmic framework which can be applied to different optimisation problems with relatively few modifications to make them adapted to a specific problem. The classical, meta-heuristic *global search* algorithms, typically used for inverse problems in oceans acoustics, iteratively update the model by introducing random variations of the control parameters (the elements in the vector γ) and in doing so can move uphill in the cost function to escape from local minima. However, for complex environments and models or for large (possibly correlated) control parameters sets they

require a huge number of modelling runs and are relatively inefficient especially moving downhill, e.g., near convergence and for correlated parameters.

The adjoint approach is a complement or alternative to the traditional inversion methods in the sense that it provides a mechanism of optimisation that is directly based on and controlled by the underlying physics of a shallow water waveguide, provides gradient information (i.e., produces corrections to the respective model inputs that caused the mismatch between the observations and model predictions). It belongs to the category of *local* methods since it is gradient-based and significantly reduces the number of modelling runs.

In the thesis various realistic shallow water scenarios were considered, demonstrating the feasibility of the approach. These scenarios were either based on the geometry and conditions during the 1994 Yellow Shark (YS'94) sea trial, for which detailed ground truth is available, or on recent experimental data from the MREA/BP'07 trial (see the next section). A performance analysis is presented, when balancing the number of hydrophones of the vertical receiver array against the broad-band, multiple-tone approach. Convergence of the solution is considered, i.e., does the final iterative solution for the sound speed profile or the geoacoustic parameter set approach the ground truth, and in how many iteration steps was this solution achieved? Also, in many cases a clear hierarchy was found where some parameters (depending on the scenario) converged before others. Also, in one scenario, the shallow water time-variability of the sound speed profile over 48 hours was studied, demonstrating the tracking potential of this inversion approach.

Since validation with experimental data has always been an important component in the RNLNC REA research program, two recent sea trials will be described briefly.

The sea trials

The RNLNC was involved in two recent sea trials: Saba'06 (RNLN only) and MREA/BP'07 (multi-national initiative). Both were organized in close cooperation with the Netherlands Hydrographic Office (NHO), and were carried out with participation of a highly modern ship from the Hydrographic Service of the Royal Netherlands Navy, HNLMS *Snellius*, shown in Fig. 4.

The Hydrographic Survey Vessel (HSV) is equipped with an impressive sensor and systems suite: a hull-mounted Search Light Sonar (FURUNO CSH5), a towed high resolution high speed Side-Scan Sonar (KLEIN 5500, 455 kHz), GPS navigation (Thales Aquarius 02, dGPS, EGNOS/WAAS, LRK), a towed magnetometer (Marine Magnetics SeaSPY), a moving vessel Sound Velocity Profiler (BOT - MV P100, SV, T&P), a Single Beam Echosounder (Kongsberg Simrad EA 600, 38, 12, 200 kHz), a Multibeam Echosounder (Kongsberg Simrad EM 3000D, 300 kHz), a Navigational Echosounder (Kongsberg Simrad EN 250, 38 kHz), a single sweep system (Seatools Ultra Short Baseline System Sonardyne), and miscellaneous datalogging and processing equipment (QINSy and ISIS Sonar SSS). The following sections will provide an overview of the experimental set-up.

Figure 4. HNLMS Snellius, hydrographic survey vessel (length: 82 m*, beam: 13.1* m*, draught: 4* m*, displacement: 1875* tons*, speed: 12* kts*, built: 2003, propulsion: diesel-electric 1250* kW*, crew: 18)*

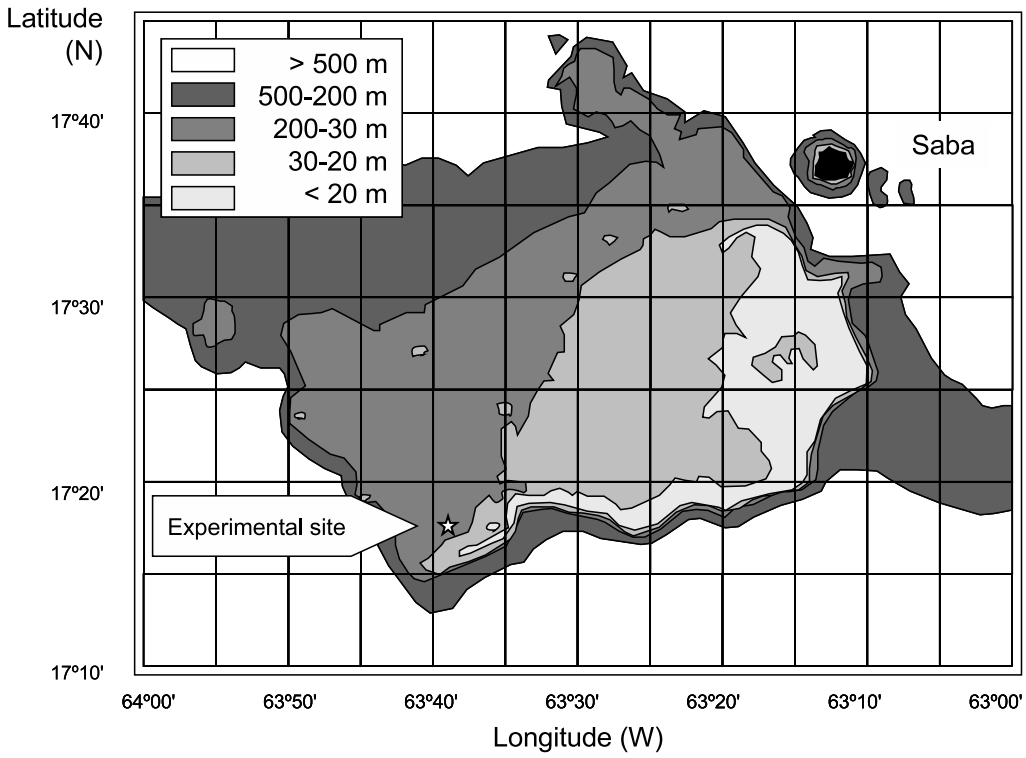


Figure 5. The geographic location of the island of Saba (top right) and the position of the Saba bank (light grey areas to the southwest of the island).

Saba'06

During the spring of the year 2006, hydrographic survey vessel HNLMS *Snellius* ran an extensive bathymetric survey on the Saba bank, a large submerged atoll located in the north-eastern Caribbean, see Fig. 5. The survey provided an excellent opportunity for a number of small-scale geoacoustic experiments in a shallow water environment. The feasibility of a rapid deployment of ocean-acoustic sensors and equipment was demonstrated for the purpose of an environmental assessment of the area southwest of the small volcanic island of Saba.

The aim of the Saba bank environmental experiments was to use a sparse setup with as few as four or five receivers. This approach reduces the large quantities of data that are recorded with dense arrays. It therefore significantly reduces the time that is needed to pre-process the data and start an inversion process that will yield the geoacoustic properties of the seafloor and sub-seafloor.

The environmental impact was kept to a minimum by exploiting the hydrographic ship as a sound source of opportunity. It was moving away from a light, sparse vertical array deployed from a rubber boat at anchor. During the morning of April 24, HNLMS *Snellius* sailed along the array in a cooperative mode on a pre-defined track with a constant speed and bearing, yielding systematic logging of accurate DGPS positions that allows for a proper reconstruction of the experimental geometry. Five tones from the diesel generator (115.5, 209.4, 269.1, 329.1 and 706.8 Hz) were selected for the geoacoustic inversion.

The acoustic array receiver data were recorded on board a small rubber boat on a digital multi-channel recorder. The pressure and temperature in the water column were measured from the rubber boat using a thermistor string and later combined with salinity data to obtain sound velocity profiles. Collected data was transferred to commercial laptop computers and processed on board of the HSV. The experiment demonstrated that a small scale REA campaign can be launched and that the geoacoustic inversion process can be completed within a 24-hour timeframe.

MREA/BP'07

The purpose of the MREA/BP'07 sea trial, in April-May 2007, south of Elba Island in the Mediterranean Sea, was a multi-disciplinary experimental effort that aimed at addressing the Battlespace Preparation (BP) concept [6]. The focus was on the establishment of an integrated 4D (3-dimensional space and time) Recognized Environmental Picture (REP) of a shallow water environment in support of two types of maritime operations: Anti-Submarine Warfare (ASW) and Amphibious Operations. For this purpose, several standard and advanced techniques of environmental characterisation covering the fields of underwater acoustics, physical oceanography and geophysics have been combined within a coherent scheme of data acquisition, processing and assimilation. Details are given in the MREA/BP'07 Sea Trial cruise and data reports [3, 4].

The BP'07 sea trial was part of a broader, multi-national Maritime REA (MREA) initiative that NATO Undersea Research Centre (NURC) coordinated in the Ligurian Sea in 2007, see Fig. 6. The experiment has benefited from the oceanographical and bathymetric surveys that have been conducted by the Italian Navy (Instituto Idrografico della Marina) during and after the BP'07 time frame. In addition, external efforts, mainly in oceanographic modeling and satellite remote sensing techniques with NRL Stennis Space Center (NRLSSC), SHOM, GHER, ULB/LOCEAN/LAMFA, MIT/Univ. Harvard, INGV, ARPA and Meteo France have contributed to the BP'07 experiments.

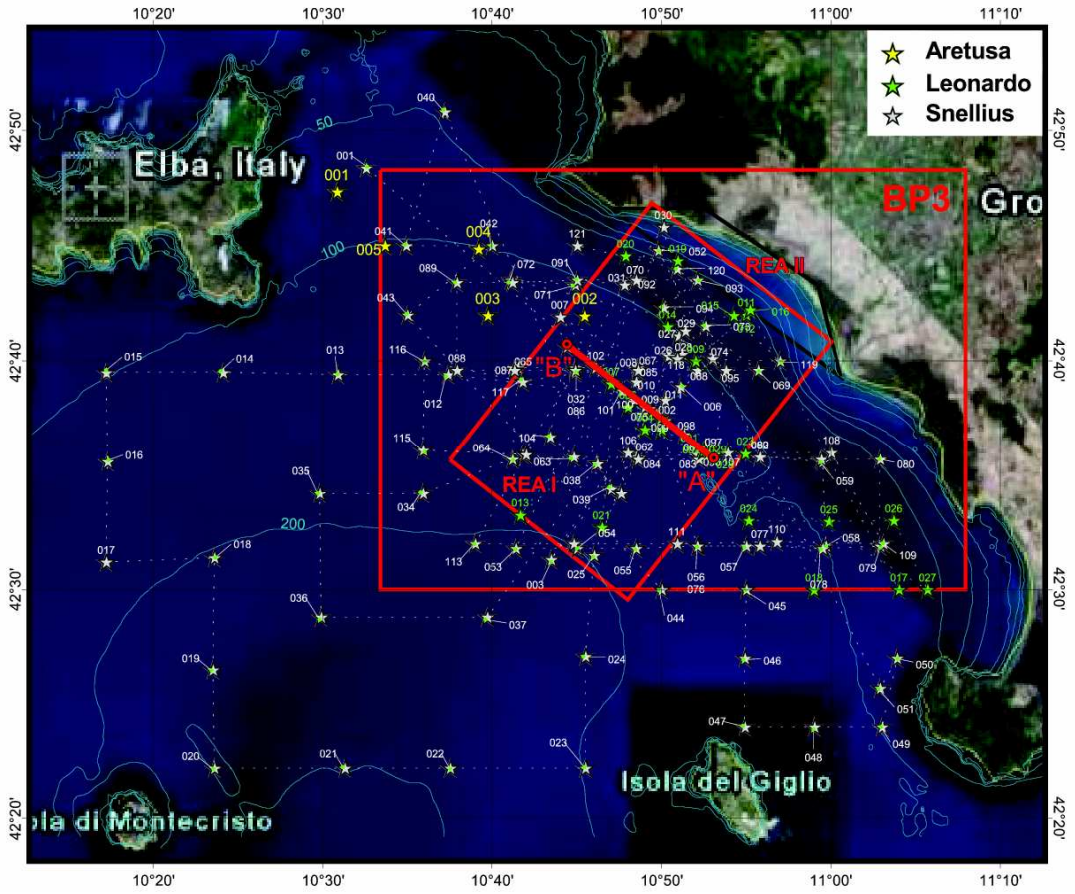


Figure 6. Geographic location of the MREA/BP'07 sea trial. The boxes REA I and REA II are the two main experimental areas. The transect A-B where part of the MREA/BP'07 geoacoustic inversion runs were carried out is the same as for the Yellow Shark '94 experiment. CTD locations are indicated (see [4] for details).

The following vessels were involved: NRV *Leonardo* (NATO), HNLMS *Snellius* (NL) and ITS *Galatea* and *Aretusa* (IT). In addition to these assets, the Marine Mammal Mitigation policy was applied during the acoustic experiments with the support of the Centro di Richerche sui Cetacei (CE.TU.S.) and its RV *Krill*.
Here is an overview of the type of measurements and models used during the trial:

- Seabed characterisation. This included a number of systems on board HNLMS Snellius: GPS, yielding National Marine Electronics Association (NMEA) position strings in the WGS84 ellipsoid and GMT, bathymetry with a Kongsberg Maritime EM3000-D multibeam echo sounder, seismic survey with Uniboom broadband geophysical source from NURC and an EdgeTech X-Star full spectrum digital sub-bottom profiler from TNO (Netherlands Organisation for Applied Scientific

Research), side-scan sonar imagery (Klein 5500), seabed classification with the bi-frequency (12 kHz and 36 kHz) Kongsberg Maritime single beam echo sounder (SBES) EA 600, superficial sediment samplings with a Hamon grabber.

- Estimation of geoacoustic properties. These included both active and passive runs. During the active runs NRV Leonardo deployed a sound source emitting LF (300–800 Hz, ping duration 3 s or 5.8 s) and MF (800–1600 Hz, duration: 1 s or 5.8 s) chirp and multi-tone signals. Data were recorded with a rubber boat or on two drifting buoys with sparse vertical line arrays deployed from HNLMS Snellius, at typical distances of 1–2 km. On the passive runs NRV Leonardo acted as a sound source of coincidence, with typical CPA's of 100-300 m. On two days RNLN AUV REMUS was used for a passive run.
- Water column properties. These were both measured and modelled.
    - For the in situ measurements, two types of Conductivity, Temperature, and Depth (CTD) sensors were used, with the second providing fluorimetry, sea water clarity and equipped with a rosette for water sampling. HNLMS Snellius deployed a Moving Vessel Profiler (MVP) free falling temperature sensor. In addition, NRV Leonardo deployed two thermistor strings. A Datawell directional waverider was deployed approximately in the middle of transect AB. Remote sensing data was also provided: NRLSSC, together with NURC, delivered AVHRR data from the NOAA12, NOAA14, NOAA15, NOAA18 satellites. The full period of the experiment was covered by those data. SST and cloud analysis were made available on the GEOS server.
    - The modelling efforts during the sea trial involved three main objectives: net-centric oceanographical forecasts linked to adaptive sampling strategies and super-ensemble predictions by NRLSSC and NURC at two resolutions (2 and 0.6 km grid size), real-time oceanographical forecasts onboard HNLMS Snellius by TNO (a 2-way nested high resolution HOPS model in the areas BP1, resolution 0.6 km, and BP3, resolution 0.3 km), and wave forecasts by NRLSSC/NURC. In order to support those efforts, global models and/or forcing fields were made available by MREA'07 partners: ALADIN wind forcing by SHOM, MFS/OPA oceanographical forecasts (~ 7 km resolution) by INGV.

Both trials have yielded enormous quantities of high-quality experimental data. At sea preliminary, short term data processing was performed, demonstrating the REA concept and providing the operational command with a 4D REP.

Results and discussion

The following sections will present a number of examples of water column and geoacoustic inversion.

Two adjoint-based inversion examples with synthetic and YS'94 data

The first example presents the sound pressure field for the WAPE model with NLBC in a shallow water environment (see Fig. 7). The synthetic true field is calculated for an isospeed water column with $c = 1520$ m/s, water depth $H = 135$ m, on a 512×512 grid area, over a hard reflecting bottom (sand, $c_b = 1600$ m/s, $\alpha_b = 0.5$ db/λ and $\rho_b = 1.8$ g/cm^3).

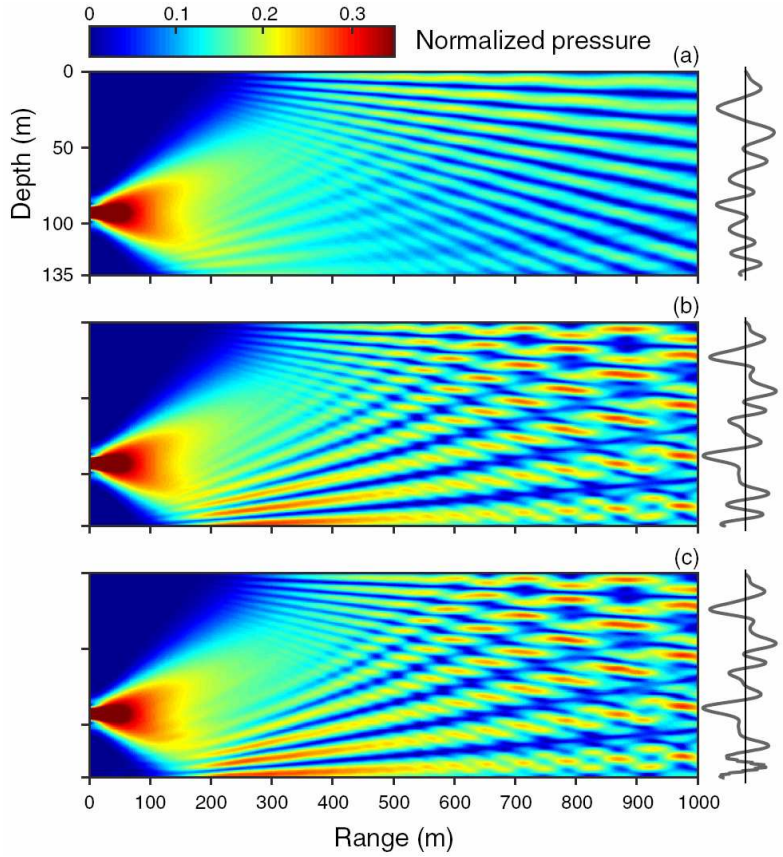


Figure 7. Acoustic pressure field, obtained with the WAPE and optimal NLBC control. Source is positioned left at 93 m *depth, with source frequency f = 500* Hz. *Vertical receiver array (512 elements, equispaced over the water column) is positioned at range 1000* m. *Left: initial guess (a), true value for synthetic data (b), and (c) inverted acoustic pressure field after 16 iterations. Right: imaginary part of the complex acoustic pressure field at range R = 1000* m *[12].*

Observe how, with a very limited number of iterations (initial guess for the bottom properties: clay, $c_b = 1505$ m/s) many detailed features in the propagating field are resolved. Integrated errors over the receiver array have been determined (not shown here).

The second example demonstrates a combined water column and geoacoustic inversion, using YS'94 shallow water (depth 113.1 m) experimental data and a multi-frequency approach (7 source frequencies: 200, 250, 315, 400, 500, 630 and 800 Hz). Measurements were done with a vertical receiver array (32 hydrophones, with 2 m spacing between 37.2 and 99.2 m depths) at $R = 1.5$ km from the source. The result in Fig. 8 shows the quality of the estimation process (compare the final estimated values with the ground truth), and illustrates the parameter hierarchy: the compression speeds in bottom, sediment layer and water column start to converge before the attenuation and density

Both examples demonstrate the capability of the adjoint-based inversion approach. It can both resolve propagation phenomena in the water column and do an efficient combined search for a set of environmental parameters, as was demonstrated in the second example.

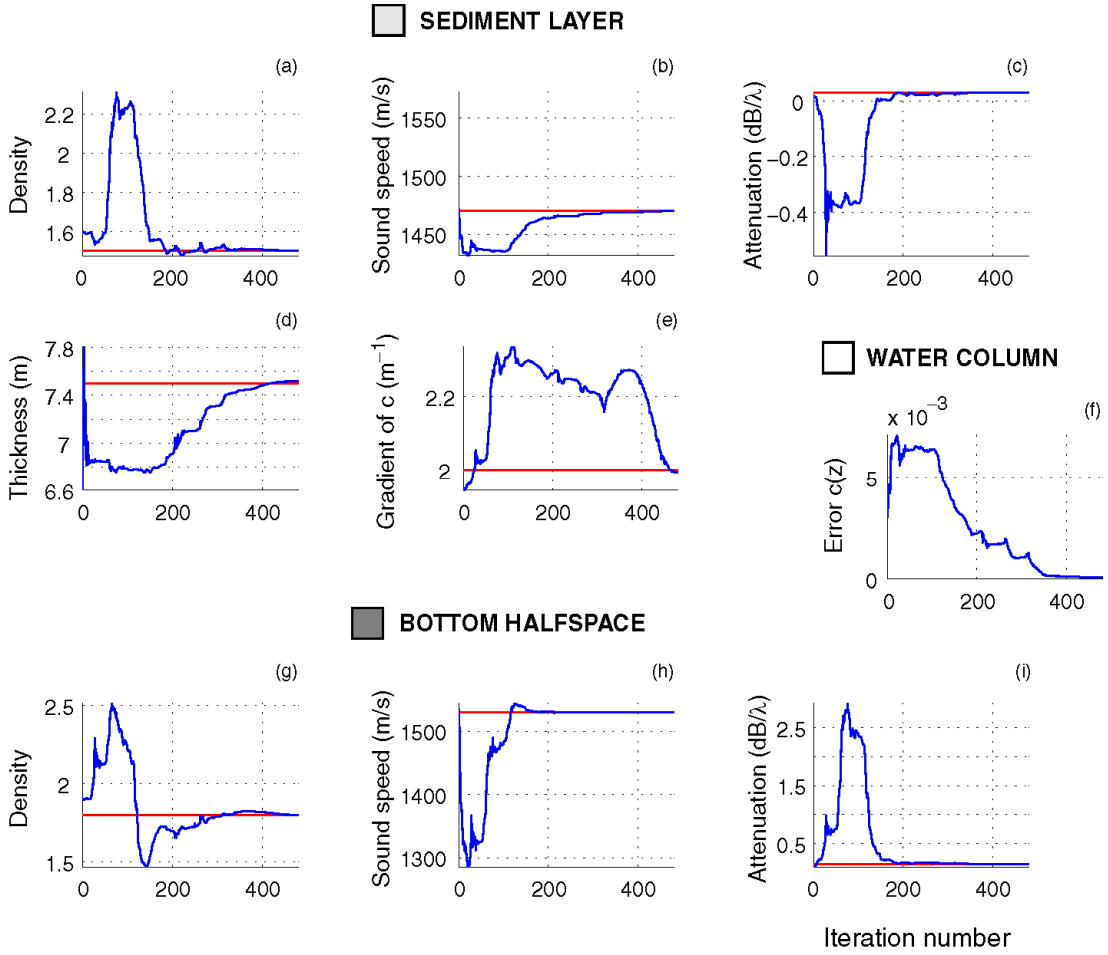


Figure 8. Results of geoacoustic inversion combined with simultaneous correction for an uncertain sound speed profile for the 32-element vertical receiver array and 7 source frequencies. Evolution of the estimated geoacoustic parameters vs. iteration number is shown (3 upper plus 2 centre left plots for the sediment layer, 3 lower plots for the bottom halfspace), together with the depth-integrated error of the water column sound speed profile (centre, right plot). Ground truth for the geoacoustic parameters is shown as red lines.

Geoacoustic inversion with Saba'06 experimental data

In [9] the search process, as part of the geoacoustic inversion, is based on a genetic algorithm. The hydrographic survey vessel acted as the sound source of opportunity and 5 tones from the ship generator noise on the receiver hydrophones (4 hours of recording on April 24, 2006) were used for the inversion of both the experimental geometry and the geoacoustic parameters in the halfspace bottom. Replica data at the receiver were obtained with the Kraken-C acoustic propagation model.

The result is shown in Fig. 9 as a set of posterior probability density distributions. The genetic algorithm settings were: per parameter 40 individuals per generation, crossover rate 0.1, mutation rate 0.1, and 2000 cells per forward run. The results indicate that the experimental geometry is estimated reasonably well, with some error on receiver array position and tilt (position verification with DGPS measurements of the set-up). The estimated values of the halfspace bottom parameters are less representative of the ground truth (a sandy sediment layer over calcareous rock sub-bottom); therefore a second environmental model was used for a more refined geoacoustic inversion run (results not shown here).

Geoacoustic inversion with MREA/BP'07 data

The next example, taken from the cruise report, serves as another demonstration of the passive geoacoustic inversion capability of the the REA concept. NRV *Leonardo* acted as a source of opportunity; sound in the frequency range 0–2 kHz was recorded over a 10 minute period and used as the basis for geoacoustic inversion. Nine parameters were initially guessed with limited a priori knowledge (estimation intervals divided into 40 samples). The final estimated parameter set is shown in the lower half of Fig. 10, and the most significant parameters are well-estimated. The values closely match those obtained 10 years ago under well-controlled conditions using a broadband controlled source, a small number (2–4) of hydrophones and model-based matched filter processing (time reversal).

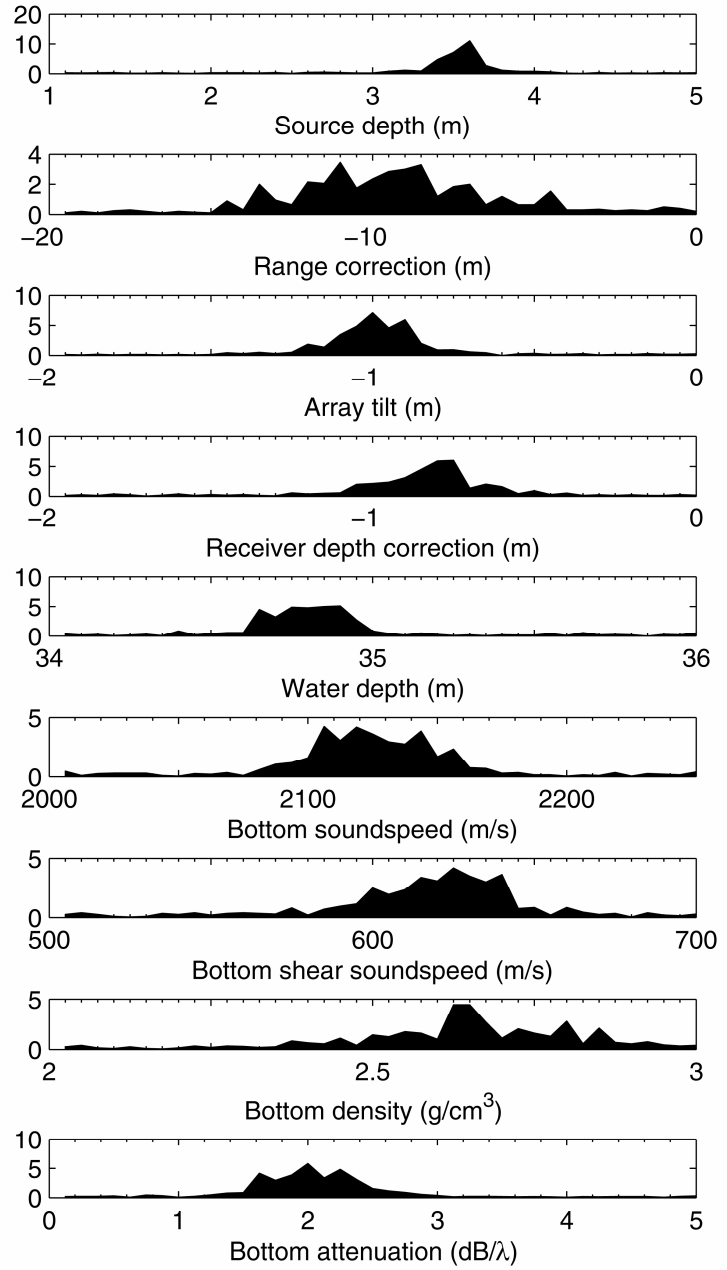


Figure 9. Posterior probability density distributions of the experimental geometry and the geoacoustic parameters during the Saba'06 sea trial. Estimates are based on 20 runs with a genetic algorithm.

The final example, taken from [12], demonstrates the tracking capability, monitoring the range average sound speed changes over a 48 hour period during the MREA/BP'07 trial. The sound speed profile (SSP) is modelled using a set of three Empirical Orthogonal Functions (EOFs) that account for 99% of the sound speed variability in the water column. The inversion process based on the acoustic measurements yields the evolution of the three weighting coefficients μ_1, μ_2 and μ_3 for the EOFs. These are compared with a detailed prediction, calculated with the NRL Naval Coastal Ocean Model (NCOM) that uses a limited set of SSP measurements. The results are shown in Fig. 11, where the upper plot shows the 48 hours NCOM prediction, and in the lower there is the evolution of the weighting coefficients that represent the SSP. Note the afternoon effect (left plot, at time intervals 10–20 and 36–46 hrs), and the excellent agreement between the prediction and the inversion results when both are reconstructed with 3 EOFs.

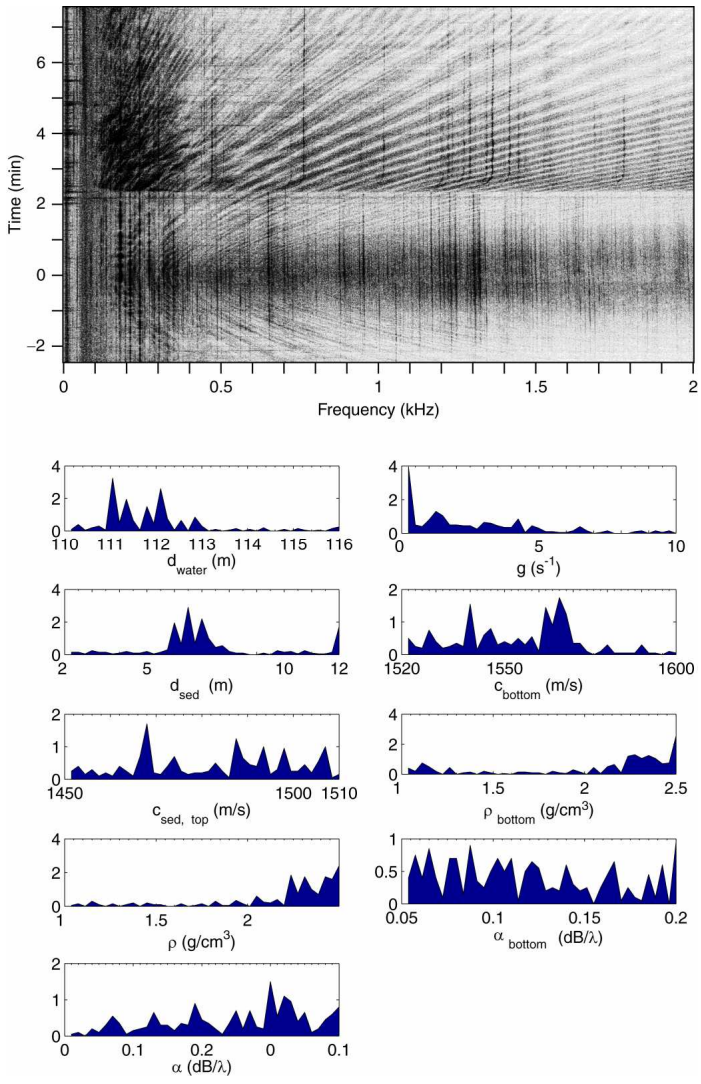


Figure 10. Preliminary results of passive geoacoustic inversion at station ST07 along the AB transect using NRV Leonardo as a sound source of opportunity. Top: spectrogram of Leonardo self-noise over a 10 minute period in the frequency range 0-2 kHz. *Frequencies, ranges and depths selected for the inversion are respectively [226.2, 452.6, 486.1, 582.9, 698.5, 948.5, 1163.4, 1239.3]* Hz; *[0.689, 0.706, 0.723, 0.740, 0.758, 0.775, 0.792, 0.809, 0.827]* km; *[19.04, 24.01, 28.98, 33.95]* m. *Bottom: a posteriori distribution of the estimated geoacoustic parameters: water depth, sediment layer thickness, compression speed, density, attenuation, speed gradient, bottom compression speed, density, attenuation [4].*

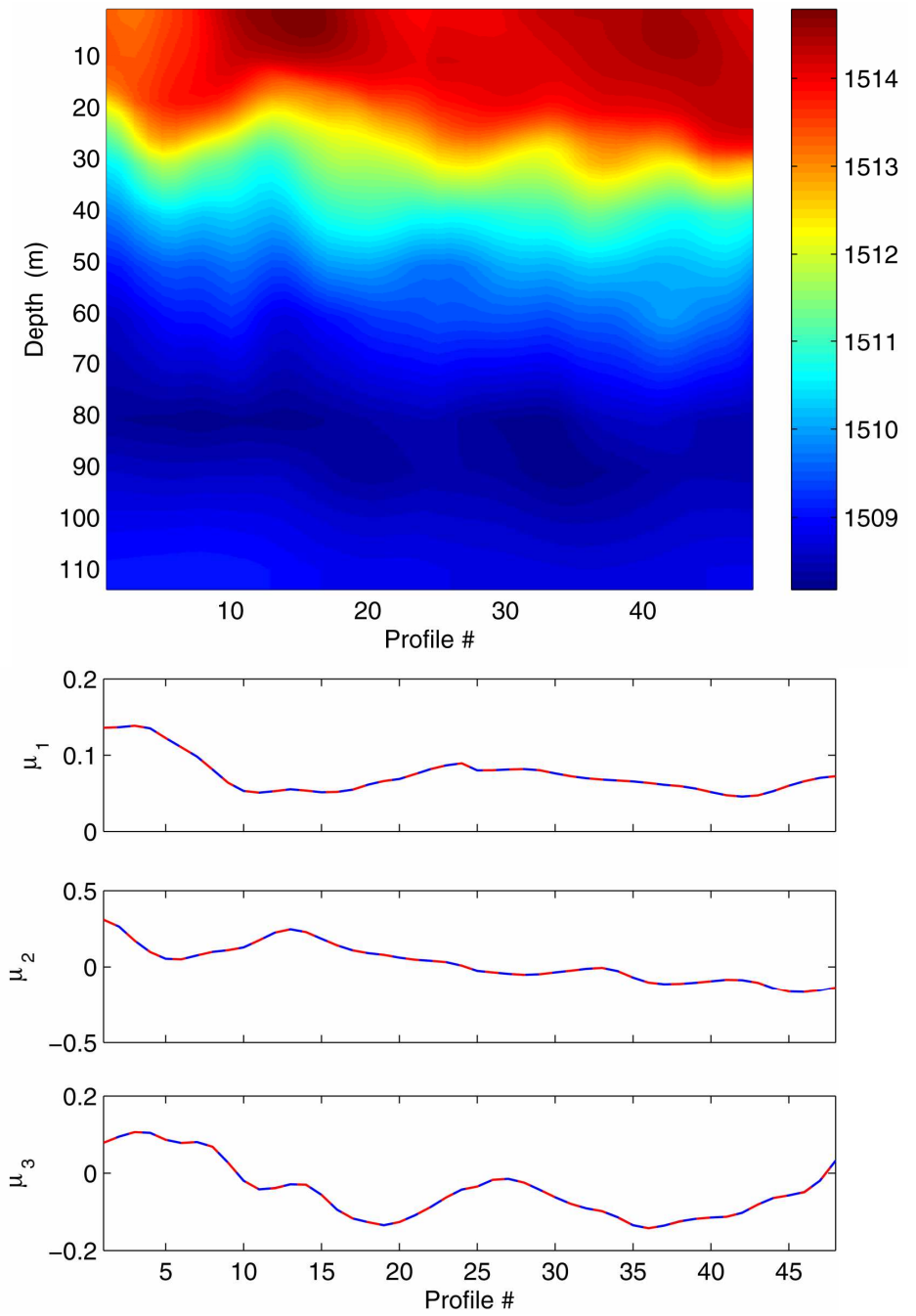


Figure 11. Temporal variability of the range-average sound speed profiles over a 2-day period (1 hr intervals starting on April 28, 2007 at 00:00h) along the MREA/BP'07 transect. Top: the NCOM predictions. Bottom: the evolution of the 3 EOF weighting functions (red dashed lines show the inversion results, blue lines the reconstructions based on the NCOM predictions). Note the afternoon effect in the upper plot and the excellent agreement in the lower plot.

Conclusions

This paper gives an overview of a 6-year RNLNC research effort into the subject of REA in shallow water areas. The research focused on demonstrating a REA concept, based on a sparse set of receivers, the use of both a controlled sound source and sources of opportunity (such as passing surface ships), and a number of optimisation schemes, based on global search and adjoint modelling. This scheme yields water column and bottom properties in an iterative process (for ocean acoustic tomography and geoacoustic

inversion, respectively), with a parameter search based on the underlying physics of the acoustic waveguide for minimized computational load (the adjoint-based approach). The REA concept investigated in this work is suitable for covert operations.

In order to validate the REA inversion approach the RNLNC has participated in two recent sea trials, Saba'06 and MREA/BP'07. Onboard near real-time processing demonstrated the capability of this concept. Both trials have resulted in high-quality and well-documented data sets. The MREA/BP'07 trial was designed for Battlefield Preparation and demonstrated that a combination of in situ measurements, remote sensing and modelling could provide the operational commander with a Recognized Environmental Picture in support of Anti-Submarine Warfare or Amphibious Operations.

A discussion on segmentation should follow the acoustic patch concept put forward in the MREA/BP'07 cruise report: after the hydrographic survey with a ship or an AUV range-independent geoacoustic inversion will be done for patches of the operational area. Separating the full spatial domain into smaller patches requires further research.

Already, the research team has produced a respectable amount of scientific output: journal and conference papers including invited papers and a PhD thesis. A basis for further research in an international context has been created, that could be pursued in the near future.

References

[1] J.-P. Hermand. Broad-band geoacoustic inversion in shallow water from waveguide impulse response measurements on a single hydrophone: Theory and experimental results. *IEEE J. Oceanic Eng.*, 24(1):41-66, January 1999.

[2] J.-P. Hermand and P. Gerstoft. Inversion of broad-band multitone acoustic data from the YELLOW SHARK summer experiments. *IEEE J. Oceanic Eng.,* 21(4):324-46, October 1996.

[3] J.-P. Hermand, MREA/BP'07 Ocean Acoustic Data Report, NLDA-CSD-2007-01, December 2007.

[4] J.-P. Hermand and J.-C. Le Gac. NURC – a NATO Research Centre BP'07 Cruise Report. NURC-CR-2007-04-1D1 / NLDA-CSD-2007-02 Report, December 2007.

[5] J.-P. Hermand, M. Meyer, M. Asch and M. Berrada. Adjoint-based acoustic inversion for the physical characterization of a shallow water environment. *J. Acoust. Soc. Amer.*, vol. 119, pp 3860-3871, June 2006.

[6] J.-C. Le Gac and J.-P. Hermand, BP'07 Test Plan, Technical document NURC-TP-2007-03-1D1, 2007.

[7] J.-C. Le Gac, Y. Stéphan, M. Asch, P. Helluy and J.-P. Hermand. A variational approach for geoacoustic inversion using adjoint modeling of a PE approximation model with non local impedance boundary conditions. In A. Tolstoy, Y.C. Teng and E.C. Shang, editors, *Theoretical and Computational Acoustics 2003*, pp 254-263. World Scientific Publishing, 2004.

[8] A.V. van Leijen and J.-P. Hermand. Geoacoustic inversion with ant colony optimization. In *Proceedings of the 8th European Conference on Underwater Acoustics,*

ECUA, pp 515–20. Algarve Technological Research Centre and University of Algarve, 2006. Carvoeiro, Portugal, 12–15 June 2006.

[9] A.V. van Leijen, J.-P. Hermand and M. Meyer. Geoacoustic inversion in the north-eastern Caribbean using a hydrograhic survey vessel as a sound source of opportunity. In: Journal of Marine Systems. Elsevier 2008.

[10] J.L. Lions. *Optimal Control of Systems Governed by Partial Differential Equations*. Volume 170 of: A series of comprehensive studies in mathematics. Springer Verlag, New York, 1971.

[11] M. Meyer and J.-P. Hermand. Optimal nonlocal boundary control of the wide-angle parabolic equation for inversion of a waveguide acoustic field. *J. Acoust. Soc. Am.*, 117(5):2937–48, May 2005.

[12] M. Meyer. The adjoint method of optimal control for the acoustic monitoring of a shallow water environment. Ph.D. Thesis, Université libre de Bruxelles, December 2007

[13] P. Oonincx, A. Auger-Ottavi, L. le Proux de la Riviere and J.-P. Hermand. Gabor Feature vectors to detect changes amongst time series: a geoacoustic example. Conference Proceedings, IEEE Eusipco, 2006.

[14] M.B. Porter. The KRAKEN normal mode code. SACLANTCEN Memorandum SM-245, SACLANT Undersea Research Centre, La Spezia, Italy, November 1991.

[15] S. Thiria, F. Badran and C. Sorror. *YAO: Un logiciel pour les modèles numériques et l'assimilation de données*. Rapport de recherche. LOCEAN, Paris, France, 2006.

From the Lab to the Sea, Acoustic Sensing in Uncertain Environments

Vincent van Leijen

1990. *Somewhere above the Atlantic Ocean a lone maritime patrol aircraft is on an ASW mission. Directed by SOSUS intelligence the crew is ordered to monitor a designated area with a field of sonobuoys. The first buoy to hit the water is an expendable bathythermograph (XBT). The device samples the temperature profile in the water column and the operator quickly derives a sound speed profile. After the propagation conditions of underwater sound are reviewed for tactical consequences a pattern of sonobuoys is dropped with favourable spacing and depth settings. It does not take long before a contact emerges on one of the outer buoys.*

2000. *An expeditionary force of various surface ships is about to enter a coastal area. To predict the performance of various acoustic sensors, the water column is sampled with a bathythermograph. Details about the local bottom conditions are unknown and the sonar performance model is then run with global parameters from an environmental database. As a result a mine hunting operation takes twice the time that was actually needed to clear the mines because of non-optimal sonar settings. Meanwhile a bottomed submarine remains effectively hidden in the reverberation, waiting for the main force to close in.*

2010. *An amphibious force is to land on a beach that has been selected from satellite imagery. A discrete campaign of rapid environmental assessment then reveals the presence of a muddy sediment layer. Mud is ideal sediment for self-burying mines and means that the beach is not accessible for heavy armoured vehicles. With the secretly gathered information a new area is selected and the amphibious operation unfolds itself as an unopposed landing.*

Introduction

When expeditionary forces enter shallow or confined waters, the environment has a great influence on the performance of platforms, sensors and weapon systems. For this reason, environmental knowledge is regarded as one of the key factors in making decisions on the course of action and asset allocation [1]. The examples above illustrate how the right level of battle space information enables effective operational planning and mission execution [2]. For naval oceanography the main objective is to provide forces with a competitive advantage over adversaries by exploiting the current and future state of the environment. The Royal Netherlands Navy (RNLN) possesses various sensor performance models and tactical decision aids for its combat systems. Many environmental input parameters can be provided in advance by the Netherlands Hydrographical Office (nautical charting) and the METOC office of CODAM (environmental briefing dockets and databases). Some parameters are measured or sampled at sea, such as weather conditions, water temperature and underwater ambient noise. For expeditionary operations it is likely that *a priori* knowledge about the environment is limited and outdated. Therefore there is a need for tools that enable hydrographers or naval oceanographers embedded with the forces to collect and validate environmental information at sea.

Environmental information for naval warfare

Each mission type has its own operational need for environmental information in terms of data accuracy and spatial and temporal resolution [2]. In Anti-Submarine Warfare (ASW) it is crucial to know how well sonar performs. Environmental information enables

the prediction of acoustic detection ranges on submarines and surface ships. For the open oceans, the propagation of sound is determined by depth, temperature and salinity only. Shallow waters are often characterized as an unpredictable and complex environment. For sonar, the performance is determined by many factors, such as tides, currents, wind, rain and reflections from the sea surface and complex bottom structures. The essential data for propagation modelling is often incomplete, and therefore the daily predictions of sonar performance are seldom close to reality. In addition, water conditions and sound speed profiles change during the day due to temperature changes and weather conditions. Mine Counter Measures (MCM) also depend on various oceanographic factors [3]. The bathymetry (charted water depth) and the acoustic properties of the medium determine how well mine hunting sonar will perform. Acoustic detection of mines is limited by sea bottom reverberation. A rough estimate of the sediment type is sufficient to indicate the underwater visibility and the likelihood of mine burial, but coastal mechanisms of river outflows and sediment transport call for repeated observations. In amphibious operations the shallow water bathymetry determines how close to the coast support ships and landing craft can safely get. Important information about the beach, such as trafficability and the slope, can be found with an autonomous underwater vehicle during high tide. In general, the characterization of the sediment and bathymetry for amphibious purposes permit a rough level of detail.

It is easily overlooked that the shallow character of the littorals can also be exploited. A rough approximation of the underwater battle space is already valuable for a tactical exploitation of the environment (TEE). A submarine can tactically exploit the reverberant properties of the sea bottom or be positioned to benefit from the directionality of ambient noise. TEE concerns easy rules of thumb and can do with rough estimates about the environment, as in "active sonar performs better in down slope direction then up slope". Environmental knowledge with a high level of detail enables passive source localisation with techniques known as Matched Field Processing (MFP). The advantage of MFP over conventional Doppler arithmetic is that the latter requires movement of the target and information about the zero-frequency and MFP does not. On the other hand MFP depends on a propagation model that operates on accurate environmental data. The technical character of MFP further calls for a highly skilled and well-instructed operator.

Various levels of battle space information can be obtained with a campaign of rapid environmental assessment (REA). The aim is then to measure, analyse and evaluate relevant properties of the environment in order to establish a recognized environmental picture (REP). The intention is that forces have a shared awareness of the battle space and that they have it in time. Since 2004 the RNLN operates two hydrographic survey vessels HNLMS Snellius and HNLMS Luymes. These modern ships are fitted with an extensive sensor suite for digital charting and further tasks of military hydrography [4]. For covert REA the navy may call upon Special Forces and submarines of the Walrus class, as was demonstrated during the exercise Joint Caribbean Lion (2006). Like many other navies within NATO, the RNLN is still in transformation from a blue water force to an expeditionary brown water force. Currently not all important environmental data for shallow water operations can (rapidly) be gathered.

Acoustic sensing in shallow water

The environmental factors that impact acoustic sensing capabilities are manifold. Shallow bathymetry and underwater obstacles may hinder the use of long towed arrays. The presence of divers or marine wildlife may call a halt to mid or low frequency sonar transmissions. Coastal ambient noise includes an abundance of directional sound sources with man made or natural origins. The focus of this paper is on those parameters that influence sound propagation, or more specific: the transmission loss due to sea bottom interaction. The water column is usually characterized by measuring conductivity (to estimate salinity) and temperature as function of depth (CTD sampling). Some empirical formula, e.g. in [5], is then used to calculate the sound speed profile. In deep water the propagation of sound is determined by this profile only; in shallow water many more parameters are involved. Various definitions can be given for shallow water [6]. From an acoustical point of view shallow water is found "when each ray from the source, when continued long enough is reflected at the bottom" [7, p9]. Another definition is "a water depth in which sound is propagated to a distance by repeated reflections from both surface and source" [8, p172]. To be practical, shallow waters are often said to be on the continental shelf and bordered by the 200 m contour line. Unlike the water column, the sea bottom cannot rapidly be characterized by insertion of some sampling device. Nevertheless, sound waves easily propagate in and out of marine sediments. Received signals can then be analysed with geoacoustic inversion techniques to backtrace acoustic properties of the ocean bottom from the spatial and temporal structure of sound pressure fields. Experiments for seabed assessment utilize a sound source and a receiver array for a one-time observation at sea of bottom reflected sound. A geoacoustic inversion process is then initiated to find a parametric description of an environmental model in terms of sediment layering properties and geoacoustic parameters such as sound speed, density and attenuation.

REA as a research project

The Rapid Environmental Assessment (REA) project at the Netherlands Defence Academy aims to understand the nature and impact of environmental conditions on the propagation of sound in shallow waters and sedimentary bottom types [9]. As such, the project aims for the development and validation of acoustic remote sensing systems and inversion methods. The result is a reliable and rapid environmental assessment of shallow water areas in support of various mission types. The question for this article is: what acoustic information about the seabed can be obtained from bottom-reflected shipping noise? The feasibility of geoacoustic inversion with non-traditional sound sources will be studied with data from two sea trials. During Saba'06, a Caribbean survey of the NL Hydrographic Office (NLHO) in 2006, small-scale experiments in a remote and isolated area were conducted from hydrographic survey vessel HNLMS Snellius [10]. The trials demonstrated a rapid deployment of sensors and equipment and resulted in a well-documented acoustic dataset. A unique achievement is that geoacoustic inversion was performed while the team was on board and an environmental debrief was provided, all within 24 hours. The BP/MREA'07 sea trials of 2007 were a much bigger effort [11]. Together with the NLHO, the NATO Undersea Research Centre (NURC) and various other institutions a shallow water area in the Mediterranean Sea was surveyed with a multitude of sensors. The overall aim of the trials was to demonstrate the concept of naval

battle space preparation by providing a recognized environmental picture (REP). The dynamic and coastal area includes deeper water (200 m), very shallow water (30 - 10 m), a harbour approach and the beach. The multi sensor approach makes it possible to validate results of geoacoustic inversion experiments with non traditional sound sources under various circumstances.

Covert REA

The preparation of some remote coastal area with an overt REA campaign is in obvious conflict with the concealed nature of submarine and amphibious operations. Therefore environmental assessment in support of military operations will often be a discrete endeavour. Covert assessment of the sea bottom calls for clandestine deployment of sound sources and receiving sensors. The REA project studies various ways in which signals with geoacoustic information can be received. Receiving sensors can be inserted in denied areas by acoustic-oceanographic buoys and drifters that exploit the local currents [12]. A drifting buoy field covers a large area and is not hindered by the presence of mines, yet radio transmissions can be intercepted. During the scientific experiments Saba'06 and MREA'07, data was also gathered with a sparse vertical array deployed from a rubber boat. The concept can easily be translated to an operational context when acoustic-oceanographic sensors are deployed and recovered by Special Forces. The feasibility of this concept has recently been demonstrated with covert hydrographic reconnaissance during exercise Joint Caribbean Lion. More information about oceanography and Naval Special Warfare can be found in [13]. Front-line units such as autonomous underwater vehicles (AUV's) and submarines are already fitted with sensors for intelligence, surveillance and reconnaissance (ISR). Typical but sensitive intelligence missions can easily be extended with an environmental component to make dual-use of ISR sensors [2]. The approach also provides a capability to make dual use of *past* intelligence missions. In this case archived sonar data from ill-documented areas can be analysed again, but now for environmental purposes.

Sound sources of opportunity

For a thorough assessment of bottom properties acoustic signals are required with low frequencies that penetrate deep into the bottom. Shipping sounds are also low, with frequencies from 50 Hz up to 2 kHz. One of the reasons to launch a REA campaign is to aid in the prediction of passive acoustic detection ranges of ships and submarines. The conventional method relies on active sonar transmissions. There are however some practical down sides to the active approach. The high power consumption of low frequency systems limits the endurance of remotely deployed systems such as drifters, buoys and autonomous underwater vehicles [2]. And assessment with loud transmissions and low frequency is also more of an overt approach. An alternative is to utilize sound sources of opportunity. A military motive to do so is that (counter) detection is avoided and environmental assessment can be done in a discrete manner. Another motivation is that the method inflicts a minimal impact on divers and marine wildlife [10].

Coastal waters allow for a high concentration of human activities and as a result shallow waters are a noisy environment. With the right sensors there are many ships that can act as a sound source of opportunity. At some distance from the coast there is merchant traffic in designated shipping lanes, augmented by fishing vessels and offshore suppliers. Closer to the coast there are the ferries and the recreational boats. In times of military conflict various types of naval vessels may patrol coastal waters.

Figure 1. Sound sources of opportunity used so far in the REA project: HNLMS Snellius, NRV Leonardo, REMUS AUV and recreational boats

The REA project has lead to geoacoustic inversion with cooperative surface ships, unmanned underwater vehicles and even uncooperative recreational boats; the platforms are pictured in Fig. 1. For the Saba bank, geoacoustic inversion with received shipping noise from HNLMS Snellius revealed a very thin layer (15 cm) of sandy sediment over a sub-bottom of calcareous rock [10]. The BP/MREA'07 sea trials featured experiments with various sound sources of opportunity. When opportunities occurred, these sources behaved as planned, as in the experiments with self noise from HMLMS Snellius and NRV Leonardo [11]. During a particular run that focussed on the self noise from the relative quiet REMUS AUV [14] there was much interference from the weekend traffic. But then these recreational boats turned out to be fantastic sources of opportunity [15] and demonstrated the strength of the inversion method in using non-cooperative sound sources for a rapid and reliable characterization of the local sediment. In the following case study an AUV is used to assess the environment. The resulting environmental model is then demonstrated to enhance acoustic sensing capabilities with matched field source localization for one of the recreational boats.

Case study: geoacoustic inversion with an AUV

In a particular experiment during MREA07 a REMUS autonomous underwater vehicle was programmed for a mission in shallow water. An area was selected near the local harbour of Castiglione della Pescaia, Italy with a locally nearly flat bottom and a water depth of 33 m. The self noise of the vehicles was received on a sparse vertical array and used to invert sea bottom properties [14]. The general geometry of the experiment is pictured in Fig. 2.

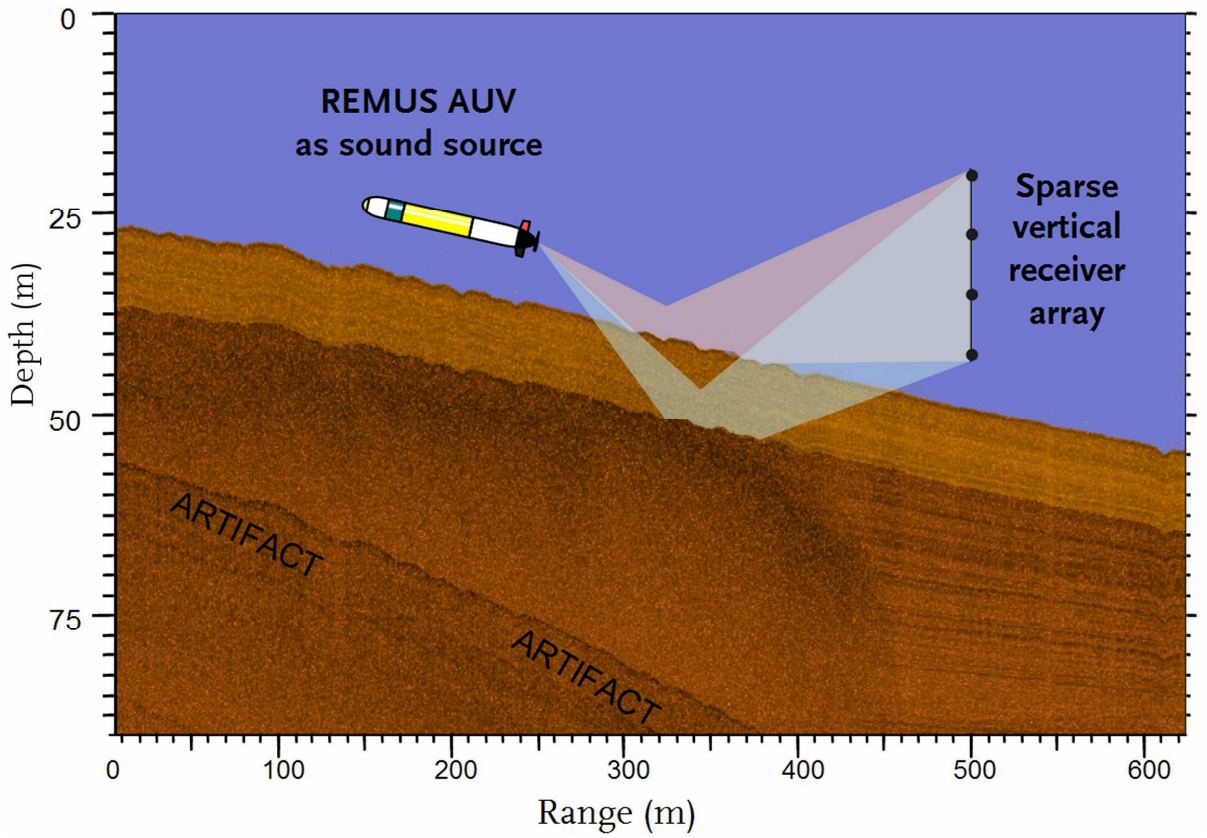


Figure 2. Concept of an inversion experiment where a REMUS AUV acts as a covert sound source of opportunity for geoacoustic assessment. Seismic profiling of the sea bottom by X-Star (TNO) on CD12500 line.

Methodology

Inversion is a search process for unknown acoustic parameters by comparison of observed underwater sound with replica data. A schematic overview of the process is given in Fig. 3.

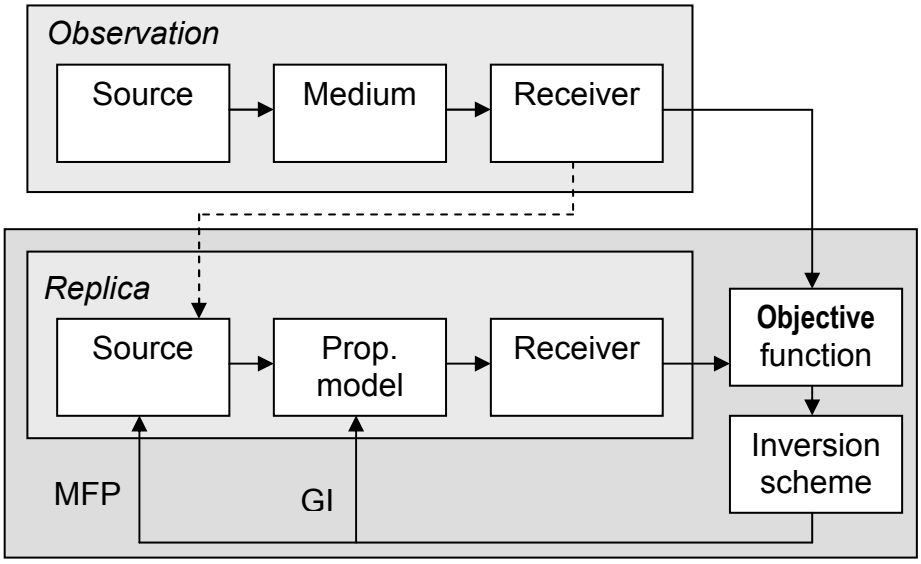


Figure 3. Diagram with the main components for Matched Field Processing (MFP) and Geoacoustic Inversion (GI)

Inversion begins with observations at sea when underwater sound is recorded. Further observations concern the experimental geometry and environmental data such as CTD samples to find the sound speed profile in the water column. Replica data with predicted propagation loss can be created with a propagation computer model. When a correct model of source position (MFP) and the underwater environment (GI) are input, the theory says that the output of the propagation model will be in perfect match with the observed data. Therefore in an iterative modelling process many input parameters are tested until a best fit is found. The mismatch is expressed by an objective function Φ. This is very often a Bartlett processor that cross-correlates data from a number of sensors [16]. The search strategy to minimize Φ in order to find an optimal solution is determined by an inversion scheme.

LOBSTER inversion toolbox

To carry out the inversion, a LOBSTER toolbox [17] has been developed at the NLDA (the Low-frequency Observation Based Sonar Toolbox for Environmental Reconstruction). This object-oriented Matlab code interfaces with variants of the KRAKEN [18] and MMPE [19] (third party) propagation models and offers a number of objective functions. The real innovation of the code is the support of inversion with acoustic particle velocity [20] and the number of included metaheuristic search strategies. Apart from conventional metaheuristics such as Simulated Annealing [16, 21] and the Genetic Algorithm [22], implementations of Differential Evolution [23] and Ant Colony Optimisation [24, 25] are included.

Geoacoustic inversion results

In the experiment, the AUV was programmed to run at its maximum speed and a ball bearing began to resonate. This proved to be highly beneficial as 8 stable tones were selected from a frequency range of 850 Hz to 1350 Hz. The phones in the sparse and vertical receiver array were at depths of 15, 20, 25 and 30 m. The applied inversion scheme was Differential Evolution with a population size of 50. The optimizer was configured to run for 40 iterations with a differential factor of 0.6, a crossover rate of 0.8 and a total of 1.6×10^4 calls to the KRAKEN propagation model. The used distances between source and receiver were less than 100 m [14].

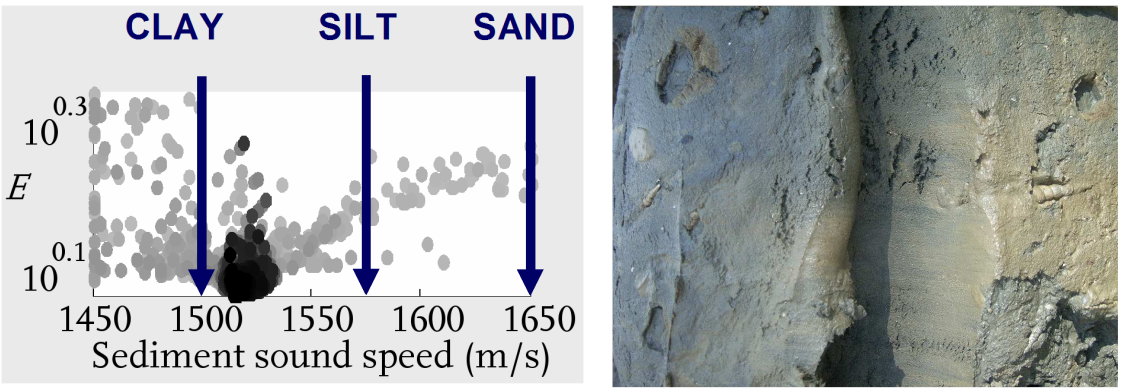


Figure 4. Convergence plot for the sediment sound speed. The markers for clay, silt and sand are from [26], the grab sample of silty clay was taken with a Van Veen grabber.

The dominant acoustic parameter turned out to be the sound speed of the sediment top layer. The 1520 m/s result obtained is characteristic for 'silty clay' [26] and corresponds with the grab sample from the sea bottom, both can be seen in Fig. 4. The seismic profile in Fig. 2 clearly shows the presence of a sub-bottom. In the inversion results however this sub-bottom was hardly perceived. A logical explanation, beside the low source level of AUV self noise, is that sound from the direct path and reflections from the sediment have considerable less propagation loss than the sub-bottom reflections. The direct path can be avoided by utilising downward reflection due to negative gradients in the sound speed profile. In this case geoacoustic inversion becomes more efficient with data from surface sources at greater distance from the receiver array.

Enhanced acoustic sensing

During the experiment there were many recreational boats that left Castiglione della Pescaia. Fig. 5 shows how one of these boats is localized with matched field processing for five tones from the inboard diesel engine, and given two different environmental models. When bottom properties from a military environmental database such as ASRAP are used, with a rough spatial resolution, the method fails to correctly identify the source position. MFP with the bottom model from the AUV inversion resulted in Fig. 5b with one clear spot at the surface and 920 m away from the receiver. This example clearly illustrates how proper environmental information enhances acoustic sensing capabilities.

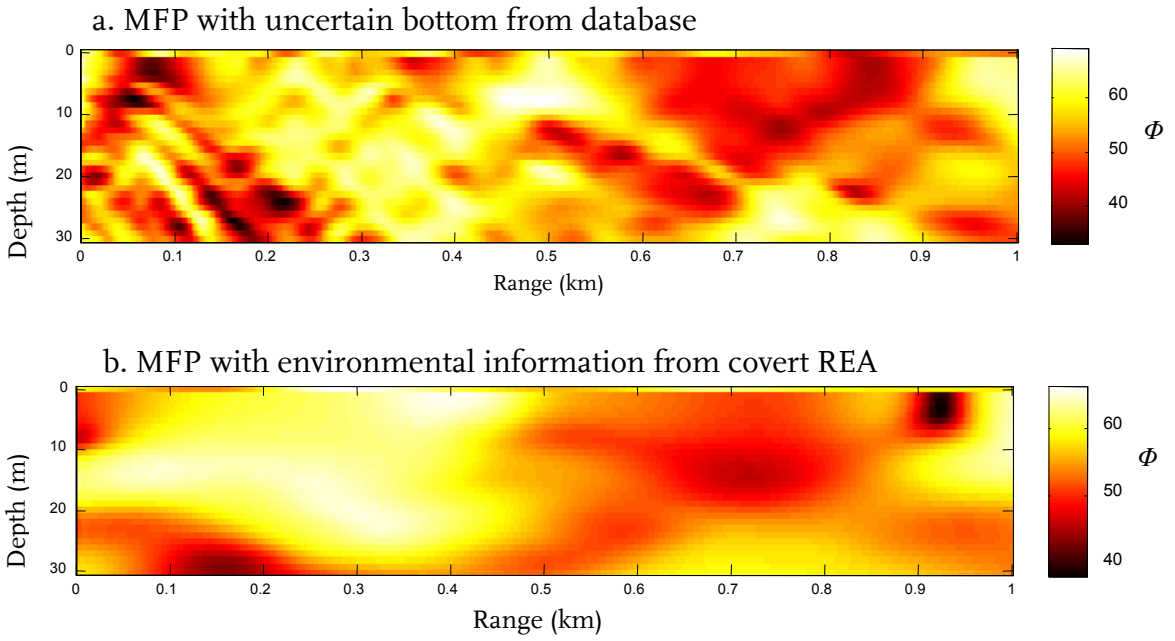


*Figure 5. The benefit of environmental information for source localization with matched field processing. Pictured is the mismatch surface for depth and range. The engine noise of the recreational boat should be found just below surface, and is identified with minimal mismatch Φ, denoted with the colour black. The upper image (*a*) is based on uncertain bottom properties drawn from databases such as ASRAP and does not give a clear solution. The lower image (*b*) based on the covert REA mission with the AUV has one clear (black) detection of the recreational boat at the surface and 920 m from the receiver.*

Discussion

Underwater acoustic sensing is a battle of the decibel. Quietening of submarines and increased ambient noise in coastal areas have resulted in a general decrease of acoustic detection ranges. For passive sonar in shallow water significant gains are possible when sensors with a vertical aperture are combined with modern signal processing techniques [27]. Real-time, environmentally adaptive algorithms may combine a track-before-detect approach with time-reversal algorithms in order to focus acoustic waves. Coastal ambient noise is highly directional in bearing and azimuth and this is where adaptive beamforming with arrays of directional vector sensors [28] can contribute even more. For passive sonar, environmental adaptive algorithms provide cleaner displays and easier track identification. The potential for active sonar is strong mitigation of reverberation. For expeditionary missions, relevant oceanographic data is often undersampled in space and time. Therefore, and to further adapt deep-water procedures for the littoral zone, the logical addition to XBT sampling of the water column is to assess seabottom properties with geoacoustic inversion techniques, as the U.S. Naval Oceanographic Office (NAVOCEANO) already practices [2]. The required resolution and acceptable level of environmental uncertainty depend on the range of mission types that naval forces fulfil. Significant advances in acoustic sensing are possible, yet they come with a price. Apart from the integration of dedicated shallow water sensors and environmentally adaptive processing, education and operational training remain a key factor. Acoustic sensing has never been easy, and a lack of education can easily degrade sensor performance. Then again, when the skilled hands of a 'techno sailor' are provided, major improvements in sonar performance are still possible.

Conclusions

The aim of this paper is to find out what acoustic information about the seabed can be obtained from bottom-reflected shipping noise. The feasibility of geoacoustic inversion with non-traditional sound sources has been studied with data from two sea trials. During the experiments on the Saba bank (2006) the concept was demonstrated with a short REA campaign in a remote and isolated area. With the MREA sea trials of 2007 in the Mediterranean Sea, the covert battle space preparation concept was further experimented with and complemented by a multi-sensor survey of various bottom types and water depths to further validate geoacoustic inversion methods. If there is one dominant parameter found that characterizes the sea bed in a shallow water area, it must be the sound speed in the upper sediment layer. With all inversions described here, the sediment sound speed was quickly found. This acoustic property prevails in its influence on the propagation of sound and it also identifies what material the seabed is composed of. Even a rough seabed characterization is highly beneficial for mine countermeasures as it suggests what mines can be deployed and indicates the possibilities and likeliness of mine burial. When visual and acoustic sensing capabilities are known it is possible to hunt mines in a time-efficient way. One step further is to use geoacoustic inversion to provide a full environmental model in support of antisubmarine warfare. When in situ data of high accuracy is input to a sonar performance model (such as Almost), instead of rough database estimates, the predicted detection ranges are guaranteed to be much closer to reality. It was further shown that geoacoustic inversion enables reliable remote

sensing capabilities with matched field processing techniques. The proposed use of true sound sources of opportunity, such as ferries, recreational boating or military patrol boats, provides the navy with the capability of discrete rapid environmental assessment of remote and denied areas.

References

[1] -, *Leidraad maritiem optreden, de bijdrage van het Commando Zeestrijdkrachten aan de Nederlandse Krijgsmacht* (Den Helder, Commando Zeestrijdkrachten) (2005).

[2] National Research Council, Committee on Environmental Information for Naval Use, *Environmental information for naval warfare*, National Academies Press, Washington D.C., 2003.

[3] National Research Council, Ocean Studies Board, *Oceanography and mine warfare,* National Academies Press, Washington D.C., 2000.

[4] Janssen Lok, J., "Military hydrography technology ventures into unchartered waters," *Jane's International Defence Review* 38, pp 42-46 (2005).

[5] Medwin, H., "Speed of sound in water for realistic parameters," *J. Acoust. Soc. Am.* 58, pp 1318 (1975).

[6] Vego, M.N., *Naval strategy and operations in narrow seas,* Frank Cass, London, 1999.

[7] Brekhovskikh, L.M., *Fundamentals of ocean acoustics*, 3rd ed. Springer, New York, 2003.

[8] Urick, R.J., *Principles of underwater sound*, 3rd ed. McGraw-Hill, New York, 1983.

[9] Absil, F., and Hermand, J.-P., *KIM Research Project: Shallow Water Rapid Environmental Assessment with a Netcentric Acoustic-Oceanographic Buoy System*, internal document, 29th August, Den Helder (2005).

[10] Leijen, A.V. van, Hermand, J.-P., and Meyer, M., "Geoacoustic inversion in the north-eastern Caribbean using a hydrographic survey vessel as a sound source of opportunity," *Journal Marine Systems* (2008) (accepted).

[11] Le Gac, J.-C., Hermand and J.-P., Demarte, M., *BR'07 Cruise Report* (Den Helder, Netherlands Defence Academy, Report NLDA-CSD-2007-02) (2007).

[12] Hermand, J.P., Boni, P., Michelozzi, E., Guerrini, P., Agate, M., Borruso, A., D'Argenio, A., Di Maio, D., Lo Iacono, C., Mancuso, M. and Scannavino, M., "Geoacoustic inversion with drifting buoys: EnVerse 1997-98 experiments," in *Proceedings of the Workshop on Experimental Acoustic Inversion Methods for Exploration of the Shallow Water Environment* (A. Caiti, J.-P. Hermand, S.M. Jesus, and M.B. Porter, eds.), pp263-286, Kluwer Academic, 2000.

[13] National Research Council, Ocean Studies Board, *Oceanography and naval special warfare, opportunities and challenges,* National Academies Press, Washington D.C., 1997.

[14] Leijen, A.V. van, "An autonomous underwater vehicle as sound source for geoacoustic inversion," *J. Acoust. Soc. Am.* 122, p 2951 (2007) (abstract).

[15] Leijen, A.V. van, "Geoacoustic inversion with recreational boat noise," *J. Acoust. Soc. Am.*, (2008) (abstract).

[16] Tolstoy, A., *Matched field processing for underwater acoustics,* World Scientific, Singapore, 1993.

[17] Leijen, A.V. van, LOBSTER *design document* (Netherlands Defence Academy, Report NLDA-CSD-2008-02) (2008).

[18] Porter, M., The KRAKEN *normal mode program* (La Spezia, NATO SACLANT Undersea Research Centre, Report SACLANT-SM-245) (1991).
[19] Smith, K.B, "Convergence, stability, and variability of shallow water acoustic predictions using a split-step Fourier parabolic equation model," J. Comp. Acoust. 9, pp 243-285 (2001).
[20] Leijen, A.V. van, Hermand, J.-P. and Smith, K.B. , "Geoacoustic inversion based on both acoustic pressure and particle velocity," *J. Acoust. Soc. Am.* 120, p 3355 (2006) (abstract).
[21] Kirkpatrick, S., Gelatt, C.D., and Vecchi, M.P., "Optimization by Simulated Annealing," *Science* 220 (4598), pp 671-680 (1983).
[22] Gerstoft, P., "Inversion of seismoacoustic data using genetic algorithms and a posteriori probability distributions," *J. Acoust. Soc. Am.* 95(2), pp 770-782 (1994).
[23] Snellen, M., Simons, D.G., and Moll, C. van, "Application of differential evolution as an optimisation method for geo-acoustic inversion," Proceedings of the Seventh European Conference on Underwater Acoustics, ECUA 2004, Delft, The Netherlands (2004).
[24] Leijen, A.V. van and Hermand, J.-P., "Geoacoustic Inversion with Ant Colony Optimisation", Proceedings of the Eighth European Conference on Underwater Acoustics, 8th ECUA, pp 515-520, edited by S. M. Jesus and O. C. Rodríguez, Carvoeiro, Portugal, 12-15 June (2006).
[25] Leijen, A.V. van and Hermand, J.-P., "Geoacoustic Inversion and Uncertainty Analysis with MAX-MIN Ant System," Proceedings of the Fifth International Workshop on Ant Colony Optimization and Swarm Intelligence, pp 420-427, Belgium, September 4-7, Springer, 2006.
[26] Jensen, F.B., Kuperman, W.A., Porter, M.B. and Schmidt H., *Computational ocean acoustics*, American Institute of Physics, New York, 1994.
[27] National Research Council, Panel on Undersea Warfare, *Technology for the United States Navy and Marine Corps, 2000-2035, becoming a 21st-century force, Vol. 7 Undersea warfare*, (Washington D.C, Naval Studies Board. Committee on Technology for Future Naval Forces. Commission on Physical Sciences, Mathematics and Applications, National Academy Press, 1997.
[28] Smith, K.B. and Leijen, A.V. van, "Steering vector sensor array elements with linear cardioids and nonlinear hippioids," *J. Acoust. Soc. Am.* 122(1), pp 370-377 (2007).

Ad Hoc Networks of Cooperative Robots
A First Impression of a New Research Project

Raymundo Hordijk & Theo Hupkens

Isaac Asimov's Three Laws of Robotics

1. *A robot may not injure a human being, or, through inaction, allow a human being to come to harm.*
2. *A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.*
3. *A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.*

Introduction

Robots can be used in a wide variety of scenarios, including hostage situations and search and rescue missions. They are particularly suitable to deal with dangerous tasks such as the investigation and disposal of explosive materials. The use of robots reduces risk to human soldiers, especially in urban warfare. Therefore, the US military spends some 340 million dollars every year on ground-based robots [Source: United Press International 2007].

There are many definitions of what is called a robot. These definitions range from "machines that can perform complicated tasks automatically or by remote control" to "devices that are capable of performing a number of human tasks". Although robots that are human-like or soldier-like may be very useful, and probably will participate in fighting on battlefields in the near future, our current research focuses on robots that are *not* human-like. Simple robots that move by wheels can in most cases do simple jobs equally well as complicated and expensive human-like robots.

The well known Asimov's Three Laws of Robotics may be very useful for general purpose robots, but for military applications we want to replace these laws by the following laws:

Three Laws of Robotics in Military Applications

1. *A robot may only injure a human being if that human being poses a threat to "us" (the owner of the robot) or our allies. [Note, however, that international laws forbid robots to use arms autonomously.]*
2. *A robot must obey the orders given it by the proper authorities.*
3. *A robot must protect its own existence as long as such protection is needed to fulfil his current assignment.*

We define robots as devices that are able to perform certain tasks *autonomously*, that are able to *communicate* with other robots and that are able to *build a certain model of their environment*, based on their own sensory observations and on observations obtained from other robots. The tasks that robots perform are based on their model of the environment and on orders the robots have been given by the proper authorities. Industrial "robots" that assemble cars for instance, do not fit this definition.

Ad hoc networks

One might think that robots that fit the above definition are very complex robots. It is our intention, however, to make the robots as simple as possible, and to incorporate any desired complex problem-solving capabilities by means of ad hoc networks that are formed by a (possibly large) number of robots. The minimum requirements for the (mobile) robots are that they must be able to accept an order from the proper authorities, to sense their environment, to find the place where they are needed, to fulfil a simple dedicated task, to report the success (or failure) in achieving the task and to share their observations with the other robots that are involved in the mission. A robot will have one type of sensor only or a suite of sensors that can easily be combined.

Compared to multipurpose robots, there are several advantages of using multiple dedicated robots that are specialized in a small number of tasks only:

- such robots can be small, cheap, robust and consume little energy;
- if another sensor is needed other robots can be brought into the scene. There is no need to modify any of the existing robots;
- if a robot is lost, only a few sensors are no longer available;
- some robots may need special protection, for instance against chemical agents. There is no need to protect all other robots;
- some tasks may be inherently too complicated for a single robot to accomplish.

A severe disadvantage is that these simple robots must be able to communicate. This makes them vulnerable because the communication signals may be intercepted or disturbed. For this reason the robots or groups of robots must be capable of working on their own for a longer period of time to avoid extensive long-range communication. So it is evident that the units must be able to act autonomously.

Cooperative robots

Several definitions of "cooperative robots" are possible (see eg. [CAO et al., 1997]). We adopt the definition of [Barnes and Gray, 1991] "joint collaborative behaviour that is directed toward some goal in which there is a common interest or reward". Collective behaviour of robots is not the same as cooperative behaviour. In our view robots *decide to* behave collectively if that is needed to accomplish their current mission.

The scientific challenge is to design a system that is flexible enough to process data from all kinds of specialized robots, where the configuration may change all the time and even new robots may come into play. These new robots may have sensors that were even not known at the time of development of the system. This behaviour is very similar to computer networks (for instance the Internet), so a number of problems involving robots that appear in or disappear from a scene already have been solved in network theory. For an extensive overview and in-depth analysis of networks and structures of networks see [Newman, 2003]. The current research focuses on this aspect of the multi-robot systems. The networks of cooperating robots should try to build a common ontology, based on the observations of the robots. (An ontology is a model of the world; in this case a model of

the direct environment in which the robots operate). This is a challenging problem, because different robots may have sensors that measure completely different aspects of the real world, for instance some robots may detect chemical weapons while other robots may detect magnetic anomalies. Although this contribution concerns material robots, many mechanisms, such as coordination [Storms and Grant, 2006], are the same for networks of software robots (or better: software agents).

A few examples of recent military applications

Today, there are already many military applications for autonomous robots. There is a growing interest in cooperative robot systems. A well known underwater "robot" is the REMUS (Remote Environmental Monitoring Units) which is an autonomous unmanned submarine. The REMUS can carry a variety of sensors to meet the mission requirements. More than one REMUS can be used, all with different sensors if desired. In the near future, the REMUS will be equipped with technology that will allow the submarines to communicate with each other using underwater acoustic modems. A new philosophy of the US Navy and many other navies is to develop lots of cheap unmanned undersea vehicles (UUVs), because with many cheap UUVs it is not so bad if a few are lost during a mission. There is a growing interest in cooperative UUVs as well [Wernli, 2000].

Unmanned ground vehicles (UGVs) are already used for mine clearance and urban reconnaissance, but usually they are operated from distance ("teleoperation"). However network-centric autonomous ground vehicle systems are in development and already at the demonstration stage (e.g. see [Committee on Army Unmanned Ground Vehicle Technology, 2002]).

This year, a special issue of the International Journal of Robust and Nonlinear Control [Rasmussen and Schumacher, 2008] appeared, filled with papers on cooperative unmanned aerial vehicles (UAVs) in a military context. Combinations of cooperative UGVs and UAVs may prove to be very useful in unknown hostile environments.

An in-depth study of collaborative core technologies used in networks of autonomous robots together with many possible military applications can be found in a report of [Singh and Thayer, 2001].

Our current research

Within our research, we want to develop and test new algorithms and paradigms rather than constructing a completely new operational system. Therefore, it is not necessary to build full scale robots. Instead, our current research uses miniature robots and computer programs that emulate robots. Many of the problems that may occur with real, full scale robots can be solved and tested by simulations and by using miniature robots. By limiting ourselves to small scale robots and simulations, we are able to obtain many of the desired results much faster and cheaper than would be possible with full-scale robots. The current research started at the end of 2007, so we are not able to present scientific results yet. However, we have already implemented a working system for the determination of

mutual distances and orientations of the robots. This localization system is described in more detail after the next section.

The next step in the development of the system hardware will be the addition of sensors that can detect obstructions. Moreover, one of the robots will be equipped with a stereo-vision system so it will be able to explore the 3D world in its immediate neighbourhood.

Educational relevance

This research project is very well suited to be used for bachelor thesis projects of our Military Systems and Technology education. In particular the project assignment that is scheduled for the third year. Over the past two years, we have already had two groups of students working on a project involving cooperative robots. They used miniature mobile robots, so-called Boe-Bots ("Board of Education Robots"), which are software compatible with the so-called ARobots we use. The students equipped their robots with several sensors and communication provisions. They were able to develop software in several computer languages to make it possible for the robots to communicate wirelessly with another robot, with a personal computer notebook and even with a handheld computer. In the future, very interesting project assignments are possible, for instance projects where as part of a strategic scenario Boe-Bots must try to disturb the mission of the ARobots or try to help the ARobots. These kinds of "strategic games" are particularly interesting if the two groups do not know each others intentions. During these projects student can gain experience on Artificial Intelligence, Wireless Communication, microprocessors and in writing realistic computer programs.

Localization system

At the moment, we are building a number of test robots (see Fig. 1). These robots are equipped with a system to determine the position of the robots by means of sound. This is very similar to the determination of the position of submarines by sonar. Although in many cases a GPS system may be available, we must consider the possibility that this is not the case at the battlefield. Our tests with the model robots will often be conducted indoors, so we certainly cannot use GPS. GPS signals are too weak to penetrate buildings and standard GPS is insufficiently accurate for use with small robots outdoors.

The reason for the use of sound instead of radio waves, is purely because with the current state of the art in electronics it is impossible to obtain the time-resolution that is needed for centimetre-resolution using cheap and small electronics. When larger robots in the open field are used, it will be no problem to use radio waves, because for most real-world applications a position accuracy of several decimetres will be sufficient. Furthermore on larger robots the antennae can be placed at a larger mutual distance, thus increasing the time differences between the arrivals of the waves. In fact, radio waves in many ways will be simpler to use, for instance because very much higher update frequencies can be used and the speed of electromagnetic waves is much more constant than that of sound.

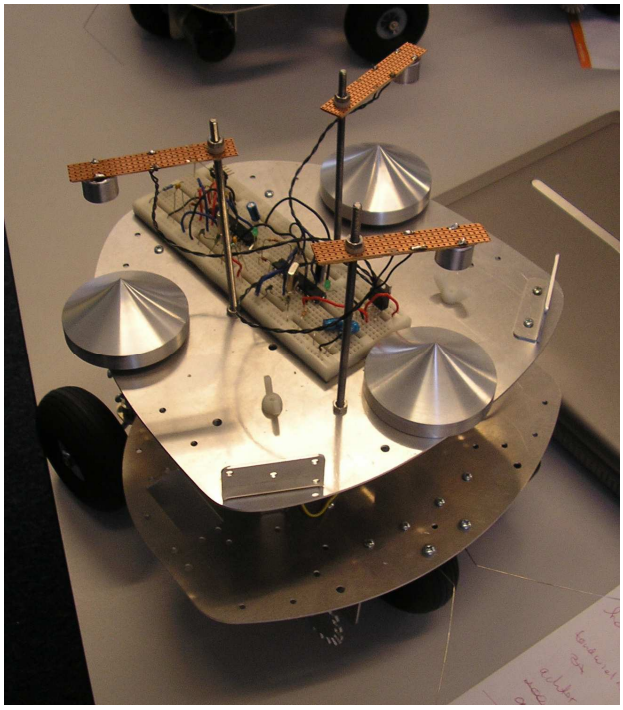


Figure 1. One of our – partially assembled – robots with three parabolic "ears"

A serious disadvantage in our employment of sound signals, compared to radio signals, is that sound can only be used if the robots move strictly on a plane surface. This is because most of the sound intensity produced by cheap 40 kHz transducers is confined to a cone of about 30° across, so many transducers would be needed to direct the sound intensity into all directions. This would make the system very complex and too energy-consuming. It is rather easy to direct the sound into all directions within one (horizontal) plane by using a reflective parabolic cone (see Fig. 2). Tests have shown that with this simple provision the sound intensity is still enough to be used up to about 10 m, which is about the maximum distance between the robots we will be using in our indoor test environments.

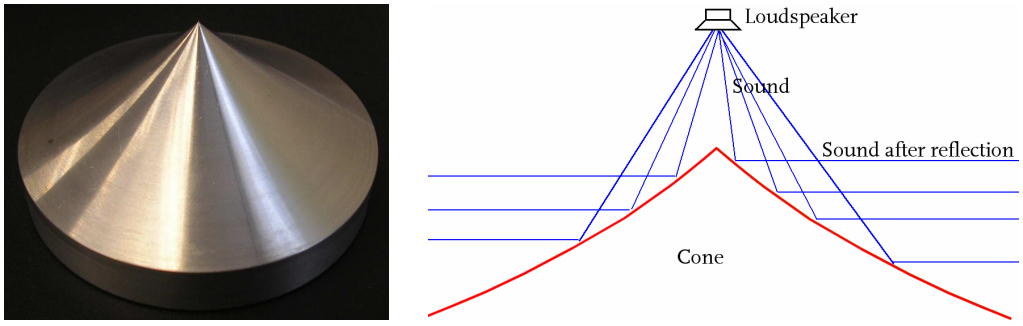


Figure 2. Left: parabolic cone made of hard alumina, used to spread the sound sideways. Right: principle of operation.

In the real world, there are several ways a robot can determine its absolute position, all of which may not be available when needed the most. If GPS is available things are simple, but the GPS can be jammed by the opposing force at any time or the signal may be lost when the robots for instance enter a dense wood. The position can also be deduced from the visual environment, for instance by cameras carried by the robots, or provided by an unmanned air vehicle. Sometimes it may be possible to make use of radio beacons as well.

We shall start developing a system that consists of an accurate sound beacon and "ears" on the robots to determine the position and heading with respect to the beacon or to another robot. The robots cannot determine their absolute position, but they can determine the relative position of other robots or beacons by the ultrasound system. If one of these objects is able to determine its absolute position all robots can calculate their absolute position as well.

The principle we use for the determination of the relative position is based on the successful Maxelbot Trilateration Project of the University of Wyoming (see for instance [Heil, 2004]). In short the system works as follows: a robot broadcasts a radio signal, containing one or several codes. The codes can be used for instance for an identification code or to address other robots. One of the robots reacts by immediately sending an ultrasonic beep signal. The first robot measures the differences between the arrival times of the beep at his three ears and the time the radio signal was broadcasted. With simple triangulation the first robot can find the position of the second robot. Then the first robot sends this position to the second robot, together with the radio signal. Although this robot now knows where it is according to the first robot, he still measures the position of the first robot with respect to itself.

Since both robots measure all positions with respect to their own coordinate system, the coordinates that the two robots find for each other's position will be completely different. However, in the ideal case, these coordinates should describe the same *vector* in space, apart from a minus sign (see Fig. 3). So, by exchanging the measured position of the other robot in their own coordinate system, both robots are able to determine the position of each other's coordinate system. If we take the positive X-axis always along the symmetry axis in the direction of the front every robot can make a fair guess about the heading of the other robots. There may be some small deviations because of inaccuracies and because of the small difference between the moments the measurements take place. Preliminary experiments show that an accuracy of about 1 cm is feasible using cheap electronic components.

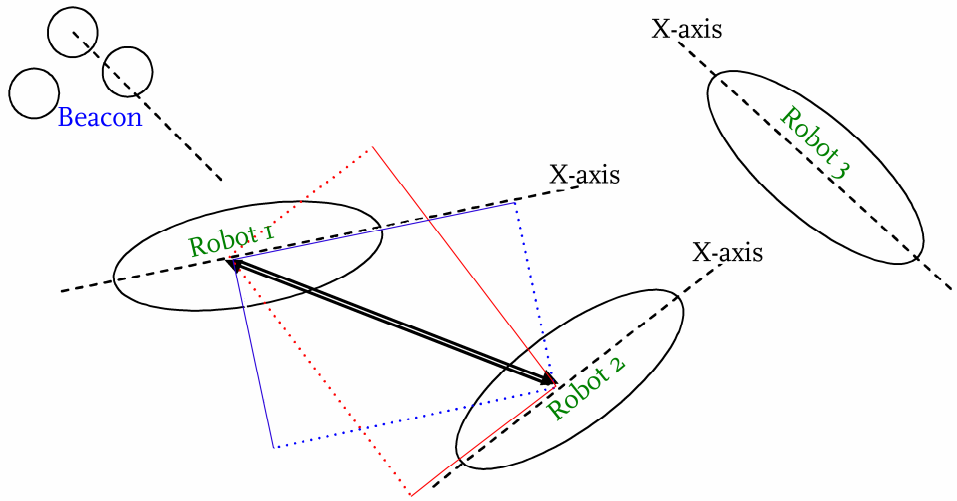


Figure 3. The dashed lines indicate the X-axis of each object. The relative coordinates of Robot 1 determined by Robot 2 (red solid lines) have no relation to the coordinates of Robot 2 determined by Robot 1 (blue solid lines). However knowing that the resulting position vectors must be exact opposites (black arrows), they both can calculate the direction of the X-axis of the other robot.

Now that the robots know not only their own position, but also each other's coordinate system it is possible to:

1. Recalculate everything to a generic coordinate system. This might be the coordinate system obtained from a beacon, or they may use their own system as a reference.
2. Robots that cannot hear the beacon, but can hear other robots may be able to know their position with respect to the beacon if there is a chain of robots, where each robot can hear its neighbour or the beacon. With an accuracy of 1 cm and a distance between the robots of about 10 m, simulations show that about 100 m away from a beacon the robots will still be able to know their absolute position (that is the position relative to the beacon) with an acceptable accuracy, provided the chain is available and works properly. In practice this will only be reliable if more robots are present in the neighbourhood of the chain. Once the chain is broken the robots may never be able to find the other robots again, because they then do not have a clue about their absolute position.

Conclusions

The research on cooperative robot systems of the Combat Systems section is still at a preliminary stage. A system for the determination of the relative position of robots has already been built and tested. With this system, together with computer simulations it will be possible to test many scenarios and principles quickly and without the need for expensive devices. Next, the focus of this research project will shift to the formation of ad hoc networks of cooperating robots and to technologies to share non-compatible information (from different sensors) between robots. At the same time we must implement methods to share a common model of the environment between robots that participate in ad hoc networks.

In the near future many interesting and challenging educational projects involving students from the bachelor-level degree programme Military Sciences and Technology can be done in cooperation with this research project.

Acknowledgements

The authors want to thank prof. dr. T.J. Grant for many stimulating discussions.

References

Barnes, D. and Gray, J. (1991) Behaviour synthesis for cooperant mobile robot control. *Proceedings of the International Conference on Control*, pp 1135–1140.

Cao, Y.U, Fukunaga, A.S. and Kahng, A.B. (1997). Cooperative Mobile Robotics: Antecedents and Directions, *Autonomous Robots,* 4, pp 1–23.

Heil, R. (2004). *A trilaterative localization system for small mobile robots in swarms.* Master's thesis, University of Wyoming, Laramie, WY.

Committee on Army Unmanned Ground Vehicle Technology, National Research Council (2002), *Technology Development for Army Unmanned Ground Vehicles*. The National Academies Press.

Rasmussen S.J. and Schumacher C.J., eds. (2008) *International Journal of Robust and Nonlinear Control, 18(2), Special Issue on Cooperative Control of Unmanned Aerial Vehicles*, Wiley InterScience.

Newman, M.E.J. (2003). The Structure and Function of Complex Networks. *Society for Industrial and Applied Mathematics (SIAM) Review* 45 (2), pp 167-256.

Singh, S. and Thayer, S. (2001) *ARMS (Autonomous Robots for Military Systems): A Survey of Collaborative Robotics Core Technologies and Their Military Applications*, tech. report CMU-RI-TR-01-16, Robotics Institute, Carnegie Mellon University.

Storms, P.P.A. and Grant, T.J. (2006) Agent Coordination Mechanisms for Multi-National Network Enabled Capabilities, *Proceedings 11th ICCRTS Coalition Command and Control in the Networked Era*

Wernli, R.L. (2000) Low cost UUV's for military applications: Is the technology ready? Space and Naval Warfare Systems Center San Diego. *Defense Technical Information Center,* Report Number: ADA422138.

Surface and Air Picture Compilation with Multiple Naval Radar Systems

Umesh Ramdaras[*] & Frans Absil

Introduction

Recent advances in Information and Communication Technology have had their effect on the architecture and the concept of operations for military systems. Increased connectivity of sensor, weapon and command and control (C2) systems is an enabler for *Network Centric Warfare* (NCW) [Alberts et al., 2000], [Cebrowski and Garstka, 1998]. The NCW concept, essentially a *distributed system*, may be depicted as a layered set of three grids: a sensor, shooter (i.e., weapons) and information grid, with the grid nodes representing individual military systems. The links within and between the grids represent the connectivity.

Introducing such a concept will have an effect on military operations; this effect is denoted as Network-Enabled Capabilities (NEC). Key NEC characteristics and processes are:

- multi-sensor data fusion, i.e., using observations from a multitude of sensor systems to compile an integrated operational picture;
- increased situational awareness, i.e., a better understanding of the operational picture in terms of military threats and own capabilities;
- information superiority. Through increased connectivity and higher link bandwidths (enabling higher data transfer rates) all parties in the network should have faster and better knowledge of the current battlefield status than an opponent who has no or a less capable network.
- A more rapid C2 loop, also indicated as the Object-Orient-Decide-Act loop [Boyd, 1992]. Based on the previous processes and with increased connectivity between Command, Control and Communication (C3) systems, the commander should be able to increase the pace of decision making and keep the momentum on the battlefield. This includes quick assessment of the outcome of the military effect.
- Modernisation of the command hierarchy, indicated by terms such as self-synchronisation and delegated authority (see [Alberts et al., 2000]).

These developments will obviously affect maritime operations. Typical surface ships such as corvettes, frigates and cruisers may contain a suite of sensor systems, dedicated to a specific warfare domain. Radar systems will search, detect and track air and surface objects, while sonars are listening for underwater sources. Typically, electro-optic systems (video, infrared, night vision, etc.) are used near the sea-air interface. Sensor systems may also play a role in the fire control process, when deploying weapon systems. The role of the combat system designer is to integrate the on-board sensor, weapon and C3 systems (the hardware) and implement military capability in the system architecture through a Combat Management System (CMS, i.e., the software).

[*] Umesh Ramdaras is with the Combat Systems Department of the Netherlands Defence Academy as well as with the International Research Centre for Telecommunications and Radar of the Delft University of Technology.

Coordinated system deployment is already an issue on a single ship. Modern, electronically steered radar systems, such as the Thales SMART-L and the APAR on-board of the RNLN Air Defence and Command Frigate have the capability to rapidly switch between radar modes or functions, and to adapt the radar settings for each of the modes or functions. Maximising the benefit from such advanced radar systems requires optimisation of their deployment and settings, a process known as *sensor management*. Also coordination between ship, sensor and communication systems is required, e.g., to prevent interference for systems with overlapping operational frequency ranges.

Moving towards an NCW architecture the coordination between multiple ships, or between ships and other platforms, such as combat aircraft, helicopters or Unmanned Aerial Vehicles (UAVs), will require attention. In a distributed system the optimisation process is extended over multiple platforms, each potentially equipped with multiple sensors. Obviously, the degree of complexity increases. Limited coordination between multiple platforms already has been achieved with communication and data link systems (e.g., Link-16, Link-22), where various data types (message, voice, video, etc.) at the tactical or strategic level are shared between units. Also, the concept of Cooperative Engagement Capability, distributing raw radar data between ships, has been tested at sea [Johns Hopkins APL, 1995], [Sijtsma, 1995].

In order to realise the capability of a NCW systems concept the coordination between various naval units will have to be increased. The CMS may have to extend its functionality over multiple platforms, and sensor management will have to be applied across ships. The Netherlands Defence Academy (NLDA) has taken up this topic in a research program.

Automatic sensor management was investigated in a collaborative research initiative, called the STATOR (Sensor Timing And Tuning on Object Request, 2003-2005) project [STATOR, 2005]. This paper gives an overview of a second PhD thesis research project, started in 2004 as a contribution to the previous activities; sensor coordination is extended to a group of moving platforms. A network of maritime radar systems used for air and surface picture compilation will be considered. Sensor management may be divided into *sensor selection* and *sensor localisation*. The outcome of the sensor selection process is the appropriate sensor for doing an observation, while sensor localisation will position the platforms such, that they can best deploy their sensor capabilities in the near future. Terminology like 'appropriate' and 'best deployment' imply an *optimisation process*, minimising a *cost function* that acts as the driver mechanism for sensor selection and localisation.

With a properly working sensor selection process, global sensor deployment (for the entire sensor suite in the network) can be optimised. Suppose that for a given target scenario one is able to identify the best sensor to observe that target within a certain planning horizon, i.e., a number of time intervals ahead; in the meantime the other sensors might be used for other tasks, reducing overlapping observations and redundant sensor measurements. Current practice in maritime operations is space (e.g., allocating search areas or sectors to specific ships in a task group) and time domain separation in

the planning stage of an operation. An adaptive and near real time sensor allocation mechanism would mean a significant step forward towards implementing NEC, and would make better use of the distributed sensor resources.

This work will be limited to the task of *target tracking*, as part of the operational picture compilation process. Target tracking means that a sequence of sensor observations will be used (not necessarily from the same sensor) to estimate the target *state vector*, i.e., a set of attributes characterising the target. These attributes may include target position, speed and course (heading), and manoeuvres (accelerations). The target state is time-dependent and therefore during the tracking process the state vector will be continuously updated.

The cost function, i.e., the decision metric for sensor management, is derived from the target state vector. It is a measure of the accuracy in the target state estimate and will contain certain elements from the *state error covariance matrix*. Which elements will be considered in the cost function may depend on the sensor task, or the stage of a military operation. For a long range surveillance task (e.g., in the range 100 – 200 km) one is not interested in a highly accurate estimate of target speed; neither is the target altitude highly relevant, so the elevation angle need not be estimated with high precision. A sufficiently accurate range and bearing angle will do. If the target is incoming (which by itself is the outcome of the estimation of the radial velocity component between target object and sensor; an approaching target implies a positive closing speed) a more accurate position, speed and heading estimate will become relevant. If at some point the target turns out to be a threat, the exact position in 3-dimensional space must be known at each time instant. Before deploying countermeasures one has to make sure that the target is within the operational envelope of defensive weapons. A guided weapon will need a good estimate of the relative geometry between target and intercepting missile; a gun fired will require a highly accurate estimate of the predicted hitting point of the projectiles (extremely high target position, speed and heading accuracy). In general, the cost function should be a variable, mission-related driver of the sensor selection process.

Sensors obviously have different performance characteristics. Radars may or may not determine the relative target radial velocity component through the Doppler shift measurement. The measurement accuracy might be different for elevation and azimuth angle. One sensor might outperform another in measuring a specific target characteristic. A sensor need not yield an observation, every time it is pointed (looking) at the target; in practice the detection probability is smaller than one and there will be missed detections. The varying sensor performance characteristics have to be incorporated into the sensor management process.

Selecting a sensor from the sensor grid to perform a task could be done in different ways. It might be carried out randomly or with a preference for a certain sensor. In both cases a sensor is selected without taking its suitability into account. On the other hand, sensor selection could be based on prior knowledge or actual performance. In the first case the knowledge, gained from experience in similar situations in the past or from experts, is translated into knowledge networks (e.g., Bayesian Networks [Yilmazer and Osadciw, 2004], Fuzzy Logic [Molina López et al., 1995], etc.). In the second case sensor selection is

based on the instantaneous performance measures of the sensors (e.g., the error covariance matrix).

The performance-based sensor selection algorithm (SSA), presented in [Ramdaras and Absil, 2006], compares sensors with respect to the *best expected performance*. For a specific scenario, at each time step within a planning horizon, the sensor with the relevant best expected target attribute accuracy does the observation. E.g., for a good target position estimate sensor selection will be determined by comparing the positional variance. Sensor performance evaluation is based on the Modified Riccati Equation (MRE). In [Boers and Driessen, 2006a, 2006b] it is shown that the best achievable error performance of the optimal state estimation filter, the Cramér-Rao Lower Bound (CRLB), has an upper bound determined by the solution of the MRE for the class of systems with probability of detection smaller than one. Besides, using the MRE yields a reduced computational load, compared to the CRLB.

In order to investigate the benefits of this MRE SSA, it has been compared to a selection algorithm based on the *trace* (diagonal elements) of the *updated predicted error covariance matrix* (TRACE SSA) [Chhetri et al., 2003]. In [Ramdaras and Absil, 2007a] results of the comparison of the MRE SSA with a random sensor selection (RSS) and a fixed sensor selection (FSS) scheme are included, while [Ramdaras and Absil, 2007b] presents simulation results for a set of different performance-based selection criteria. In all of these the effect of reduced detection probability of detection was taken into account.

However, remember that sensor selection may also be bounded by external factors that have to be taken into account. During a military operation rules of engagement will be in effect (e.g., prohibiting transmissions in certain bearing sectors). There may be criteria of physical nature (e.g., radar horizon, weather conditions) that limit the selection procedure.

This paper will give an overview of the sensor selection research topic. Results from an extensive MATLAB® computer simulation approach will be presented, assuming a single, non-moving platform with multiple sensors (the extension to multiple platforms is discussed at the end of the paper) and a single combat aircraft target. MRE based sensor selection strategies for different sensor selection criteria will be considered. Comparison between the MRE SSA and RSS, FSS and TRACE SSA will demonstrate the usability and benefits of the MRE SSA. Sensor selection strategies and performance evaluation will be discussed for several planning horizons and for various values of the detection probability.

State estimation

State estimation can be done with the Kalman Filter (KF) [Kalman, 1960], [Bar-Shalom and Fortmann, 1988], a first order recursive algorithm that will yield the minimum mean squared state estimate error for a linear state transition and observation model and assuming zero-mean Gaussian state vector and noise terms. The KF is used in many radar tracking applications as target state estimator.

Although extensions to the KF exist, the Particle Filter (PF) [Gordon et al., 1993] has become a popular method for stochastic dynamic estimation problems. This is due to the fact that it is possible to design for any nonlinear and non-Gaussian dynamic estimation problem an accurate, reliable and fast recursive Bayesian filter [Ristic et al., 2004]. The PF is population-based: at each instant of time the target state is represented by a set of particles, the particle cloud, that will move according to a state transition model, describing target motion. Examples of such target models are the constant velocity straight line trajectory and the horizontal turn. The set of particles can be drawn from any probability density function, thereby relaxing the Gaussian state vector and noise condition. The state transition model need not be linear, which makes the PF more generally applicable than the KF. As the observation from the sensor comes in, the processing will update the state vector for each individual particle. The target state vector and state error estimate are based on the ensemble-averaged statistics of the particle cloud. Like the KF, the PF is an iterative algorithm, that will yield state estimate updates at each new measurement.

The simulation in this work uses a 2-dimensional *x, y* orthogonal coordinate system. There are two sensors, S^1 and S^2, co-located at the origin on a single, non-moving platform. The data will be processed in the discrete time domain $t_k = kT$, where $T = 1$ s is the time interval between data points.

The target state vector $\mathbf{x} = \begin{bmatrix} x & \dot{x} & y & \dot{y} \end{bmatrix}^T$, where x and y are the target position and $\dot{x}$ and $\dot{y}$ are the target speed components (^{T} indicates the transpose). Measurement data are in polar coordinates (range r, Doppler $\dot{r}$ and bearing ϕ). Target motion is represented by a state-space process model. The process equation is given by

$$\mathbf{x}_k = \mathbf{F}\mathbf{x}_{k-1} + \mathbf{G}\mathbf{u}_{k-1} + \mathbf{v}_{k-1}, \tag{1}$$

where **F** is the state transition matrix, **G** is the input transmission matrix, **u** is the (optional) control input vector and **v** is the additive process noise, a zero-mean Gaussian process with covariance matrix $\mathbf{Q} = \mathbf{G}\mathbf{G}^T$ (probability distribution $p(\mathbf{v}) \sim N(0, \mathbf{Q})$). The matrix **F** relates the target state at the previous time step $\mathbf{x}_{k-1}$ to the current state $\mathbf{x}_k$. Here, target motion is described with a constant velocity process model. The matrix **G** relates the control input (e.g., evading manoeuvres in case of manned aircraft) to the target state.

The measurement model equation for both sensors is given by

$$\mathbf{z}_k^j = \mathbf{h}^j(\mathbf{x}_k) + \mathbf{w}_k^j, \tag{2}$$

where $\mathbf{z}^j$ is the measurement vector for sensor j ($j = 1$ or 2), $\mathbf{h}^j$ is the observer function and $\mathbf{w}^j$ is the measurement noise, a prior known, zero-mean Gaussian process with covariance matrix $\mathbf{R}^j$ (with probability distribution $p(\mathbf{v}) \sim N(0, \mathbf{R}^j)$). In this equation $\mathbf{h}^j$ relates the target state $\mathbf{x}_k$ to measurement $\mathbf{z}_k^j$. Furthermore, the probability of detection $0 \le p_d^j \le 1$.

In this work two different sensors will be defined. Sensor S^1 yields range (r), Doppler ($\dot{r}$) and bearing (ϕ) information

$$\mathbf{z}_k^1 = \begin{bmatrix} r_k \\ \dot{r}_k \\ \phi_k \end{bmatrix}, \ \mathbf{h}^1(\mathbf{x}_k) = \begin{bmatrix} \sqrt{x_k^2 + y_k^2} \\ \dfrac{x_k \dot{x}_k + y_k \dot{y}_k}{\sqrt{x_k^2 + y_k^2}} \\ \arctan \dfrac{y_k}{x_k} \end{bmatrix}, \ \mathbf{R}^1 = \mathrm{diag}\left(\begin{bmatrix} \sigma_r^2 & \sigma_{\dot{r}}^2 & \sigma_\phi^2 \end{bmatrix}\right), \tag{3}$$

whereas S^2 provides range and bearing only

$$\mathbf{z}_k^2 = \begin{bmatrix} r_k \\ \phi_k \end{bmatrix}, \ \mathbf{h}^2(\mathbf{x}_k) = \begin{bmatrix} \sqrt{x_k^2 + y_k^2} \\ \arctan \dfrac{y_k}{x_k} \end{bmatrix}, \ \mathbf{R}^2 = \mathrm{diag}\left(\begin{bmatrix} \sigma_r^2 & \sigma_\phi^2 \end{bmatrix}\right). \tag{4}$$

Note that the sensors have different Jacobian matrices $\mathbf{H}^j$ [Bar-Shalom et al., 2001], that may be considered as a sensitivity measure. They contain a set of gradients that indicate how a measurement component will change with a variation of a state vector component, and these will be a deciding factor in the sensor selection.

The PF state estimator consists of a sequence of processing steps: particle cloud initialisation, particle cloud propagation (the prediction step) and measurement update (weighted update after the measurement). Updating the particle cloud requires a resampling process and this step is skipped in case of a missing detection ($p_d < 1$).

In Fig. 1 the predicted and resampled particle clouds are depicted for $t = 1$, 15, 33 and 55 s. Observe the large-sized initial particle cloud at $t = 1$ s in the lower right corner. The wide predicted particle clouds become narrower after a measurement update and resampling. Since the sensors are positioned at the origin, one could infer some knowledge from the shape of the particle cloud about the measurement accuracies of the sensors. E.g., for $t =$ 15 and 55 s a sensor with a good range accuracy, but poor bearing accuracy yields the measurements.

Target state ($\hat{\mathbf{x}}_k$) and accuracy ($\hat{\mathbf{P}}_k$) are calculated with the first and second statistical moment (i.e., mean and covariance, respectively) of the particle cloud.

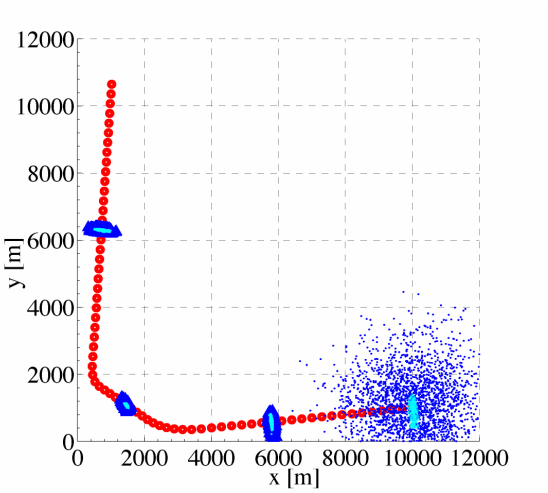


Figure 1. Predicted and resampled particle clouds of a particle filter state estimator. The red dotted line indicates a true target trajectory with states every 1 s. Note that the wide predicted particle clouds (blue dots) become narrower resampled particle clouds (cyan dots) after a measurement update. The clouds are plotted for t = 1, 15, 33 and 55 s with t = 1 s in the lower right corner. See Fig. 3 for more information about the target trajectory.

Sensor selection

The sensor selection algorithm (SSA) in [Ramdaras and Absil, 2006] is based on the Modified Riccati Equation (MRE) given by

$$\hat{\mathbf{P}}^j_{k+1|k} = \mathbf{F}\hat{\mathbf{P}}_{k|k-1}\mathbf{F}^T - p^j_d\mathbf{F}\hat{\mathbf{P}}_{k|k-1}\mathbf{H}^{j^T}\left(\mathbf{H}^j\hat{\mathbf{P}}_{k|k-1}\mathbf{H}^{j^T} + \mathbf{R}^j\right)^{-1}\mathbf{H}^j\hat{\mathbf{P}}_{k|k-1}\mathbf{F}^T + \mathbf{Q}, \tag{5}$$

where $\hat{\mathbf{P}}^j_{k+1|k}$ is the expected performance at time step $k+1$ for sensor j, $\hat{\mathbf{P}}_{k|k-1}$ is the predicted state error covariance matrix (that can be computed using the covariance of the particle cloud after the PF prediction step), p^j_d is the probability of detection and $\mathbf{H}^j$ is the Jacobian of the measurement matrix using the state estimate for the linearization process. One may observe that in (5) the sensor properties are included in p^j_d and the measurement accuracy $\mathbf{R}^j$. Also, for every sensor the instantaneous target state, and therefore the current geometry is represented in this equation by $\mathbf{H}^j$.

Criteria for sensor selection are based on considering specific elements from $\hat{\mathbf{P}}^j_{k+1|k}$ and minimising a cost function C^j. In [Ramdaras and Absil, 2006] the sensor selection criterion is the best expected target position accuracy (i.e., minimum positional variance in x and y, as expressed by σ^2_{xx}, σ^2_{xy} and σ^2_{yy} and therefore the cost function is defined as

$$C^j = \det\begin{pmatrix}\hat{\mathbf{P}}^j_{k+1|k}(1,1) & \hat{\mathbf{P}}^j_{k+1|k}(1,3)\\ \hat{\mathbf{P}}^j_{k+1|k}(3,1) & \hat{\mathbf{P}}^j_{k+1|k}(3,3)\end{pmatrix} = \det\begin{pmatrix}\hat{\sigma}^2_{xx} & \hat{\sigma}^2_{xy}\\ \hat{\sigma}^2_{xy} & \hat{\sigma}^2_{yy}\end{pmatrix}. \tag{6}$$

In [Ramdaras and Absil, 2007b] four alternative selection criteria are considered: best expected heading, range, Doppler and bearing accuracy.

In that case the cost function is expressed as [Zwaga and Driessen, 2005]

$$C^j = E[\mathbf{H}^c(\hat{\mathbf{x}}_{k|k-1})\hat{\mathbf{P}}^j_{k+1|k}\mathbf{H}^c(\hat{\mathbf{x}}_{k|k-1})^T], \tag{7}$$

where $\hat{\mathbf{x}}_{k|k-1}$ is the predicted target state vector and $\mathbf{H}^c$ is one of the following cases:

$$\mathbf{H}^c = \begin{cases} \left[\begin{array}{cccc} 0 & -\dfrac{\dot{y}}{\dot{x}^2+\dot{y}^2} & 0 & \dfrac{\dot{x}}{\dot{x}^2+\dot{y}^2}\end{array}\right] & \text{for heading,} \\ \left[\begin{array}{cccc} \dfrac{x}{r} & 0 & \dfrac{y}{r} & 0\end{array}\right] & \text{for range,} \\ \left[\begin{array}{cccc} \dfrac{\dot{x}}{r}-\dfrac{x^2\dot{x}+xy\dot{y}}{r^3} & \dfrac{x}{r} & \dfrac{\dot{y}}{r}-\dfrac{xy\dot{x}+y^2\dot{y}}{r^3} & \dfrac{y}{r}\end{array}\right] & \text{for Doppler,} \\ \left[\begin{array}{cccc} -\dfrac{y}{r^2} & 0 & \dfrac{x}{r^2} & 0\end{array}\right] & \text{for bearing/azimuth.} \end{cases} \tag{8}$$

Here, $r = \sqrt{x^2+y^2}$ and for convenience $\hat{\mathbf{x}}_{k|k-1} = [x_{k|k-1} \;\; \dot{x}_{k|k-1} \;\; y_{k|k-1} \;\; \dot{y}_{k|k-1}]^T$ is written as $\hat{\mathbf{x}}_{k|k-1} = [x \;\; \dot{x} \;\; y \;\; \dot{y}]^T$. For quantities that are directly measured, such as range, Doppler and bearing, $\mathbf{H}^c$ will contain the corresponding row from $\mathbf{H}^j$.

The optimal sensor $\hat{j}_k$ at time step k is selected by minimising the cost function as

$$\hat{j}_k = \arg\min\{C^j\}, j = 1,2,\ldots,j_{\max}. \tag{9}$$

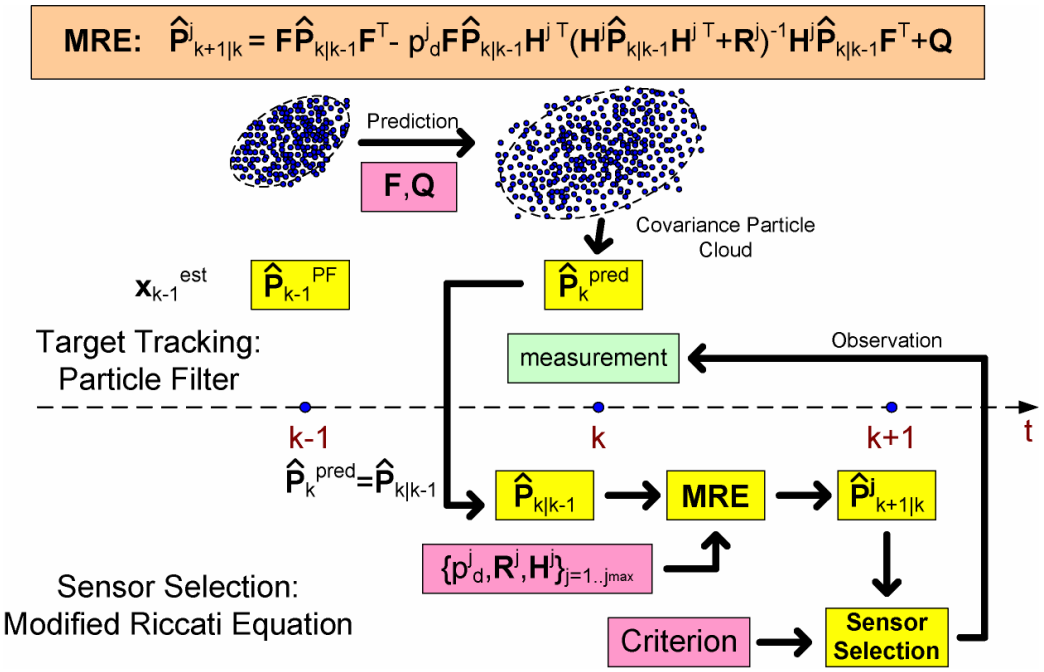


Figure 2. A schematic representation of the MRE SSA. The upper half of the figure represents the target tracking algorithm, while the lower part depicts the SSA (see explanation in the text).

In Fig. 2 the MRE SSA is shown in diagram. The upper half of the figure represents the target tracking algorithm, while the lower part depicts the SSA. The particle cloud at time step $k-1$ (see upper left) is progressed in the PF prediction step into a somewhat wider cloud at the current time step (see upper right). From this cloud $\hat{\mathbf{P}}_{k|k-1}$ is computed and used as input for the MRE equation together with the sensor properties (see upper red input block in the lower half of the figure). This yields an expected performance for each sensor. Based on the selection criterion (8) a sensor is selected to perform the measurement at time step k (measurement update step PF).

Simulation parameters

The computer simulation is based on a planar geometry with an area of 12×12 km^2 and a low-flying air target. The target trajectory consists of a closing and opening leg at zero altitude and constant flight speed of 300 m/s and two lateral manoeuvres (see Fig. 3): a 3g horizontal turn between $t = 22 - 28$ s and a 9g horizontal turn between $t = 38 - 40$ s. Two sensors are positioned on a non-moving platform at the origin. They represent two radar systems with a different measurement accuracy

$$\mathbf{R}^1 = \mathrm{diag}\left(\begin{bmatrix} 600 & 100 & 2.47\times10^{-4} \end{bmatrix}\right),\ \mathbf{R}^2 = \mathrm{diag}\left(\begin{bmatrix} 1000 & 2.47\times10^{-6} \end{bmatrix}\right).$$

Sensor S^1 yields good range and Doppler measurements ($\sigma_r = 24$ m and $\sigma_{\dot{r}} = 10$ m/s), while the bearing accuracy is poor ($\sigma_\phi = 0.9°$). Sensor S^2 yields no Doppler measurements (two elements in the matrix only), poor range information and high bearing accuracy ($\sigma_r = 32$ m and $\sigma_\phi = 0.09°$).

For the prior assumed probability of detection three cases are considered:

- Case 1: $p_d^{1,2} = 1$ (no missed detections),
- Case 2: $p_d^1 = 0.90$, $p_d^2 = 0.85$ (S^1 better than S^2) and
- Case 3: $p_d^1 = 0.85$, $p_d^2 = 0.90$ (situation reversed).

State estimation for target tracking is done with the PF. The particle cloud contains $N = 2500$ particles and for the resampling process Kitagawa's deterministic resampling algorithm [Doucet et al., 2001] has been used to get rid of the outliers relative to the observation. The PF update and resampling steps are skipped in case of a missing observation for a sensor with $p_d < 1$.

The performance of the sensor selection schemes is determined with simulated data over 70 s ($k_{\max} = 70$). The MRE SSA is based on one of the five sensor selection criteria: best expected position, heading, range, Doppler and bearing accuracy. In the case of fixed sensor selection (FSS) either S^1 (FSS-S1) or S^2 (FSS-S2) will be used along the entire target trajectory, in the case of random sensor selection (RSS) an uniformly distributed random variable will decide on which of the two equally probable sensors will do the next observation. The TRACE SSA is based on the minimum *trace* of the updated predicted error covariance as presented in [Chhetri et al., 2003].

The analysis includes the number of lost tracks, the quality of the state estimate and the sensor selection strategy. The results are ensemble-averaged over 1000 runs. Besides, the sensor selection strategy is evaluated for a planning horizon of $M = 1$ and $M = 6$ time steps ahead (for MRE SSA).

Simulation results

In the first part of this section MRE SSA is compared with RSS and FSS for $p_d^{1,2} = 1$. In the second part the MRE SSA is compared with the TRACE MRE for $p_d^{1,2} = 1$ and the third part describes two cases with $p_d < 1$ to demonstrate the benefits of the MRE SSA.

Comparison between MRE SSA, RSS and FSS for $p_d^{1,2} = 1$

The true target track and estimated target state are given in Fig. 3 for the three selection schemes and Case 1. Here state estimation with the MRE SSA is based on the best expected heading performance and is evaluated for a planning horizon of $M = 1$ and $M = 6$, while estimation with RSS and FSS implies $M = 1$. Besides, for FSS two options are considered: FSS-S1 (S^1 only) and FSS-S2 (S^2 only). Given the values of the measurement accuracies and due to the careful tuning of the PF there were no lost tracks.

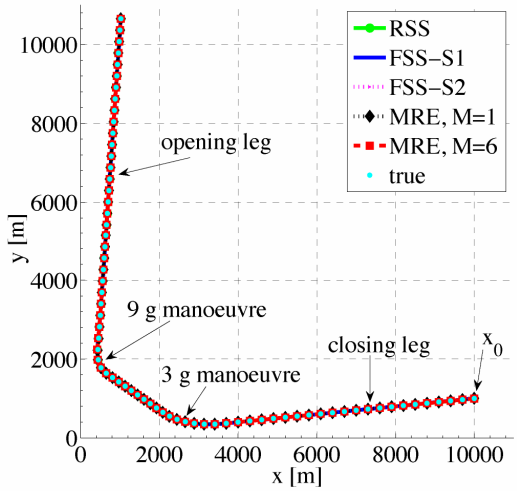


Figure 3. The true and estimated target state (average over 1000 runs). Sensor selection with the MRE SSA is based on the heading criterion. The planning horizons are $M = 1$ and $M = 6$; $p_d^{1,2} = 1$.

In the upper and middle plot of Fig. 4 the sensor selection strategy for Case 1 (MRE SSA based on the heading criterion), expressed as the number of times a sensor was selected, is depicted for a planning horizon of $M = 1$ and $M = 6$, respectively. S^2 is selected in the majority of cases. There is a clear change in sensor priority during the central part of the track, where S^1 with the additional Doppler information becomes more important. For $M = 6$ the change in sensor selection preference occurs after time instant $t = nM + 1$; note the changes after $t = 31$ s ($n = 5$) and $t = 43$ s ($n = 7$) in the middle plot. The lower plot shows the sensor selection strategy in case of random selection, and confirms the equal likelihood for both sensors. The trivial case of either FSS-S1 or FSS-S2 throughout is not plotted.

Whereas sensor selection based on best expected heading performance demonstrates sensor switching during the central part of the track, three alternative MRE sensor

selection criteria (best expected range, Doppler or bearing performance) show a different pattern. With these criteria and $p_d^{1,2} = 1$ there is no switching: the radar with the most accurate (i.e., the smallest) appropriate measurement error covariance term in $\mathbf{R}^{1,2}$ will be selected throughout the entire trajectory for both planning horizons $M = 1$ and $M = 6$ (not shown here). E.g., sensor selection with the range criterion gives preference to S^1, because $\sigma_{r_1}^2 < \sigma_{r_2}^2$.

In Fig. 5 the difference between the true and estimated target heading is given for the different selection schemes for Case 1: note the grouping of the MRE results with either FSS-S2 or FSS-S1. Sensor selection based on MRE SSA with the best expected heading accuracy criterion for $M = 1$ gives better results than RSS and FSS. Comparing MRE SSA for $M = 1$ with MRE SSA for $M = 6$ we may observe a difference in heading accuracy between $t = 3 - 8$ s, $t = 28 - 32$ s and $t = 42 - 45$ s. This corresponds to the difference in sensor selection strategies (see the upper and middle plot of Fig. 4). During the closing and opening leg FSS-S2 performs comparable to MRE SSA, during the central section there is agreement with FSS-S1. The RSS level lies between the lowest and highest target heading accuracy level. Note the humps in all curves between $t = 22 - 28$ s and $t = 38 - 42$ s due to the two lateral manoeuvres. The figure confirms that the actual performance is in agreement with the expected performance; it has been verified for the other selection criteria (not shown here).

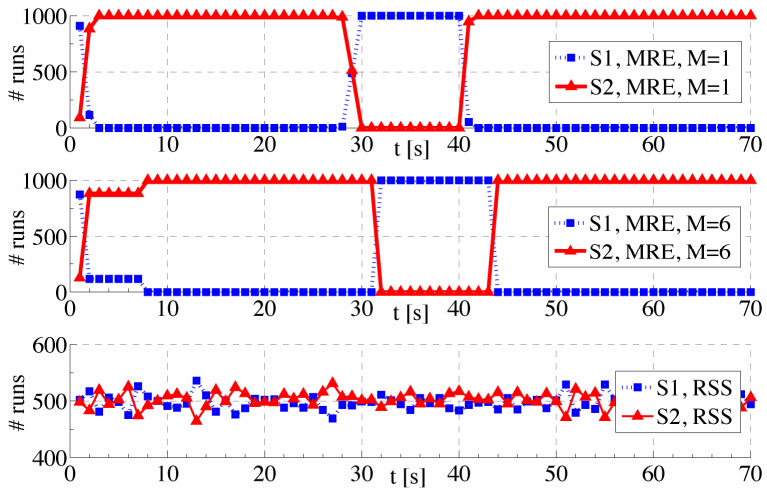


Figure 4. The sensor selection strategies with the MRE SSA based on the best expected heading accuracy (upper and middle plot) and with the RSS scheme (lower plot); $p_d^{1,2} = 1$. Note the dominance of sensor 2 in the closing and opening leg for MRE SSA; random selection is confirmed in the lower plot.

Comparison between MRE SSA and TRACE SSA for $p_d^{1,2} = 1$

Using the *trace* of the error covariance matrix for sensor selection as given in [Chhetri et al., 2003] implies an addition of dissimilar quantities (positions and speed) and available information in the error covariance matrix is not fully used. The comparison is done for Case 1 only (no missed detections) and the TRACE SSA was slightly modified: the *trace* of only the position elements in the predicted error covariance matrix is used, see (6).

In the upper plot of Fig. 6 the sensor selection strategy for the TRACE SSA is presented for Case 1 and $M = 1$, while the middle plot shows the selection strategy for the MRE SSA with the best expected position accuracy criterion. Observe that the MRE algorithm always

starts with sensor S^2 for $M = 1$, while the other algorithm always starts with S^1. For the closing and opening leg the sensor selection strategies are comparable, choosing sensor S^2 for the majority of cases; there the combination of highly accurate bearing plus somewhat poorer range accuracy yields the best position estimate. For the MRE SSA the additional Doppler information becomes more important during the central track section, as shown by the increase in S^1 use between $t = 30 - 50$ s. For the TRACE SSA the period of switching is longer ($t = 20 - 60$ s) and there is not a distinct preference for either S^1 or S^2. Between $t = 23 - 52$ s the selection strategy is comparable to the RSS strategy. The lower plot shows the selection strategy for the MRE SSA with $M = 6$ and the best expected position accuracy as criterion. One may notice the impact of extending the planning horizon from $M = 1$ to $M = 6$ during the central track section ($t = 30 - 50$ s): the preference for S^1 becomes more outspoken.

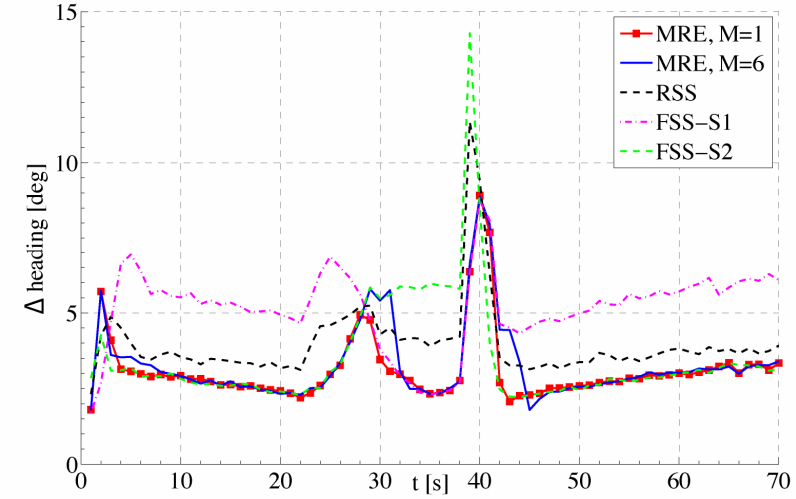


Figure 5. The difference between true and estimated target heading for three sensor selection schemes. For MRE SSA the planning horizons are $M = 1$ and $M = 6$; $p_d^{1,2} = 1$. Note how the performance of MRE SSA approaches that of a fixed sensor in different sections of the trajectory.

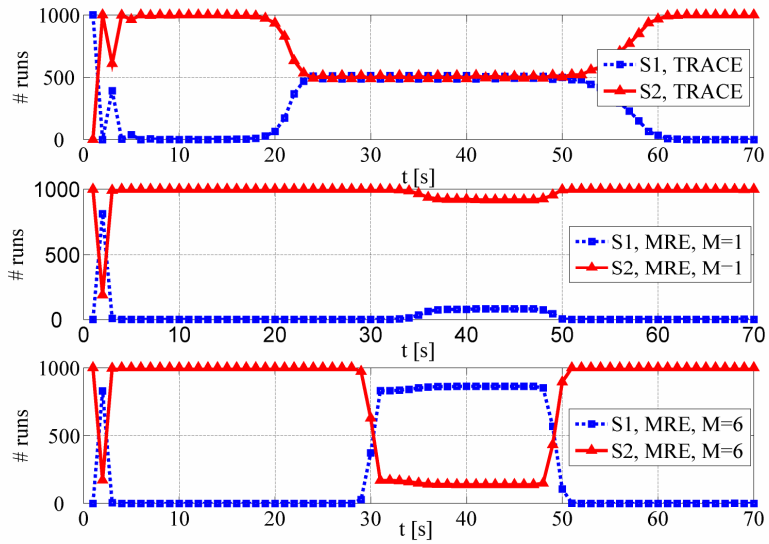


Figure 6. The sensor selection strategy of the TRACE SSA for a planning horizon of $M = 1$ (upper plot) and the MME SSA based on the best expected position accuracy, $M = 1$ and $M = 6$ (middle and lower plot); $p_d^{1,2} = 1$. During the closing and opening legs the sensor selection is comparable, while during the central section TRACE SSA is random.

Case 2 and 3 (lowering the detection probability)
In Figs. 7–11 the combination of selection criteria and p_d cases is deliberately chosen to test the flexibility and performance of the MRE SSA in a situation where either S^1 or S^2 currently yields the best accuracy.

The selection strategies for Case 1 (Fig. 4) and Case 2 for the opening and closing legs are comparable when the best expected heading criterion is used. Sensor 2 is selected in the majority of cases (see Fig. 7). There is a clear change in priority of sensors during the central part of the track, where S^1 with the additional Doppler information becomes more important. In Fig. 7 the sensor selection strategies for Case 2 are depicted for a planning horizon of $M = 1$ and $M = 6$. The lower plot shows the number of observations as a function of time. This is a check on the actual p_d during the simulations; note the central hump, due to the preference for S^1 with corresponding higher p_d.

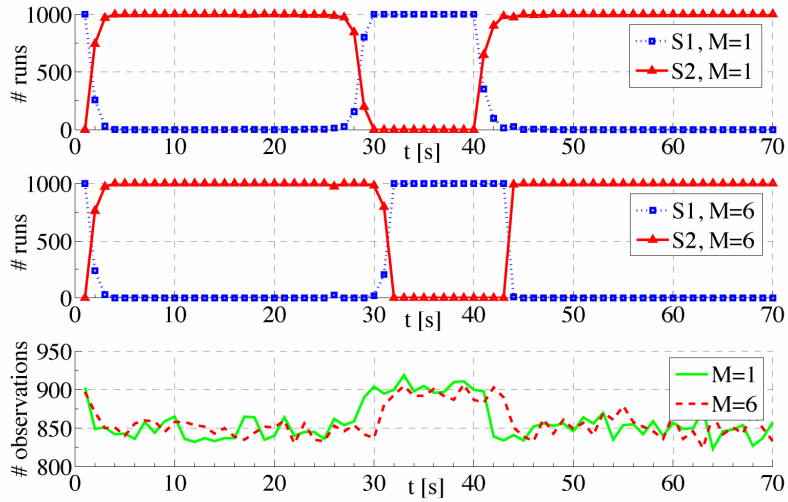


Figure 7. The sensor selection strategy based on the best expected heading accuracy (upper and middle plot) and the number of detections (lower plot) as a function of time, $P_d^1 = 0.90$ and $P_d^2 = 0.85$.

In Fig. 8 the sensor selection strategy of the MRE SSA with again the best expected heading criterion is depicted for Case 3 to compare the reversed situation of the p_d for the same planning horizon. Observe the complete dominance of S^2 during the entire track. The earlier preference for S^1 at $t = 1$ s and in the central part ($t = 30 - 40$ s) has completely disappeared. Obviously the effect of the swapped p_d-values outweighs the additional information obtained with S^1 during the target turns. The actual number of observations (lower plot) confirms $p_d^2 = 0.90$.

In Fig. 9 the selection strategy for Case 3 is presented for MRE SSA with the best expected range accuracy as selection criterion and RSS (upper and middle plot). The planning horizon is $M = 1$. Observe in the upper plot the complete preference for S^1 during the entire track except for $t = 1 - 2$ s. For $t = 3 - 70$ s the selection strategy is the same as FSS-S1. Although $p_d^1 < p_d^2$, the better range accuracy achievable with S^1 dominates the sensor selection strategy. In the lower plot the actual p_d's once again confirm the a priori assumed values during the simulations. Note the change from $p_d = 0.90$ to $p_d = 0.85$ for the MRE selection scheme during $t = 1 - 2$ s. This corresponds with the selection strategy during this part of the track. For RSS the average $p_d \approx 0.875$.

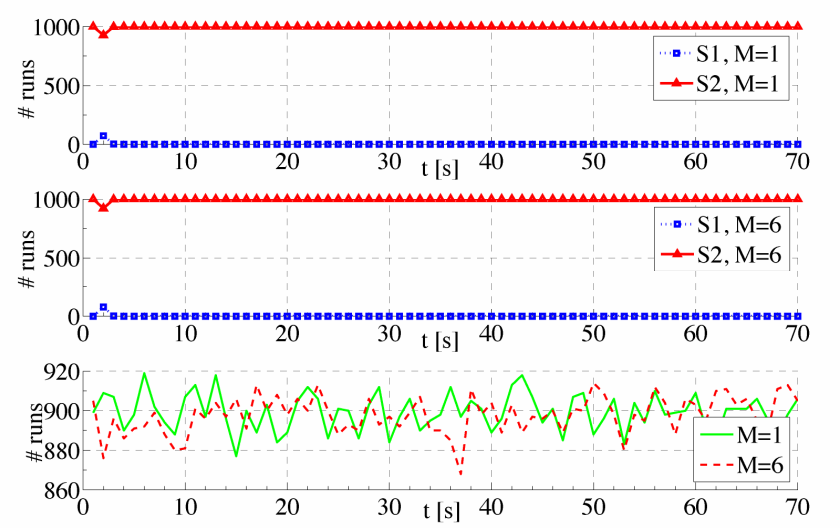


Figure 8. The sensor selection strategy based on the best expected heading accuracy (upper and middle plot) and the number of detections (lower plot) as a function of time, $P_d^1 = 0.85$ and $P_d^2 = 0.90$.

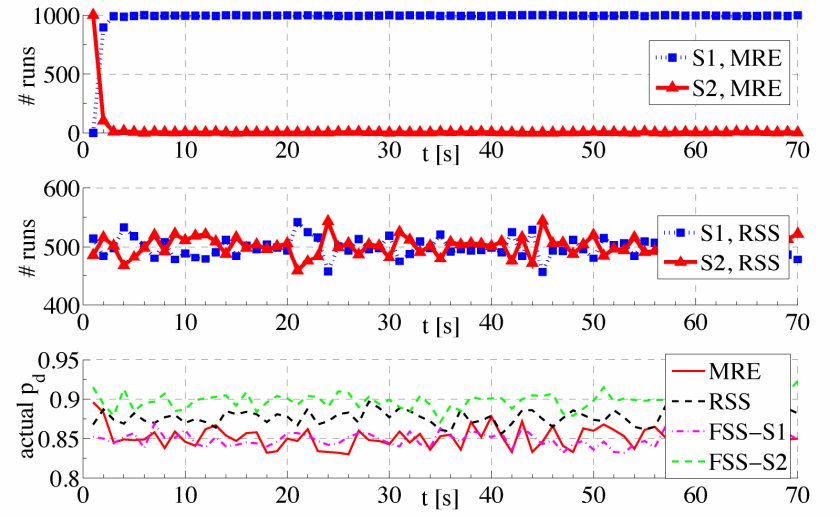


Figure 9. The sensor selection strategies with the MRE SSA (range criterion, $M = 1$) and the RSS scheme (upper and middle plot), and the actual P_d's (lower plot); $P_d^1 = 0.85$ and $P_d^2 = 0.90$.

In Fig. 10 the difference between the true and estimated target range is given for the MRE SSA based on five sensor selection criteria for Case 3 and $M = 1$: best expected position, heading, range, Doppler and bearing accuracy. As expected, sensor selection based on the MRE SSA with the range criterion performs better than the other selection criteria. There is one exception: MRE SSA with the Doppler criterion performs better during $t = 1 - 10$ s due to two effects: for $t = 1 - 2$ s the selection strategy based on the Doppler criterion has a solid preference for S^1, which is comparable to FSS-S1 (not shown here), while for $t = 3 - 10$ s (and the other parts where the Doppler line is below the range line) statistical effects become clear, since the process noise ($\mathbf{v}_k^j$) and the measurement noise $\mathbf{w}_k^j$ are simulated every run instead of using a fixed data set during all runs and for all criteria. As expected from the values of the elements in $\mathbf{R}^1$ and $\mathbf{R}^2$, S^1 indeed yields better range accuracy.

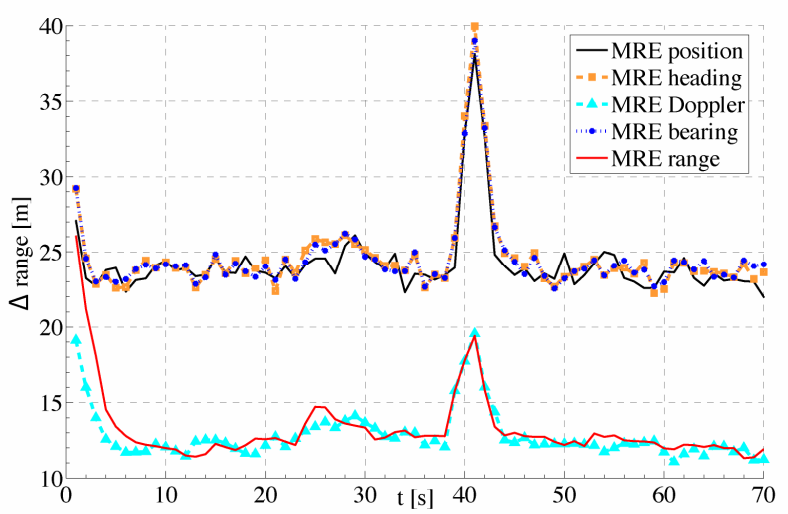


Figure 10. The difference between true and estimated target range for five sensor selection criteria with the MRE SSA. The planning horizon $M = 1$; $P_d^1 = 0.85$ and $P_d^2 = 0.90$.

In Fig. 11 the difference between the true and estimated target range is given for the three sensor selection schemes for Case 3. Observe that sensor selection based on the MRE SSA with the best expected range accuracy as selection criterion performs better than RSS and FSS-S2. FSS-S1 performs better during $t = 1 - 10$ s due to the difference in sensor selection strategy in the beginning of the track.

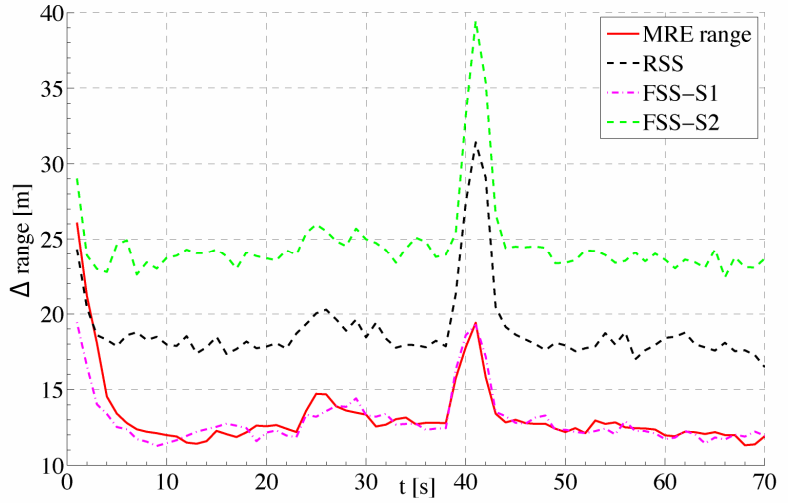


Figure 11. The difference between true and estimated target range for three sensor selection schemes. The planning horizon $M = 1$; $P_d^1 = 0.85$ and $P_d^2 = 0.90$.

The results of these two cases demonstrate the benefits and general applicability of the MRE SSA for a fairly realistic target scenario and sensor properties, such as the measurement accuracy and the probability of detection.

Conclusions, future research topics and implementation aspects

In current NLDA research on air and surface picture compilation with a network of naval radar systems a *sensor selection* approach is based on the Particle Filter target tracking technique and a minimisation approach to the cost function, derived from the Modified Riccati Equation (MRE).

In order to evaluate the benefits of the MRE sensor selection algorithm (SSA) this paper compares four different sensor selection schemes in a target tracking scenario: MRE SSA, random sensor selection (RSS), fixed sensor selection (FSS) and sensor selection based on the *trace* of the updated predicted error covariance matrix (TRACE SSA).

Several expected performance criteria were used for the MRE SSA that should yield maximum accuracy for specific target state parameters (best expected position, heading, range, Doppler and bearing accuracy). Several planning horizons (1 and 6 time steps ahead) were compared and the effect of reduced detection probability (p_d between 0.85 and 1.0) was studied. Results of the computer simulations include the sensor selection strategy (pick one from a set of two radars to do the observation) and the performance analysis (the quality of the state estimate during the target trajectory).

For the MRE SSA with the best expected range, Doppler or bearing performance as criterion, the sensor with the most accurate (i.e., the smallest) appropriate measurement error covariance term will be selected throughout the entire trajectory for both $M = 1$ and $M = 6$ planning horizons. The actual performance, expressed by the difference between the true and estimated state, is in agreement with the expected performance. These simulations serve to verify the applicability of this sensor selection algorithm for fairly realistic sensor properties and an air target scenario.

Sensor selection based on the MRE performs better than RSS. For the best expected range or Doppler performance the sensor selection strategy of the MRE SSA is comparable with FSS-S1, for the best expected bearing performance it is comparable with FSS-S2. These latter two cases were deliberately chosen to demonstrate the flexibility and quality of the MRE SSA in situations where a single sensor currently yields the best performance. For the best expected heading and position performance the MRE SSA gives better results than FSS. For the best expected position performance the MRE SSA gives better results than TRACE SSA.

For the values of p_d considered, there were no lost tracks and the quality of the state estimate is good. Although the ratio of the corresponding elements in the radar measurement error covariance matrices $\mathbf{R}^1$ and $\mathbf{R}^2$ dominates the sensor selection strategy for $p_d^{1,2} = 1$, the effect of the prior known values of p_d in the MRE is noticeable when the difference between p_d^1 and p_d^2 increases. For the MRE SSA an increase of the planning horizon from $M = 1$ to $M = 6$ has no significant deteriorating effect on track accuracy level.

Until now the research has focused on a thorough investigation of the various forms of the sensor selection algorithm. Extending the analysis to more realistic cases is straightforward for some aspects:

- Inclusion of the third spatial dimension, i.e., altitude, only means an additional term in the state and observation vectors. The selection algorithm is identical for a 3-D geometry and hardly needs software modification.
- Autonomously moving platforms can be incorporated in the selection scheme,

since the observations include relative kinematic quantities (currently both sensors are positioned at the origin).

- Multiple targets can be handled by applying the selection algorithm sequentially to each target. For each individual target the best suitable sensor will be determined over a certain planning horizon, and the selection strategy is based on target-related accuracy criteria. When simultaneous sensor claims arise from multiple targets, this has to be handled at a higher level of the sensor management architecture, where threat prioritisation and sensor scheduling take place.
- The way ahead for this work is to study the sensor localisation topic, i.e., moving the platforms around in such a way that future observations will again yield best expected accuracy for specific target state attributes, while maintaining adequate sensor coverage capability over a certain operational area. A number of localisation techniques will be implemented and tested in a computer simulation for realistic scenarios.

Next, a software architecture has to be defined, that somehow merges the sensor selection and localisation approaches. This requires definition of a certain sensor management hierarchy (localisation gets higher priority than selection, or the other way round). Another aspect of interest is the relevance of reliable communication links between the moving operational units. Research questions are: what happens if there is temporary loss of communications (no longer in line-of-sight, which is relevant in a wider area scenario where ships lie over the horizon and there is neither an aircraft nor a satellite relay station option)? How robust is the network sensor management approach? Can there be a fallback to the local optimisation process, with picture compilation at the single unit level? How is this temporary single ship approach later merged into a restarted global sensor grid optimisation?

At some point the mix of both sensor platforms and targets will have to be extended. It will be most interesting to see how the network sensor management strategy behaves in a setting with a wide range of kinematic and dynamic parameters (just think of the different speeds and manoeuvring capabilities of air vs. surface units).

If this research leads to a network sensor management strategy, that has demonstrated its capability in ample computer simulations, a most challenging next step would be to test the approach at sea. Since in the research set-up data exchange between units takes place at the plot level (i.e., target state attributes, no raw sensor data), communication bandwidth between units should not be a problem. However, at longer ranges in a more realistic exercise scenario (e.g., 200 × 200 km^2) the communication relay function will have to be realised (by either helicopter, UAV or satellite). And obviously, organising a test at sea involving multiple surface and air units and targets will be a complex and costly operation, but an unavoidable step on the way to operational deployment. As distributed system concepts with a multitude of military systems are becoming a reality, the requirement for global network optimisation and sensor grid management is evident. This research hopes to make a contribution to that end.

Designing the software architecture to coordinate the sensor grid as part of a Network Centric Warfare and realise Network-Enabled Capabilities certainly is a challenging and new subject of rewarding scientific research at the NLDA.

Acknowledgements

The authors would like to thank L.P. Ligthart and P. van Genderen, both with the International Research Centre for Telecommunications and Radar of the Delft University of Technology, and J.N. Driessen from Thales Nederland B.V., for their contribution to the scientific discussion on this subject.

References

Alberts, D., Garstka, J. and Stein, F. (2000) *Network Centric Warfare: Developing and Leveraging Information Superiority*. CCPR Publication Series, 2nd edition.

Bar-Shalom, Y. and Fortmann, T. (1988) *Tracking and Data Association*. Academic Press, New York.

Bar-Shalom, Y., Li, X. and Kirubarajan, T. (2001) *Estimation with Applications to Tracking and Navigation: Theory, Algorithms and Software*. John Wiley & Sons, Inc.

Boers, Y. and Driessen, H. (2006a) On the Modified Riccati Equation and its Application to Target Tracking. *IEE Proceedings - Radar, Sonar and Navigation*.

Boers, Y. and Driessen, H. (2006b) Results on the Modified Riccati Equation: Target Tracking Applications. *IEEE Transactions on Aerospace and Electronic Systems*, 42(1):379–384.

Boyd, J. (1987-1992) A Discource on Winning and Losing. Unpublished briefing notes, available at: http://www.d-n-i.net/richards/boyds_ooda_loop.ppt.

Cebrowski, K. and Garstka, J. (1998) Network-Centric Warfare: Its Origin and Future. *Proceedings of the Naval Institute*, 124(1):28–35.

Chhetri, A., Morrell, D. and Papandreou-Suppappola, A. (2003) Scheduling Multiple Sensors Using Particle Filters in Target Tracking. *IEEE Proceedings of the Statistical and Signal Processing Workshop*.

Doucet, A., de Freitas, N. and Gordon, N. (2001) *Sequential Monte Carlo Methods in Practice: Statistics for Engineering and Information Science*. Springer-Verlag New York.

Gordon, N., Salmond, D. and Smith, A. (1993) A Novel Approach to Non-Linear/ Non-Gaussian Bayesian State Estimation. *IEE Proceedings-F.*, 140(2).

Johns Hopkins APL (1995) The Cooperative Engagement Capability. *Johns Hopkins APL Technical Digest*, 16(4):377–396.

Kalman, R. (1960) A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME - Journal of Basic Engineering*, 82:35–45.

Molina López, J., Jiménez Rodríquez, F. and Casar Corredera, J. (1995). Fuzzy Reasoning for Multisensor Management. *Proceedings of the IEEE Conference on Systems, Man and Cybernetics: "Intelligent Systems for the 21st Century"*, 2:1398–1403.

Ramdaras, U. and Absil, F. (2006) Networks of Maritime Radar Systems: Sensor Selection Algorithm for $p_d < 1$ Based on the Modified Riccati Equation. *Proceedings of the IEEE Nonlinear Statistical Signal Processing Workshop: Classical, Unscented and Particle Filtering Methods*.

Ramdaras, U. and Absil, F. (2007a) Sensor Selection: the Modified Riccati Equation Approach Compared with other Selection Schemes. *Proceedings of the Tenth International Conference on Information Fusion*.

Ramdaras, U. and Absil, F. (2007b) Target Tracking in Sensor Networks: Criteria for Sensor Selection. *Proceedings of the IEEE Radar Conference*, pp 192–196.

Ristic, B., Arulampalam, S. and Gordon, N. (2004) *Beyond the Kalman Filter: Particle Filters for Tracking Applications*. Artech House, Inc.

Sijtsma, F. (1995) Cooperative Engagement Capability: Interessant voor de KM? *Marineblad*, 105(4):100–103 (in Dutch).

STATOR (2005) Sensor Tuning and Timing on Object Request (STATOR), Final Report. Technical Report KIM-IWW-2005-01, Royal Netherlands Naval College, International Research Centre for Telecommunications- transmission and Radar of the Delft University of Technology and Thales Naval Nederland.

Yilmazer, N. and Osadciw, L. (2004) Sensor Management and Bayesian Networks. *Proceedings of SPIE – Multisensor, Multisource Information Fusion: Architectures, Algorithms, and Applications*, 5434:238–248.

Zwaga, J. and Driessen, H. (2005) Tracking Performance Constrained MFR Parameter Control: Applying Constraints on Prediction Accuracy. *Proceedings of the Eighth International Conference on Information Fusion*, pp 546–551.

Sensor Synergetics: The Rationale of Sensor Fusion

Ariën J. van der Wal

PROLOGUE: Why study sensor fusion?

In this prologue we will introduce the subject and the motivation of this study through a simple example illustrating some of the problems associated with sensor fusion. Sensor fusion is the process of combining the individual information streams acquired with a number of sensors in order to achieve "better" situation awareness. How to combine these streams depends on both the specific goal of the observation process and, having defined a performance measure for fusion, how this measure is affected by external factors and intrinsic sensor limitations.

Let us imagine the following situation: An autonomous vehicle (a robot) is programmed to move from position A to position B. Positions A and B are located in the same horizontal plane in which the robot can move. In the plane there are a number of obstacles that have to be avoided by the robot. The robot is equipped with a number of sensors, say 3, in such a way that the robot can sense any obstacles that lay ahead in its path.

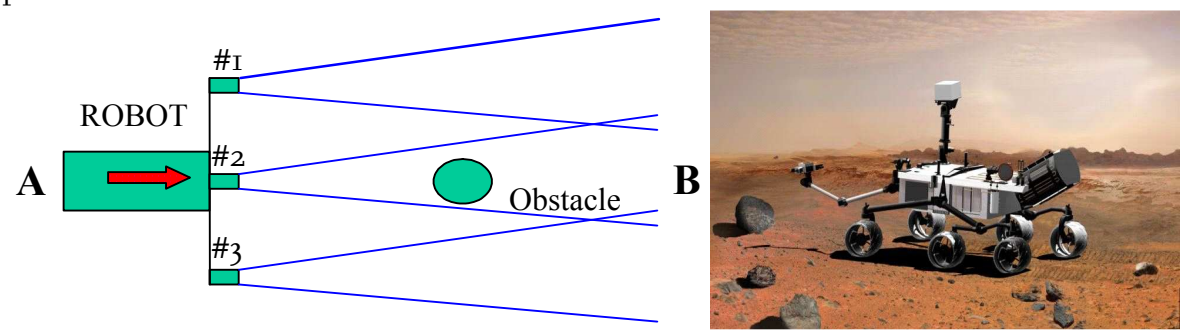


Figure 1. Top view of a robot with 3 identical forward-looking sensors moving from A to B using sensor fusion to avoid collision with an obstacle. The blue lines indicate the field of view (FOV) of each sensor. At the right an artist impression of such a robot is shown.

Now the following situation could arise: (cf. Fig. 1) Sensors #1 and #3 indicate that the way ahead is free, whereas sensor #2 "sees" an obstacle, say a tree, ahead. Basically there are two fundamental choices possible for a rudimentary sensor fusion mechanism: a linear and a non-linear scheme. The first would consist of a superposition (= linear combination) of the partial decisions made by the individual sensors, viz. "If no obstacle is sensed, continue route along original path, else change heading". A joint decision based upon the partial decisions with this rule and using the principle of a majority vote would result in a collision with the tree, since two out of three sensors arrive at the conclusion to continue with the original heading.

Also in the case that a layered (i.e. *delayed*) decision scheme would be used, problems could arise, as illustrated by the following: Suppose that the partial decisions of the sensors #1 and #2 are derived via the rule: "If sensor #1 indicates no obstacle and sensor #2 detects an obstacle, then turn left, else continue route along original path", and

similarly a partial decision is derived for sensors #2 and #3, we are faced with the situation where the partial decisions of the pairs (1,2) and (2,3) are to be combined. If we use superposition, again we end up with an undesirable result, because linear combination of the decisions "turn left" and "turn right" effectively results in steering straight ahead, thereby colliding with the obstacle.

Only if we combine the partial decisions in a non-linear way, the combination of three sensors could help to avoid a collision with the obstacle and therefore adds value to the robotic observation system, e.g. by postulating the sequence of following rules:

R1: If all three sensors do not detect an obstacle, move ahead.
R2: If sensor #1 (#3) sees an obstacle turn right (left).
R3: If sensor #2 sees an obstacle, throw a coin and decide to turn right (left) in case of heads (tails).

With this very simple example in mind it will be evident that linear combinations of sensor inputs have only limited use for making decisions. For this reason we direct our attention to *non-linear* combinations of partial decisions. In the present article we first discuss why effective sensor fusion necessarily needs to be a non-linear process and next sketch the use of partial and soft decisions and the way to aggregate these via non-linear schemes, in such a way as to arrive at useful and meaningful final decisions, based upon the available raw information from individual sensors. The mathematical formalism of fuzzy logic provides a versatile and adequate means to formally describe sensor fusion.

Introduction

Numerous research papers have been published dealing with the application of multi-sensor data fusion, also referred to as distributed, or "network enabled", sensing combined with high-level fusion, especially in the domain of military observations [1-6]. Although intuitively appealing, one may conclude that data fusion has not yet brought about the expected breakthrough. Several explanations for this can be given, such as the particularity of the application domain, the limited availability of general methods for fusion, and finally the quality of the primary 'raw' sensor data. Another problem may be the unrealistic expectations of the virtues of the synergy of multiple sensors.

In the absence of a general way to approach the subject, many ad hoc experiments and simulations have been published. In the following we will shortly review the history of fusion, define sensor fusion as a field of research in its own right, and next discuss the problem of how to model sensor fusion and suggest some directions for answering some of the pertinent questions in this field using concepts from soft computing. Especially the use of fuzzy measures looks promising as a way to model the sensor fusion process quantitatively.

Historical overview

Historically the idea of sensor fusion is not new: As early as in the sixties multi-radar trackers have been in use by the military for air traffic control and air defence. Multi-

sensor data fusion seeks to combine information generated by multiple sensors or multiple samples from one sensor to achieve goals that would be very hard or impossible to achieve with a single sensor, or a single sample. From the point of view of efficiency, scheduling, accuracy, and redundancy it seems intuitively obvious that several sensors should be 'better' than a single sensor.

Nowadays data fusion is a well-accepted method for making superior inferences in the field of industrial automation (e.g. for controlling a power plant, an oil refinery, a cement kiln; for a review on industrial applications, see e.g. [7, 8]), or even a nuclear reactor [9, 10], and for carrying out real-time pattern recognition in industry using a variety of sensors. Especially since the advent of soft-computing methods, such as fuzzy logic, data fusion has become a widely accepted successful fusion technology in industry. We note however that the success of such methods is primarily due to their ability to model human behaviour or expertise in supervisory control. Sensor fusion also endeavours to mimic cognitive processes in humans by absorbing the signals of the human observation system, i.e. our five senses, from the real world and integrate, or 'fuse', these signal streams to arrive at a coherent picture of our environment. As such, sensor fusion is concerned with lower abstraction levels and much higher information rates. It requires therefore faster response than the data fusion used in supervisory control systems. This forms also the key problem in applying soft computing methods to this field: in controlling complex industrial or organizational processes over relatively long timescales human operators have accumulated ample experience over the years. In contrast, there is only limited insight in the way a human being builds an environmental picture, his awareness, from continuous multi-sensate observations. It may therefore be a useful approach when studying sensor fusion to have a close look at how the human cognitive system works. Although cognition is still far from understood in detail, a few global characteristics are apparent: the human recognition system consists of a massively parallel processor that merges vague, qualitative inputs and a priori knowledge, acquired by learning from experience into a more or less consistent picture of the environment. It consists of a large number of hierarchically ordered decision processes running concurrently, simultaneously inferencing on the same set of input data at different levels of granularity, both in feature space and in space-time. We will not discuss these points in detail here, since they are outside the scope of this article.

Although sensor fusion is important to virtually all phenomenological sciences and engineering disciplines, most research has been done in the field of *defence* research. One reason for this can be understood as follows. In *analytical* approaches, e.g. in a physics experiment, the measured quantities or interactions are often so small that the experimental setup has to be designed in such a way as to make sure that the desired quantity or effect is optimally measurable. If the quantities to be measured are small, the experiment is repeated many times and ergodicity and statistics are used to arrive at average values with low standard deviation. Especially in the case where one tries to prove or disprove the correctness of a theoretical model, this often is a good approach. A final point to note here is that – apart from intrinsic physical real-time aspects – such experiments very often can be repeated many times and that real-time constraints are not a bottleneck.

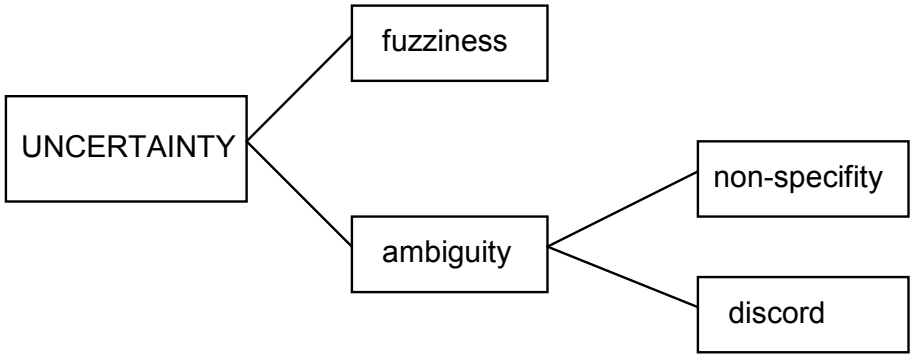


Figure 2. Taxonomy of different types of uncertainty

In engineering approaches the use of sensors is more *synthetic*, as illustrated e.g. in the field of factory automation. Here one deals with a well-defined problem such as the quality control of products on a manufacturing line, e.g. checking the soldering joints on a printed circuit board with an automated vision system. This problem certainly has real-time aspects, but the optimization can be done offline and the observation circumstances, like in the physics experiment, can be optimized offline, e.g. by testing the best combination of sensors, the proper cameras and illumination and parallel operation with more than one quality control station if the speed of production requires so.

In contrast, in military observation systems we deal with a situation that is far less comfortable than the situations described above: generally speaking it is necessary to assess in real-time an often complex situation, that almost certainly is outside one's complete control. Handling such observations requires the modelling of *uncertainty*. Apart from the ordinary problems such as noise and clutter, radar and electro-optical sensors operate also under adverse weather and atmospheric conditions, without any possibility to improve the circumstances of the experiment, or to repeat the experiment, under strict real-time constraints, with sometimes enormous consequences of false classification and even more serious penalties for non-detection. In addition, by the nature of the military métier, most targets of interest move at high speeds, try to actively or passively avoid detection or mislead sensors by jamming or using decoys and are designed in such a way as to present a minimal scattering cross section to commonly used sensors and thus to be virtually invisible ('stealth').

Under such circumstances it is clear that doing military observations invariably implies the modelling of uncertainty. Classically this is often done by applying statistical methods, notably Bayes' theorem to formulate a (multi-) hypothesis testing problem [11]. It is however clear that statistical uncertainty can only model part of the uncertainty. The different types of uncertainty, whose measures are now well established in classical set theory, fuzzy set theory, probability theory, possibility theory and evidence theory [11] are schematically summarized in Fig. 2. The breakdown distinguishes *fuzziness*, or vagueness due to a lack of definite and sharp conceptual distinctions and *ambiguity,* the situation where we are dealing with one-to-many relationships in the information obtained from sensors, yielding *non-specifity* in the case that the data leaves two or more alternatives

unspecified, or even *discord*, i.e. disagreement in choosing from among several alternatives.

Methods that explicitly deal with ambiguity and partially overlapping hypotheses such as Dempster Shafer theory [12, 13] and the application of belief functions instead of probability densities have become popular. Of more recent date is the application of general fuzzy measures [14]. The difficulty inherent to making accurate observations in military applications and the lack of measurement statistics are the prime motivations to improve single sensor observations by merging (partial) inferences/conclusions from one sensor with inferences from the another one. Recent history shows that the nature of military operations changes rapidly: Although sensors are vital to the success of any military mission, it becomes at the same time much more difficult to interpret these observations. This can be explained by the introduction of stealth technologies (radar), the subtleties of 'peacekeeping' missions compared to classical, full scale warfare scenarios, and finally the complexities and greater vulnerability of navy vessels operating close to shore ('littoral warfare' or 'brown water operations'). Finally it should be noted that there is a genuine need to fuse sensor-generated information, at least at the higher levels of command and control: the throughput of the man-machine interface being the limiting factor. Although new sensors have been developed (e.g. GPS) and accuracy and resolution in space and time of most existing sensors have greatly increased over time, the bandwidth of the man-machine interface has *not*. The situation of having to deal with more information than one can process in a certain time is not dissimilar to the situation where a *lack* of information exists. Both situations involve taking decisions in the presence of uncertainty and would benefit from intelligent data reduction techniques, such as sensor fusion.

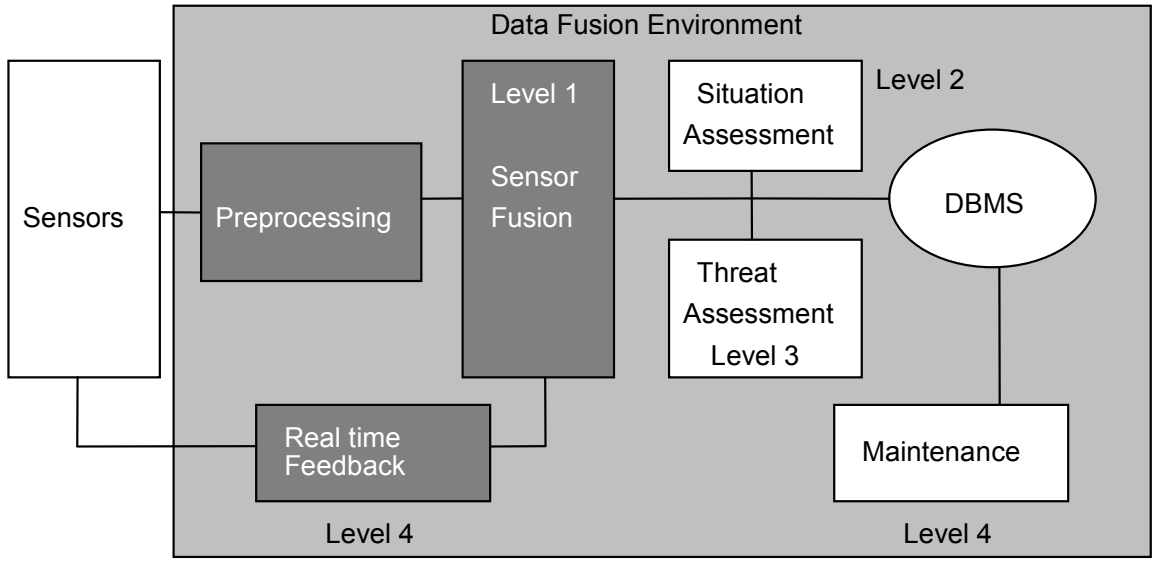


Figure 3. The JDL process model for data fusion distinguishes 4 levels of data processing. The darker areas indicate the scope of this paper. The data base management system (DBMS) provides the environmental information.

Sensor fusion vs. data fusion

Following the definition of the functional model of the data fusion process that is widely accepted in the military research community, as e.g. presented by Hall [15], and using the terminology as agreed by Joint Directors of Laboratories (JDL) of the Data Fusion Subpanel, multi-sensor data fusion is defined by the JDL as *"A continuous process of dealing with association, correlation and combination of data and information from multiple sources to achieve refined entity position and identity estimates, and complete and timely assessments of resulting situations and threats, and their significance."*

In this paper we restrict multi-sensor fusion to level 1 processing, with basic processes: data alignment, association and correlation, positional and identity fusion, complemented by the real-time part of level 4 processing ("maintenance") (Fig. 3). The reason for including part of level 4 perhaps seems strange at first sight, but is immediately apparent when maintenance is interpreted as the assessment of the status of each sensor *at short time scales* in order to keep it optimally tuned. Apart from optimizing individual sensors the monitoring of sensor performance makes it possible to perform *'sensor management'*, i.e. to optimize groups of sensors, which is e.g. important for military observation systems where a large number of sensors co-operate in a coherent way and where part of the sensors may be damaged during operation (cf. a phased array radar system). The short-time sensor management system contributes therefore directly to the *robustness* of a system.

Early sensor fusion can be viewed as a two-step process, a direct fusion step followed by a complimentary fusion step (Fig. 4). The direct fusion process acts immediately on *raw* sensor data, after a possible preprocessing stage for alignment. This type of fusion is in practice limited to combining signals from similar sensors. In the next stage, in the complementary fusion process, very different types of sensors can be fused. In this stage it is possible that a considerable data reduction is achieved and that the information can be represented as a vector in feature space. Features such as range, position, orientation, effective cross section, shape, colour, etc. are extracted from the different sensors and combined in qualitative or quantitative ways. Combination of information from complimentary sensors can thus be seen as augmenting the dimensionality of the feature space. After this fusion step all sensor information has been fused and next one needs to combine feature vectors with existing, a priori information about the environment, collected from previously measured data or intelligence. This more abstract fusion process is typical for levels 2 and 3 of the JDL model and will not be considered here.

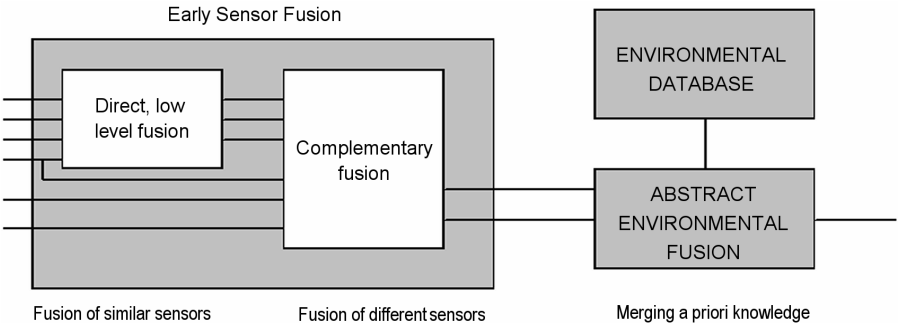


Figure 4. Global information flow in sensor fusion: the level of abstraction in the fusion process increases to the right, whereas the data rate increases to the left reflecting the data reduction caused by the fusion process

Sensor fusion

Motivation for applying sensor fusion

The general motivation why multi-sensor fusion is pursued is generally answered in terms of "to improve the combined observations of different sensors and thus create a better situational awareness". From the operational point of view possible benefits include greater user friendliness because of data reduction, a greater robustness of the resulting observation system, higher reliability due to a higher plot rate and therefore a better performance of e.g. a tracking algorithm, and thus a better observation, even under adverse circumstances. Scientifically however one immediately faces difficulties with this formulation of the motivation, because it introduces subjectivity: How can one measure "situational awareness"? An even more complicated question to be answered is *what* exactly can be improved in conventional, single sensor observation and *how* different sensors can benefit from each others measurements. Answering these two questions is by no means trivial and they can in fact only be answered by considering a specific application. Before doing so, it is worthwhile to take a moment and review the various abstraction levels, methods, hardware and software implementation methods of observation systems and the different ways that they can interact and co-operate. The basic difficulty is the absence of a unique theoretical framework for objectively combining the information streams generated by the various sensors. The difficulty lies essentially in adequately describing the information content in each sensor stream. The amount of *useful* information in a data stream created by a sensor is of course dependent on the ultimate goal of the total of observations. Whenever this goal cannot be formulated in a clear, transparent and unambiguous way, it will be extremely difficult to develop synergy between the sensors and to compare the performance of the various fusion algorithms.

Benefits and limitations of sensor fusion

We first note that most of the benefits quoted in literature, see e.g. Waltz [16], are benefits that are exclusively associated with the presence of *multiple* sensors; they are *not* the benefits of sensor fusion *per se*. Most of these benefits are qualitatively and intuitively immediately clear. Globally we can distinguish three types of benefits:

- The first type of benefit of multiple sensors is an extended spatial, temporal, or spectral coverage of the observed phenomenon.
- A second type of benefit follows from statistical arguments: multiple sensors increase the measurement accuracy and from this an increased confidence may be derived, or at least a reduction in the number of hypotheses about the targets and thus an improved detection rate, c.q. a shorter detection time. Only in the last two cases a sensor fusion step is needed.
- Finally multiple sensors create overlap in observations and thus redundancy. If this redundancy is properly exploited in the system design, the maintenance (level 4) module will optimize the sensor scheduling and will result in the graceful degradation of system performance if sensors breakdown.

A quantitative aim of sensor fusion is to improve the accuracy of the observation, e.g. the position of a target. An example of this is e.g. the combination of a forward-looking IR sensor (FLIR) and a radar sensor. The inaccuracy in azimuth and elevation of the radar sensor is compensated by the more accurate measurements of the FLIR sensor, while the pulsed radar accurately determines the range. This example illustrates how a radar can initially detect a target, because of its wider field of view. Subsequently the FLIR can be cued using the inaccurate coordinates of the radar to initiate the FLIR measurement. Together they determine a small region of interest around the target, so that the benefit of fusion is an improved estimate (or reduced uncertainty) of the position of the target.

An interesting attempt to illustrate in a quantitative way the virtues of sensor fusion is described in [17]. In the article an odd number N of identical sensors are fused with the aim to classify an observed phenomenon following a majority vote rule. The sensors are assumed to be statistically independent and the a priori probabilities are taken equal to $1/N$, corresponding to the principle of maximum entropy, equivalent with a minimum of a priori knowledge. Although the example is an idealized model case of identical, independent, unbiased sensors, all following the same statistics, using binary classification and a majority vote scheme as fusion aggregator, a number of qualitative results are worth mentioning here:

- Fusing data from multiple sensors (each with an individual probability of correct detection c.q. classification of less than 0.5) results in a decrease in performance in going from a single sensor classification to the multiple sensor fused result.
- If the individual sensors are very accurate (probability of correct detection larger than 0.95) sensor fusion cannot significantly improve the results of the inference process.
- The relative improvement in performance of an N sensor fusion process over single sensor performance increases as a function of N levelling off at about $N = 10$. Adding more, identical sensors does not pay off beyond this number (cf. Fig. 5).
- The maximum relative improvement of N sensor performance for $N \rightarrow \infty$ compared to a single sensor is of the order of 15-25 %, depending on the fusion scheme. The maximum marginal gain is reached if the single-sensor probability of correct detection is in the range between 0.60 and 0.75.

Of course the numbers mentioned above should be treated with care because they depend on the type of aggregation operator chosen to represent the fusion process. Moreover these conditions refer to the simplified case of *identical* sensors, i.e. same positioning, calibration statistics, biases, sampling rates, bandwidths, sensitivities, dynamic behaviour and the same measured quantities. If a new type of sensor is added to the sensor suite, the dimensionality of the observation is increased and a substantial increase in information content may be expected, depending on fusion objective.

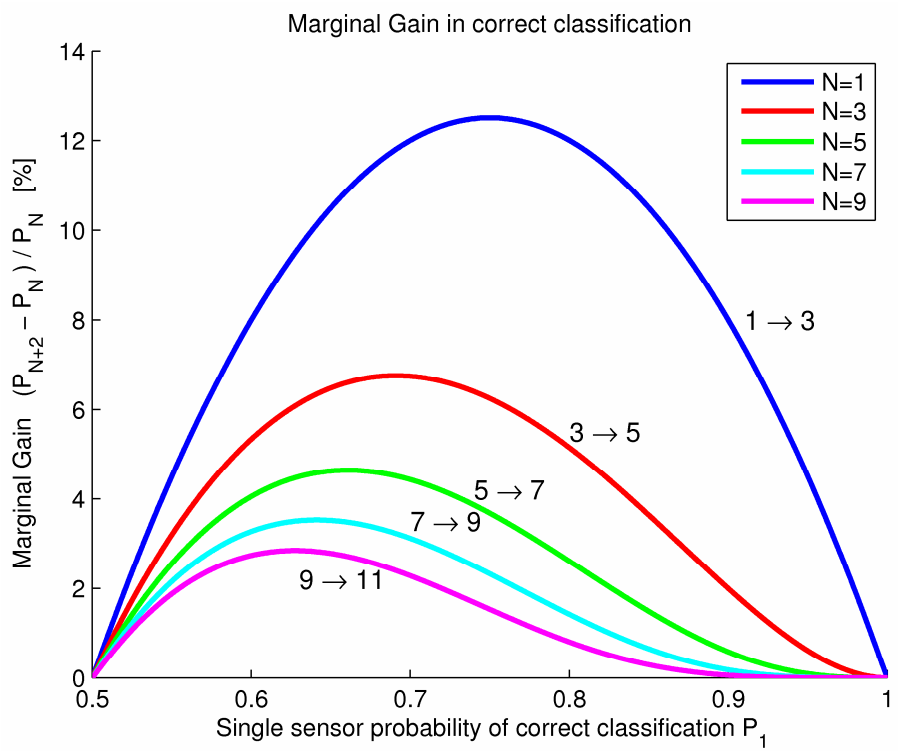


Figure 5. Marginal gain in the probability of correct classification by adding sensors in going from 1 to 3, 3 to 5, until 9 to 11 sensors as a function of the single-sensor probability of detection P, as calculated in [17]

In another recent study [18] the dependence of decision reliability on different fusion strategies for the case of 2 different sensors (i.e. sensors that are not identical) has been investigated based upon simple Boolean-type operators, along the lines of a probabilistic detection system (two hypotheses that are mutually exclusive and span up the whole universe of discourse). In a way this study extends the validity of the rules formulated above to more subtle fusion strategies than the 'majority vote', e.g. fuzzy decision strategies [19]. Also from these results we may conclude that for fusion to be successful the single sensor detection probability must be higher than 0.50 and that, in the case of two-sensor fusion, increases in detection performance of maximum 11% (from a single-sensor probability of detection P_1 of 0.80 to a two-sensor fused probability of detection P_2 of 0.90), or 17% (an increase from P_1 = 0.80 to P_2 = 0.941) have been calculated, depending on the fusion method.

Although these results do in no way preclude a substantial increase in the combined sensor performance brought about by a suitably chosen sensor fusion strategy, these numbers indicate that if measurements from identical, hard decision-making sensors are combined under the assumption of statistical independence, the marginal gain in performance will be limited to a percentage of a few tens, say 10-20%.

Sensor fusion: synergy

A general concept that is intimately connected to the idea of fusion is that of *ergodicity*, i.e. the concept that the outcome of an observation is unique, independent of the fact that one makes a series of consecutive measurements with one system, or that one makes N parallel setups and combines the N different outputs at one time. In the macroscopic physical world the concept of ergodicity generally is assumed to hold without exceptions,

although in microscopic physics some examples have been found in spinglasses, where ergodicity does not hold. For a review the reader is referred to [20]. Throughout this paper we will assume that ergodicity holds. However it should be pointed out that this statistical principle is sometimes difficult to apply in the real world, such as in military observations, because of the fact that we are mostly dealing with *single,* non-repetitive isolated events.

The fusion of sensor observations, i.e. the combination of observations obtained from the same sensor at different times (temporal fusion), or the combination of simultaneous observations taken by a number of equivalent sensors (repetition), or the combination of sensor observations with a priori information obtained from previously measured data, is the focus of attention in the present work. The key concept in fusion is how to take advantage of *non-linearity*.

We are not so much concerned with increasing the accuracy of an observation by repeating a measurement a number of times and thus reduce the statistical variation of the average. Rather it is our objective to extract additional information out of this data set (reduction of information) by correlating (*not* superimposing) the measurement with observations from other sources. Moreover in the case of fast moving targets it is generally impossible to obtain a sufficient number of samples to apply statistics.

Two different types of measurements are of interest in military observation systems:

1. Determination of presence, position, orientation and speed of a target.
2. Identification and recognition of a target (type, identification friend vs. foe etc.).

Although identification clearly is an entirely different characteristic of a target compared to its position and speed and although the latter can generally be determined at much larger distances than those at which identification can the accomplished with reasonable confidence, identification can help improve the accuracy with which speed can be determined and vice versa. In particular the identification of a target may be helped through a wealth of observations, whereas establishing the position and speed of a target can only be accomplished by the few sensors. The identification of a target will be accomplished more easily, because of the higher dimensionality of the 'feature vector', provided that a good database of properties is available. An example is in underwater acoustics where non co-operative target recognition of vessels by means of their acoustic signature is standard practice.

The basic problem in recognition is to exploit the high dimensionality and representing data in such a way that differentiation between various possibilities becomes easier. Therefore the task of sensor fusion is the combination of, possibly incompatible, measurements and to try and construct from these an improved environmental picture.

In this modelling one needs to include the confidence level of the new measurement, as well as *how* to combine this information with the already existing picture. Various schemes have been proposed in the past: e.g. Bayes' rule of combination from statistics, belief measures, and Dempster Shafer theory. Although most of these methods have a sound theoretical basis, their application sometimes lacks theoretical justification or simply yield non-intuitive results in specific situations. This makes it difficult to compare results that are obtained with different methods.

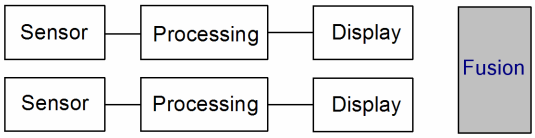
A systematic approach to Level 1 processing

In this section we will outline a practical approach to sensor fusion in military observation, following the theoretical framework referred to as Level 1 data fusion in the JDL model. Two generic tasks are of importance in almost all observation processes, viz. detection and recognition. Despite the fact that different sensors and methods are commonly used to accomplish these tasks, it is obvious that the successful completion of one task will almost certainly have a positive effect on the other one. If however the identification and recognition tasks are considered to be "hard" decisions that result from the independent processing of separate sensors, interaction of the two processes can only take place *after* the first process has reached a decision. In executing more complex tasks we have already indicated in the section where we discussed generic sensor fusion that it may be more advantageous to allow for partial, delayed, or "soft" decisions which may offer a way to separate the various goals and thus allow us to break down a complex task into a hierarchy of relatively simple decisions. Partial or "soft" decisions can be combined at earlier moments in the processing chain without discarding too much information and thus offer a method for applying early fusion of sensor streams. Before discussing in more detail how soft computing methods can be used to achieve sensor fusion, we will first review the signal processing steps that are necessary to benefit from sensor fusion.

The classical way in which fusion is applied is by transforming a physical measurement into some hard decision (e.g. a 'plot', 'track', 'identity', etc.) that is communicated to the user via the man-machine interface, generally an optical display. The fusion process of the information shown on a number of different displays then takes place *in* the mind of the operator, who assesses the situation and makes a threat analysis. All these fusion processes take place in the human mind, *after* the sensor signals have been processed completely (Fig. 6a).

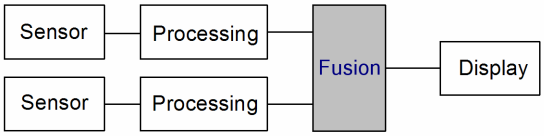
A first step towards true multi-sensor fusion is 'late' fusion (Fig. 6b): the construction of a special, goal-oriented architecture that fuses on the level of images, with the goal to enhance the image (e.g. combining IR and visible light images using some false colour scheme), or to fuse the consecutive plots of moving targets into a single track by taking into account some type of assumed target dynamics, or the fusion of tracks generated by different sensors (e.g. two radars or a radar and an electro-optical sensor).

a



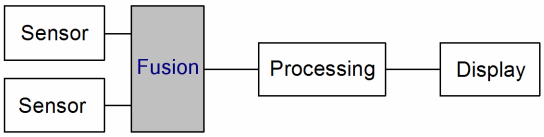
Fusion of information in human mind

b



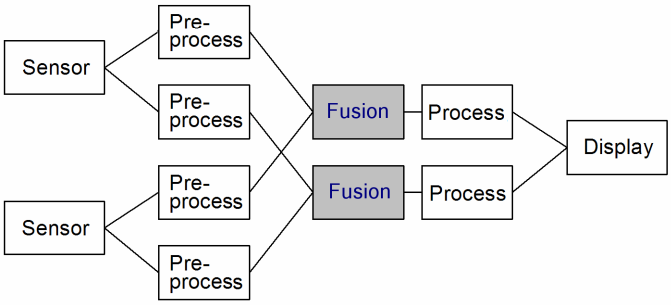
Late fusion: e.g. Image fusion, Track fusion

c



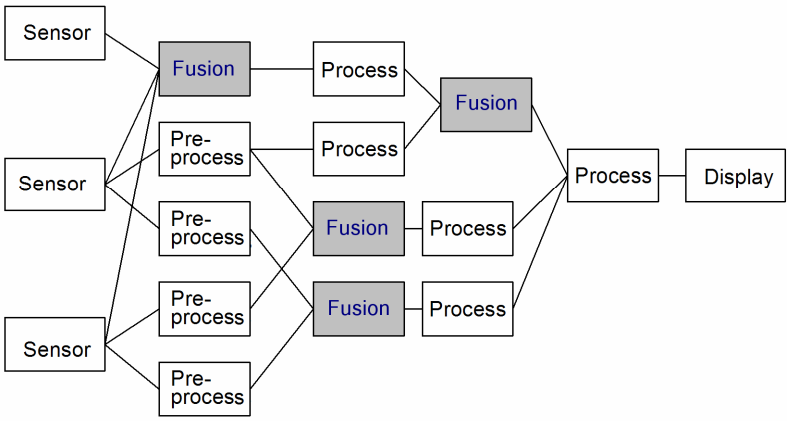
Early fusion with similar sensors, e.g. SAR

d



Early fusion with different sensors and with different fusion objectives

e



Distributed multisensor fusion: concurrent early and late fusion on different levels

Figure 6 a-e. Different fusion architectures starting from fusion in the human mind (a) and late fusion (b), via early, multi-sensor fusion with the same (c) and with different sensors (d) to the general case of multi-sensor multiobjective fusion on different levels concurrently (e)

The results are qualitatively better, especially if the sensors are on different platforms, or, alternatively, under difficult circumstances (adverse weather, jamming) when one sensor can be supplemented by the other. In the case of *early* fusion with many sensors of the same type, one can process directly on the raw data without much preprocessing, possibly except for bias removal. In this sense one may view SAR (synthetic aperture radar) also as a 'temporal' fusion process, combining different samples of single sensor to simulate a kind of multi-sensor phased array with a much larger virtual aperture than that of the actual sensor (Fig. 6c).

The early fusion process that we would like to discuss is schematically given in Fig. 6d: a collection of interacting sensors, each contributing to one or more decision trees of positional c.q. feature declarations. Each tree represents a hierarchical decision process, building up from fast, low level, noisy decisions, up to well-founded, abstract decisions that require some time to take. The sensor fusion architecture outlined in Fig. 6e illustrates the ultimate goal of sensor fusion: mixed early and late fusion processes on different levels of abstraction, each subgoal benefiting from its own support set of sensors and with the results presented on a single display to the human observer in a representation that is highly informative, indicating at the same time alternatives, as well as the associated confidence levels.

We have already noted that application of multiple sensors in general can be beneficial to the accuracy of an observation, even though a low resolution sensor directs a sensor with a higher resolution ("cueing") without actually sensor *fusion* occurring. For making hard detection statements explicit knowledge of the clutter is required, and in order to improve single-sensor detection capabilities, one needs to carefully analyze the model assumptions that have gone into the system design. New, more detailed clutter models corresponding to the state of the art in sensor technology may be needed for the detection of harder targets. On top of that layer of processing sensor fusion may be applied. But it should be noted that sensor fusion can never make up for poor clutter modelling.

For any sensor fusion process to be successful, one has to properly prepare the raw signals originating from the single sensors. From a system point of view one has to determine the stage at which fusion has to take place (ranging from 'early' to 'late') in relation to the goal that has to be achieved by the fusion process. Once this has been decided the first step in preparing the sensor signals for fusion is *alignment*, to guarantee that the fields of view (FOV) have maximum overlap. Early fusion can only be successful if there is an *overlap* between the FOVs. The second step is the proper correlation of the same objects in the FOV of the sensors. This task can in practice be quite laborious when many targets are observed simultaneously (large FOV; scanning sensors). In addition attention has to be paid to the optimization of performance of the isolated sensor by *removal of the biases* in the observed quantities, before any of these preprocessing transforms and associations can be carried out.

Fusion

As we have seen in the previous section, it is worth to select the proper sensors and it is also necessary to spend sufficient effort in the preparation of the signals before they can

be fused. It should be noted that fusion is *not* a magic way to reduce the quality or the price of a sensor and still get the same observation performance. Making accurate observations always requires a great deal of study, modelling and experimentation; the successful application of sensor fusion initially means more work than applying isolated sensors.

Depending on what the goal of the fusion process is, there are different time scales to consider:

- The maximum sampling rate, limiting the highest instantaneous bandwidth of a signal, which is essentially a measure for the sensor resolution: in a pulsed radar it is the range resolution and in a camera system it is the transverse 2D spatial resolution.
- The frame rate. This rate is important in extracting information from a time sequence of camera images, e.g. with optic flow analysis. By using the temporal correlation of objects in the pictures when the platform is moving, it is possible to make estimates of the distance of each of these objects.
- The batch processing rate: If no batch signal processing is performed as in the case of Kalman filtering, one is not forced to make a priori assumptions and may therefore be more accurate than Kalman filtering in the early stages of the signal processing chain. A disadvantage is that batch processing is considerable costlier than Kalman filtering in terms of processor (CPU) time.

Although we will here focus only on sensor fusion, many more aspects in modelling need to be considered, before an attempt to set up sensor fusion should be made. We mention here only a few:

- Clutter modelling: the type of statistics, correlation times and correlation lengths.
- Construction of an 'a priori' environmental data base, necessary to make (partial) decisions on identity and position.
- The modelling of the target dynamics if the target is moving and its significance with respect to improving classification.

If the observables of the fused sensor suite are mutually 'orthogonal', complementary fusion will invariably yield more information then each of the separate sensors can provide. It is therefore conceptually the simplest way to demonstrate the benefits of fusion in practice. In this context one could make an analogy between a single-sensor observation of the real world as a (stochastic) *projection* of the real world onto a sensor observation space. In this analogy, fusion can be seen as (partially) reconstructing the real world, representing it as the direct product space of all observation spaces of sensors that participate in the fusion suite. Effectively the dimensionality of the observation space increases by adding up the dimensions of complementary sensor spaces. Adding sensors of the same type through the Ergoden hypothesis basically improves the statistics of the observation in the particular sensor space. However the *dimensionality* of the single-sensor space does *not* increase by adding more sensors of the same kind.

In case that the observables of the sensor suite are *not* 'orthogonal', the fusion process can increase the speed with which the accuracy or resolution of the observation is achieved by

acting as a smart scheduler, via cueing. The shorter response time is realized by first determining areas of interest via the sensor with the lower resolution or accuracy and the largest FOV, and then focusing attention on these areas using the high accuracy sensor, instead of scanning the entire area with a high accuracy sensor with a small FOV. In addition this type of sensor fusion increases the robustness: if the cued sensor fails or is jammed, the other sensor can take over, although with lower resolution or accuracy *(graceful degradation)*.

An example of multi-sensor fusion with different sensors is the combination of a radar measurement and an optical image: if an airplane is observed by radar, the range from observer to the airplane is accurately measured, while azimuth and elevation can only coarsely determined. In contrast, an optical measurement provides azimuth and elevation with relatively high accuracy, whereas the uncertainty in range is high. Combination of the two sensor types can considerably diminish the absolute uncertainty in the position of the airplane in 3D space, which is a natural consequence from the two complementary measurement principles.

Finally we remark that from a system theoretical point of view we can express the expected effect of the fusion process symbolically as: $S_1 \oplus S_2 \geq S_1 + S_2$,where $\oplus$ represents the fusion operator and S_i is a quality measure associated with sensor i.

Fusion with fuzzy aggregation operators as a way to reduce complexity
In our study we have concentrated on the synergy of sensors at the signal level ("*early fusion*"). Although this does by no means preclude the use of a priori information or taking into account any human-generated inputs and feedback, our study focuses specifically on sensing, because the signals are not yet distorted or corrupted by signal analysis operations, and because there is relatively little room for subjectivity. The attractiveness of this approach is of course that by operating close to the primary sources of information, one expects to be able to significantly enhance the detection and recognition processes by applying sensor fusion.

A sensor fusion system that receives raw signal inputs from all sensors retains control over all primary sensors and has, at least theoretically, a number of advantages over secondary (level 2 and higher) fusion. Apart from the larger information content of raw information, it should however be noted that each fusion step requires a certain processing time and in early fusion it is effectively the slowest sensor in the fusion suite that determines this latency, even if we neglect the execution time for the fusion process itself. In addition the latency is increased because fusing information from autonomous, asynchronous and dissimilar sensors requires synchronization. A designer of a sensor fusion suite should be aware of this and take precautions to ensure that the pile-up of latencies does not degrade the real-time performance of the overall fusion system, or jeopardize the quality of the fusion process, e.g. by constructing a deficient situation awareness picture.

It is relatively straightforward to illustrate these ideas by the improvement of operation of a target tracker during the observation of a manoeuvring target in cluttered areas. A variety of different sensors can be used to generate plot reports and these can be

combined on the basis of a simple confidence criterion that is based on the presence of clutter for a particular sensor. In [21] this has been illustrated. However, though a useful idea, this example basically supports the idea of increasing robustness by increasing the number of different sensors. Our goal is more ambitious: we would like to improve the quality of the single-sensor conclusions in such a way that

$$P(S_1 \cup S_2) \geq P(S_1) + P(S_2),$$

or if this condition appears to be too strong, at least

$$P(S_1 \cup S_2) \geq \max(P(S_1), P(S_2)),$$

where $P(S_i)$ indicates the performance or 'added value' (the effective added information accumulated over time) of stream S_i, measured by sensor i. From this formulation it is clear that in order to model sensor fusion, we will need *nonlinear* operators.

The earliest attempts to combine measurements from multiple sources are by Bayes [22]. He introduced the notion of conditional probability $Prob(A|B)$, defined by:
$Prob(A|B) = \frac{Prob(A \cap B)}{Prob(B)}$ i.e. the probability of A, given that event B has occurred. This definition is easily extended to n observations obtained by n sensors. There are a number of difficulties connected with the application of the Bayesian sensor fusion formula:

- difficulty of assigning a priori probabilities;
- complexity when there are multiple hypotheses and/or multiple conditional events;
- requirement that hypotheses have to be mutually exclusive and exhaustive;
- absence of uncertainty modelling.

In trying to find an appropriate way to model fusion and take advantage of the nonlinearity of the process, Dempster and Shafer (DS) created a generalization of Bayesian theory that allows the incorporation of uncertainty by using (overlapping) probability *intervals* and uncertainty modelling to determine the likelihood of hypotheses based on multiple evidence [23]. The essential generalization of DS theory is that not all hypotheses need to be mutually exclusive as in the Bayesian theory. In DS fusion evidence is assigned both to single and more general propositions, instead of assigning directly a probability to hypotheses as in Bayesian theory.

Noting that belief and plausibility measures are both examples of Sugeno's [24] λ-fuzzy measure g_λ, the question arises whether it is possible to combine the intuitive ideas on sensor fusion and the properties of g_λ. We will show that in contrast the basic probability assignment in DS theory, fuzzy g_λ measures can indeed be utilized for the problem under consideration. We will take a closer look at this in the following and propose to view the multisensor fusion process in terms of a synergy between (sets of) sensors that are grouped in such a way as to support a certain decision or hypothesis. Instead of attempting to make a decision (detection or classification) in one step, either by a single sensor, or by a linear combination of a group of sensors, it is proposed to combine supporting evidence for a hypothesis in a hierarchical way by building a tree structure

that combines at the lowest level clusters and in the next levels aggregates the outputs of several initial clusters in superclusters and so on. At each level in the tree decisions need to be made from different sources with different weights. This is conveniently modelled by the fuzzy λ-measure g_λ ($0 \le g_\lambda \le 1$). In particular we have in the absence of relevant information towards the classification/detection goal: $g(\varnothing) = 0$ and $g(A) \le g(B)$ if $A \subseteq B$. This coincides with the intuitive feeling that if the evidence support is larger (i.e. if we observe the same scene with more sensors), that then the information content should also increase. In addition the following property holds for all $A, B \subset X$ with $A \cap B = \varnothing$:

$$\exists \lambda > -1 \quad g(A \cup B) = g(A) + g(B) + \lambda\, g(A) g(B) \ .$$

This again supports the intuition that adding more independent data ($A \cap B = \varnothing$) co-operates towards an increase in confidence about the final decision. In addition both intuitive features about the fusion of two independent sensors are reproduced, viz.

$$\lambda \ge 0 \qquad g(A \cup B) \ge g(A) + g(B)\,,$$

i.e. fusion is more than superposition and

$$-1 < \lambda \le 0 \quad g(A \cup B) \ge \max(g(A), g(B))\,,$$

implying that even if the level of confidence is larger than -1 (as reflected by the negative λ) then it may still be fruituitous to apply sensor fusion. In the event that λ = 0, i.e. the case where all sensors have the same importance and completely cover the universe of discourse, the degree of importance g_λ towards the final decision becomes additive and coincides with the definition of a probability measure.

Following ideas put forward in [25], sensor fusion may also be modelled using the concept of fuzzy integration. For a review on the role of fuzzy integrals in the framework of multiple criteria decision-making see [26]. A fuzzy integral may be interpreted as an aggregation functional of subjective evidence, where the subjectivity is expressed in the fuzzy measure, and integration is defined over measurable sets [27]. In contrast to normal (Lebesque) integrals, fuzzy integrals are *non-linear* functionals. It is exactly this nonlinearity and the possibility to include a fuzzy measure g_λ that is attractive in the context of fusion. Formally Sugeno's fuzzy integral is defined in the following way: Let X be a set of elements (e.g. sensors, features or classifiers) and let h(x): X→[0,1] denote the confidence value belonging to element x∈X (e.g. the class membership of data determined by a specific sensor (classifier)), then the *fuzzy integral* of h(x) over a subset E of X with respect to the fuzzy measure g can be calculated. The evaluation of the fuzzy integral may be interpreted as evaluating the degree of agreement between objective evidence h(x) and the expected observation outcome (the hypothesis). We will not discuss the properties of this fuzzy fusion operator here, but note that it is ideally suited to combine information from different sources *without* having to deal with the combinatorial explosion.

A similarity between fuzzy fusion (FF) aggregation and the way DS theory fuses data from different sources is that both make use of fuzzy measures: DS uses the belief measure exclusively, whereas the FF operator uses the g_λ measure. For $\lambda \geq 0$ this measure is equivalent to the DS belief measure. The conceptual difference between both methods is twofold:

1. The frame of discernment (the universe of discourse) is different in both methods.
2. There exists a clear separation of objective and subjective uncertainty in the case of FF.

We will illustrate these points in the following: For the FF scheme the frame of discernment contains the information sources (the sensors) related to the hypothesis under consideration, whereas in DS theory the universe of discourse contains *all* possible hypotheses. In combining the different information streams, the fuzzy fusion aggregator fuses all sources according to their relative a priori importance as well as to the degree to which each sensor supports the hypothesis under consideration. In contrast, the fusion process in DS theory associates with each knowledge source a belief function that is defined over the power set of the set of hypotheses and combines these in the fusion process. The evaluation of Dempster's rule of combination therefore has exponential complexity $O(2^N)$, where N equals the number of hypotheses under consideration. In contrast, in FF one fuzzy integration has to be calculated, which implies that g_λ has to be calculated nN times, where n is the number of sensors. The evaluation of the fuzzy integral can then be carried out in $O(n)$ steps. The second advantage of fuzzy aggregation is that both the weighting with the degree of support by which a sensor supports a certain hypothesis, as well as the weight of importance of a certain sensor, reflecting a subjectivity or an a priori confidence in the particular sensor, are explicitly modelled.

We therefore conclude that the formalism of fuzzy measure theory offers an opportunity to model the process of sensor fusion in a natural, intuitive and adequate way, allowing arbitrary sensors to be fused and allowing different ways to weigh various combinations of observations. As an example of the application of the belief measure g_λ consider three sensors, labelled 1-3 with belief measures 0.1, 0.3, and 0.2, respectively. If sensor 2 ("the most decisive sensor in support of the hypothesis under consideration of the three sensors") is combined with one of the other two sensors (1 or 3), the combined evidence as reflected by g_λ must be larger than the sum g_2+g_3. This is indeed the case, as follows from the definition of g_λ with $\lambda = 3.109$: $g_{23} = 0.687 > g_2 + g_3 = 0.3 + 0.2$ and also $g_{23} > g_{21} > g_{13}$, since $0.687 > 0.493 > 0.362$, which we would also expect intuitively. In addition we have that $g_{123} = 1$ and $g_\emptyset = 0$, thus having consistency within the powerset of the three sensors.

Conclusions

In this article we reviewed the added value of sensor fusion in military observation systems. Sensor fusion is motivated by the expected qualitative and quantitative improvement of observations and thus of situation awareness. We have focused on early sensor fusion and found that the performance enhancement due to extending one sensor to a suite of identical sensors and assume a majority vote is limited to a few tens of a percent. Early sensor fusion offers the best perspective to maximally benefit from

multiple sensor observations, but at the same time demands extensive data acquisition efforts. The real-time constraints and the need to (re)use intermediate fusion results in different decision processes suggest that in early fusion soft decisions are more effective than hard decisions. Finally it can be concluded that from the theoretical point of view key concepts of fuzzy logic, such as fuzzy belief and plausibility measures and Sugeno's fuzzy integral, provide us with suitable mathematical tools to combine soft decisions and describe the type of synergy behaviour expected of a fusion process.

References

[1] J. Llinas and E. Waltz, Multisensor Data Fusion, ArtechHouse, Norwood, MA. 1990.

[2] D. Hall and R. Linn, "A Taxonomy of Multi-Sensor Data Fusion Techniques", Proc. 1990 Joint data fusion symposium, vol 1, pp 593-610, 1990.

[3] C.B. Weaver, Ed., "Sensor Fusion", Proc. of SPIE, vol 931, Orlando, FL, 1988.

[4] P.S. Schenker, Ed., "Sensor Fusion, Spatial reasoning and scene interpretation", Proc. of SPIE, vol 1003, 1988.

[5] C.B. Weaver, "Sensor Fusion II", Proc. of SPIE, vol 1100, Orlando, FL, 1989.

[6] P.S. Schenker, "Sensor Fusion II, Human and Machine Strategies", Proc. of SPIE, vol 1198, Philadelphia, PA, 1989.

[7] A.J. van der Wal, "Application of fuzzy logic control in industry", Fuzzy Sets Syst 74, pp 33-41, 1995.

[8] A.J. van der Wal, "Fuzzy Logic: Foundations and Industrial Applications", ed. D. Ruan, Kluwer, Chapter 14, 275-311 (1996).

[9] D. Ruan, Z. Liu, L. Van den Durpel, P. D'hondt, and A.J. van der Wal, "Progress of Fuzzy Logic control applications for the Belgian Nuclear reactor BR1", Proceedings EUFIT '96, Aachen vol 2, 1237-1241 (1996).

[10] A.J. van der Wal and D. Ruan, "Controlling the output power of a nuclear reactor with fuzzy logic", JCIS 97, vol 1, 136, 1997.

[11] G. J. Klir and B. Yuan, Fuzzy sets and fuzzy logic, Prentice Hall, New Jersey, 1995.

[12] A.P. Dempster, "Upper and lower probabilities induced by a multivalued mapping", Ann. Math. Statistics, 38, pp 325-339, 1967.

[13] G. Shafer, A mathematical theory of evidence, Princeton University Press, Princeton, 1976.

[14] K. Leszczynski, P. Penczek, and W. Grochulskki, "Sugeno's fuzzy measures and fuzzy clustering", Fuzzy Sets Syst., vol 15 pp 147-158, 1985.

[15] D.H. Hall, Mathematical techniques in multisensor data fusion, Artech House, Boston, 1992.

[16] E.L. Waltz, "Data fusion for C3I: A tutorial", Command, control, communications and Intelligence handbook, EW Communications Inc, Palo Alto, CA, pp 217-226, 1986.

[17] P.J. Nahin and J.L. Pokoski, "NCTR plus sensor fusion equals IFFN", Proc. IEEE Trans. Aerospace Electronic Systems, Vol AES-16, pp 320-327, 1980.

[18] B.V. Dasarathy, "Fusion strategies for enhancing decision reliability in multisensor environments", Opt. Eng. 35 (3), pp 603-616, 1996.

[19] L.A. Zadeh, "The linguistic approach and its application to decision analysis", in Y. C. Ho and S. K. Mitter, eds., Directions in large scale systems, Plenum Press, New York, pp 339-370, 1976.
[20] A.S. Mikhailov and A.Yu. Loskutov, Foundations of Synergetics II; Chaos and Noise, Springer, Berlin, 1996.
[21] E. Sviestins, "Multi-sensor tracking for air traffic control and air defense", ATC Systems, pp 10-16, 1995.
[22] T. Bayes, "Essay towards solving a problem in the doctrine of chances", Philos. Trans. Royal Soc. London, vol 53, pp 370-418, 1763.
[23] S.C.A. Thomopoulos, "Sensor integration and data fusion", J. Robotic, Syst. Vol 7(3), pp 337-372, 1990.
[24] M. Sugeno, "Fuzzy measures and fuzzy integrals: A survey", in Fuzzy Automata and Decision Processes, North Holland, Amsterdam, pp 89-102, 1977.
[25] J.M. Keller, H. Qiu, and H. Tahani, "The fuzzy integral and image segmentation", Proc. NAFIPS June 1986, pp 324-338, 1986.
[26] M. Grabisch, "On the equivalence classes of fuzzy connectives-The case of fuzzy integrals", IEEE Trans Fuzzy Syst.3(1), pp 96-109, 1995.
[27] W.F. Pfeffer, Integrals and measures, Marcel Dekker, New York, 1977.

Fuzzy Logic Assisted Helicopter Flight Control

Ariën J. van der Wal

Introduction

Flying a helicopter is a task that requires a great amount of experience and skill. This is due to the strong coupling that exists between the six degrees of freedom, resulting in the 12 dimensions needed to describe the dynamics of a helicopter. Therefore, to perform even a simple flight movement, such as "go up" or "go down", the helicopter pilot has to carefully adjust more than one control simultaneously.

The use of small-sized helicopters as UAV for professional applications is rapidly increasing. Examples of such applications include military reconnaissance and police surveillance, for movie filming, surveying, etc. However the strong nonlinear coupling among the degrees of freedom and the amount of experience and skills required for safe flight control of small-size remote-controlled helicopters, make it attractive to develop a flight assistant that aids inexperienced operators in flying a successful mission. This is even more true for small helicopters, because of their smaller inertia and the associated smaller time scales of the dynamics involved.

This motivated the present research and development of an intermediate intelligent agent that is capable to navigate a small helicopter safely using elementary commands that are given by a non-expert user. This means that anyone can set a flight path via a user interface, by giving elementary commands (e.g. "go up", "go down", "hover", "go forward"). The intelligent agent must take all the necessary actions that the experienced pilot would take to control the helicopter and ensure the implementation of the desired flight path within a safe flight envelope. The architecture of the agent-helicopter system is schematically depicted in Fig. 1.

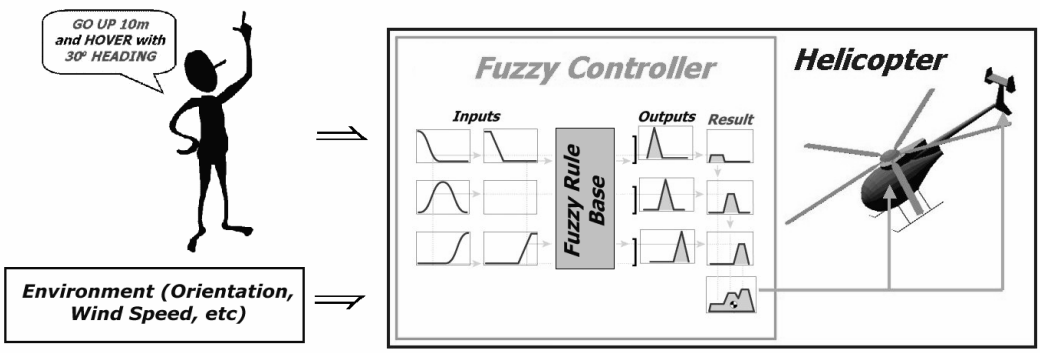


Figure 1. A schematic view of the user-agent-helicopter architecture

Guidance systems have been developed and implemented for model helicopters by different research groups, e.g. Linköping University [1], Swiss Federal Institute of Technology in Zurich [2] and Tokyo Institute of Technology [3]. Within the present paper the fuzzy logic approach will be investigated. The decision to use fuzzy logic for the implementation of the helicopter controller was based on the ability of fuzzy systems to model and absorb human experience and actions, even in the presence of uncertainty.

This research is conducted in order to verify the ability of a fuzzy controller to encapsulate a helicopter pilot's experience and actions. This means that within this work, the capability of fuzzy logic to control systems with strongly coupled degrees of freedom and dynamics will also be investigated. Although M. Sugeno of the Tokyo Institute of Technology has conducted the first work in this field [3] already in 1995 and has developed an autonomous helicopter using fuzzy logic control, it is very difficult to locate any specific publication with details on the implementation of the actual fuzzy logic controllers.

As a first step, we designed the hovering control. This is motivated by the fact that take-off and hovering at relatively high altitude are the first lessons that a real helicopter pilot takes. In order to be able to implement a helicopter movement, first a mathematical helicopter model has been implemented [4-7]. The fuzzy logic controller was designed to encapsulate the experience and knowledge of the pilot in order to take off and make the helicopter hover at a user-specified altitude and heading. Additionally, an interactive Graphical User Interface (GUI) has been developed so that a user can define the desired altitude and heading that the helicopter should reach. Also, within this GUI the user can directly manipulate the helicopter's controls and thus fly the helicopter model manually.

For the development of the mathematical model and the fuzzy logic controller and the implementation of the Graphical User interface, the modelling platforms of Matlab, Simulink and the Matlab fuzzy toolbox were used. This work sets the basis for further development of an ensemble of fuzzy controllers that will be able to perform all of the actions that a helicopter pilot can take. This should ultimately lead to an autonomous flight controller for unmanned helicopters. Therefore the reliability, robustness and safety of such system must be determined.

Fuzzy logic control

The use of soft-computing theory, such as fuzzy logic covers a broad scope, ranging from theoretical work in e.g. the foundations of quantum mechanics [8], to industrial applications in pattern recognition and sensor fusion (for a review see: [9-11]), mission-critical applications [12], and nuclear reactor control [13]. Modelling and simulating human knowledge and intelligence has been an active area of research over the past decades. There are various examples of procedures for which the relation between the inputs and the outputs of a system is only qualitatively known and therefore control cannot be achieved with conventional methods. Still, experienced operators manage to efficiently control such processes without having precise knowledge of the underlying physics or mechanics. In practice, the user consciously or subconsciously uses rules that he has learned and which he constantly updates. These rules are the result of experience acquired from learning in the real world.

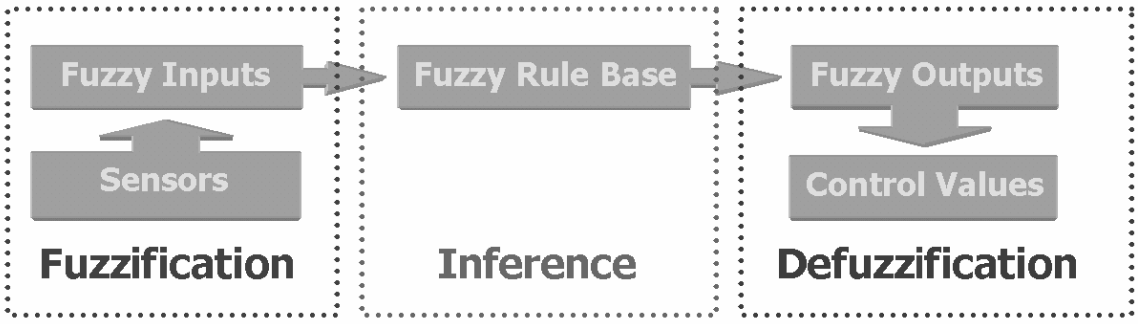


Figure 2. The three main parts of a fuzzy controller: real-world inputs have to be converted to fuzzy numbers, so that the fuzzy rule base (the expert system) can infer with the inputs and generate the fuzzy outputs. The last are defuzzified in the final stage to yield real-world values.

The aggregate of all the rules that describe how a process could be controlled, results in "intuitive skill-based models".

Fuzzy controllers make use of such experience models. The formulation of the control rules is not analytic, instead they are expressed in linguistic form. The basic problem in the design of a fuzzy interface is the representation of such an experience model or expert system in a concise and computationally treatable way. We distinguish different parts of our controller. The basic parts of a generic fuzzy model are displayed in Fig. 2.

In general, there must be a mechanism that is capable of translating numerical (measured) values from the various sensors into fuzzy concepts ("fuzzification"). In addition we must have a mechanism to take fuzzy decisions on the basis of the expert knowledge as stored in linguistic rules ("inference"). Finally we must translate the fuzzy output commands (decisions) into real-world control values ("defuzzification").

Each fuzzy variable, e.g. *InputX,* is characterized by a set of Membership Functions (MFs). Membership Functions may partially overlap with each other. They assign numerical values to linguistic values via weights $\mu_{MF} \in [0,1]$, e.g. $\mu_{MF}(InputX) = 0.85$. In fuzzy set theory MFs are a generalization of the characteristic functions in classical set theory. With the help of MFs it is possible to define operations on sets, such as the complement, union ($\cup$), and intersection ($\cap$).

The inference process translates fuzzy inputs into fuzzy outputs using a rule base that defines the structure of the controller. It makes decisions (i.e. activates output MFs) on the basis of the actual fuzzy input values, i.e. the activated input MFs. Fuzzy rules can be activated simultaneously and are of the following form:

IF *InputX* **is** *MF1* **AND** *InputY* **is** *MF2* **THEN** *OutputZ* **is** *MF3*

The activation σ_j of a rule R^j can be calculated from the fuzzy input values *InputX* and *InputY* (or any other inputs that may exist) in the following way as the (fuzzy) intersection of the relevant input variables:

$$R^j : \sigma_j(InputX, InputY) = \mu_{MF1 \cap MF2}(InputX, InputY) \qquad (1)$$

The activation σ_j of a rule R^j can finally be expressed by:

$$R^j : \sigma_j\left(InputX, InputY\right) = \min\left(\mu_{MF1}\left(InputX\right), \mu_{MF2}\left(InputY\right)\right) \tag{2}$$

It should be noted that in contrast to *boolean* logic set operations, *fuzzy* logic set operations, e.g. intersection or union, are not uniquely defined. In Eq. (2) we have chosen to implement the intersection of fuzzy sets as the *minimum* of their MFs. Next the weight for the output membership function(s) for each of the activated rules is calculated (in this example the weight of MF3 of variable *OutputZ*). The final weight ξ_{MF3} applied to the fuzzy controller output MF is determined by aggregating the output weights σ_j from all rules:

$$\xi_{MF3} = \max_{All\ rules}(\sigma) \ . \tag{3}$$

The final step is to defuzzify the output function to produce a numerical value. There are many methods to implement the inference and defuzzification steps; the most common way is to determine the final value with a simple calculation of the centre of gravity (COG) of the surface below the final output membership function, Eq. (4).

$$\hat{y}_{COG} = \frac{\sum_{Output\ MFs} \xi^v \mu_v}{\sum_{Output\ MFs} \mu_v} \ . \tag{4}$$

Overview of helicopter dynamics and flight control

Helicopter flight is a complicated task due to the strong nonlinear coupling of the various degrees of freedom of the helicopter. The work of a helicopter pilot is therefore more difficult than that of a pilot of a fixed-wing airplane. From the military perspective we note that a combat aircraft pilot can only devote a part of his time to controlling the platform, because this is just one aspect of the mission. The extra difficulties associated with flying a combat helicopter are also reflected in its standard crew of two, vs. only one pilot in a jet fighter.

The six degrees of freedom of a helicopter are: up-down and yaw (z-axis), right-left and pitch (y-axis), forth-back and roll (x-axis). The coordinates x, y, and z are fixed to an inertial system in space. As a consequence the state of a helicopter can be represented as a point in the phase space spanned by 6 coordinates (three for position and three for attitude) and 6 velocity components (three each for translation and rotation).

In order to fly and control the helicopter, the pilot has to simultaneously operate three different helicopter controls, which manipulate the angle of attack of the main and tail rotor blades. The prime role of the main rotor is to provide the lift force that allows the helicopter to hover and fly. During flight the main rotor maintains a constant angular velocity and is controlled by two conventional helicopter controls, named the Collective and the Cyclic. The tail rotor produces lateral thrust in the same way as the main rotor of the helicopter does. It changes the amount of thrust that is produced by changing the angle of attack of the tail rotor blades. The tail rotor is connected with the main rotor

through a gearbox and therefore also has a constant angular velocity that depends on the gear ratio of the gearbox. The tail rotor is primarily needed to counteract the top axis moment that is exerted on the helicopter body by the movement of the main rotor blades. The secondary effect of the tail rotor is to enable the helicopter to rotate about the main rotor's shaft axis (yaw). The working conditions of the main and tail rotor define the dynamic behaviour of the helicopter. The helicopter controls can thus be divided into two groups, the controls responsible for the manipulation of the main rotor (collective and cyclic), and the ones that are responsible for the tail rotor (tail pedals).

The collective control is responsible for providing the lift of the helicopter. It consists of a hand-operated lever that can be raised or lowered and this position is linearly linked to the angle of attack of the main rotor's blades and the throttle of the engine to keep the angular velocity of the blades constant. The collective control changes the angle of attack of all the blades of the main rotor simultaneously. The higher the lever is lifted, the steeper the angle of attack of the helicopter blades and the more lift force is produced and the more power is delivered by the engine. The cyclic is also a hand-operated control, which is positioned in front of the pilot and can be moved in any horizontal direction (forth-back, left-right and combinations). The cyclic controls the lateral and longitudinal translation of the helicopter and it changes the angle of attack of each rotor blade individually. This allows the helicopter to move in any horizontal direction. The tail pedals allow the pilot to change the angle of attack of the tail rotors blades. In this way, they control the amount and the direction of the tail thrust and therefore the heading of the helicopter body and its yawing movements.

Hovering

First we implemented hovering at a certain altitude with a given heading. Even this elementary action is complicated, since in order to reach a certain height a well-defined thrust of the main rotor is required. The main rotor thrust is strongly coupled with the angular momentum produced about its shaft axis and therefore influences the heading. It is commonly observed at helicopter take-off that the helicopter slightly rotates about the main rotor shaft axis (yaw), before the pilot can stabilise and bring the heading back to the initial heading. Similar phenomena are observed when the pilot tries to change direction while having low forward speed. The difference in the starboard and portside contribution to the lift force due to cyclic control command not only results to a change in direction but also to loss of height, because additional thrust is needed to compensate for the inclination and subsequent reduction of the effective rotor blade surface. When using the tail rotor trying to compensate the yaw torque, the result is an excess of force in the direction, for which the tail rotor is meant to compensate, that will tend to make the helicopter drift sideways. Pilots tend to compensate for this effect by simultaneously applying a little cyclic pitch, but designers also help the situation by setting up the control rigging to compensate ("trimming"). The result is that most helicopters tend to lean to one side when hovering and often touch down consistently on the same wheel first. Hovering in a helicopter requires experience and skill. The pilot adjusts the cyclic to maintain the helicopter's position over a point on the ground. The pilot also adjusts the collective to maintain a fixed altitude (especially important when close to the ground). Finally, the pilot adjusts the foot pedals to maintain the direction that the helicopter is pointing. External disturbances (e.g. wind) further complicate the hovering manoeuvre.

Helicopter modelling and simulation

For the design of the fuzzy controller the use of a competent mathematical helicopter model was required. A mathematical model for a helicopter that was designed by the Aviation Department of M.I.T. [4] was selected for this work. This model was found competent enough for the present study as it describes the dominant behaviour and the coupling among the degrees of freedom of a helicopter, without taking into account secondary flight dynamic effects that only insignificantly contribute to the overall behaviour of the helicopter. The helicopter dynamics can be derived by solving the Newton Euler equations of motion, three for the translational and three for the rotational degrees of freedom:

$$\frac{du}{dt} = (vr - wq) - g\sin\theta + (X_{mr} + X_{fus})/m \tag{5}$$

$$\frac{dv}{dt} = (wp - ur) - g\sin\varphi\cos\theta + (Y_{mr} + Y_{fus} + Y_{tr} + Y_{vf})/m \tag{6}$$

$$\frac{dw}{dt} = (uq - vp) - g\cos\varphi\cos\theta + (Z_{mr} + Z_{ht})/m \tag{7}$$

$$\frac{dp}{dt} = qr(I_{zz} - I_{yy})/I_{xx} + (L_{mr} + L_{tr} + L_{vf})/I_{xx} \tag{8}$$

$$\frac{dq}{dt} = pr(I_{xx} - I_{zz})/I_{yy} + (M_{mr} + M_{ht})/I_{yy} \tag{9}$$

$$\frac{dr}{dt} = pq(I_{yy} - I_{xx})/I_{zz} + (N_{mr} + N_{vf} + N_{tr})/I_{zz} \tag{10}$$

where:

- m is the mass of the helicopter;
- u, v, and w are the translational velocities along x, y and z axis;
- p, q, and r are the angular velocities along x, y and z axis;
- X, Y, and Z are the forces applied along x, y and z axis;
- L, M, and N are the moments along the x, y and z axis respectively;
- φ, θ, and ψ are the angular displacements about the y, x, and z axis, respectively;
- g is the acceleration of gravity;
- I_{jj} are the moments of inertia along the j-th axis (the I-tensor is diagonal in x,y,z).

Fig. 3 shows the position where the forces and moments are applied on the helicopter, as well as the direction of the resulting velocities and rotations.

The inputs of the mathematical helicopter model are the control commands (Collective, Cyclic and Tail Pedals Value) and the outputs are the speeds and displacements (translational and angular) for each of the axis (x, y and z).

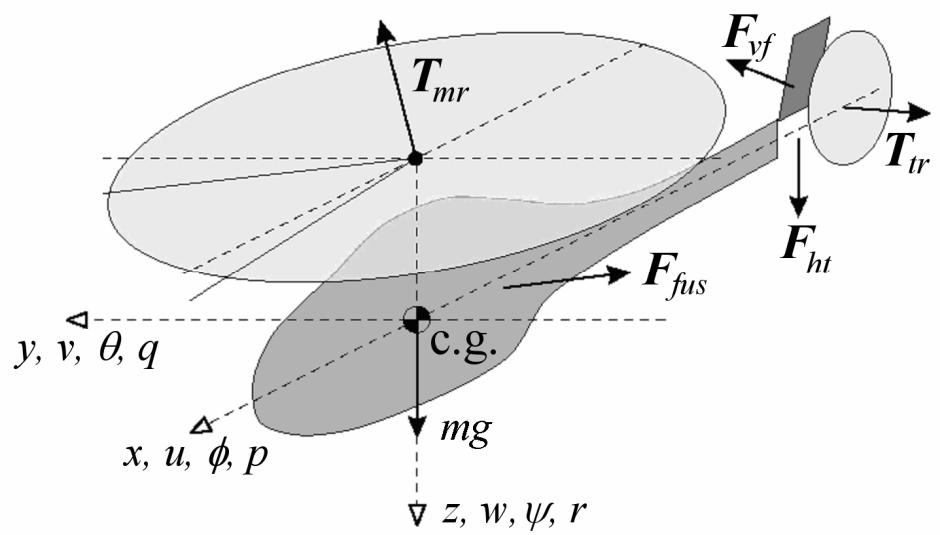


Figure 3. Coordinate system and forces (F) and moments (T) acting on a helicopter. The various subscripts that accompany the forces and moments are: $( )_{mr}$ for main rotor, $( )_{tr}$ for tail rotor, $( )_{fus}$ for fuselage, $( )_{vf}$ for vertical fin and $( )_{ht}$ for horizontal stabilizer.

Fuzzy Logic Controller

The role of the fuzzy logic controller (FLC) is to carry out the user's commands and translate them into actions of the helicopter. That, in a real life situation could be translated as a helicopter passenger that tells the pilot what actions the helicopter should perform. The "passenger" (user) does not need to know what actions the "pilot" (i.e. the FLC) has to take in order to correctly and safely carry out the required commands. The requirements for the design of the fuzzy controller are to control lift-off, vertical position and hovering with certain heading. Since this movement involves only the vertical position and orientation of the helicopter, cyclic commands will not be investigated in this paper and therefore will be assumed to be "zero". Therefore, the pilot's knowledge and experience to be modelled by means of fuzzy logic is limited to the use of the collective and the tail pedals. Two separate fuzzy controllers have been developed to perform the pilot's actions, one for each of the conventional helicopter controls.

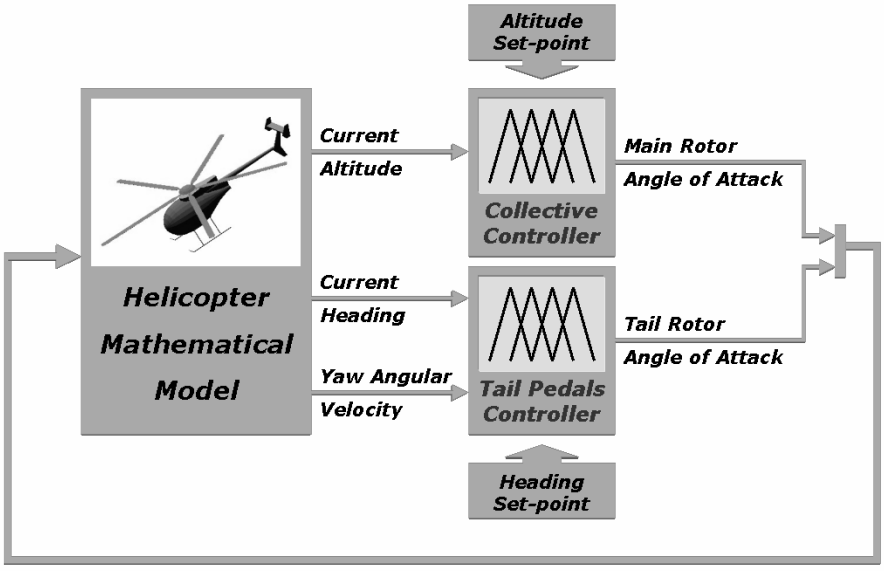


Figure 4. Architecture of the fuzzy controller and its interconnection with the mathematical model of the helicopter

The two controllers together compose the Controller system. As is displayed in the Fig. 4, the inputs of the controller are the actual altitude (vertical displacement along the z-axis), heading and yawing angular velocity of the helicopter model, as well as the set point values specified by the user. The output of the control system directly sets the angles of the main rotor collective and tail rotor pedal controls. The fuzzy logic controller was designed and tested using the fuzzy logic toolbox of Matlab.

Main rotor collective controller: Altitude

The main rotor control system consists of a fuzzy logic controller (Fig. 5) that controls the main rotor collective command according to the required vertical displacement. The output of the "Fuzzy Altitude Controller" is incremental, as schematically indicated by the delay feedback loop, labelled "memory". The limiter placed after the output ensures that the output value will not exceed the actual physical limits of the helicopter model. The memory loop provides the possibility to have different output values for the same input conditions, since different hovering altitudes require different angles of attack on the rotor blades. Therefore integration via the memory loop is required to distinguish between the several altitude hovering positions. The inputs of the controller are the "altitude error", which represents the difference between the current and the required altitude of the helicopter model, and the "altitude rate of error", the rate at which this error changes.

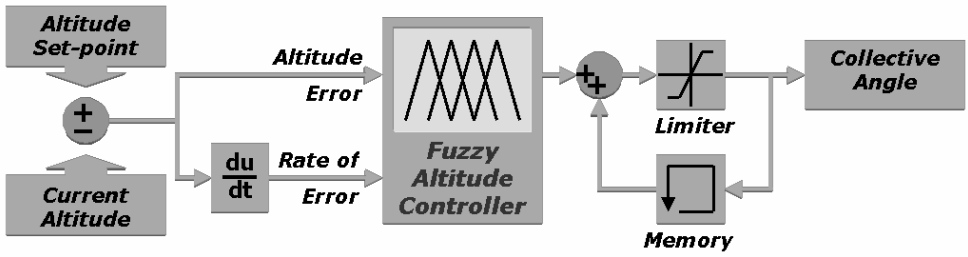


Figure 5. Altitude controller architecture

The controller has the structure of a fuzzy PD (proportional and differential) controller. The input "altitude error" consists of 5 membership functions, labelled {BNE, SNE, NoE, SPE, BPE}. These are displayed in Fig. 6. In determining the range of the fuzzy membership functions, scientific judgement on magnitude of the altitude error has been taken into account. The "altitude rate of error" input, which represents the rate of the error input, also consists of 5 membership functions, labelled {BN, SN, ZA, SP, BP} that have been determined experimentally by manual flying the helicopter model, that the maximum "altitude rate of error" values assumed, are within the range of [-10, 10] m/s. The collective angle output variable of the main rotor collective fuzzy controller consists of 7 membership functions, labelled {BNT, NNT, NT, ZT, PT, NPT, BPT}.

The rules for the altitude control are straightforward. The helicopter pilot increases the collective angle when he wants to gain altitude, and decreases it when he wants to lose altitude. The collective command is kept at a certain angle when the pilot wants to hover. Each altitude has a different hovering angle as air density and temperature greatly influence the lift produced by the main rotor blades at constant speed of rotation. Taking

these rules of thumb as a basis for our design, a set of fuzzy rules was determined (Table 1). From the structure of Table 1, the nonlinear relation between the output and the two inputs is apparent.

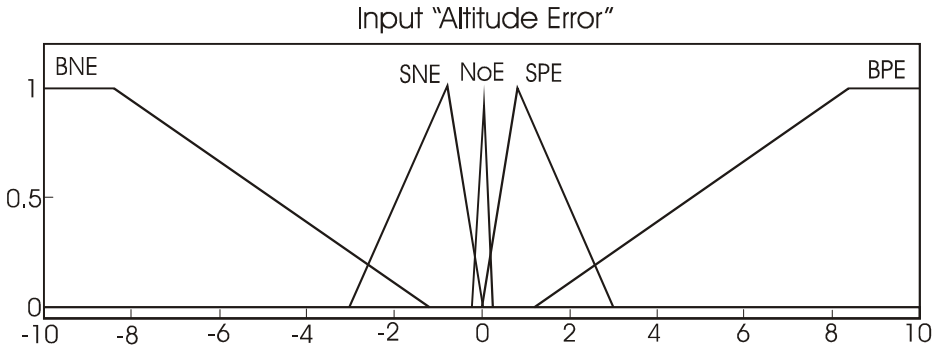


Figure 6. Membership functions for Altitude Error Input [m]: Big Negative Error (BNE), Small Negative Error (SNE), No Error (NoE), Small Positive Error (SPE), Big Positive Error (BPE)

Table 1: Collective fuzzy controller rule base. The table displays the activated output membership function according to possible input membership function combinations. A typical rule (see highlighted cell) is: **IF** (AltError **is** NoE) **AND** (AltRateError **is** ZA) **THEN** CollectiveOutput **is** ZT

	Altitude Rate of Error Input				
Altitude Error Input	**BN**	**SN**	**ZA**	**SP**	**BP**
BNE	BNT	NNT	NT	NT	ZT
SNE	NNT	NT	NT	ZT	ZT
NoE	NNT	NT	ZT	PT	NPT
SPE	ZT	ZT	PT	PT	NPT
BPE	ZT	PT	PT	NPT	BPT

Pedals controller for tail rotor: Heading

The tail rotor control system consists of two fuzzy logic controllers (Fig. 7), one to control the yawing angular velocity of the helicopter and the other to control its heading (angular position). The need for the two controllers arises from the fact that we have two different control objectives, corresponding to two different control regimes. The first control objective has to do with safety and staying within the operational flight envelope. The second control objective is maintaining the desired heading. We note that high angular velocities about the z-axis can produce instability of the helicopter system. Once the velocity of the helicopter is controlled and does not introduce any instability factors into the system, it is possible to implement the positioning control for obtaining the required heading. Helicopter pilots use a similar approach. They also make the helicopter rotate

with constant (low) yawing angular velocity until they stabilize the helicopter in a certain heading.

As can be seen in Fig. 7 the fuzzy heading controller has a similar structure as the fuzzy altitude controller. While trying to achieve the desired heading, special care must be taken as to prevent the helicopter from obtaining high yaw velocity since this may yield instability. The fuzzy yaw controller is responsible of keeping the yaw velocity of the helicopter within the required limits for safe and stable aviation. If the yaw velocity of the helicopter is normal (within the flight envelope), the output of the fuzzy yaw controller is very small or zero. In this case the heading controller takes over and is responsible for the value of the angle of attack of the tail rotor's blades.

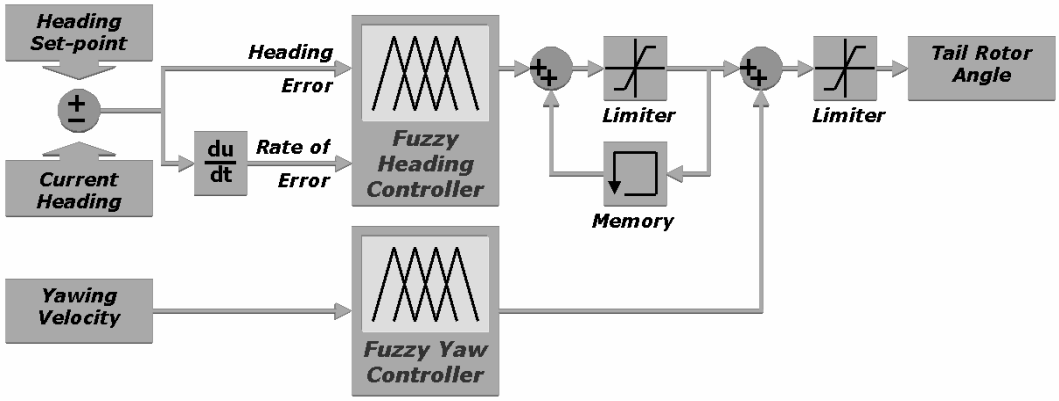


Figure 7. Heading Controller Diagram

The fuzzy yaw controller is responsible for the control of the yawing angular velocity. It consists of one input and one output. If the yawing velocity becomes large, this introduces the risk of instability of the system, the controller takes actions to oppose the current movement and reduce the velocity to within the required margins. The input of the controller is the yawing velocity, which represents the angular velocity of the helicopter model about its z-axis and is represented by 3 MFs, labelled {NegYaw, NormYaw, PosYaw}. It was determined from experiments with the helicopter model, that with a maximum yawing velocity of -1 to 1 rad/sec, it is possible to control the helicopter, whereas outside this flight envelope the control of the tail angular velocity becomes very difficult and this renders the system unstable. Therefore the allowed velocities are the ones that exist within the membership function of "NormYaw". The range of allowed values ("support") of NormYaw is [-1,+1] rad/s and defines the flight envelope. The output of the controller is the "Tail Rotor Angle", which represents the angle command that is passed to the tail rotor blades and consists of the 3 membership functions, labelled {NegOut, NoOut, PosOut}.

The control commands that can be given to the helicopter model's tail rotor angle of attack varies from -28.6° to 28.6°. These limits are prescribed by the limitations of the actual helicopter model. These limits also apply for the output values of the fuzzy heading controller and integration scheme of Fig. 7. While trying to obtain the required heading, it is crucial to simultaneously control the yawing speed of the helicopter to avoid instability. Due to the control approach chosen, the parallel fuzzy controller responsible for the

yawing speed needs to be able to numerically override the commands of the heading controller. Therefore the numerical output of the yaw rate controller must be at least twice the output of the heading controller after the integration scheme (Fig. 7). The fuzzy rule base for the control of the yawing velocity of the helicopter is quite simple. When the velocity is very negative, the controller applies positive angle to the tail rotor in order to counteract it. On the other hand, when the yawing velocity is very positive, the controller applies negative angle to the tail rotor. When the yawing velocity is between the desired limits, the controller does not apply any force. Taking these empirical rules as a base, the following set of fuzzy rules has been determined:

1. If (YawSpeed is NegYaw) then (TailAngle is PosOut)
2. If (YawSpeed is NormYaw) then (TailAngle is NoOut)
3. If (YawSpeed is PosYaw) then (TailAngle is NegOut)

The fuzzy heading controller is responsible for the control of the yawing angular displacement (i.e. the heading). It consists of two inputs and one output. The inputs of the controller are the "heading error", which represents the difference between the actual and the desired heading of the helicopter model, and the "heading rate of error", which is the rate of change of this error. The input "heading error" consists of the 7 MFs {NegOut, Neg, SmNeg, Zero, SmPos, Pos, PosOut}. For the range of the fuzzy membership functions, human judgement (of both the pilot and the flight engineer) on magnitude of the heading angles and heading was taken into account. The "heading rate of error" input, which represents the rate of the error input, also consists of 5 MFs, labelled {BigNeg, NegError, ZeroRate, PosError, BigPos}. It was determined from experiments with the helicopter model, that by manually flying the helicopter, the maximum heading rate of error values that were developed are within the range of [-300, 300] °/sec.

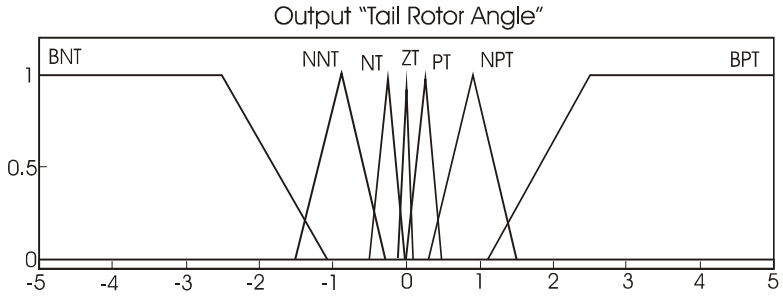


Figure 8. Tail Rotor Angle output from heading position control [°]. With membership functions: Big Negative Tail Angle (BNT), Normal Negative Tail Angle (NNT), Negative Tail Angle (NT), Zero Thrust (ZT), Positive Tail Angle (PT), Normal Positive Tail Angle (NPT), Big Positive Tail Angle (BPT).

The "tail rotor angle" output variable consists of the seven membership functions displayed in Fig. 8. The rules for the heading angle control are straightforward. When the helicopter heading error is very big on the positive side and it is growing even bigger, then the angle of attack of the tail rotor must get a value that will help it counteract and reduce the error. The opposite occurs when the helicopter's tail error is becoming smaller. Then the pilot takes actions to counteract the movement and make the yawing angular velocity equal to zero when the heading angle error is becoming small. Generally, the action of the helicopter pilot is to keep a constant speed while yawing and taking suitable counteracting measures to the movement only when the pilot needs to maintain

a heading. Taking these empirical rules as a base, a set of fuzzy rules has been determined (Table 2).

Experiments and results

To estimate the performance and quality of the fuzzy logic approach to the helicopter aviation problem, a number of tests have been conducted. The tests have to prove the ability of the fuzzy controller to perform helicopter take-off and landing as well as to hover at several altitudes with different headings. Therefore the altitude and heading time response characteristics as a function of time are of importance for each of the tests. In this section all three actions (take-off, hovering and landing) are investigated. For each of the tests, two plots are presented. The top plot contains the characteristic of the obtained altitude, whereas the bottom plot contains the characteristic of the heading of the helicopter model.

Take-off and hovering at various altitudes and headings

The first test was to determine the ability of the controller to make the helicopter to take off, and to change altitude and heading according the user commands. The controller should be able to manipulate the pitch angle of the rotor blades in such way that the helicopter body will not start revolving about the main rotor axis and that the helicopter model will be able to reach a certain altitude with a desired heading as quickly as possible without interfering with the safe flight envelope. For this test the initial conditions of the model helicopter are starting from the ground (0 m) with zero heading (0^o). The initial set-point for the fuzzy logic controller was to bring the helicopter to an altitude of 4 m with a heading of 10^o (Fig. 9 movement to point A).

Table 2: Tail rotor fuzzy heading controller rule base. The table displays the activated output membership function according to possible input membership function combinations.

	Heading Rate of Error				
Heading Error	**BigNeg**	**NegError**	**ZeroRate**	**PosError**	**BigPos**
BigNeg	BNT	NNT	NNT	NT	ZT
Neg	NNT	NT	NT	ZT	PT
SmNeg	NNT	NT	ZT	ZT	PT
Zero	NT	NT	ZT	PT	PT
SmPos	NT	ZT	ZT	PT	NPT
Pos	NT	ZT	PT	PT	NPT
BigPos	ZT	PT	NPT	NPT	BPT

As can be seen from Fig. 9, the helicopter model is gaining altitude and reaches the desired altitude without overshoot. Due to the extra moment that is produced from the increase in thrust on the main rotor, small fluctuations (overshoot at t = 0.1, 2.2, 2.7 s and undershoot at t = 0.7, 1.3, 3.7 s) appear at the required heading until the helicopter stabilises its altitude. Then the tail rotor control takes additional action and stabilizes the heading of the helicopter at the required value. Next the controller was instructed to bring the model helicopter to different altitudes with different headings. The following commands were given to the fuzzy controller for implementation:

- Raise altitude to 6 m and simultaneously change the heading to -3° (Fig. 9, point A to point B).
- Maintain the altitude of 6 m and change the heading to 16° (Fig. 9, point B to point C).
- Change the altitude to 13 m and maintain the heading of 16° (Fig. 9, point C and on).

In Fig. 9 the resulting trajectories are presented. The helicopter model ascends quickly and reaches the desired altitudes without any noticeable overshoots. The heading shows some fluctuations as before, in terms of overshooting and undershooting, due to the changes of the imposed moment from the main rotor.

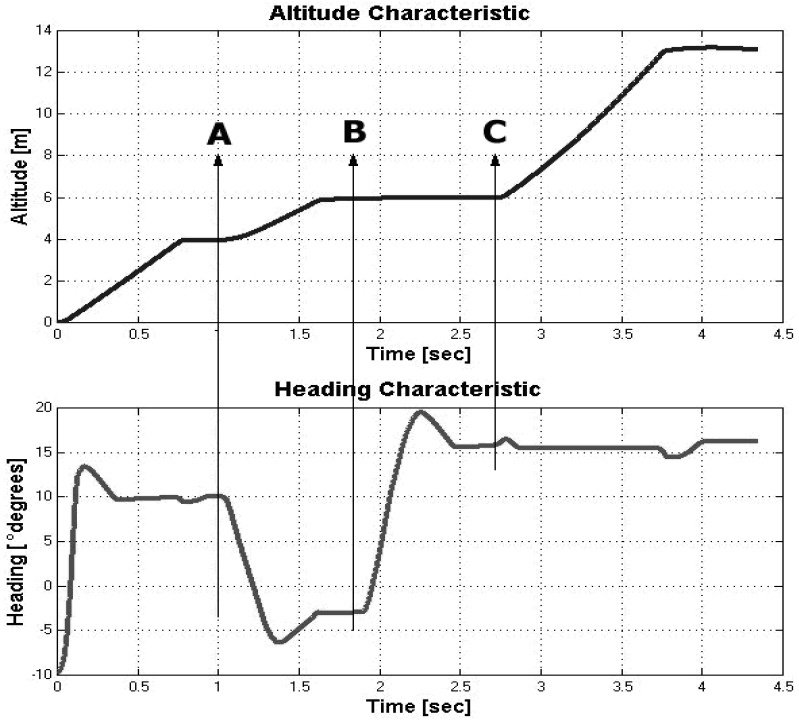


Figure 9. Results of hovering: Changing the altitude and heading of the helicopter. A, B, and C correspond to stable hovering with setpoints (Altitude [m],Heading [°]) of (4,10°), (6, -3°), and (6,16°), respectively.

Landing

In the second test the ability of the controller to safely land an initially hovering helicopter was investigated. For this test the initial conditions of the helicopter model were, starting at an altitude of 10 m with a heading of 16°. The setpoint for the fuzzy logic controller was to bring the helicopter down to an altitude of 0 m with a heading of 10°. The resultant trajectory is presented in Fig. 10. The helicopter model descends with gravitation until a point where the helicopter is increasing its throttle in order to drastically decrease its vertical speed and make a smooth landing. The smoothness of the landing is well observed when zooming in as shown in Fig. 11, starting after 5 seconds of flight time. Approximately at 0.1 m the helicopter changes its descending speed by more than an order of magnitude and the landing follows a smooth trajectory towards a soft touchdown. In the real world several phenomena could take place at that point (wind gusts, sudden change in wind direction, turbulence, etc.) and therefore it could be more advantageous for the controller to initiate a smooth landing earlier in the descend. We also note that in real life helicopter pilots generally prefer to maintain a small forward speed during landing approach in order to avoid the helicopter landing in its own "downwash", i.e. the air that is forced down by the main rotor during the creation of lift.

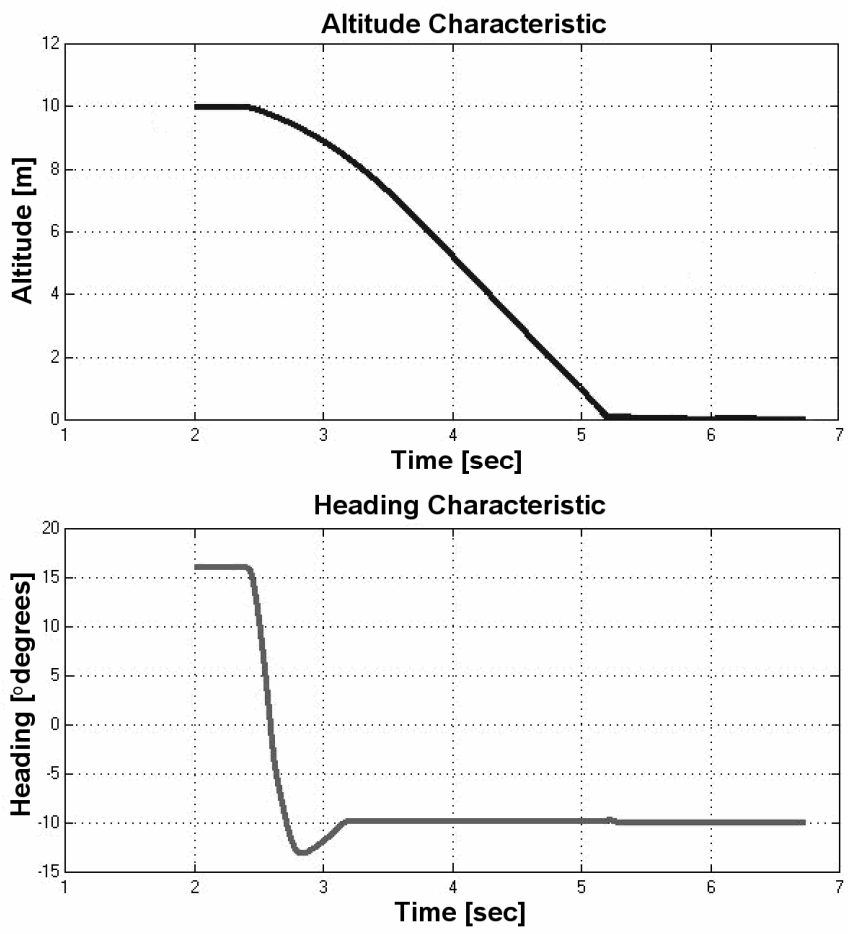


Figure 10. Flight trajectory of a helicopter landing: Altitude and heading as a function of time. Note that the typical timescale of the heading controller is of the order of 0.5 s, whereas the altitude controller changes much slower. This is explained by a combination of two effects: In the first place elementary physics limits the flight dynamics via mass and moment of inertia and in the second place one must follow the requirements set by the flight envelope. The last is implemented in the controller models by inserting limiters in the controller outputs.

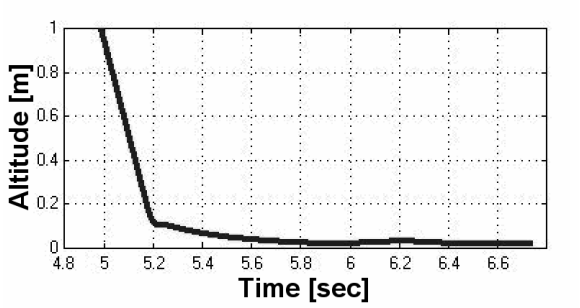


Figure 11. Detail of the altitude vs. time graph (Fig. 10) showing a soft landing by reducing the vertical speed at t=5.2 s with a factor of more than 10

Conclusion

We have demonstrated in the present work that fuzzy logic controllers are capable of controlling two of a model helicopter's coupled degrees of freedom. The incorporation of both scientific knowledge and the helicopter pilot experience into a fuzzy rule base has been experimentally demonstrated by successful take-off, hovering with defined heading and controlled soft landing. The system is open for extensions that could further enhance its performance (supervisory fuzzy controller, learning control, speed control). Developments to also control the other degrees of freedom of the helicopter are necessary in order to fly a fully autonomous aerial vehicle. Finally, it should be understood that this work describes laboratory-scale experiments and that in order to apply these controllers for real-life UAV missions ICAO certification must be obtained for the system.

References

[1] E. Skarman, "A Helicopter Control System", Linköping Electronic Articles in Computer and Information Science, Vol.4, No. 15, 1999.

[2] J. Chapuis, C. Eck, M. Kottman, M. Sanvido, O. Tanner, "Control of Helicopters", ETHZ Swiss Federal Institute of Technology - Measurement and Control Laboratory, Zurich, 1999.

[3] M. Sugeno, H. Winston, I. Hirano, S. Kotsu, "Intelligent control of an unmanned helicopter based on fuzzy logic", American Helicopter Society/51st Annual Forum Proceedings, Houston, 1995, pp. 791-803.

[4] V. Gravilets, B. Mettler, E. Feron, "Nonlinear Model for a Small-Size Acrobatic Helicopter", AIAA 2001-4333, August 2001.

[5] G. Padfield, "Helicopter Flight Dynamics", Blackwell Science Ltd, 1996.

[6] M.R. Spiegel, "Theory and Problems of Theoretical Mechanics", McGraw-Hill Book Company, 1967.

[7] G. Sposito, "An Introduction to Classical Dynamics", John Wiley & Sons, 1976

[8] A.J. van der Wal, "The role of fuzzy set theory in the conceptual foundations of Quantum Mechanics: an early application of fuzzy measures", (invited), Foundations and Applications of Possibility Theory, Vol. 8 of Advances in Fuzzy Systems, World Scientific, Singapore, 234-245 (1995).

[9] A.J. van der Wal, "Applications of fuzzy logic in industry", (invited), Proc. FLINS '94, Mol, 91-95 (1994).

[10] A.J. van der Wal, "The potential of fuzzy logic applications in industry", (invited), Fuzzy Logic Foundations and Industrial Applications, Kluwer, Boston, 275-312 (1996).
[11] K.M. Passino and S.Yurkovich, "Fuzzy Control" Addison Wesley, Menlo Park, CA (1998).
[12] A.J. van der Wal, "The importance of soft-computing networks for mission systems: a tutorial", AGARD, NATO RTO-MP-3, Monterey CA (1998).
[13] D. Ruan and A.J. van der Wal, "Controlling the power output of a nuclear reactor with fuzzy logic", Int'l Journal of Information Science 110, 1/4,151-177 (1998).

Finding Moving Objects in Video Recordings

Theo Hupkens

Introduction

Today, many military platforms are equipped with electro-optical video systems. Examples are: the Mirador and Sirius onboard of frigates, thermal imager and CCD day vision camera on the Fennek Reconnaissance Vehicles, Day TV and FLIR for the Apache attack helicopters, et cetera. The four new Dutch Ocean Patrol Vessels will be equipped with the multi-spectral, high resolution Gatekeeper system.

These platforms have in common that the video camera is constantly moving. Although human observers are very well able to observe threats or unusual situations on the monitors displaying the video recordings, they become bored and fatigued and less observant when nothing happens for some time. Therefore automatic image processing and pattern recognition systems must be developed, which take over the task of monitoring the video output of the surveillance systems. The development of such systems is not easy because of the difficult situations that are common in military operations: adversaries that try to be as invisible as possible by wearing combat clothing; night-time registrations using infrared cameras or image intensifiers suffering from severe noise; abrupt changes of the camera orientation; a constantly changing sea-background with barely visible swimmers or small boats, and so on. The motion estimation method that is the subject of this study can deal with most of these situations.

Video surveillance systems produce a continuous stream of images. A single image from a video sequence is often called a frame. However, if we want to emphasize the image properties we shall still use the word image. A first step in automatic pattern recognition is segmentation of an image into separate objects. To do so, it is necessary to find out which pixels of an image belong to a certain object. Then this object can be separated from the background and from any other objects. This process is called segmentation. There are several cues that can be used for this purpose, for instance colour differences or texture differences. Segmented regions can be used for further pattern recognition analysis. This paper describes segmentation based on motion. A group of (connected) pixels that move together is assumed to belong to one object. One advantage of using motion is that an object that consists of different parts is detected as one object, whereas if for instance colour differences are used the same object may be detected as several smaller segments which may have to be put together by sophisticated algorithms. Another advantage of the method is that after the motion is estimated, it becomes possible to correct for the changes due to the motion and then average the corrected frames in order to improve the signal to noise ratio.

In this paper, a brief description is given of the original motion estimation method described by [Odobez and Bouthemy, 1995]. This method is very well suited for the estimation of the camera motion. Experimental results obtained with this method from real infrared and colour videos are presented. The results are very accurate, even when noisy videos are used. The same method can be used to estimate the motion of separate

objects as well, by using only regions that move differently than according to the camera motion. This extended method is described in detail and the results when applied to synthetic and real sequences are discussed. After having found the motion parameters of an object, the exact location and the shape of that object as it appears in each video frame is known. Therefore, the extended method can be used for segmentation based on motion. We shall give several examples of this.

Motion parameters estimation for infrared and visual light video sequences

First, we need to define what we mean by "motion". In the simplest motion model, motion is described by two parameters: the velocity in the horizontal and that in the vertical direction. This two-parameter motion model describes a translation in the plane of the image. Several different motions are possible. For instance, an object might move away from the camera. This is almost the same as an object that is becoming smaller in time. A similar effect is obtained when the video camera zooms out. Three parameters are needed for an object that moves away from the camera or for a zooming camera. When an object rotates within the plane of the camera, for instance due the "rolling" of a ship or from an aeroplane that is filming an area while it is making a turn, the object's motion is a combined rotation and translation. Another possible motion would be that the object rotates away from the camera. When using cameras with a frame rate of 25 or more frames per second, this kind of rotation will hardly change the appearance of the object (it will be seen from approximately the some viewpoint), but will look like a slight change of the length to height ratio. All of these motions, and a few more, can be described by just six parameters.

Outline of the method

For the estimation of the motion of any object, an often-used approach is to try to find similar areas in succeeding frames. However, in practice finding similar areas when objects are rotating or changing their appearance is not easy. Therefore, we use a different approach, which is more suitable for the motion model we use. We use local changes of the intensity (that is the *gradient* of the intensity plot) together with intensity changes between successive frames. There exists a known relation between the gradient of the intensity in the neighbourhood of a certain pixel and the change in intensity of that pixel between two frames; this relation also depends on the size of the shift between the two frames. In the simplest model, the shift is proportional to the velocity. Because both the intensity changes and the gradients are known, we are able to calculate the corresponding shift. This is illustrated in Fig. 1. The lower pictures show the intensity of a horizontal line, at two points in times (i.e. for two separate frames of the video recording). Because the motion can just as well take place in the vertical direction, we need the gradient both in the horizontal and vertical direction.

The method starts with a least squares estimate of the parameters of the initial motion of the whole image, based on the relation between the gradients and intensity changes between two succeeding frames. We use high-resolution images only, which tend to be large in terms of numbers of pixels. For these large images, a somewhat different approach is used: in order to avoid excessive computation times, images at several reduced resolutions are used. Decreasing the resolution corresponds to resizing the

image to a smaller size. From every image several resized images are calculated; the first image is resized to 50% (in both directions); the second one to 25% and so on (see Fig. 2). The method starts with a guess of the initial motion of the whole image, at the smallest size that is used.

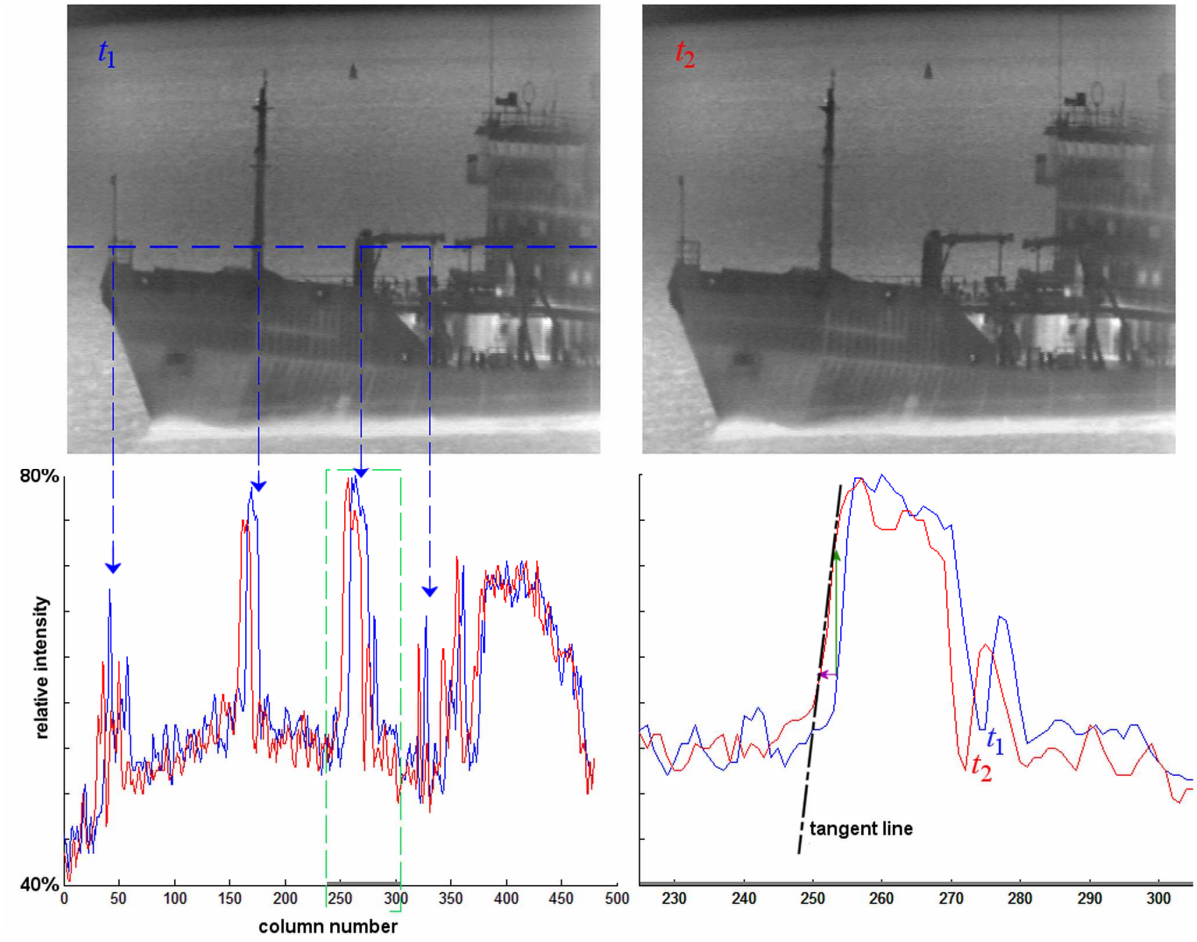


Figure 1. Upper images: two frames at times t_1 and t_2 taken from an infrared video recording. Under left: the intensity at a certain horizontal line for both frames. The vertical elements clearly show up in the intensity plots (blue arrows). Right: a magnified part of the intensity plot, showing the relation between the intensity difference between the two images at some pixel (green vertical arrow), the slope of the curve and the displacement (purple arrow). A larger velocity would result in proportionally larger intensity changes.

The initial estimate will be refined in succeeding steps, using the difference between the motion of every pixel and the estimated overall motion. Pixels that do not fit the current overall motion are called outliers. In the following iterative steps, the outliers will still be used for the determination of the global motion, but will have a smaller weight in the calculations. Due to these weights, after each iteration step the motion model will fit the overall motion better while the outliers will deviate more and more from the motion model. So the weights will increase for non-outliers and decrease for outliers. Fig. 3 illustrates this principle by means of an analogue. After several iterations, the changes in the motion parameters will become less than a preset threshold. At that moment, the weight function will be changed in such a way that pixels that move only *slightly* differently than most other pixels will be considered outliers as well. When after several iterations the motion parameters hardly change anymore, the image with reduced size

will be upsized by a factor of two and the process starts again, using the motion parameters from the last iteration (some parameters have to be adjusted to the new resolution). The process is repeated until a stable solution for the image at its original size is reached. For more details of this method, see [Odobez and Bouthemy, 1995 and Hupkens et al., 2000].

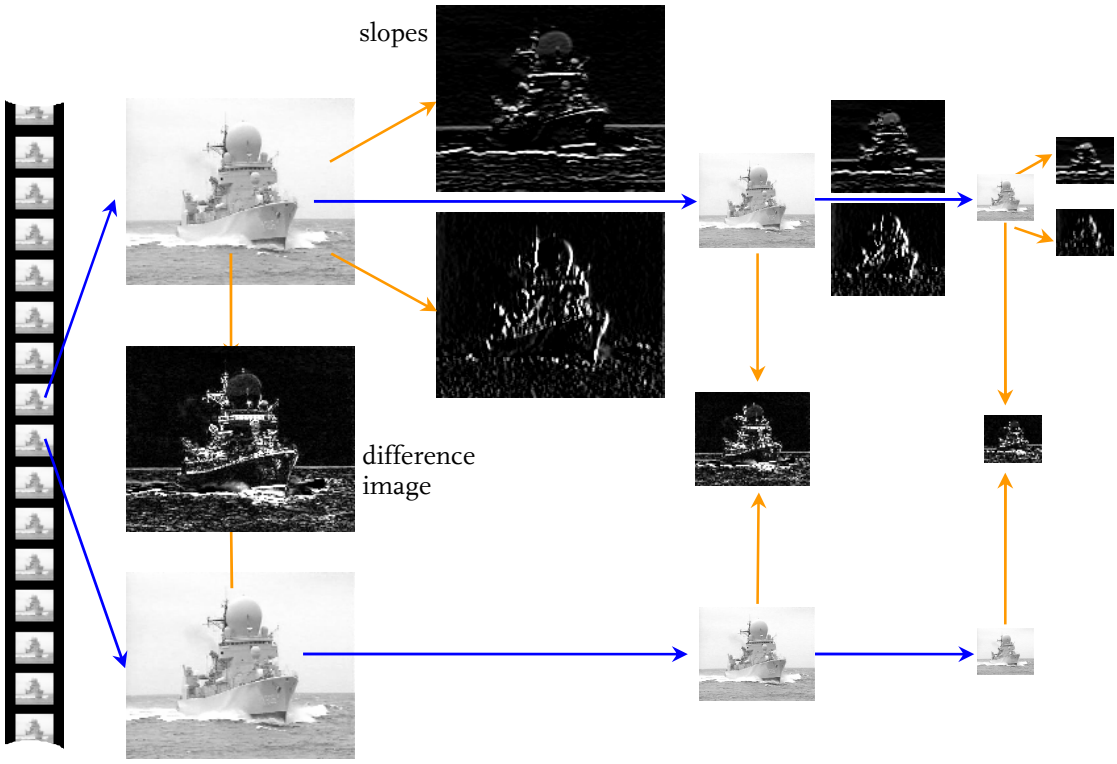


Figure 2. Generation of images to be used with the multi-resolution method. From left to right: from two succeeding frames three images are built: the difference image, an image that contains the horizontal slopes of the first frame and an image that contains the vertical slopes (black = no slope: white is steep slope). The original frames are downscaled by a factor of two and from these images again the difference and slope images are calculated. This process is repeated until the images are small enough for a fast convergence; usually three levels are perfect. The images are analyzed from right to left.

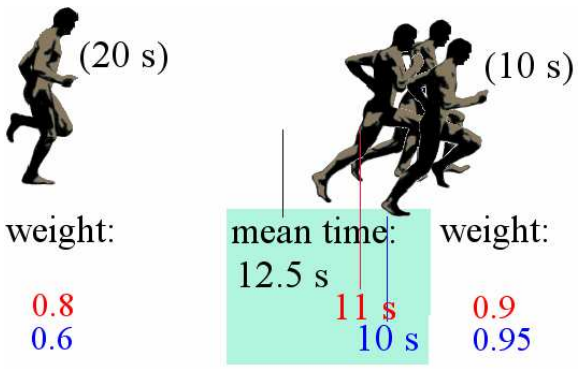


Figure 3. Illustration of the outlier principle: First, the four runners are seen as one group with an estimated 100 m time of 12.5 s. Then the weights are changed and new times are calculated (blue rectangle). After the third iteration, the slowest runner is exposed as an outlier or rather "outrunner". This example also illustrates the segmentation principle: after the third iteration, it has been established that the three fastest runners belong to one group. Then the algorithm will start searching for the next group by running the same algorithm on the outliers. The algorithm will find a second group, which – in this example – is the slowest runner.

New in our approach is that we keep the weights, so after the dominant motion has been found, we can simply go on using the same method on all outliers.

During the video recording, the overall illumination might change. Since it is essential for the method that the illumination does not change, we need to estimate the illumination changes and correct for it. We assume that the illumination changes proportionally to the intensity of each pixel, so we need one extra parameter. During the iterative loops, the intensity of the second frame is gradually changed until it matches that of the first frame. This kind of global intensity change may result from for instance the sun disappearing behind the clouds, although there will always be local differences such as shadows and glittering. Anyway, it appears that a single, global correction works well for all sequences used in this study. It is possible though to include different types of global illumination changes (see for instance [Kim et al., 2005]). It should be noted that the illumination changes usually are very small when two successive frames of a video sequence are used (typically less than 0.2 %). However, in practice it may be necessary to use frames with a much larger time distance, for instance if the expected motions are very slow. If the global illumination changes by even a few percents, inclusion of the illumination change factor is crucial, because otherwise the method might not converge or find wrong motion parameters.

Whenever colour sequences were used, the same method was applied, but with the intensity replaced by a triplet consisting of the primary colour components (red, green and blue). We use only one set of motion parameters (not three sets), which is calculated for the colours together. One might be tempted to think that using three sets of parameters (one set for each primary colour) would improve the segmentation process, because the colour would act as a cue as well. However, any real colour seldom is a pure primary colour (apart from the fact that the red, green and blue colours are additive in contrast to paint colours which are subtractive), so these three colours almost always contain mainly the same information.

Linear motion models, such as described above, have proved to be very useful and robust for motion estimation (see [Fuh and Maragos, 1991] or [Torr and Murray, 1993]), motion segmentation (see [Bouthemy and Rivero, 1987]) and tracking (see [Meyer and Bouthemy, 1992]).

Experimental results: background motion

First, the original method (without the extension for finding multiple objects) was used to see whether background or camera motion could be estimated reliably for real infrared or colour video recordings. All results described in this paper were obtained with identical thresholds and weight function parameters and all were using three resolution levels. The quality of the obtained motion parameters was judged by visual inspection of the *displaced frame differences*; this is a picture of the difference of two successive frames, of which one is corrected for its motion relative to the other frame. Hence, if the difference picture shows regions containing pixels that are non-zero, those regions do have a different motion.

An example involving rotation is shown in Fig. 4. The frames were extracted from a moderate quality AVI movie. Despite the poor quality of some of the intermediate frames, the correct motion is found and the average of 16 frames that are corrected for their relative motion is a sharp image (see Fig. 5). Therefore, the method correctly finds the (irregular) movements of the camera.

Figure 4. Left: first frame of a sequence. Right: sixth frame of the same sequence: blurry images like this one are typical for compressed movies.

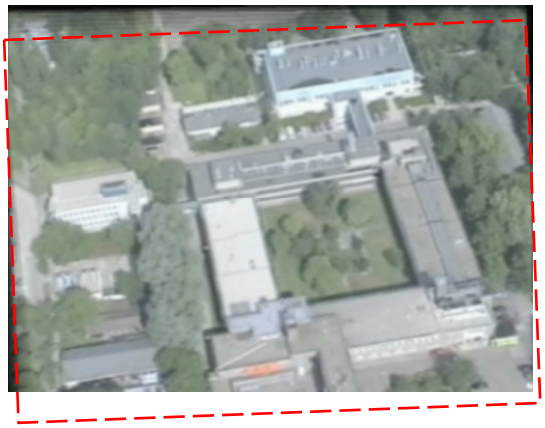


Figure 5. Average of 16 frames. The dashed rectangle shows the borders of the last frame of this sequence after correcting for the observed motion. The borders of all separate frames after correction for the estimated motion are just visible at the upper part of this figure (assuming the printing quality is good enough).

Fig. 6 shows an example of the results of a zoom sequence, taken with a camera on board of a ship. Again, the correct "motion" parameters were found, as can be seen in the averaged image after correction. Fig. 6 also shows a similar example, but now the object is approaching the camera, resulting in an apparent zoom. During the video recording, the vehicle changes its direction a little but also rotates in the plane of the image (compare Figs. 6d and 6e), but these motions are included in the model, so the correct motion is found. In fact, this sequence lasted until the vehicle almost reached the camera. Therefore, in the last frame only the grille of the vehicle was visible. Still the average of all frames was sharp at the position of the grille.

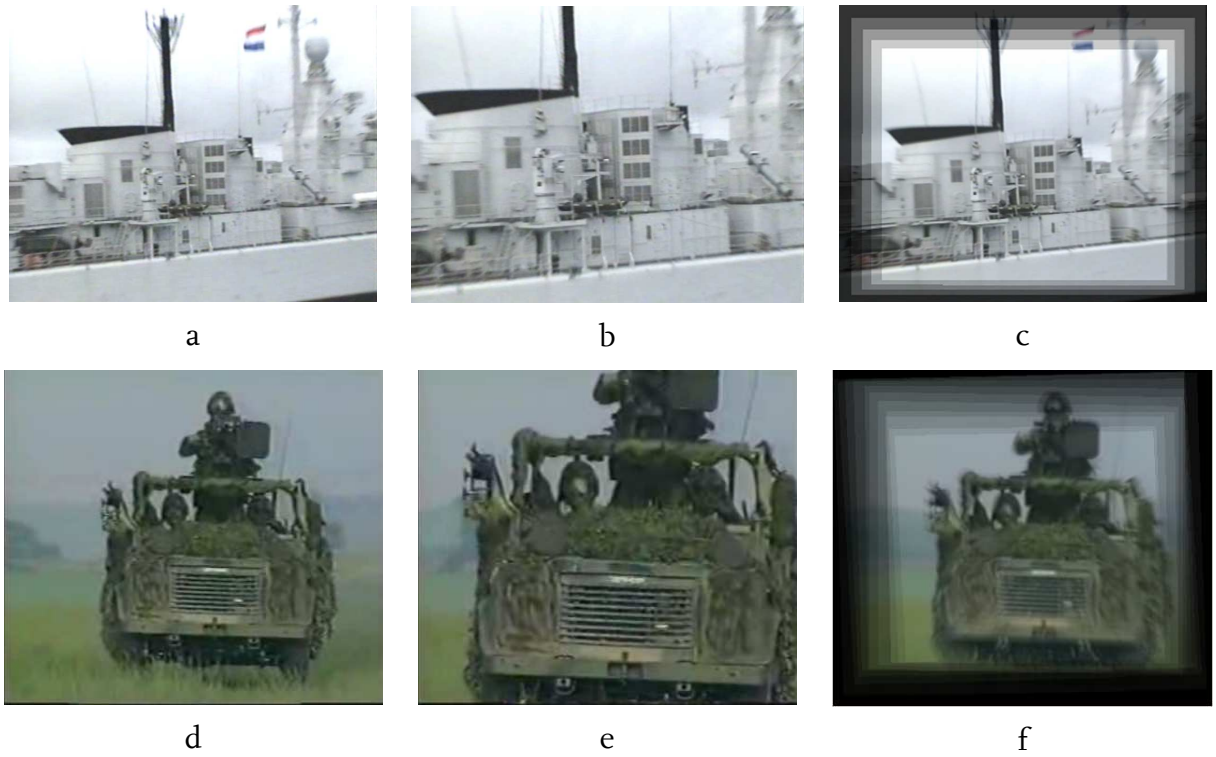


*Figure 6. (*a, b*): first and last frame of the zoom sequence. (*c*): average of five frames, after correcting for the estimated zoom. (*d, e*): first and 11th frame, taken from a long video (source: YouTube). (*f*): averaged frames after correction for the motion of the vehicle.*

The motion detection method was tested further with extremely noisy, realistic images. One result, obtained from an old infrared video recording of the HNLMS Tydeman is shown in Fig. 7. All displaced frame differences for the HNLMS Tydeman contained hardly any structure, so the image moves as a whole and its motion was correctly estimated. Obviously, this "motion" results from camera movements, so in fact it reflects the inverse camera motion. The motion parameters were also estimated by measuring the position of the visible lights, which were mounted on the ship, in successive frames. Both estimates agreed to within experimental error. Even though only six frames were averaged after correction for the motion, the resulting image clearly shows a reduction of the noise (see Fig. 7). From these experiments, it can be concluded that the motion detection method works very well and can be used to correct for camera movements for instance with the purpose of averaging frames.

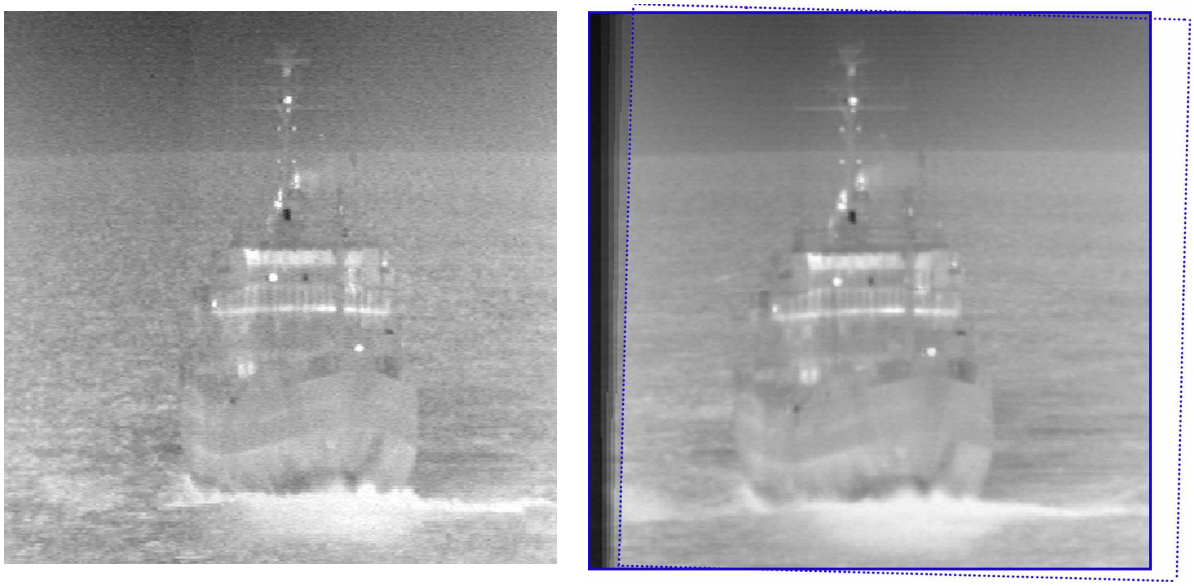


Figure 7. Left: noisy infrared image. The ship rotates by about 0.25 degrees and it 'moves' by 10 to 20 pixels between successive frame due to a moving camera. Right: average of six successive displaced frames. The dotted rectangle indicates the contour of the sixth displaced frame.

Segmentation on the basis of motion

Theoretical description

After the last iteration in the method described above, all the weights of the pixels are known, so it is easy to determine which pixels contribute to the final motion estimate. So in principle we are able to segment the images into areas that move differently. In practice, this is not so straightforward however. Due to noise and other irregularities (such as small sea waves), many isolated pixels belonging to the background will not fit well with the model, and pixels within the foreground will often fit by chance. There are several ways to improve this situation (see for instance [Odobez and Bouthemy, 1994]), one of which will be discussed in this paper.

Since we want the method to find objects autonomously, we need an exact criterion for which pixels 'belong' to the background. Since the weights after the final iteration tend to be either very small (≈ 0) or large (≈ 1), as will be shown later in Fig. 13b, a threshold of 0.5 seems to be a good choice. However, if we simply exclude all points that have a weight below this threshold, many isolated pixels and many pixels that in reality belong to the already found background will still be included. Isolated pixels cannot be used anyway, because the gradient is not defined at these points. There is also a principal problem: pixels can fit several motion models at the same time. This happens for instance in areas with a constant colour, or in areas that have a regular pattern. Pixels that lie in constant areas will be automatically assigned to the background object, whether or not this is correct. Methods to solve this kind of ambiguous situations are beyond the scope of this paper.

In the present study the decision whether or not to exclude a pixel from further calculations, is based on the weights of the 5×5 neighbourhood of that pixel. If 12 pixels out of these 25 pixels have a weight above 0.5, the central pixel is excluded from further calculations. Although with this approach good results were obtained for the images used in this study, it can certainly be improved and therefore should be the subject of further studies.

In order to have a flexible system, we implemented the method such that several motion models could be used:

- two parameters for translation only;
- three parameters for translation and rotation over a small angle;
- three parameters for translation and zoom;
- four parameters for translation, rotation and zoom;
- four parameters for translation and asymmetric zoom;
- six parameters for any motion that can be described by a linear motion model.

It is also possible to use a combination of these models: for instance, the two-parameter model can be used if it is likely that the camera movement will result in a pure translation. It can then be followed by the complete six-parameter model if the other objects are likely to have a more complicated motion.

Experimental results on segmentation

Synthetic sequences

The extended method was tested on a number of synthetic sequences, in order to get a feeling for the accuracy of the method. The synthetic images consisted of moving backgrounds with several differently moving objects. A typical example of such a synthetic image set is shown in Fig. 8. Here three objects rotate together about the same axis. The background of the second image is shifted by 1.75 pixels horizontally and 2.75 pixels vertically. Furthermore, the intensity of the background of the second image was deliberately lowered by 6.0%. After the first run of the algorithm, the motion parameters for the background were found to within 0.1%. The estimated global illumination factor was -6.3%. Note that the method can find a translation over a non-integral number of pixels. From the results of this and other synthetic and real sequences (not shown here), it can be concluded that the method can be used to estimate sub-pixel movements as well. The weights at the end of the first run are shown in Fig. 9a. From the weights, areas are calculated that do not fit the motion model (Fig. 9b); these areas can be used directly as an indication of the segments. After the second run of the algorithm, using only these areas, the affine parameters of the rotating rectangles are found. The correct values were found to within 1%. Fig. 10 shows the averages after correcting the second image for the estimated motion.

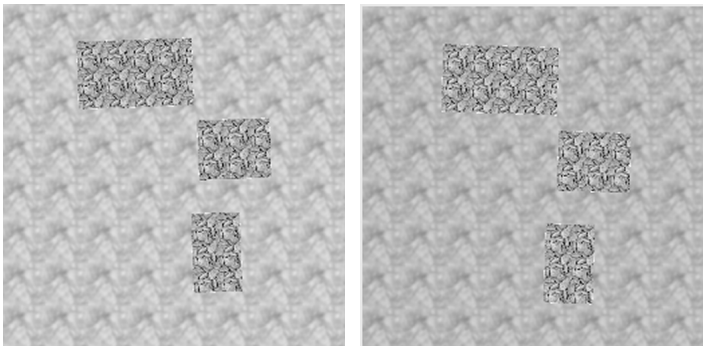
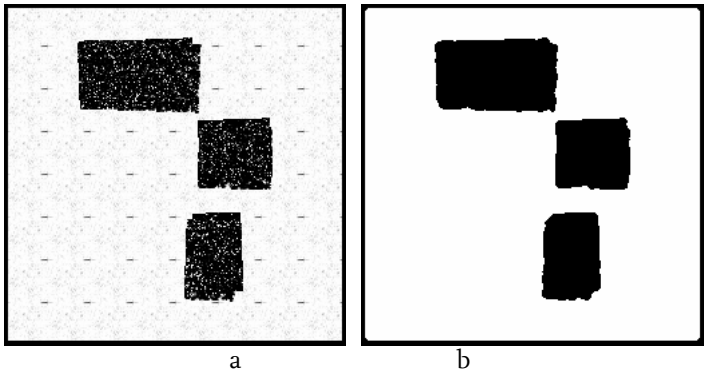
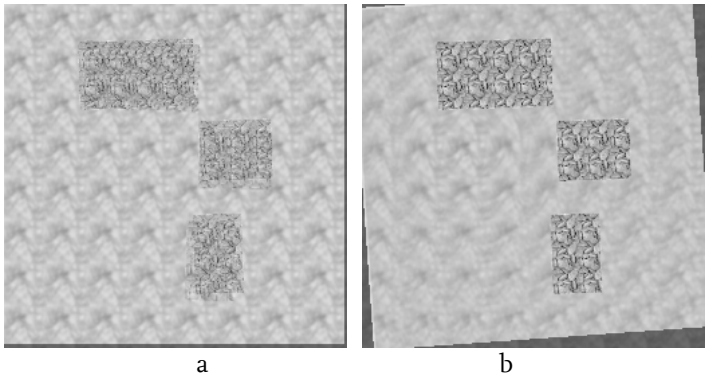


Figure 8. Synthetic frames, the rectangles are rotated clockwise over 4 degrees about a point close to the lower left corner



a b

Figure 9. (a) Weights after first run (black = 0; white = 1). (b) The 5 × 5 neighbourhood requirement ensures that loose pixels are not used in the second run; only the black rectangles are used in the second run.



a b

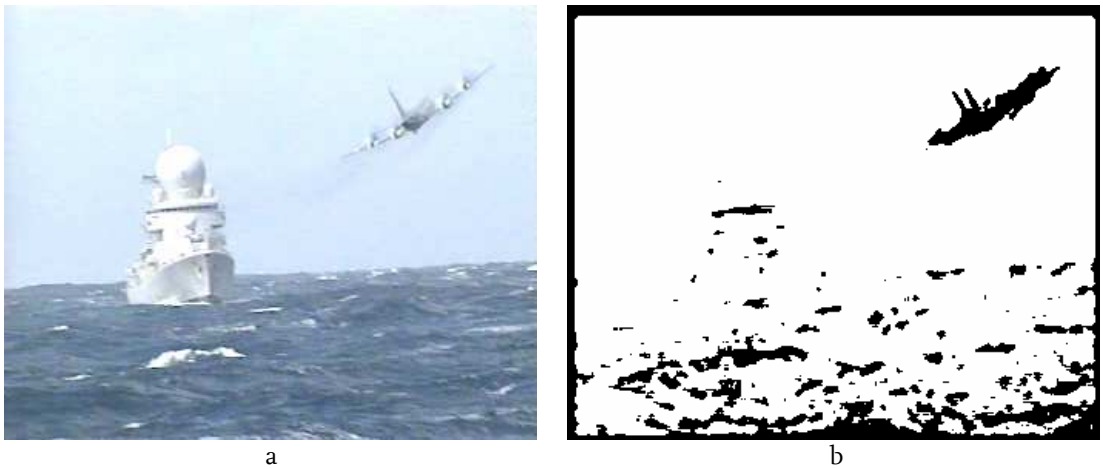
Figure 10. (a) Average after correcting the second image (Fig. 8b) for the obtained background 'motion', so the background is sharp. (b) Average after correcting the image for the obtained motion of the three rectangles, so the foreground is sharp.

Real sequences

Next, the algorithm was tested on several real sequences. An example is shown in Fig. 11, which contains two objects moving at different speeds. In Fig. 12 the displaced frame differences, corresponding to the example of Fig. 11, are shown and in Fig. 14 the displaced frame averages are shown. Sometimes *averaged* frames are preferred to show the results, because if an averaged frame is not perfect, this is noticed immediately by the human eye. On the other hand, the difference between two frames may be very small even if an imperfect correction is applied, so the displaced frame difference not always gives a clear view of the quality of the motion estimate. Averaged frames have the added advantage that they can be calculated for any desired number of frames. The results for

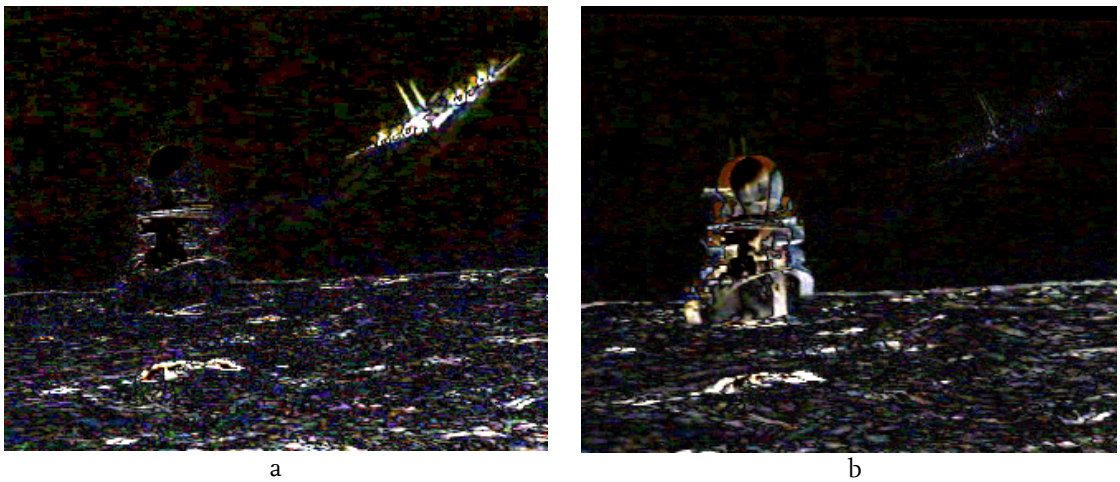
the sequence of Fig. 11 as shown here were obtained with a four-parameter (translation, rotation and zoom) model.

With this example, inaccurate results were obtained with the six-parameter model. In this case both the airplane and the white wave that is visible at the foreground move with respect to the background. Since the airplane and the wave move in opposite directions this is approximately the same as a rotation of both objects about a point somewhere between the two objects. In such cases, very many pixels are required to find a correct solution, even if there is not much noise. With a limited number of (possibly noisy) pixels, many more ambiguous solutions are possible. If the motion parameters must be used directly for some application, this may cause a problem, but if we only need the result of the motion (for instance for calculating averaged images, as illustrated before), this is less important. However, if one knows beforehand that certain motions are not very likely, it is better to use a motion model with as few parameters as possible. If for instance 25 or 50 images per second are recorded, it is very unlikely to have a large rotation component.



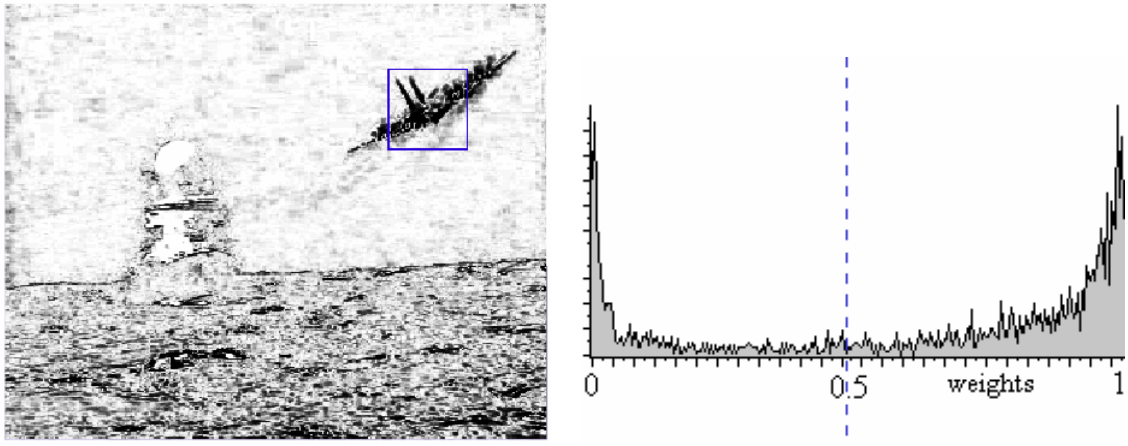
a b

Figure 11. (a) One frame taken from a sequence; the aeroplane is approaching the frigate (b) Black: pixels that are used for further calculations, based on the weights after the first run (black = 1; white = 0).



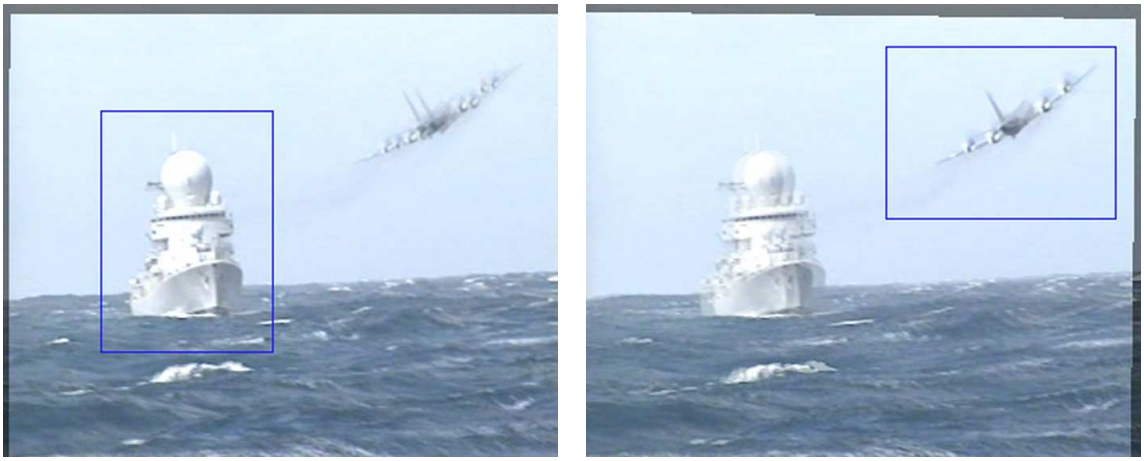
a b

Figure 12. Displaced frame differences (black = no difference; all intensities are multiplied by 4 to make the differences more clearly visible. (a) After the first run, the estimated motion is that of the largest object, so the aeroplane is visible. (b) After the second run, the estimated motion is that of the aeroplane.



a b

Figure 13. (a) Weights after the first run (black = 0, white = 1). (b) Histogram of the weights in a small rectangular area around the aircraft.



a b

Figure 14. Displaced averages of two frames after first (a) and second (b) motion estimate. The rectangles indicate the objects of which the motion parameters were found, so these objects should be sharp.

With the restricted motion model we obtained very good results for this particular sequence. We agree with [Wu and Kittler, 1990] that it is far from clear whether more complicated motion models will yield better results in practice. With many of our test sequences good results were obtained if *only* translations were taken into account. In such cases, models using four parameters almost always gave somewhat better results than the six-parameter model. One example of such a situation is shown in Fig. 15a. The camera tries to follow the two men. It is very likely that the camera motion is a simple translation. Indeed, when a two parameter motion model is used, the algorithm converges rapidly and it finds the correct background motion. In this particular case a much slower six-parameter motion model yields about the same results, because there are no ambiguous situations and because a clear structure is present in the background. During the second run of the method, the motion of the two men should be found. Since the men are moving freely, any apparent motion can be possible, so now a six-parameter affine model is used. We did not compensate for intensity changes this time.

Of course a more advanced model is needed to account for all details of the movements involved. The results are so good that in the movie that is made of the frames corrected for the motion of the foreground object, the two men are quite sharp and you can even see the hand of one of the men moving. We cannot show this in print, so instead the average of all frames, after compensation for the estimated motion of the two men is shown in Fig. 15d.

Figure 15.
(a) *One of the original frames.*
(b) *All pixels that move with equal speed (black).*
(c) *Average of six frames after compensation for the motion of the background. The background is very sharp. At the underside of the image you can see how much the individual frames had to be shifted to obtain the compensation.*
(d) *Average of six frames after compensation for the motion of the foreground "object".*

Men in camouflage uniforms caught in the act

The motion estimation method is particularly suited to find objects that do have a clearly visible structure but are still hardly visible because of the resemblance to the structure of its surroundings (for instance men in camouflage uniforms against a background that resembles the camouflage pattern). Figs. 16 and 17 show two examples. From two succeeding frames the location of the moving objects are already found. For every new frame, the contours of the moving objects will become more certain (see Fig. 16d). In both cases, the two men can be found by other techniques, not based on motion, because the structures of their uniforms differ considerably from the background. However, when the background changes due to camera movements, many standard methods to find moving objects, which are often based on differences between succeeding images, will fail [Nascimento and Marques, 2006; Boult et al. 2001]. Using synthetic images, we have been able to show that our method works very well, even when the objects have exactly the same structure as the background. It is then impossible to see the objects, unless they move. If they move, humans are able to see the exact shape of the moving objects immediately. The method described in this paper will also find the exact shape after analyzing just a few frames.

Figure 16 a, b*: two succeeding frames. Source: YouTube.* c*: weights after first run (black: does* **not** *fit the background motion).* d*: the combined result of three succeeding frames; the contours of the right man are already clearly visible; the left man is less visible (not shown in this image) because he hardly moves.*

Figure 17. Above: two succeeding frames. Source: YouTube. Below, left: weights after first run (black: does not fit the background motion). Below, right: black indicates the presence of objects that move with respect to the background.

Real time calculations

It is evident that for most military applications all processing must take place in real time or near real time. Although the simulations were not optimized for speed, on a 3 GHz personal computer the current implementation needs only a few seconds for the calculations of images of about 512 × 512 pixels. Therefore, it may be expected that with optimizations and special purpose hardware near real time applications will be possible.

Conclusions

The motion estimation method as described by Odobez and Bouthemy gives good results even for noisy sequences. The method yields correct parameters for camera movements. We have shown that the method can be applied repeatedly, while excluding pixels that were assigned to a motion region in the previous estimation step. The weights after each final motion estimate can then be used to find the contours of the regions where the object is located (segmentation). Experiments with synthetic as well as with real infrared sequences show very good results. The extended method is able to find objects moving at different speeds. A possible application is finding moving, camouflaged objects. The method is able to find the motion parameters of the separate objects, so it becomes possible to make averages where one of the objects is sharp, even if the object is moving, to suppress noise. However, one may have to restrict the model when ambiguous motions are possible. It is suggested that in future work the method is extended by adding a mechanism to consider only contiguous areas, in order to avoid artefacts due to ambiguous situations. In principle, this method can be repeated until all objects that move differently are found, but in practice the method will work only with rather large, well-structured objects.

Acknowledgements

The infrared sequences were kindly supplied by Dr. P.B.W. Schwering of the Electro-Optics Group of TNO Defence, Safety and Security in The Hague, The Netherlands. The visible light video sequences (Figs. 4, 6a, 6b, 11 and 15) are courtesy of the "Audiovisuele Dienst Defensie" (Dutch Defense-AudioVisual service).

References

Boult, T.E., Micheals, R.J., Gao, X and Eckmann, M. (2001) Into the Woods: Visual Surveillance of Non-Cooperative and Camouflaged Targets in Complex Outdoor Settings, *Proceedings of the IEEE*, pp 1382-1402.

Bouthemy P. and Rivero J.S. (1987) A hierarchical likelihood approach for region segmentation according to motion-based criteria, *Proceedings of the 1st Int. Conf. In Comp. Vision,* pp 463-467.

Fuh C.S. and Maragos, P. (1991) Affine models for image matching and motion detection, *Tech. Report CICS-P-280*, Center for Intelligent Control Systems.

Hupkens Th.M., Vos M. de, Patras I. and Hendriks E.A. (2000) Segmentation Based on Successive Robust Motion Estimates, *Proceedings of the sixth annual conference of the Advanced School for Computing and Imaging*, pp 237-244.

Kim Y-H, Martinez, A.M. and Kak A.C. (2005) Robust motion estimation under varying illumination, *Image and Vision Computing* 23 (4), pp 365-375.

Nascimento, J.C. and Marques, J.S. (2006) Performance Evaluation of Object Detection Algorithms for Video Surveillance. *IEEE Transactions on Multimedia 8(4)*: 761-774.

Meyer F. and Bouthemy P. (1992) Region-based tracking in an image sequence, *Proceedings of the 2nd Europ. Conf. In Comp. Vision*, pp 476-484.

Odobez J.M. and Bouthemy P. (1994) Detection of multiple moving objects using multiscale MRF with camera motion compensation. *Proceedings of the 1st IEEE ICIP,* vol. 2, pp 257-261.

Odobez J.M. and Bouthemy P. (1995) Robust Multiresolution Estimation of Parametric Motion Models, *J. Visual Comm. Image Representat.*, 6(4), pp 348-365.

Torr P.H.S. and Murray D.W. (1993) Statistical detection of independent movement from a moving camera, *Image and Vision Computing* 11(4), pp 180-187.

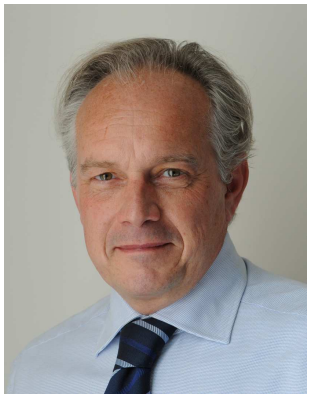
Wu S.F. and Kittler J. (1990) A differential method for simultaneous estimation of rotation, change of scale and translation, *Signal Processing: Image Communication* 2, pp 69-80.

About the Authors

Name	***Prof. dr. ir. F.G.J. Absil***
e-mail	*fgj.absil@nlda.nl*
Position	*Full Professor of Combat Systems and Military Technology*
Research interests	*Sensor networks and signal processing, Target tracking algorithms, Weapon systems, Missile guidance and control, Missile defence, C4I systems*
Current research	*Sensor management in networks, Modern missile guidance, Rapid environmental assessment*

Name	***Dr. ir. F. Bolderheij***
e-mail	*f.bolderheij@forcevision.nl*
Position	*Head Planning and Decision Support Department CAMS-Force Vision*
Research interests	*Computer sciences, Sensor technology, Command and Control, Decision Support*
Current research	*Sensor management, Decision support*

Name	***Prof. dr. T.J. Grant, CEng***
e-mail	*tj.grant@nlda.nl*
Position	*Full Professor of Operational ICT and Communication*
Research interests	*Process models of C2, Architectures and decision support for C2, Intelligent planning and scheduling*
Current research	*OODA-based architectures for C2, Agile planning using BPR and CE, Modelling adversarial interactions, Cultural and organizational influences on C2.*



Name	***Dr. ir. J.-P. Hermand***
e-mail	*jp.hermand@nlda.nl*
Position	*Associate Professor*
Research interests	*Geosciences, Ocean acoustics, Marine geophysics, Model-based signal processing, Inverse problems, Communication, estimation and classification theory*
Current research	*Development of acoustic inversion methods for REA based on a network of acoustic and oceanographic sensors*

Name	***Ir. R.R. Hordijk***
e-mail	*rr.hordijk@nlda.nl*
Position	*Assistant Professor of Computer Science*
Research interests	*Robotics*
Current research	*Networks of cooperative robots*

Name	***Dr. Th.M. Hupkens***
e-mail	*thm.hupkens@nlda.nl*
Position	*Associate Professor of Electro-Optical Systems*
Research interests	*Artificial intelligence, Pattern recognition*
Current research	*Cooperative mobile robots, Automatic detection of moving objects in video recordings, Feature extraction from digital images*

Name	***Drs. J.M. Jansen***
e-mail	*jm.jansen.04@nlda.nl*
Position	*Associate Professor of Command & Control*
Research interests	*Implementation and applications of declarative programming languages, C4I systems*
Current research	*Dynamic workflow systems, Applications for military planning and crisis management*

Name	***Dr. ir. R.H.P. Janssen***
e-mail	*rhp.janssen@nlda.nl*
Position	*Assistant Professor of Mathematics and Operations Research*
Research interests	*Stochastic processes, Network science, Decision analysis*
Current research	*Networks and agent based modelling, Robust planning*

Name	***Dr. L. Koene***
e-mail	*l.koene@nlda.nl*
Position	*Assistant Professor of Ammunition and Ballistics*
Research interests	*Non-lethal weapons, Terminal ballistics, Forensic science*
Current research	*Non-lethal weapons*

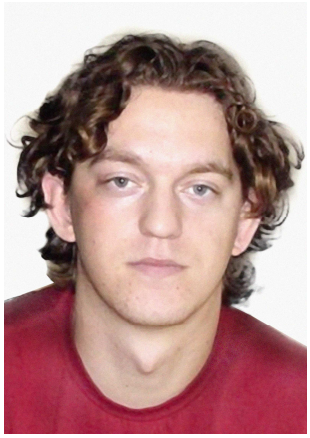
Name	***Drs. A. V. van Leijen***
e-mail	*a.v.van.leijen@forcevision.nl*
Position	*Head Planning and Decision Support Department CAMS-Force Vision*
Research interests	*Passive sonar, Rapid environmental assessment, Metaheuristic optimization*
Current research	*Geoacoustic inversion with sound sources of opportunity*

Name	***Dr. H. Monsuur***
e-mail	*h.monsuur@nlda.nl*
Position	*Associate Professor of Mathematics and Operations Research*
Research interests	*Network science, Game theoretic modelling, Decision analysis, Social network analysis*
Current research	*Networks and agent based modelling, Robust planning*

Name	***D.M. Ooms MSc, Capt NL Navy (ret)***
e-mail	*dm.ooms.02@nlda.nl*
Position	*Associate Professor of C4ISR*
Research interests	*C2, C4I architecture, NEC/NCO, Operational use of mobile networks*
Current research	*NL Defence C4I architecture*

Name	***Ir. U.D. Ramdaras***
e-mail	*u.ramdaras@nlda.nl*
Position	*PhD student, subject: Networks of maritime radar systems for air and surface picture compilation*
Research interests	*Network centric warfare, Sensor management, Target tracking, Particle filtering techniques*
Current research	*Sensor selection and localisation algorithms for target tracking in sensor networks*

Name	***Prof. Drs. Dr. L.J.M. Rothkrantz***
e-mail	*ljm.rothkrantz@nlda.nl*
Position	*Full Professor of Sensor Systems*
Research interests	*Intelligent systems, Speech recognition and synthesis, Multimodal information fusion and smart cameras*
Current research	*Multimodal information fusion, Intelligent systems, Speech recognition and synthesis, Facial expression recognition and synthesis*



Name	***Ir. J.C. Stevens***
e-mail	*c.stevens@nlda.nl*
Position	*PhD student, subject: Intelligent multimodal information fusion*
Research interests	*Cognitive modelling, Human perception, Gestalt theory, Emergent systems, Artificial intelligence*
Current research	*Designing and implementing an autonomous, adaptive and context sensitive surveillance system based on a novel computational model of human audiovisual perception.*

Name	***Ir. E. J. Trottemant***
e-mail	*e.j.trottemant@tudelft.nl*
Position	*PhD student subject: Advanced missile guidance laws*
Research interests	*Optimal and robust control, Convex programming, Game theory, Linear matrix inequalities, Model predictive control*
Current research	*Robust minimax strategies for missile guidance*

Name	***Drs. M.P.A. van de Ven***
e-mail	*mpa.vd.ven@nlda.nl*
Position	*Assistant Professor of Mathematics and Operations Research*
Research interests	*Military operations research, Data analysis*
Current research	*Search and detection*

Name	***Dr. ir. A.F. Vermeulen***
e-mail	*af.vermeulen@nlda.nl*
Position	*Associate Professor of Guided Weapon Technology*
Research interests	*Missile guidance, Control engineering*
Current research	*Missile interceptor guidance methods, Optimal evasive manoeuvres*

Name	***Dr. A.J. van der Wal***
e-mail	*aj.vanderwal@nlda.nl*
Position	*Associate Professor of Signal Analysis and Artificial Intelligence*
Research interests	*Modelling in physics, Solid state physics, Analogue electronics, Quantum computing, Real-time control and soft Optimization*
Current research	*Fuzzy logic control, Artificial neural networks and evolutionary computing*