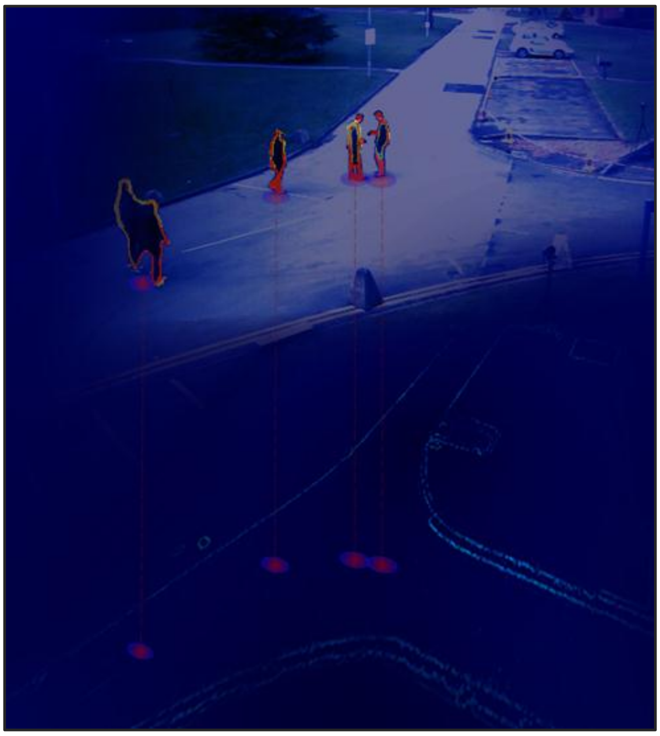
Laboratory for Geo-information and Remote Sensing

Thesis Report GIRS-2015-08

Geospatial video tracking for enhanced situational awareness

Loek van Beusekom



April 2015

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Geospatial video tracking for enhanced situational awareness

Loek van Beusekom

Registration number 89 01 05 063 120

Supervisors:

Dr.ir. RJA van Lammeren (WUR)
Richard Joseph (CGI)

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands

April 2015
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2015 -08
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

It reflects common sense that ‘everything that happens, happens somewhere in space and time.´
However, it is exactly that implicitness to our understanding of the processes of our world that shows
and underlines its critical importance.

(Goodchild 2005)

Abstract

Even in this age where sensors are abundantly available, situational awareness (SA) is far from certain. The objective of this study is therefore to investigate the potential of geospatial video tracking application to enhance SA. Two research questions were formulated focusing respectively on the performance and on potential use cases of geospatial video tracking. To test the performance both a single and multi-camera video tracking application were developed and evaluated based on the CLEAR method (Bernardin and Stiefelhagen 2008). For evaluation purpose the Performance Evaluation of Tracking and Surveillance (PETS) 2009 dataset was used. The found accuracy for the multi-camera video tracking application is 65% lower than the standard single camera tracking application. Found tracking precision of the multi-camera video tracking application is 8.5% higher compared to the single video tracking application. Secondly an expert survey was performed to gather more insight in potential use cases for geospatial video tracking. The performed expert survey clearly indicates numerous applications in which geospatial video tracking could contribute. Overall it can be concluded that geospatial video tracking has a significant potential, but needs serious improvements before it can be used in practice. A number of improvements are provided for further research.

Foreword

This thesis is written for the completion of the MSc program Geo-Information Science at Wageningen University. This MSc program focusses on the integrated use of earth observation techniques (Remote Sensing) and Geographic Information Systems (GIS) for problem-solving within the environmental discipline. The current developments within the GIS domain, specifically the fields of real-time GIS and object based classification have however significantly increased the potential of GIS within numerous sectors outside its traditional field of application. Therefore this study embarks on a journey to explore how GIS can contribute within the field of video tracking, which is a fast developing topic within the domain of computer vision.

This study was performed from March 10th till October 10th 2014, and was facilitated by the Dutch GEO-ICT department at CGI, which provided guidance and facilities during the whole process. During this journey I received the help and support of many people, and I would like to extend my sincere gratitude for this. Specifically towards Robert Voûte (Practice leader Geo-ICT) and Kerijn van der Burgt (Practice Manager Geo-ICT) for providing this opportunity and their inspiring vision on the topic. I also would like to thank my supervisors Ron van Lammeren (WUR), Bram Driessen (CGI) and Richard Joseph (CGI) for their guidance and critical feedback. Additionally I would like to thank all experts who participated in the survey for their time and insight; this really helped me develop a more complete view on the potential of GIS within the video tracking domain. Furthermore it is a great pleasure to thank everybody from the Geo-ICT department including all other interns for their hospitality and support. You really made me feel welcome and at home.

Table of Contents

List of abbreviations

GI(S)	Geo information (system)
fov	Field of view
fp	false positives
miss	Misses
mme	Mismatches errors
MOTA	Multiple Object Tracking Accuracy
MOTP	Multiple Object Tracking Precision
Pdf	probability density function
POM	Probability occupancy map
PTZ	Pan–tilt-zoom
SA	Situational awareness
sdv	Standard deviation

1 Introduction

Understanding what is happening can be of vital importance. This is especially relevant in domains where poor decisions may lead to serious consequences for example during police or military operations. This state of understanding, often referred to as situational awareness (SA), has to do with being aware of what is happening in the vicinity in order to understand how information, events, and one's own actions will impact goals and objectives, both immediately and in the near future (Wikipedia 2014). The concept of situational awareness (SA) has become a key concept in many fields of research including emergency response, cyber security, law enforcement and certain business lines.

With the current state of sensor technology often more than enough data is available, but the challenge remains to extract important information and present it to the user in an understandable and timely manner. To aid the process of sorting immense streams of available video data frequently advanced computer vision techniques are used like video tracking. This can be used to detect and track different objects, and automatically identify specific situations, for example the gathering of an angry mob, or the occurrence of a major incident. Because often not a single camera or sensor is used, integration of different sensors in a single system is becoming increasingly important. By integrating multiple sensors, objects can be tracked automatically over multiple screens or pinpointed with a greater accuracy. Advanced sensor and communication technology has resulted in a situation in which decisions are increasingly made from distant location like a remote control room. Extreme examples are the military drones operating in the harsh environment of the Middle East remotely controlled from an air-conditioned complex in Nevada. This development however puts additional strain on the used technology to provide a complete and understandable image of the situation to ensure maximum SA.

The gathered information during crisis situations is often spatio-temporal in nature, understanding the spatial and temporal dimensions is therefore regarded as an essential part of SA (Luokkala and Virrantaus 2014).Currently however many of the information systems used in disaster management and other sectors with a high focus on SA lack effective integration of the spatio- temporal dimension (Aubrecht, Fuchs, and Neuhold 2013). Often just a simple (printed) map is used in combination with an incident log file, which drastically limits obtaining and sharing of SA (Luokkala and Virrantaus 2014). In many other fields Incorporation of the spatio temporal dimension is done by integration of a Geographic information system (GIS). These systems are explicitly developed for the management, analysis and display of data with spatial and temporal components. GIS is therefore not only a useful tool for representing relevant information in a understandable way, but also for advanced analysis like pattern recognition and the integration of other information sources based on their location or vicinity. This is specifically relevant in relation to the constant development towards more complex and bigger camera systems. Especially because recent advances in real-time GIS can provide a (near) real-time overview of the situation and perform all kinds of analysis on the fly. The concept of geospatial video tracking can potentially revolutionise the way in which complex technology or video aided operations are managed and organised. The main objective of this study will therefore focus on investigating the potential of geospatial video tracking to enhance situational awareness.

1.1 problem definition and scope

Imagine a big control room like the one shown in Figure 1, where one or more walls are filled with monitors. Each of these monitors is displaying the live video feed from a different surveillance camera. When a major incident occurs, for example a natural disaster or a clash between rival supporter groups, each of these monitors will show a piece of the puzzle. It is however unlikely that one single screen will give a complete overview of the situation. None of the people in the control room will be able to watch all screens simultaneous and therefore no one will be completely aware of what is going on. Based on this the problem statement was defined as following: ‘Without integration of available camera images within a common plane, situational awareness is limited, resulting in sub optimal decision making.’

To enhance situational awareness during such crisis situations, in this thesis the concept of geospatial video tracking is introduced. A Geospatial video tracking application can be used to automatically extract relevant information from the available video streams, and project it on a single plane (i.e. map). Not only could this facilitate easier understanding of the current situation. It can also function as a basis for further analysis (e.g. trend detection) because relevant information for the different cameras is combined in single view with a common reference (location).

Traditionally the term geospatial was only used in reference the spatial dimension. Today however the line between the spatial and temporal dimension is often faded. Because this is an exploring study and not at all intended as a way to restrict the potential of geospatial video tracking, the term geospatial will be used in the broadest sense of the word. Therefore in this paper the terms geospatial and spatio temporal are used interchangeably.

Many camera types are available and depending on the situation they can be applied in different ways. Often simple stationary cameras are used, but many other types are available ranging from pan–tilt-zoom (PTZ) to 360° cameras with night vision (IR) capabilities. To optimise coverage different (types of) cameras are often applied within a system. In such a system for example an overview camera is combined with a number of zoomed-in or PTZ cameras to provide detail without losing overview. Often these systems (temporally) have blank or weak spots (no or low camera coverage) which make it harder to continuously track a moving object. Numerous studies are devoted to the optimisation of camera type, location, viewing angle and methods for object tracking with sparse camera coverage. Although these issues are very relevant in the context of multi-camera video tracking this study only focusses on object tracking in areas covered by multiple fixed type cameras. The term single and multi-camera video tracking is in this report therefore used as a reference to the number of cameras covering a specific object or point, rather than the number of cameras used in a specific surveillance system.

Figure 1 Control room of Rio de Janeiro, currently the biggest control room in Latin America

1.2 Objective and research questions

The objective of this study is to investigate the potential of a geospatial multi-camera tracking application to enhance situational awareness.

From this objective the following research questions are derived:

- **RQ1. What is the tracking performance of a multi- versus single video tracking application?**
  Performance is the key when applying video tracking for enhancing situational awareness. By combining different overlapping camera views object detection and therefore location estimation and tracking performance can potentially be increased compared to a single camera video tracking application. To objectively investigate the performance and possible points of improvement of a multi camera video tracking application, a comparable single camera version is required. Because no suitable single camera video tracking application or usable benchmark data is available, both a single and multi camera video tracking application will be developed based on identical components. Additionally both applications will be evaluated based on their tracking accuracy and precision.

- **RQ2. What are potential use cases for geospatial video tracking to enhance situational awareness?**
  Geospatial video tracking could be applied in many situations. Each situational however has its own specific requirements. To identify currents gaps in technology a survey will therefore be performed with experts in the sectors security & law enforcement, military and consumer behaviour focussing on the potential use cases and their specific requirements.

1.3 Reading guide

In the following chapters the answers on the research questions will be given. Doing so chapter two will continue with a more in depth literature review on situational awareness, GIS and video tracking. In chapter three the proposed method is described, continued by chapter four on the design of both video tracking applications. In chapter five the results from the tracking performance evaluation are described and in chapter six the most important remarks from the expert survey are summarised. The report ends with a discussion of the followed approach and concluding remarks in relation to the resulting outcome.

2 Review chapter

From the start this study has been conducted with a GIS perspective. To acknowledge this approach the term geospatial video tracking is introduced, which is a reference to the geospatial part of GIS and also indicating a wider perspective towards current challenges within the video tracking domain. As far as known by the author the term geospatial video tracking has not been used in any scientific article or other publication. However in an article on port security the concept of geospatial surveillance was introduced, which has some similarities with the concept of geospatial video tracking ((Olson 2013). The combination of GIS with video tracking is rather sporadic in literature. In these exceptional cases these two fields are combined, the purpose is however often very different from the goal which is pursued in this study. For example In Bradford et al. (2011) the authors combine GIS and video tracking to provide geographical context to the target feature space specifically to improve tracking performance by minimizing the search area in which a target is likely to be located. There are however many studies, specifically in the field of security and law enforcement, which (maybe unintentionally) do use simple geospatial/GIS components (e.g. maps) in their video tracking application to describe the movement of the detected objects (Fleuret et al. 2008; Lee, Romano, and Stein 2000; Park and Trivedi 2006). However the link with GIS is limited, and only a few (Zwahlen, Yahr, and Berven 2012) go into the broader advantages GIS has to offer (spatial and temporal queries, use of multiple layers etc.).

In relation to situational awareness, GIS is gaining increasing interest in literature, specifically in the context of disaster management and early warning systems (ESRI 2008; Jiang et al. 2012; Luokkala and Virrantaus 2014). However these studies focus heavily on manually collected or satellite data, and less on continuous data streams like surveillance cameras.

Only a single article published in 2006 brought together video tracking, GIS and situational awareness (Park and Trivedi 2006). In this article the authors presented an application to automatically detect and classify moving objects in different surveillance cameras (video tracking), show them on a plot and estimate their trajectory with the purpose to locate and warn for potentially unsafe traffic situations (situational awareness).

Because this study focusses on the potential of video tracking techniques and GIS to enhance situational awareness, in the following paragraphs a general overview of each of these themes will be given based on existing literature.

2.1 Situational awareness

The concept of situational awareness (SA) has become a key concept in many fields of research including military and security. The term originally comes from the world of the military fighter crew. SA is especially important in domains where the information flow can be quite high and poor decisions may lead to serious consequences. According to Luokkala and Virrantaus (2014) SA can be divided in the following levels, which are (1) perceiving the elements in the environment, (2) comprehending the current situation, and (3) projecting the future status of the situation. The First level of SA (perception) is constructed from perceiving what is happening in the environment. This can be from direct observations, or technology aided observations like with information- or actor linked display systems. The second level of SA (comprehension) has to do with understand what the perceived information means when taken together and in relation to the current goals and objectives. This step is about judging and prioritizing the importance of information and its meaning in relation to the goals. The final level of SA (forecast) consists of the ability to foresee how the situation will develop in the (near) future. Achieving this level of SA requires expertise about the operations and dynamics of the system operating in. This expertise can only be reached by devoting significant amount of time on practice to develop a set of strategies for different situations. This allows them to respond in a quick and proactive manner. Often the first level of SA is seen as the most important one because: (1) The first level serves as the basis upon which the other two levels will be built and (2) most of the reasons that prevent an actor from achieving SA are associated with the first SA level (Luokkala and Virrantaus 2014).

According to this definition SA is highly situation and objective dependent. Therefore no universal method can be defined to achieve maximum SA. However, in a study on the decision making process during crisis situations Luokkala and Virrantaus (2014) noted that most of the gathered information was spatio-temporal in nature, and that maps formed an important basis for SA, communication and decision making. They also discovered that despite its importance only simple map functions were used, and that more advanced GIS methods, such as data mining and space–time cubes, were not used in any of the cases. (Luokkala and Virrantaus 2014) In a similar context Aubrecht, Fuchs, and Neuhold (2013) described that understanding spatial patterns of (hazardous) events, as well as the geographical extent of their impacts and risk development over time is crucial for both risk reduction as well as recovery efforts. The assessment of spatial and temporal dimensions is therefore an essential part within integrated (disaster) risk management; however, according to Aubrecht, Fuchs, and Neuhold (2013) this has often been neglected in respective academic efforts. Additionally In the specific case of a surveillance scenario Snidaro, Visentini, and Foresti (2011) argue that data fusion (combination of multiple source information like databases, sensors, human input) is a necessary tool to provide flexibility to manage unpredictable events and to enhance SA. Although these observations are not universal they would definitely argue that in certain cases SA could benefit from a more integrated and spatio- temporal approach.

2.2 Geographic information system

A geographic information system (GIS) is a computer system designed to capture, store, manipulate, analyse, manage, and present all types of spatial or geographical data (Wikipedia 2015) the main advantage of GIS is that it can relate seemingly unrelated information by using location, as the key index variable. It is often stated that 80% of all data has a spatial component, which would emphasize the potential for such applications. traditionally GIS's main purpose was to aid in analysis for science and planning purposes, but with the development towards real time GIS the potential greatly increases as a decision support system and real time information platform for a wide public, including people working in public safety and disaster management (Hsu et al. 2010; Kamel Boulos et al. 2011)

Depending on the specific situation and objective, a number of GIS techniques can be applied to enhance SA. For the first level of SA, which has to do with perceiving the elements in the environment, two promising examples are geospatial sensor systems and intelligent user interfaces. As described in Olson (2013) the advantage of geospatial sensor systems is that because these sensors are aware of their location they can give the specific location of a detected incidence, and automatically direct other sensors (like pan tilt zoom cameras) or people to the right location. The integration of GI within a user interface can greatly increase the understanding of the user, by creating a single overview map on which all relevant information can be plotted, and which can function as a solid basis for integration with other information sources and layers. For the second level of SA, which is about comprehending the current situation, GIS offers methods for replaying whole or a part of an incident and provides capabilities to query for specific events based on location and other types of extracted metadata. This can be helpful when investigating specific incidents or looking at trends (Zwahlen, Yahr, and Berven 2012). The third level of situational awareness, which is about projecting the future status of the situation, can be aided by using GIS for the development of likely scenario's based on historical data and for creating simulation incidents for training purposes (Park and Trivedi 2006). These above stated examples are not in any way intended to give a complete overview of the capabilities of GIS in relation to SA, but rather as an indication of what is possible.

2.3 Video Tracking

Computer vision has fascinated humans for many years, and this interest continuous today. Not only for purely intellectual reasons, but also due to the potential of such a system (Andreopoulos and Tsotsos 2013). One of the key applications within computer vision is video tracking, which is the process of locating one or more moving object(s) over time using camera technology (Wang 2013). This is often used within fields like human-computer interaction, robotics, security and surveillance, augmented reality, traffic control, medical imaging and video editing (Maggio and Cavallaro 2011; Yilmaz, Javed, and Shah 2006).

Simple object detection under stable conditions and without object interaction is considered solved by the majority of the computer vision community, however a number of issues still make the tracking process a challenge (Andreopoulos & Tsotsos 2013). These include general tracking issues like lighting conditions, shadow, object pose, occlusion, object interaction, background change. An extensive review of these issues is given in Maggio and Cavallaro (2011) and Panin (2011).

Currently a clear development towards bigger and more advanced camera systems is visible. Advantage of such systems is that people, or objects in general, can be tracked over a broader area and that overlapping camera views can be used to limit occlusions (Yilmaz, Javed, and Shah 2006). However with such systems a number of new challenges arise related to the integration of different camera feeds and information representation to keep the situation understandable for the users (Wang 2013).

In the following sub-paragraphs the basic components of a video tracking application will be explained.

2.3.1 Basic video tracking components

Two key steps in the video tracking cycle are object detection and tracking, however a number of additional steps like pre-processing and object recognition are often included in the video tracking process. Figure 2 shows a number of components frequently included in the video tracking cycle. To keep this report concise as possible we will only go into the steps relevant to this project, namely object detection, object tracking and data fusion. A more detailed review of these additional steps is given in (Yilmaz, Javed, and Shah 2006).

In the paragraphs 2.3.3 to 2.3.5 the relevant video tracking components are explained in more detail. However the selected video tracking approach is for an important part influenced by specific object characteristics. Therefore in paragraph 2.3.2 first an introduction of the available visible features will be given.

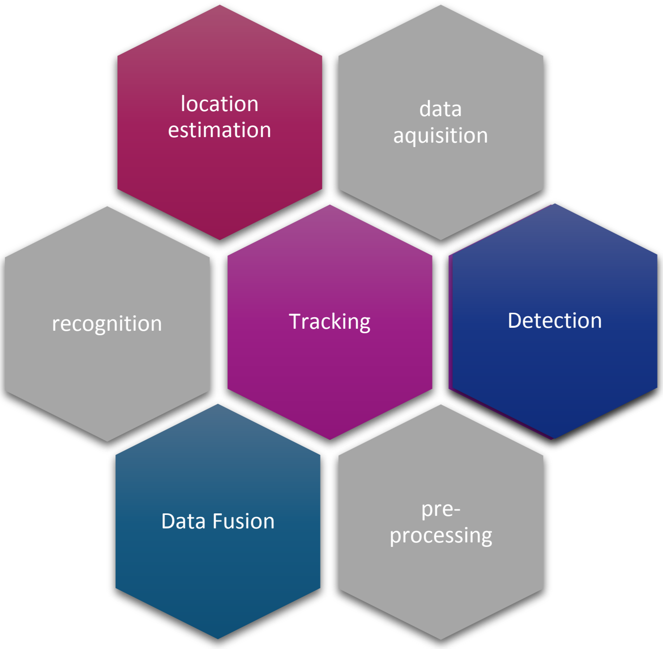


Figure 2 Common tracking components adapted from Panin (2011). Greyed out components are non-relevant for this study, and not further discussed in this article.

2.3.2 Visual Features

An important step in de development of a tracking application is the selection of suitable features to track. These are often features which gives the object its uniqueness, so that is can easily be distinguished from its surrounding. In general a combination of these visual features is used to create a robust video tracking application. In Figure 3 a number of visual features are shown.

2.3.2.1 Colour

Colour is one of the most widely used features for object tracking, especially if the object colour is stable and clearly distinguishable from the background. Good examples are number plates and faces. Because of changing object orientation, viewing direction and illumination, artefacts can occur resulting in a change of object colour. To accommodate for slight colour changes a colour range can be defined based on specific object characteristics. The human perception of colour is based three different light sensitive cones, with high sensitivity in respectively violet, green and yellow-green light. The CIE 1931 RGB colour space is a mathematical representation of the human visual perception and is specifically designed to encompass all colours the average human can see. Different other colour spaces and models are developed with the CIE 1931 RGB colour space as basis. Since the response curves of these light sensitive cones overlap to a large extent, a number of colours are undistinguishable by the human eye. By using other colour spaces, otherwise indistinguishable objects can potentially be separated and used to reduce artefacts from shading and reflection. For example HSL (hue, saturation, lightness) can be used to reduce colour changes related to surface orientation, illumination direction and illumination intensity changes. Additionally normalised RGB colour and normalised colour difference models can be used to reduce the influence from object reflection (Maggio and Cavallaro 2011).

2.3.2.2 Edges

The boundaries between objects or background often show strong variation in reflection intensity. Edge detection methods can be used to identify these boundaries. Because edges are less sensitive to illumination changes than colours, they are often used for object detection. Different types of edges are available, which can be used in a variety of ways. Internal- and Object boundary edges (contours) can be used for object recognition, and external edges are often used for background removal. To locate contours within an image different edge detection methods like canny edge or the Marr – Hildreth technique can be used (Maggio and Cavallaro 2011). These edge detection methods filter an (often smoothed) image in vertical and horizontal direction to detect consecutive pixels with high brightness differences. Additional edges can be selected above certain threshold, and thinned edges can be converted into separate features for further processing. To reduce errors during the matching process also edge direction and edge intensity can be used (Panin 2011).

2.3.2.3 Optical flow

In certain situations objects can be detected by their movement related to the background or other moving objects. This motion is not the real object movement but the so called apparent motion, which is the 2D motion over the image plane. The apparent motion can be estimated by constructing displacement vectors between pixels or features in subsequent images. Pixels or features with significantly differentiating vectors can then subsequently be identified as separate objects. For the calculation of displacement vectors brightness constancy of corresponding pixels in consecutive frames is assumed. Because the motion is based the shortest vector between corresponding reference points in sequential images, this method performs less well with low texturized or fast moving objects. Although optical flow can be used as main visual tracking feature, it is most often used as an additional step in the tracking cycle (Maggio and Cavallaro 2011).

2.3.2.4 Texture

Texture is a measure of the reflectance variation due to surface roughness and irregularity. In certain cases texture provide a very useful cue for object tracking, especially for objects with a distinctive local texture pattern. Keypoints like corners or edge crossings tend to stay distinctive under different angles and illumination. Similar to edge features, the texture features are less sensitive to illumination changes compared to colour.

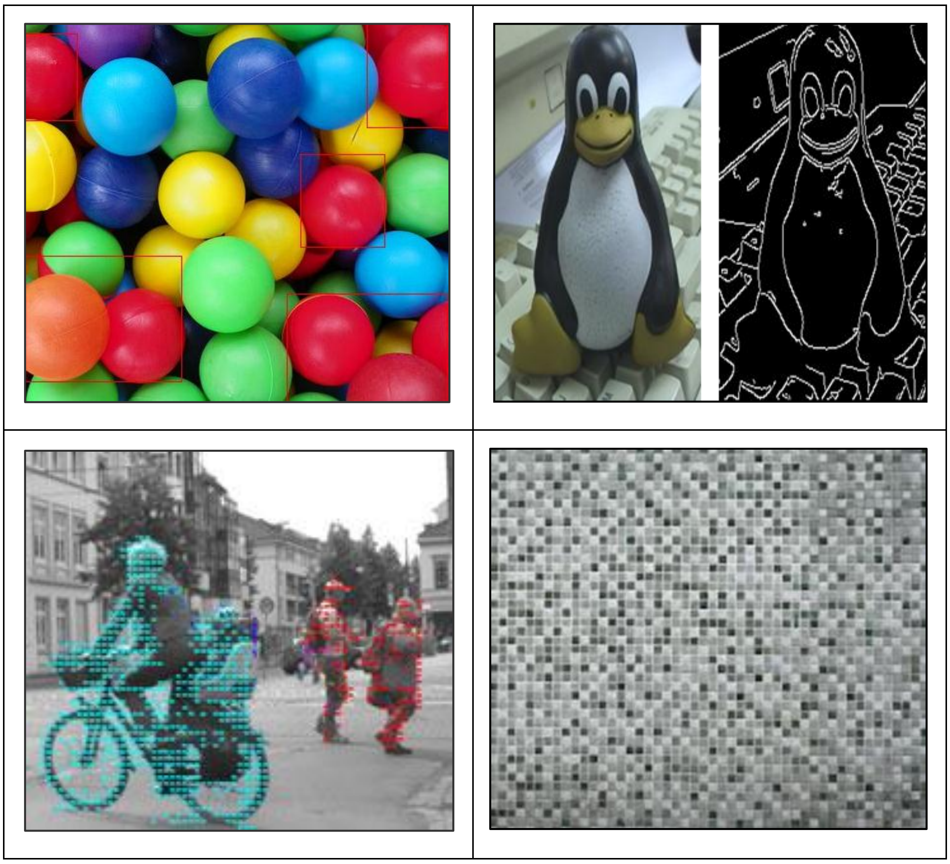


Figure 3 Visual features. From top left to bottom right: colour, edges, optical flow and texture

2.3.3 Object detection

Object detection concerns the localisation of relevant objects within an image. This can be based on object characteristics (like a colour range), movement or by comparison with an object library. In Figure 4 a number of object detection methods are shown. In paragraphs 2.3.3.1 to 2.3.3.4 different object detection methods will be described in more detail. Object detection can be performed continuously, or only when it first appears in the scene combined with a robust tracker algorithm. Often object detection is performed based on a single frame, however some detection methods use the difference between consecutive frames to minimise errors causes by image noise.

2.3.3.1 Point detectors

Point detectors focus on points which have characteristic texture pattern in their neighbourhood. These interest points, or also called key points, are amongst others used in context of motion tracking and object recognition. A desirable quality of an interest point is its invariance to changes in illumination and viewing angle. Different keypoints detection methods have been developed in the past. Simple keypoint detectors like the Moravec corner detection method identify pixels with high total intensity variation in 8 different directions within a defined box (e.g. 3x3 pixels)(Maggio and Cavallaro 2011). The Harris corner detector not only uses the total intensity variation, but additionally divides the intensity variation in an x and y component. The relation between the two components defines the type of keypoint (edge or corner). The eigenvalues of these components are rotational invariant and therefore the Harris corner detector can be used to detect keypoints which are rotated (Panin 2011).

Because in certain cases keypoint characteristics depend on the scale at which they are viewed, detection is also scale dependent. To mitigate the scale effects some keypoint detectors use a number of scaled down images to determine which keypoints are scale invariant. A well-known scale invariant detector is the Scale-Invariant Feature Transform (SIFT), but different other detectors like the Speeded Up Robust Features (SURF) are also applied. These scale invariant keypoint detectors have a high level of precision. Mikolajczyk and Schmid (2005) empirically showed that SIFT outperforms most point detectors and is more resilient to image deformations. However because of the complex detection process it requires more time compared to simpler keypoint detectors and are therefore less suitable for real-time application (Rosten and Drummond 2006).

2.3.3.2 Background subtraction

Besides a direct detection method objects can be located by removing pixels which are not classified as objects, resulting in regions likely covered by objects of interest. This detection method is based on the comparison of a background model with the current image. Depending on the amount of background variation and required precision the background model can vary between a simple empty background image to a complex background model which accommodates for global or regional intensity variation due to illumination or shadow change (Maggio & Cavallaro, 2011) A background model is often created by collecting pixel values over longer time to extract statistics parameters. With these parameters the likelihood can be calculated if a pixel belongs to the background. Not only the original images can be used for background subtraction, but also derivatives like an edge map can be used. Advantage is that these derivatives are less sensitive to illumination changes and therefore require less a complex background model.

Advanced background subtraction models like the one from Elgammal et al. (2002) not only match the current pixel with historical pixel values but also with values from neighbouring pixels. Li and Leung (2002) incorporate texture into their background model to reduce model sensitivity for illumination changes and to accommodate for shadow effects. Rittscher et al. (2000) use a Hidden Markov Model to include an additional shadow state besides commonly used back- and foreground states. A more elaborate review of background subtraction models is given in (Yilmaz, Javed, and Shah 2006). In general background subtraction tend to fail when the background changes significantly in a short period, for example when the sun becomes obscured by clouds, or when background objects like trees are moving. Another limitation is the inadequacy to distinguish between occluded objects or objects within close proximity.

2.3.3.3 Segmentation

Segmentation methods are used to divide the image in perceptually similar regions. A segmentation method consists of two components, namely a criterion which clearly identifies the object(s) and an algorithm which can efficiently achieve this segmentation. To track object, segments can be followed over consecutive images.

Simple segmentation methods use a colour range to distinguish areas which match the object colour. More advanced methods automatically define coherent regions based on a number of predefined parameters. Often used methods are Mean-Shift Clustering and active contours. The Mean-Shift method uses randomly located points as starting point to create areas with similar characteristics. The active contour method uses an energy function to find the optimal contour based on the image intensity landscape (Blake 2006). Often a likely location (I.e. the object location from previous image) is used as a starting point to create a new object area. Different energy functions can be developed for example based on image gradient (starting from the detected edges) optimal contour distance, or shapes based on a reference library. More information by active contour is given in (Blake 2006; Yilmaz, Javed, and Shah 2006).

2.3.3.4 Supervised learning

Object detection can also be performed by a supervised learning mechanism. With this technique a machine is trained to automatically distinguish certain objects base on a set of classified examples. Unlike other detection methods the selection criteria are not specifically defined by the user. Only a number of potentially interesting features is provided, which can be used by the machine to develop a suitable classification formula. The advantage of such a system is that many moderate or weak features can be combined to develop a strong classification method. many different features can therefore be used including object area, object orientation, and object appearance in the form of a density function (for example a histogram) (Yilmaz, Javed, and Shah 2006). These learning approaches include, but are not limited to, neural networks (Rowley, Baluja, and Kanade 1998), adaptive boosting (Viola, Jones, and Snow 2005), decision trees (Grewe and Kak 1995), and support vector machines (Papageorgiou, Oren, and Poggio 1998).

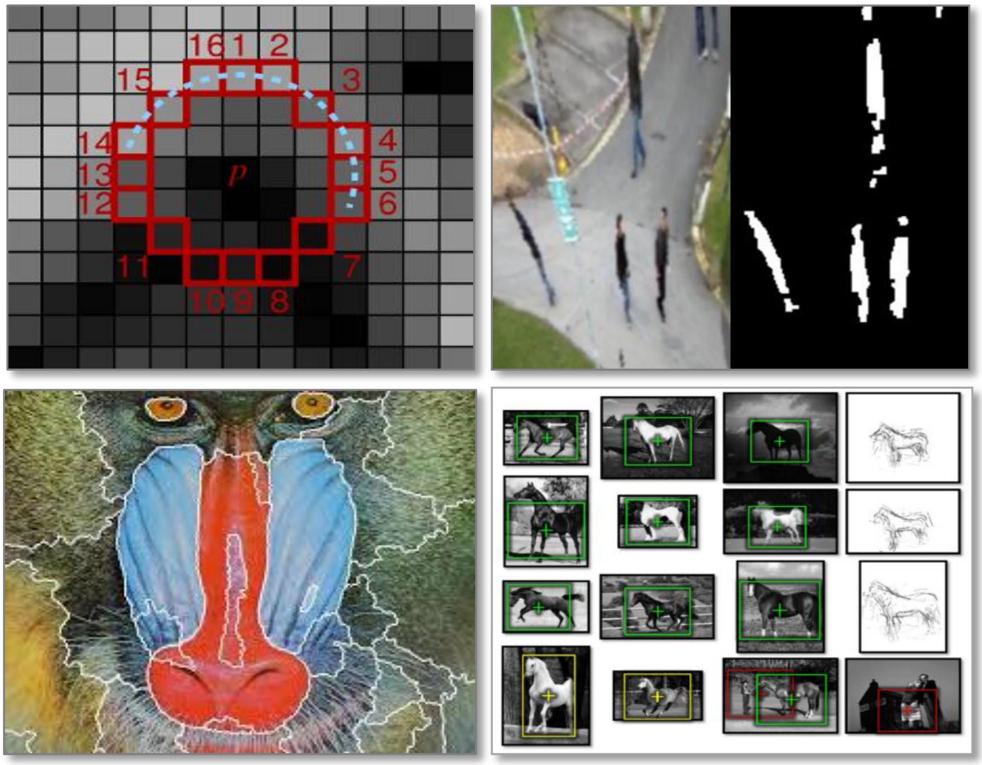


Figure 4 Object detection methods. From top left to bottom right: point detection, background subtraction, segmentation and supervised learning (Shotton, Blake, and Cipolla 2008)

2.3.4 Object tracking

Object tracking is the process of following a specific object over time by detecting its position in consecutive images. During this process each different object is given a unique ID and its object trajectory is stored into a database so it can be plotted on a map or used for further analysis. Depending on the situation and object specifications different tracking techniques can be used. In the following subparagraphs these different techniques are described in more detail.

2.3.4.1 Point tracking

Point tracking is used to track small objects, or specific points of an object based on visual characteristics (see paragraph 2.3.2 for more information on visual features). This approach requires an external mechanism to detect the objects in every frame (for example SIFT point detection). Point association is based on the previous object state which can include object position and motion. Many different methods exist for point tracking but in general they try to link points of the same object in consecutive images by minimising overall distance without neglecting model constrains (e.g. max speed), or by probability based on historical object velocity and direction. Because both methods depend on a known number of objects additional constrains are required for occlusion handling and object birth and death. In case an object consists of multiple points additional constraints can be introduced limiting the relative movement of points located on the same object (Maggio and Cavallaro 2011).

2.3.4.2 **Kernel tracking**

A kernel is a simple representation of the object shape, like for example a rectangular or ellipsoid enclosing the specific object. With kernel matching the area located within the kernel of a reference image is iteratively matched with different pieces of the image to find the area with the highest correspondence. This matching process can be based on the actual pixel values within the kernel, or for example a derivative like average brightness. The most straight forward method for template matching is the brute force method. This method divides the image in different (overlapping) blocks and matches each of these blocks to the reference object. The object is located based on the location of the block with the greatest similarity. When considering changing object shapes and scale variation brute force matching becomes a very computational exhaustive process. Therefore a number of more advanced methods have been developed. For example the mean shift procedure, which maximised the appearance similarity by searching a coherent area in the surrounding of the previous object with a matching histogram (Yilmaz, Javed, and Shah 2006). This technique can not only be used for object recognition, but is also very suitable for object tracking. for increased performance the search area can be further increased by using historical object movement to predict the most likely location (Yilmaz, Javed, and Shah 2006). A related term which is often used in literature is blob tracking. This term is often used as a reference to the cluster of pixels defining a specific object or object outline. This is specifically relevant in the context of object extraction (like background subtraction).

2.3.4.3 Silhouette tracking

Objects can have complex shapes, like the human body or hands, which are hard to describe with simple geometric shapes. In that case tracking methods based on silhouette could provide the required flexibility. Two different silhouette methods can be distinguished, namely shape matching and contour tracking. With shape matching the detected edges are compared to a reference contour to check for similarities (in the same way as kernel tracking). The contour tracking method on the other hand evolves an initial contour to its new position in the current frame similar to the (advanced) segmentation methods as described in paragraph 2.3.3.3. The most important advantage of tracking silhouettes is their flexibility to handle a large variety of object shapes. an extensive review of silhouette tracking methods is given in (Yilmaz, Javed, and Shah 2006).

2.3.5 Data fusion

The term data fusion can be used for many processes in which information or data from different sources is combined. In this study it is however specifically used for the combination of object locations derived from different (partly) overlapping video cameras. The word “partly” is added to exclude situations in which cameras have no overlapping field of view (fov), because this requires a totally different approach (non-overlapping camera tracking deals more with object re-identification and movement prediction) which is not covered in this article. More information about data fusions with non-overlapping fov is given in (Alahi et al. 2010; Wang 2013). Fusion can be done to increase the tracking extend, to improve tracking performance or both. Fusion can be performed in different ways, but most commonly an occupancy map or object angles are used.

2.3.5.1 Occupancy map

In this data fusion method the location (or location probability density) of identified objects are combined in a single occupancy map from which the most suitable object locations can be derived. The derived object location is therefore more like an average between different views. To combine images from different angels it is required to project the images onto a common plane. Often the actual ground plane is used, but this depends on the situation. In the context of GIS such transformation is called geo-referencing and is performed based on control points visible in the video images and on a reference map. For simple situations a linear transformation can be sufficient, but when the area is more complex advanced image transformations are required. Figure 5 (left side) shows the principle of an occupancy map.

2.3.5.2 Object angle

This data fusion method uses object angles from different views (with fixed location) to estimate the object location. For every detected object the angle is calculated in relation to the horizontal camera viewing angle (Kim and Davis 2006). When the object angles from multiple cameras are combined the object location can be calculated in relation to the camera positions. See Figure 5 (right side) for an example. To calculate an object location the object needs to be covered by at least two cameras. To limit (short term) occlusion issues multiple overlapping cameras can be used together with e.g. a multiple hypothesis tracker. A multiple Hypothesis tracker maintains multiple tracking correspondence hypothesis for each object over multiple frames. The final object track is based on the most likely tracking hypothesis over a specific time period. This method can be used to filter for object occlusion or other artefacts occurring in a single frame or short time interval (Yilmaz, Javed, and Shah 2006).

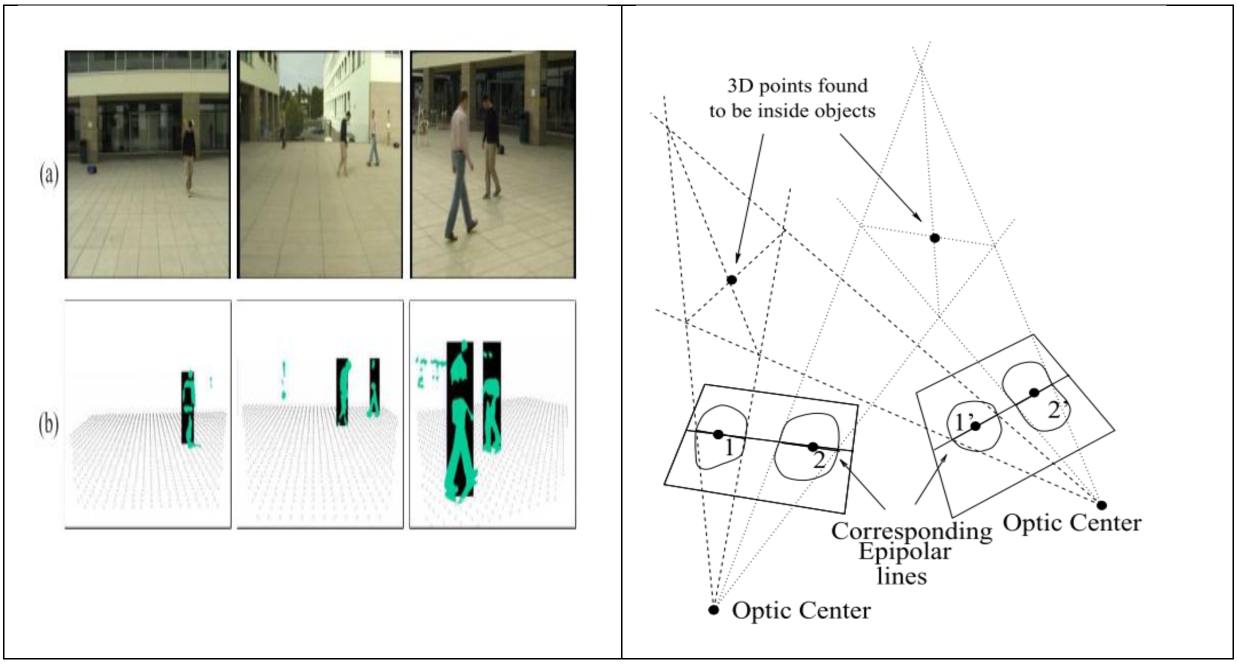

(a)
(b)
3D points found to be inside objects
1
2
1'
2'
Corresponding Epipolar lines
Optic Center
Optic Center

Figure 5 left: principle of an occupancy map (Fleuret et al. 2008). Right: multi-camera location estimation based on object angle (Kim and Davis 2006)

2.3.6 Conclusion

As described above many different approaches exist to perform video tracking. This study focuses on a simple video tracking approach which could easily be extended with a geospatial and multi-camera component. Both the single and multi camera video tracking application will be developed with components for object detection and tracking. The multi camera video tracking application will be equipped with two additionally components for data fusion and visualisation purposes. Because people often do look and dress very different from each other, they don’t have a unique visual feature which can be used for detection. To avoid the need for complex and extensive reference libraries it was therefore decided to use a background subtraction technique for object detection combined with simple kernel/blob tracking. For data fusion the occupancy map technique is selected because it is also suitable for areas which are only covered by a single camera and it also works without camera calibration. For visualisation purposes the multi camera video tracking application will be extended with an overview map in which the detected objects will be displayed on top of a map of the area (overview map). The selected methodology is in more detail described in paragraph 3.2. Because video surveillance is an important application field of video tracking, the envisaged application will also be developed with this specific objective in mind.

3 Methodology

To answer the two research question first the materials (paragraph 3.1) used are described. Next the designs of the single and multi-camera application are explained in paragraph 3.2 Paragraph 3.3 describes how the performance of both applications is evaluated in relation to the tracking precision and accuracy. Finally in paragraph 3.4 the expert survey strategy is introduced.

3.1 Materials

Different datasets are available for the performance evaluation of video tracking applications. An extensive overview of datasets is given in (Maggio and Cavallaro 2011; Vezzani, Baltieri, and Cucchiara 2013) The Performance Evaluation of Tracking and Surveillance (PETS) 2009 benchmark dataset was selected because of its easy availability. Besides it has a number of relevant scenarios, and it was recorded at the same time from 8 different angles. For this study scenario 2 of the PETS dataset was selected because it covers a rather simple situation (limited number of people, few interactions). Scenario 2 has a length of 756 frames, and a frame rate of 7 fps. In Figure 6 four example images from the PETS 2009 dataset are shown and Figure 7 gives an overview of the area, including the different camera locations.

For development and testing purpose a 64 bit Windows 7 laptop with Intel i5 M560 processor, 4 GB RAM and a NVIDIA NVS 2100M video card with CUDA support was used. The application was developed in EMGU CV (which is a C# wrapper for OpenCV) in combination with Visual Studio 2013. For image geo-referencing ArcGIS 10.2 was used. The results were evaluated in MAXTRAQ 2.5.2.1

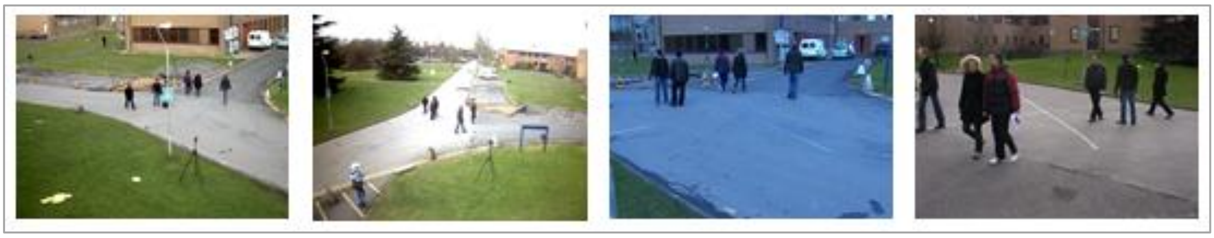


Figure 6 Example images from the PETS 2009 dataset. From left to right view 1, 2, 6 and 8 corresponding to the camera locations shown in figure *7.*

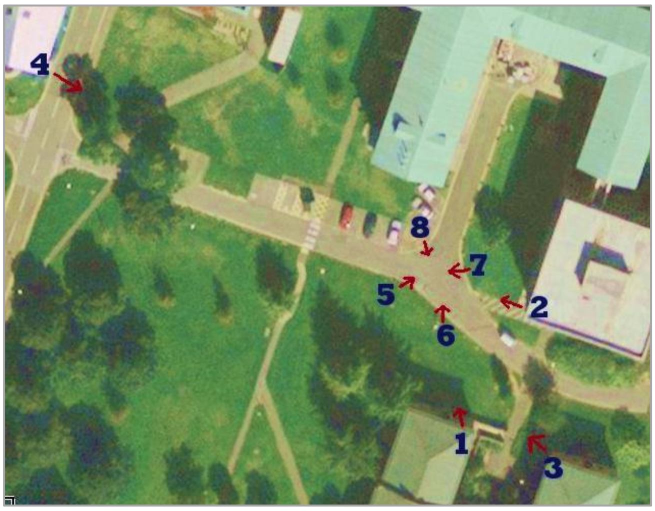


Figure 7 Camera locations from the PETS 2009 dataset

3.2 Design: single and multi-camera video tracking applications

The application development method can best be described as a spiral approach (Centers for Medicare & Medicaid Services (2008). In this approach short cycles are made in which new objectives are set, implemented and evaluated. This approach was selected because it provides high flexibility. Where possible existing techniques were used to limit the development time and reduce risks.

3.2.1 Video tracking components

To evaluate the performance of the multi-camera video tracking application a single camera version is developed based on identical detection and tracking techniques. The multi-camera tracking application is however extended with additional components for data fusion and representation (overview map). In Figure 8 an overview is given of the components used for each tracking application. To combine the images from different cameras with the multi-camera tracking application, it is required to project all images onto a common plane (geo-referencing). Ideally this step would be performed after the ground point estimation step. However no suitable operation was available within the EMGU CV library, therefore it was decided to perform this step in advance of all other operations. To keep the performance evaluation as fair as possible, it was decided to use the projected video(s) for both the multi and single video tracking application.

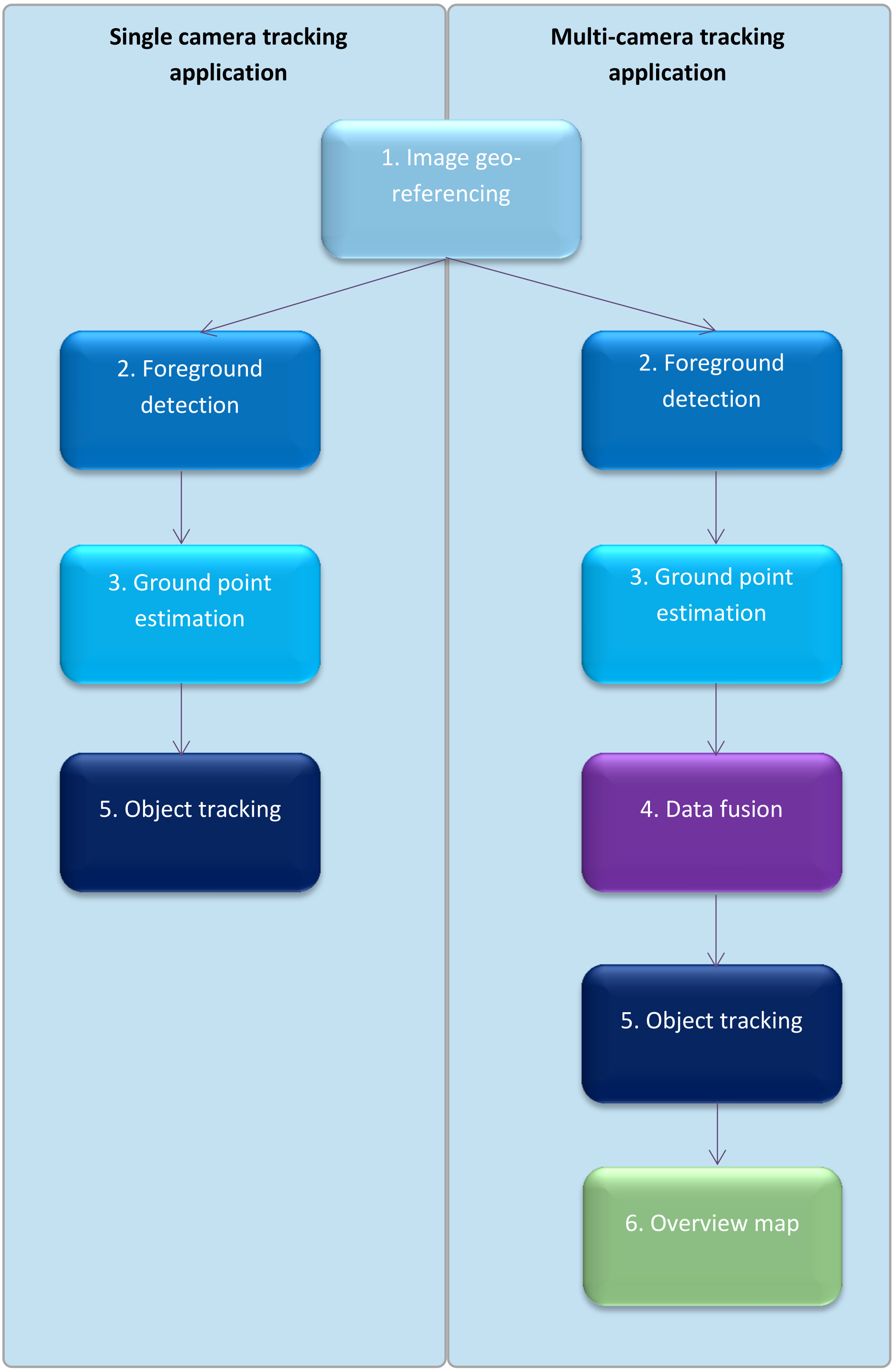


Figure 8 Components of a single- and multi-camera video tracking application. The numbers (1-6) correspond with the sections 3.2.1.1 - 3.2.1.6 which describe the different video tracking components in more detail.

3.2.1.1 Geo-referencing

Geo-referencing is the process of aligning geographic data to a known coordinate system so it can be viewed, queried, and analysed with other geographic data (ESRI n.d.). In general georeferencing consists of two steps. In the first step a number of ground control points are identified, which are locations that can be accurately identified on the object (e.g. an image or raster) and in real-world coordinates or reference map. The connection between one object control point and the corresponding reference control point is called a link. In the second step these links are used to transform the complete object to its spatially correct location. Depending on the image complexity and available reference points different transformation methods are available, ranging from linear/affine to higher order transformations/rubber sheeting. Within the used EMGU CV library only a simple linear image transformation is available. Because this didn't provide satisfactory results (see Appendix II Comparison between affine and spline transformation) it was decided to perform the geo-referencing process within the ARCGIS package in advance of the other steps. As reference map for geo-referencing a satellite image from Google maps was used. The resolution of this image is however rather low in comparison to the camera images and therefore only view 1 was directly referenced on the satellite image. The other camera views were referenced on the projected image from view 1. To increase the number of ground reference point's two different frames were merged together. In this way an image is created which shows more people, and therefore has an increased number of available ground reference points. For the actual image georeferencing process a model was created in ArcGIS 10.2.1. All views were transformed based on a spline transformation. The spline transformation is a true rubber sheeting method and optimizes for local accuracy. It is based on a piecewise polynomial that maintains continuity and smoothness between adjacent polynomials. The Spline technique transforms the source control points exactly to target control points. The pixels in between control points are interpolated (ESRI 2013). A description of the projection model is given in Appendix III ArcGIS model description

3.2.1.2 Background subtraction

Background subtraction (or sometimes called foreground detection) is applied to discriminate between foreground and background pixels by based on a background model. Any pixel which does not fit within the model (range) is classified as foreground. The step implements an algorithm described in Li et al.(2003). To remove noise and to avoid longitudinal segmentation of contours the image is dilated and eroded (both 3 times). In this process the foreground objects are first grown and then shirked at the edges. In this way foreground objects which are located close to each other are merged together and small objects (noise) are removed. To minimise horizontal merging the dilation process is performed with a vertically oriented kernel. This means that a background pixel is assigned foreground if for example a foreground pixel is present three pixels higher or lower. For a horizontal pixel this distance is however limited to 1 pixel.

3.2.1.3 Ground location

When determining object location from a viewing angle not perpendicular on the ground plane distortion can occur. To minimise this distortion object location is estimated based on the position where object and ground plane connect. In case of walking people this often corresponds to the location of the feature. In this step the ground location is estimated for each detected object. For this process the height and centre of the detected object is used.

3.2.1.4 Object tracking

To follow objects over consecutive frames object tracking is applied. Because this study focusses only on people tracking and is performed in combination with background subtraction, kernel/blob tracking is selected. Within EMGU CV different blob tracking techniques are provided but based on performance a simple blob tracking method was selected using the connected components technique. This method tracks coherent areas (detected objects) over consecutive images based on overlap with previous locations.

3.2.1.5 Data fusion

An important step in the multi-camera tracking application is data fusion. In this process a Probabilistic Occupancy Map (POM) is used to combine location data from different views. A POM is a two dimensional view of the probability distribution of an occurrence (in this instance the probability that an object is located at that specific location), and can be shown as a surface model. See Figure 9(right side) for an example. To create a POM the probability for every pixel is calculated according to the probability density function (pdf) shown in equation 1.

$$pdf = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

Where:

$\sigma = standard\ deviation$
$\mu = mean\ pixel\ coordinate(x, y)$
$x = current\ pixel\ coordinate\ (x, y)$

For this formula two input values are definition (µ) and the standard deviation (σ). For the distance between the current and mean location a distance transform raster was used. (See Figure 9 (left side)).The standard deviation is related to the accuracy of the location estimation. This is location specific and amongst other depends on the object- sensor distance and viewing angle. Because this can be a study on itself it was decided to use an overall sdv of 6 pixels, which was experimentally selected based on the best tracking performance and corresponding to the average object width. To merge multiple POMs into a combined POM, the individual POMs are multiplied by an equal weight factor and added together. The weight factor is equal to one divided by the number of combined POM's (1/n). To determine the locations with the highest object probability the pixels with a local maximum are selected (the peeks of the probability curves). This is done by pixel wise evaluation of its neighbourhood.

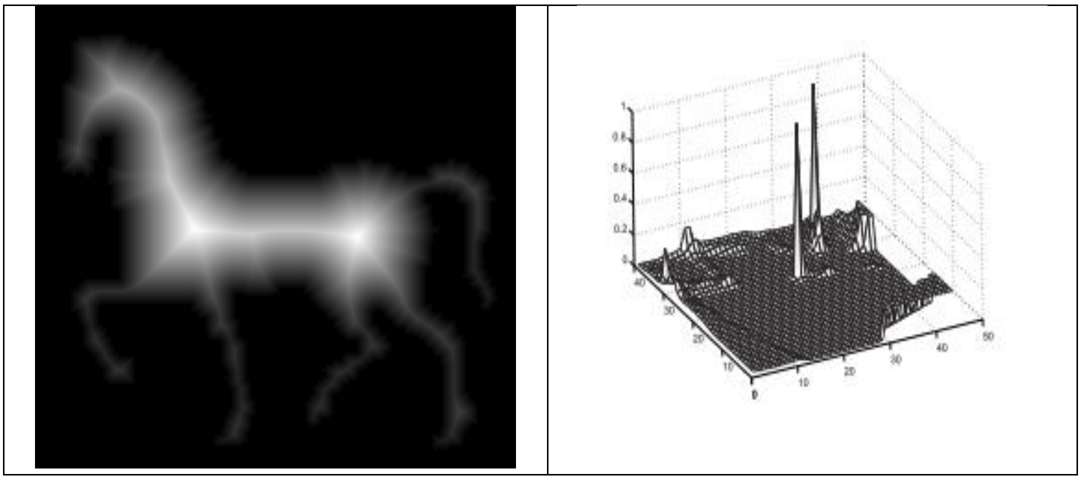


Figure 9 Distance transform (left) and Probability Occupancy map (right)

3.2.1.6 Overview map

The combined object location resulting from the data fusion step can be used to gain an enhanced understanding of the situation. It can be used or visualised in many different ways, varying from real-time mapping, pathway analysis or integrated into a complex decision support system. For demonstration purposes the multi-camera tracking application is extended with a simple geographic viewer were object locations are plotted on an overview map of the location. As base layer for the overview map a blueprint (edge map on blue background) is used derived from a Google satellite image of the area.

3.3 Performance evaluation

To evaluate both tracking applications two tests are performed focussing on tracking precision and accuracy.

3.3.1 Tracking precision

This test will evaluate how well the position of persons is estimated by the single and multi-camera tracking application. The method used for evaluation is based on the Multiple object tracking precision (MOTP) method proposed by Bernardin and Stiefelhagen (2008) with the objective to measure multiple object tracker performance in a more objective and standardised way. As shown in equation 2, MOPT gives the total estimated position error for matched object-hypothesis pairs over all frames, averaged by the total number of matches made. MOPT is an indication of the tracker's ability to estimate precise object positions. To calculate MOPT the number of matching pairs and pair distance is for each frame manually extracted in MAXTRACK, which is a video annotation and analyse program.

$$MOPT = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \qquad (2)$$

$Where:$
$c_t = number\ of\ matching\ pairs\ for\ time\ t$
$d_t^i = distance\ between\ matching\ pair\ (i) for\ time\ t$

3.3.2 Tracking accuracy

To investigate the object tracking accuracy of the developed video tracking applications the multiple object tracking accuracy (MOTA) method is used. This method was proposed by Bernardin and Stiefelhagen (2008) As shown in equation 3, MOTA shows how many mistakes the tracker made in terms of misses (miss), false positives (fp), mismatches (mme) in relation to the total number of actual objects. As shown in equation 4, 5 and 6 The MOTA can be further divided into specific error ratios for miss, fp and mme. To calculate MOTA the number of available objects and tracking errors are manually annotated in each frame. This is done in MAXTRACK and further processed in MS Excel. The definition of the different errors is given in Table 1

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \qquad (3)$$

$$\overline{m} = \frac{\sum_t m_t}{\sum_t g_t} \qquad (4)$$

$$\overline{fp} = \frac{\sum_t fp_t}{\sum_t g_t} \qquad (5)$$

$$\overline{mme} = \frac{\sum_t mme_t}{\sum_t g_t} \qquad (6)$$

Where:

$m_t = Number\ of\ misses\ for\ time\ t$

$fp_t = Number\ of\ false\ positives\ for\ time\ t$

$mme_t = Number\ of\ mismatches\ for\ time\ t$

$g_t = Number\ of\ objects\ for\ time\ t$

Table 1 Error types

Error type	**Definition**
Mismatch (mme)	When existing match is replaced by a new hypothesis*. Example: in frame 10 object 1 has been given track no 1. A mismatch occurs when in the next frame (11) another track number is assigned to the same object.
False positive (fp)	Remaining hypothesis* (no suitable object is available (anymore)) Example: all available objects have a tracker assigned, but there is still a tracker left. This can occur when for example a shadow or three is detected as an object.
Miss(m)	Remaining object (no suitable hypothesis* is available (anymore)) Example: three objects are available in a frame, but only two are detected and tracked.

* A hypothesis is a potential object identified by the tracker.

3.4 Expert survey on potential use cases for geospatial video tracking to enhance SA

In the previous paragraphs the method for development and evaluation of a geospatial video tracking application is described. This application will mainly be developed for evaluation and demonstration purposes. For the implementation in real life situations significant additional work is required. However each individual situation has its own specific requirements. Depending of these requirements the implementation of a geospatial video tracking application will be more feasible than others. To identify currents gaps in technology a survey will therefore be performed focussing on the potential use cases and their specific requirements. This survey will be performed with a number of experts in the sectors security & law enforcement, military and consumer behaviour. These sectors were selected because they are often mentioned in studies related to video tracking and SA. The names of the participants are left out for privacy purposes. Appendix I shows the survey which is used as basis for these interviews.

4 Single and multi-camera video tracking applications

In this study a single and multi camera tracking application were developed. Because the multi camera version used two additional cameras, a number of additional components are added to visualise the integration process and resulting overview. In the following paragraphs both interfaces are described in more detail. On the attached CD two movies are included showing a demonstration of the single and multi-camera tracking application.

4.1 Single camera video tracking application

In Figure 10 the interface for the single camera video tracking application is shown. The left screen shows the projected input video including detected objects. On the right a control video (from a different camera) is shown. In this control image the ground locations of the detected objects from the video screen on the left are shown. The right screen is added only for evaluation purposes. In both screens also the object id for each detected object is plotted in white.

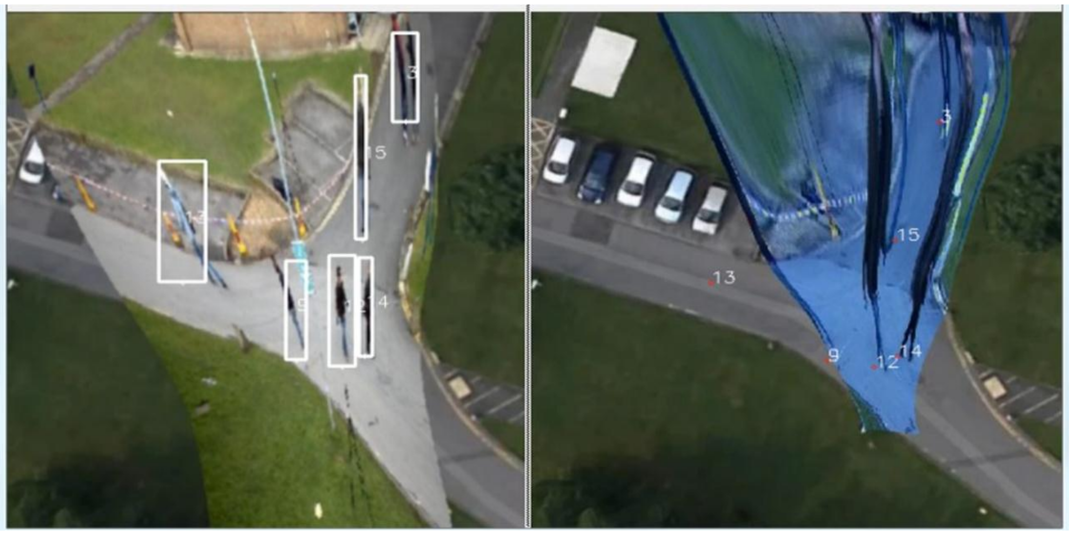


Figure 10 interface of the single camera video tracking application

4.2 Multi-camera video tracking application

Figure 11 shows the interface the multi-camera video tracking application. The interface for the multi-camera video tracking application has a number of additional components compared to the single camera version. In Table 2 each component is described.

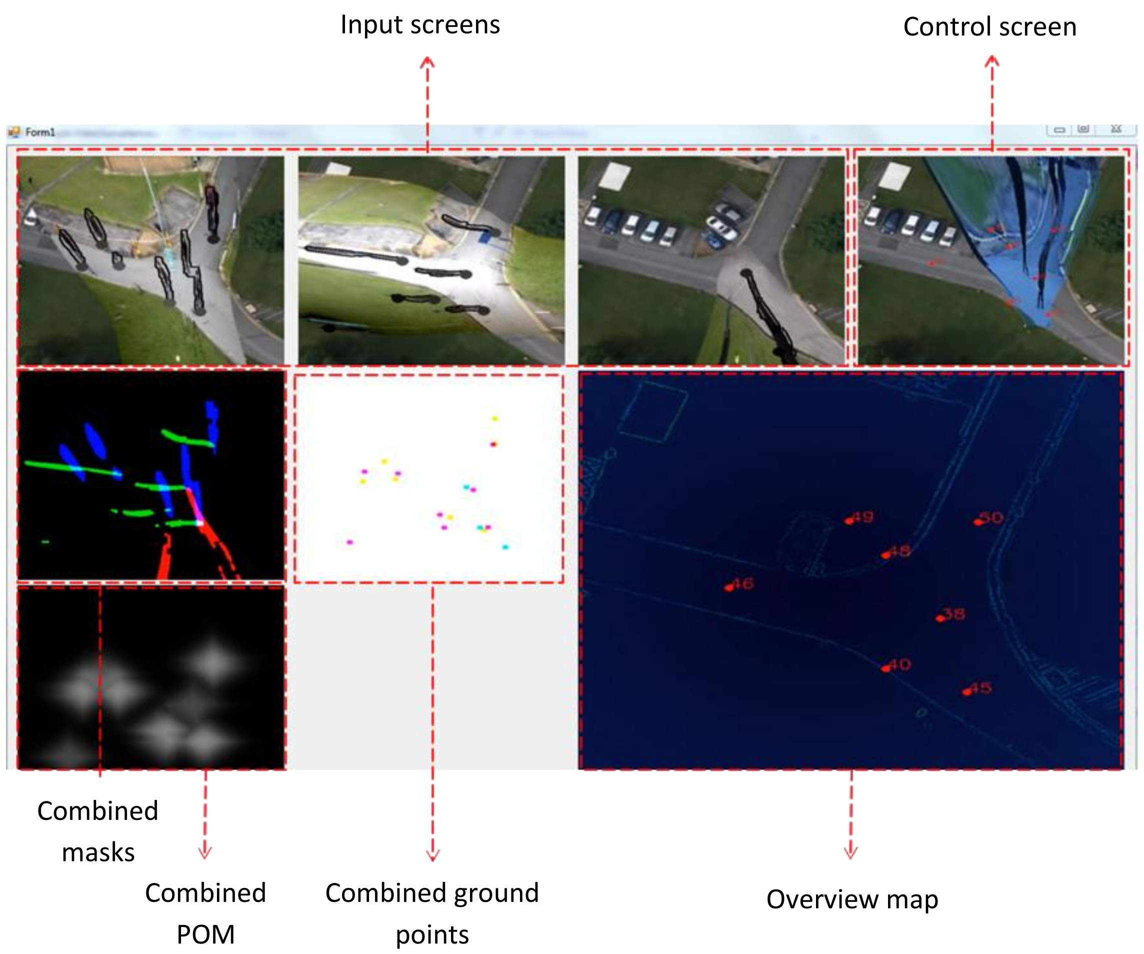


Figure 11 interface of the multi-camera video tracking application

Table 2 Components of the multi-camera video tracking application

Component	**Description**
Input screens	Shows the three input videos including detected objects
Control screen	Shows the ground location of the detected objects from the input videos (plotted on a different camera view)
Combined masks	The masks detected in each of the input videos are combined in a single overview screen. Each input screen is represented by a different colour.
Combined ground points	The object location derived from the different views is combined into a single overview screen. Each input screen is represented by a different colour.
Combined POM	From the derived object ground points a combined Probability Occupancy Map is created. This is used to determine the most likely object locations.
Overview map	Shows the tracked objects plotted on a map.

5 Tracking performance single and multi-camera tracking applications

In line with the methodology described in paragraph 3.3 the tracking performance for both the single and multi-camera video tracking applications are measured. In the following paragraphs the tracking precision (5.1) and accuracy are given (5.2).

5.1 Tracking precision

MOPT is an indication of the tracker's ability to estimate precise object positions. It gives the total estimated position error for matched object-hypothesis pairs over all frames averaged by the total number of matches made. A low MOPT indicate high tracking precision. Table 3 shows the overall MOTP for both the single and multi-camera tracking application. when comparing the MOTP for the single and multi camera tracking applications a slight increase (8.5%) in accuracy is visible for the multi camera tracking application compared to the single camera version. Figure 12 shows the mean tracking precision for each frame for both the single and multi-camera tracking application. In this figure the average distance between tracker and actual object location for each frame is shown. No clear similarities are visible between the single and multi camera video tracking applications. Also the precision for both applications varies considerably over time.

Table 3 Overall MOTP for single and multi-camera tracking application

System	total Distance	No. of Matches	MOTP
SINGLE	13153	1068	12,3
MULTI	14173	1259	11,3

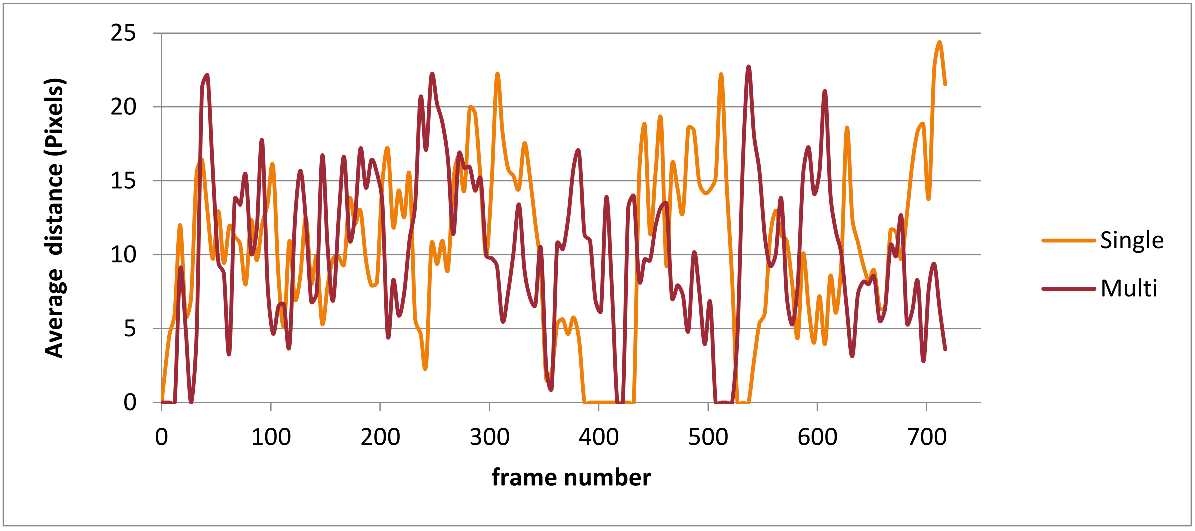


Figure 12 Mean tracker precision

5.2 Tracking accuracy

MOTA indicates how many mistakes a tracker made in terms of misses, false positives, mismatches in relation to the total number of actual objects. In Table 4 the number of matches, miss-, fp-, mme ratio and MOTA are given for both the single and multi-camera tracking application. When comparing the MOTA for the single and multi camera video tracking application the multi camera version performs considerably lower (65%) on accuracy in comparison with the single camera video tracking application. This reduction is mainly caused the high number of Fp´s for the multi camera video tracking application. The miss ratio is lower for the multi camera video tracking application in comparison the single camera version. For both versions the mme ration is relatively small (few mismatch errors). In Figure 13 the number of correctly detected objects per tracking application in relation to the number of actual available objects (ground truth) per frame is given. The difference between detected and available objects from this figure is similar to the MOTA. In Figure 14 and Figure 15 the stacked error distribution (miss-, fp-, and mme ratio) is shown for respectively the single and multi-camera tracking application. When comparing the different errors ratios for the single and multi camera versions a clear increase in fp errors for the multi camera video tracking application is visible. Additionally the different distributions are very variable and no clear relation between the single and multi camera versions is visible.

Table 4 Number of matches, miss-, fp-, mme ratio and MOTA

system	No of Matches	miss ratio	Fp ratio	mme ratio	MOTA
SINGLE	1068	0,586	0,227	0,021	0,167
MULTI	1259	0,518	0,396	0,028	0,058

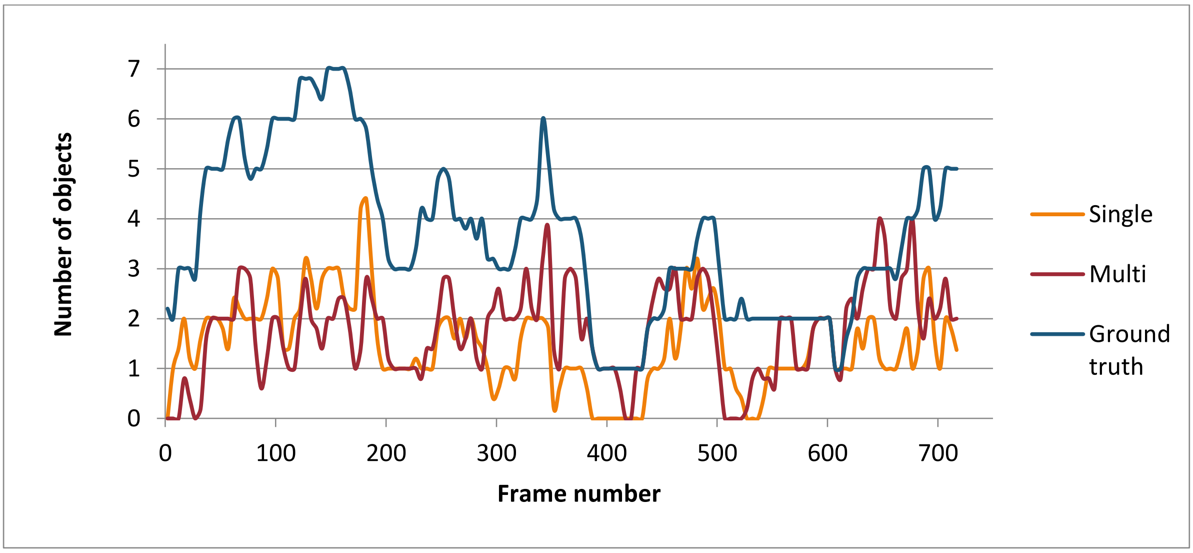


Figure 13 Number of actual (ground truth) and correctly detected objects for both the multi and single camera video tracking application

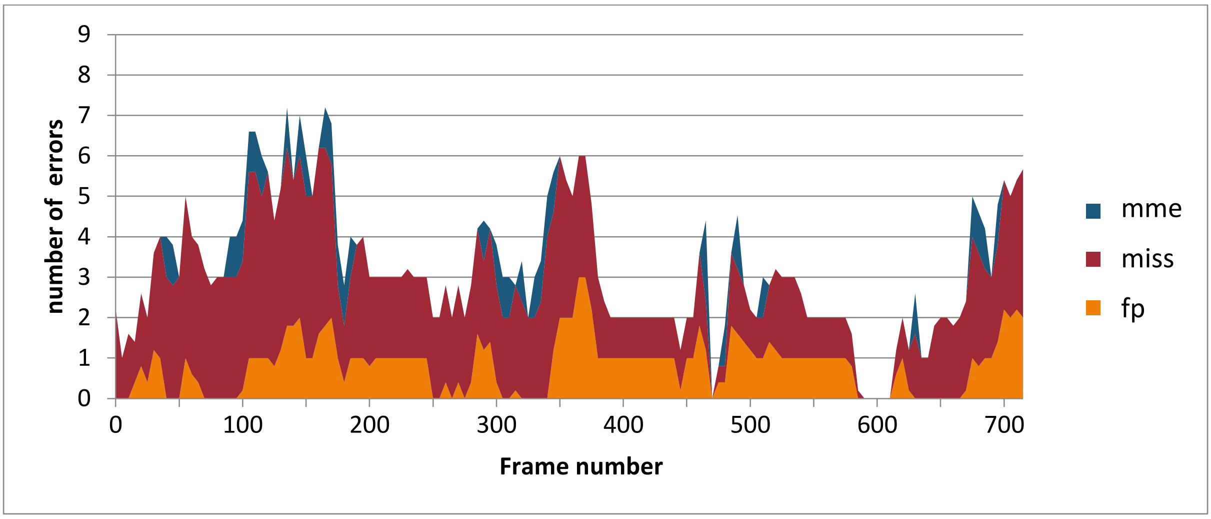


Figure 14 Stacked error distribution for single camera video tracking

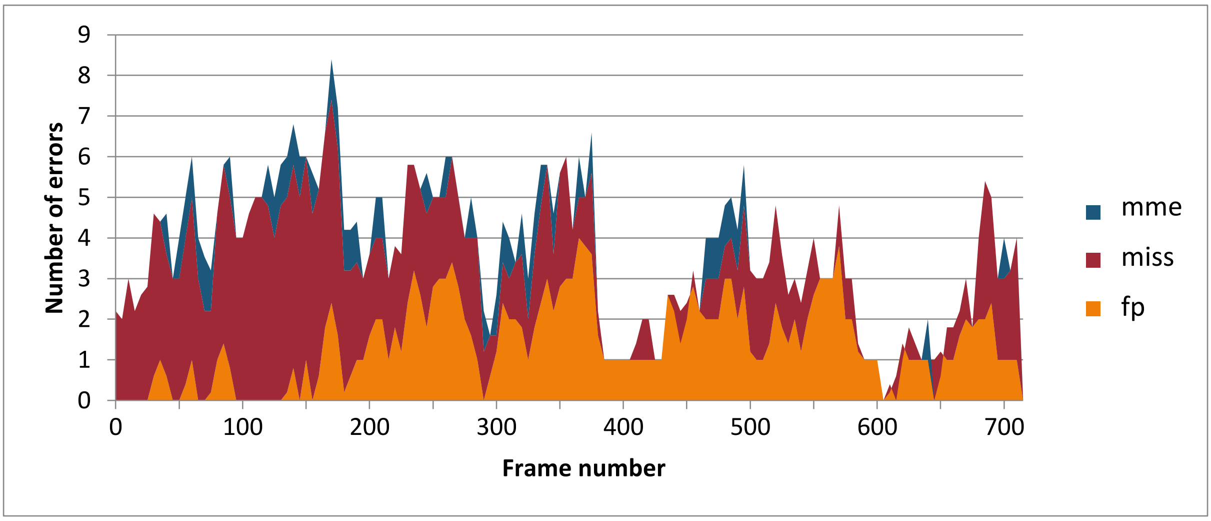


Figure 15 Stacked error distribution for multi-camera video tracking

6 Potential use cases for geospatial video tracking to enhance SA

To investigate the potential of a multi-camera tracking application for improving situational awareness, a survey was performed focussing on experts within the field of security & law enforcement, military and consumer behaviour. In the following paragraphs an overview is given of the most important remarks per sector.

6.1 Security and law enforcement

To develop more insight in the potential of geospatial tracking within the security and law enforcement sector personal interviews have been conducted with two experts. One of them is involved as a commercial ICT consultant in the safety and justice sector and in this position often closely involved with the development and implementation of new software solutions for the police and other private security companies. The second expert is a commissioner at The Hague police department and is involved with the coordination of a number of different teams within the police department, including traffic, SWAT and surveillance and protection teams.

Possibilities for video surveillance within the police force are constantly increasing following the continues improvement in camera technology. Currently further developments are however restricted by limited manual capacity (to analyse and store all recorded data). Also the link between the police and the justice department is becoming stronger, which means that more detailed and compelling evidence is expected from the police so that offenders can be convicted faster and more successfully. The potential of geospatial video tracking lays therefore in the possibility to aid in faster detection, increase understanding and by providing more extensive and reliable evidence. Additionally geospatial video tracking could assist with data reduction by automatic information filtering.

Specific applications vary from surveillance in public spaces (like shopping centres) for the detection of suspicious behaviour based on historical data, to crowd management during big events. Because gathered information should be usable as evidence in court an important factor is the reliability and accuracy of such a system. In this context geospatial video tracking can play an important part by providing a constant and solid object trajectory or corridor of the suspect. For enhanced functionality the geospatial video tracking application could be extended with the capacity for easy event replay, object recognition capabilities, and include behaviour detection or other profiling techniques to identify potential suspects.

6.2 Military services

In the military sector three personal interviews were conducted. The first interview has been performed with a senior Geo-ICT consultant. During his military service he worked as fire support officer at the Dutch army and still holds the function as reserve officer artillery. From his experience as Artillery officer he clearly emphasizes the importance of situational awareness during military operations. Specifically knowing when and where allied and opposite forces are located. An overview map showing the current situation is therefore very important, because this is still the preferred source people turn to when the shit hits the fan and decisions have to be made. Current information systems can monitor real time location of allied forces but lack the capacity to show real-time information and location of the enemy.

The second interview has been performed with a researcher in the field of ethics and military technology, with a specific focus on the current developments in the battlefield. Important findings of the research is that the use of such communication technology can lead to tunnel vision which can distract the user from the bigger picture, which is especially dangerous higher up the military hierarchy. Additionally technology can have a hidden bias towards a certain outcome.

The third interview has been performed with a Lieutenant-Colonel from the Dutch military, who involved as a researcher with the Land Warfare Centre and currently also enrolled in the master program military strategic studies at the Royal Military Academy. As researcher at the Land and Warfare Centre he is engaged with preparing the army for the battlefield of the future. For the coming years he sees a number of important changes. Because in the future more and more people will live in urban areas, (reaching 70% in 2050), the battlefield is likely to move towards these areas, probably the mega cities of the future. Additional the development will continue from large scale battles between huge armies towards small, flexible and all-round combat units, working together in multinational operations.

To make effective use of available resources including intelligence and troops, specifically in fast changing environments, a network centric approach is required. With such approach information is not only shared hierarchically, but also horizontally between cooperating teams to stimulate timely distribution of relevant data and to synchronize actions. However rather than just an information source such a system should actively provide relevant information to the user. A geographical interface is therefore an important aspect of such a system, not only for visualisation purposes, but also for filtering the information and placing it in its context. With the continuous developments of sensor technology and the change towards urban warfare sensors will play an increasingly important role. Not only because technological developments will result in smaller and cheaper sensors, but also because in future battlefields sensors systems will be already available. The key however lies in gaining control over these assets. Also the number different sensors will only increase in the future, ranging from UAV’s and personal cameras to acoustic localization- and micro sensors (which can be used to form autonomous monitoring networks). To be widely deployable for military operations the system should not only be real time and robust, but also modular (flexible/scalable based on different components which can be combined depending on the situation) and should clearly indicate the reliability of the shown information. For increased functionality the application could be extended with pattern detection and object recognition techniques.

6.3 Consumer behaviour

In the sector consumer behaviour one interview has been performed with a researcher involved with the Restaurant of the Future, which is a facility for close observation of eating and drinking behaviour. To monitor consuming behaviour the restaurant is equipped with 23 video cameras recording the food selection, movement and eating process of the participants. According the interviewed expert geospatial video tracking could contribute to an increased understanding of consumer behaviour. Multiple use cases can be thought of, but the most evident case is related to consumer movement and product interaction. For example in a restaurant or shop it would be interesting to know how long somebody is looking at a product, if he takes it, or comes back another time. This is can be a robust analysis tool especially when it can be extended or combined with gaze tracking and object recognition. Another possible advantage of such a system is that in many countries it is not allowed to store surveillance footage due to of privacy issues, but for derived products like hot zones or object trajectories such restrictions don’t exist.

7 Conclusion, discussion and recommendations

In the previous chapters the concept of geospatial video tracking is introduced together with the method for performance evaluation and an expert survey on potential use cases. Additionally in the chapters 4, 5, and 6 the results of this study are described. In the following paragraphs the followed approach will be discussed combined with a number of concluding remarks and recommendations.

7.1 Conclusion

This study arose from the notion that even in this age where sensors are abundantly available, situational awareness is far from certain. In this paper the hypothesis was therefore posed that without integration of individual cameras within a common plane, situational awareness is limited, resulting in sub optimal decision making. In line with this statement the objective of this study is to investigate the potential of a geospatial multi-camera tracking application to enhance situational awareness. To reach this objective two research questions were formulated focusing on the performance and potential use cases of geospatial video tracking.

The result from the performance evaluation clearly shows considerable performance differences between the single and multi camera video tracking applications. Besides an expected reduction in frame processing rate, the developed multi-camera application scores 65% lower on accuracy than the standard single camera tracking application. The tracking precision of the multi-camera video tracking application is 8.5% higher than the single video tracking application. In both applications the false positives (fp) cause the most errors, closely followed by the number of misses (miss). For both performance indicators no clear relation is visible between the single and multi-camera video tracking application. To identify currents gaps in technology a survey was performed with experts in the sectors security & law enforcement, military and consumer behaviour focussing on potential use cases and their specific requirements. The performed expert survey clearly indicates numerous SA enhancing applications in which geospatial video tracking could contribute considerably. Many of the mentioned applications however demand a high level of reliability, accuracy and speed, which are currently not yet available. Overall it can be concluded that geospatial video tracking has a significant potential for enhancing SA, but needs serious improvements before it can be used in practice.

7.2 Discussion and recommendations

Based on the sheer volume of recent research articles related to video tracking it can safely be said that video tracking is quite of interest within the science community. This has resulted in many intelligent approaches to tackle challenges related to video tracking, with each their specific advantages and limitations. This study focuses on the potential for integrating a geospatial component within the video tracking domain, and not on the development of an (new) video tracking method itself. To keep the performance evaluation of the multi-camera tracking application as objective as possible, it was decided to develop two applications. One single- and one multi-camera video tracking application; making use of identical components. This has resulted in a single camera video tracking application based on foreground subtraction and simple blob tracking. For the multi-camera tracking application two additional components were added for data fusion and visualisation purposes. Additionally it was decided to project the used camera footage for both applications in advance of the actual video tracking. As with any tracking method, the selected one has its limitations. Many of these are in detail described in literature, like the issues with background subtraction for the identification of individual objects located close to each other (Maggio and Cavallaro 2011). However also a number of issues are related to the specific design of this study. So has the decision to geo-reference the used video footage in advance of the actual tracking steps a considerable negative effect on the tracking performance. The selected geo-referencing method transforms the original video images onto a reference map. With this process the image quality is reduced, and because of the applied transformation objects are often harder to recognise and detect. When image projection would be applied within the tracking process this issue wouldn't occur, because then the detection process would be performed in advance of the projection step. However during the realization of this study no suitable function was available within the selected image processing library to perform real time higher order image projection. Linear projection methods were available, but didn't provide sufficient accuracy (as shown in Appendix II Comparison between affine and spline transformation). Secondly for the POM calculation a standard deviation of 6 pixels is assumed. In reality this value varies per location based on object and camera position. The incorporation of a dynamic POM model as used in Fleuret et al. (2008) could however significantly increase the performance of the object locations estimation process.

For the performance evaluation of both video tracking applications the evaluation criteria proposed by Bernardin and Stiefelhagen (2008) were used. In their article they introduce two intuitive and general metrics to allow comparison of tracker characteristics, focusing on their precision in estimating object locations and the ability to consistently label objects over time. As described in chapter 3 this method results in two indicators, namely the MOPT and MOTA. Both the number of matches (required for calculating the MOTP) and the number of misses (required for calculating the MOTA) depend on a manually selected value for the maximum allowable distance between actual and estimated object location. Although Bernardin and Stiefelhagen (2008) suggest using a logical maximum distance like the average width of an object, they aknowledge the limitations of their approach for comparing different case studies because of the arbitrary nature of this value. In our study we used a maximum distance value of 25 pixels, which is significantly more than the suggested mean object width (approx. 8 pixels). This higher value was selected so more matches would be available, resulting in a more accurate estimation of the mean tracker precision. The selected distance value can howeverconsiderably influence the resulting MOPT and MOTA values, limiting its effective use. Therefore it was decided to compare the results of the multi-camera tracking application only with the single camera tracking application, and not with results from other studies.

The performance of the multi and of the single video tracking applications are measured based on their precision and accuracy. Overall the tracking precision of the multi-camera video tracking application is 8,5% higher than the single video tracking application. A greater increase in precision was however expected since the amount of available video data was more than double for the multi-camera video tracking application. When comparing the overall tracker accuracy for both applications the multi-camera application scores 65% lower than the standard single camera tracking application. The performance difference is mainly caused by the fp ratio, which is almost twice as high for the multi-camera tracking application. A number of possible explanations can be identified for the reduction in tracking accuracy. Firstly because information from different views is used, also the noise from different views is combined, resulting in an increased number of detection mistakes causing a rise in the number of false positives. Secondly during the fusion process the detected objects are reduced from an object contour to a single point on the map. Therefore the object is harder to track, resulting in an increased number of mismatch errors. An interesting alternative way which would prevent the accumulation of noise related errors would be to use pathway matching instead of matching objects location for each single instance. With this method detected object are tracked in each view and subsequently the extracted objects pathways from each view are matched between different views to track objects over multiple cameras. This method however requires further research and testing.

To identify potential use cases for a multi-camera system a number of personal interviews were performed with experts from the field of security & law enforcement, military services and consumer behaviour. When comparing results from the three surveyed sectors it is clearly visible that although the current sensor technology provides great opportunities, they also bring new challenges. As explained by one of the interviewed experts the introduction of certain new technologies like UAV's, when not managed well can have considerable negative effects on SA. Specifically in high risk situations like a battlefield such a reduction in SA can have tremendous consequences. Therefore it is very important to look at the effects of new technologies and how they effectively can be incorporated and used within the system. The concept of GIS combined with video tracking appealed to many of the interviewed experts, and numerous potential uses cases were mentioned during the survey. Many of them require a high level of reliability, accuracy and speed, limiting the direct usability of Geospatial video tracking. However sectors with lower risk setting, like consumer behaviour could already benefit from a simple geospatial video tracking application. For example in a more generalised way (hot zone detection instead of people tracking) or in situations which don’t require real-time processing. This will not provide (as much) direct situational awareness, but can still contribute to a better understanding of the situation. Additionally this will provide a suitable basis for the development and testing of future geospatial video tracking applications.

References

Alahi, Alexandre, Pierre Vandergheynst, Michel Bierlaire, and Murat Kunt. 2010. “Cascade of Descriptors to Detect and Track Objects across Any Network of Cameras.” *Computer Vision and Image Understanding* 114(6): 624–40. http://linkinghub.elsevier.com/retrieve/pii/S1077314210000275 (April 22, 2014).

Aubrecht, Christoph, Sven Fuchs, and Clemens Neuhold. 2013. “Spatio-Temporal Aspects and Dimensions in Integrated Disaster Risk Management.” *Natural Hazards* 68(3): 1205–16. http://link.springer.com/10.1007/s11069-013-0619-9 (January 20, 2015).

Bernardin, Keni, and Rainer Stiefelhagen. 2008. “Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics.” *EURASIP Journal on Image and Video Processing* 2008: 1–10. http://jivp.eurasipjournals.com/content/2008/1/246309 (July 24, 2014).

Blake, A. 2006. “Visual Tracking: A Short Research Roadmap.” *Handbook of Mathematical Models in Computer Vision*. http://link.springer.com/chapter/10.1007/0-387-28831-7_18 (April 22, 2014).

Bradford, Brian, Eric M Dixon, Joshua Sisskind, and William D Reynolds Jr. 2011. “Target Tracking with GIS Data Using a Fusion-Based Approach.” *Proc. SPIE* 8053: 80530F – 80530F – 11. http://dx.doi.org/10.1117/12.883424.

Centers for Medicare & Medicaid Services. 2008. “Selecting a Development Approach.” *Centers for Medicare & Medicaid Services*: 1–10. http://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/XLC/Downloads/SelectingDevelopmentApproach.pdf.

Elgammal, a, R Duraiswami, D Harwood, and L S Davis. 2002. “Background and Foreground Modeling Using Nonparametric Kernel Density Estimation for Visual Surveillance.” *Proceedings of the IEEE* 90: 1151–63. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1032799.

ESRI. 2008. “Geographic Information Systems Providing the Platform for Comprehensive Emergency Management.” (October).

———. 2013. “Fundamentals for Georeferencing a Raster Dataset.” http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#/Fundamentals_for_georeferencing_a_raster_dataset/009t000000mn000000/.

———. “GIS Dictionary.” : 1. http://support.esri.com/en/knowledgebase/GISDictionary/term/georeferencing (April 14, 2015).

Fleuret, François, Jérôme Berclaz, Richard Lengagne, and Pascal Fua. 2008. “Multicamera People Tracking with a Probabilistic Occupancy Map.” *IEEE transactions on pattern analysis and machine intelligence* 30(2): 267–82. http://www.ncbi.nlm.nih.gov/pubmed/18084058.

Goodchild, MF. 2005. “Geo-Information Science for Disaster Management. Keynote at the First International Symposium on Geo-Information for Disaster Management, Delft.”

Grewe, L, and A C Kak. 1995. “Interactive Learning of a Multiple-Attribute Hash Table Classifier for Fast Object Recognition.” *Computer Vision and Image Understanding* 61: 387–416.

Jiang, Jiping et al. 2012. “A GIS-Based Generic Real-Time Risk Assessment Framework and Decision Tools for Chemical Spills in the River Basin.” *Journal of hazardous materials* 227-228: 280–91. http://www.ncbi.nlm.nih.gov/pubmed/22664261 (September 5, 2014).

Kim, Kyungnam, and LS Davis. 2006. “Multi-Camera Tracking and Segmentation of Occluded People on Ground Plane Using Search-Guided Particle Filtering.” *Computer Vision–ECCV 2006*: 98–109. http://link.springer.com/chapter/10.1007/11744078_8 (September 24, 2014).

Lee, L, R Romano, and G Stein. 2000. “Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame.” *Pattern Analysis and Machine …* (1655). http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=868678 (January 28, 2015).

Li, Liyuan, Weimin Huang, Irene Y. H. Gu, and Qi Tian. 2003. “Foreground Object Detection from Videos Containing Complex Background.” *Proceedings of the eleventh ACM international conference on Multimedia - MULTIMEDIA ’03*: 2. http://portal.acm.org/citation.cfm?doid=957013.957017.

Li, Liyuan, and Maylor K H Leung. 2002. “Integrating Intensity and Texture Differences for Robust Change Detection.” *IEEE Transactions on Image Processing* 11: 105–12.

Luokkala, Pekka, and Kirsi Virrantaus. 2014. “Developing Information Systems to Support Situational Awareness and Interaction in Time-Pressuring Crisis Situations.” *Safety Science* 63: 191–203. http://linkinghub.elsevier.com/retrieve/pii/S0925753513002725 (October 3, 2014).

Maggio, Emilio, and Andrea Cavallaro. 2011. *Video Tracking: Theory and Practice*. http://books.google.com/books?hl=en&lr=&id=v8g6_N1E-tYC&oi=fnd&pg=PT5&dq=VIDEO+TRACKING+THEORY+AND+PRACTICE&ots=F_gDIxRHA4&sig=PatW6wwYtVhi-NwFPioUp2iMg6I (April 22, 2014).

Mikolajczyk, Krystian, and Cordelia Schmid. 2005. “Performance Evaluation of Local Descriptors.” *IEEE transactions on pattern analysis and machine intelligence* 27(10): 1615–30. http://www.ncbi.nlm.nih.gov/pubmed/16237996.

Olson, Eric. 2013. “[Feature] Port Security Goes Geospatial.” : 1. http://www.maritime-executive.com/article/Feature-Port-Security-Goes-Geospatial-2013-09-17/ (May 20, 2014).

Panin, Giorgio. 2011. *MODEL-BASED VISUAL TRACKING The OpenTL Framework*. Hoboken, NJ, USA: John Wiley & Sons, Inc. http://doi.wiley.com/10.1002/9780470943922.

Papageorgiou, C P, M Oren, and T Poggio. 1998. “A General Framework for Object Detection.” In *Computer Vision, 1998. Sixth International Conference on*, , 555–62.

Park, Sangho, and MM Trivedi. 2006. “Multi-Perspective Video Analysis of Persons and Vehicles for Enhanced Situational Awareness.” *Intelligence and Security Informatics*: 440–51. http://link.springer.com/chapter/10.1007/11760146_39 (May 23, 2014).

Rittscher, J, J Kato, S Joga, and A Blake. 2000. "A Probabilistic Background Model for Tracking." *Computer Vision—ECCV 2000*: 336–50. http://link.springer.com/chapter/10.1007/3-540-45053-X_22.

Rosten, Edward, and Tom Drummond. 2006. "Machine Learning for High-Speed Corner Detection." *Computer Vision–ECCV 2006*: 1–14. http://link.springer.com/chapter/10.1007/11744023_34 (January 28, 2015).

Rowley, H a, S Baluja, and T Kanade. 1998. "Neural Network-Based Face Detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20: 23–38. http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=655647.

Snidaro, Lauro, Ingrid Visentini, and GL Foresti. 2011. "Data Fusion in Modern Surveillance." *Innovations in Defence Support Systems …*: 1–21. http://link.springer.com/chapter/10.1007/978-3-642-18278-5_1 (January 7, 2015).

Vezzani, Roberto, Davide Baltieri, and Rita Cucchiara. 2013. "People Reidentification in Surveillance and Forensics: A Survey." *ACM Computing Surveys (CSUR)* 1(1). http://dl.acm.org/citation.cfm?id=2543596 (April 22, 2014).

Viola, Paul, Michael J. Jones, and Daniel Snow. 2005. "Detecting Pedestrians Using Patterns of Motion and Appearance." *International Journal of Computer Vision* 63: 153–61.

Wang, Xiaogang. 2013. "Intelligent Multi-Camera Video Surveillance: A Review." *Pattern Recognition Letters* 34(1): 3–19. http://linkinghub.elsevier.com/retrieve/pii/S016786551200219X (March 28, 2014).

Wikipedia. 2014. "Wikipedia." http://en.wikipedia.org/wiki/Situation_awareness (May 21, 2014).

———. 2015. "Wikipedia.org." http://en.wikipedia.org/wiki/Geographic_information_system.

Yilmaz, Alper, Omar Javed, and Mubarak Shah. 2006. "Object Tracking." *ACM Computing Surveys* 38(4): 13 – es. http://portal.acm.org/citation.cfm?doid=1177352.1177355 (April 29, 2014).

Zwahlen, Heather, A Yahr, and D Berven. 2012. "SATURN (Situational Awareness Tool for Urban Responder Networks)." *… (FUSION), 2012 15th …*: 2428–35. http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6290598 (October 9, 2014).

Appendix I Expert survey

Geospatial video tracking is the process of projecting and fusing object location derived from different camera feeds on a common plain to create a single overview map showing object movement combined with other spatial information. This expert survey is performed as part of my master thesis on geospatial video surveillance for increased situational awareness. To gain more insight in the potential and possible challenges of geospatial video surveillance I would like to ask you the following questions:

1. Can you give a short overview of your professional career, and shortly summarize your professional experiences (and your role) related video tracking, network enabled decision making or virtual situational awareness etc.

   ……

   ……

   ……

   ……

2. Did you ever work with or heard about geospatial video tracking or a similar concept? Please explain.

   ……

   ……

   ……

   ……

3. What are possible cases within your field of expertise in which geospatial video tracking could contribute, and what is the main advantage of the incorporation of such geospatial component.

   ……

   ……

   ……

   ……

4. What are important requirements of the geospatial application mentioned in question 3. If more than one situation is given, please define per case.

Characteristics	**Situation 1**	**Situation 2**	**Situation 3**	**Situation 4**
Real-time				

High accuracy				
………………………………..				
………………………………..				

5. Which additional functionality could give such application additional functionality?

Characteristics	**Situation 1**	**Situation 2**	**Situation 3**	**Situation 4**
Behaviour detection				
Pathway analysis				
Object detection				
Data mining				
………………………………..				
………………………………..				

6. For the cases mentioned in answer 3, is an overview map itself sufficient, or do the original camera images still play an important role for understanding the situation.

……

……

……

……

7. In general: how many cameras/ sensors a video tracking system should contain to significantly benefit from the incorporation of a geospatial component.

……

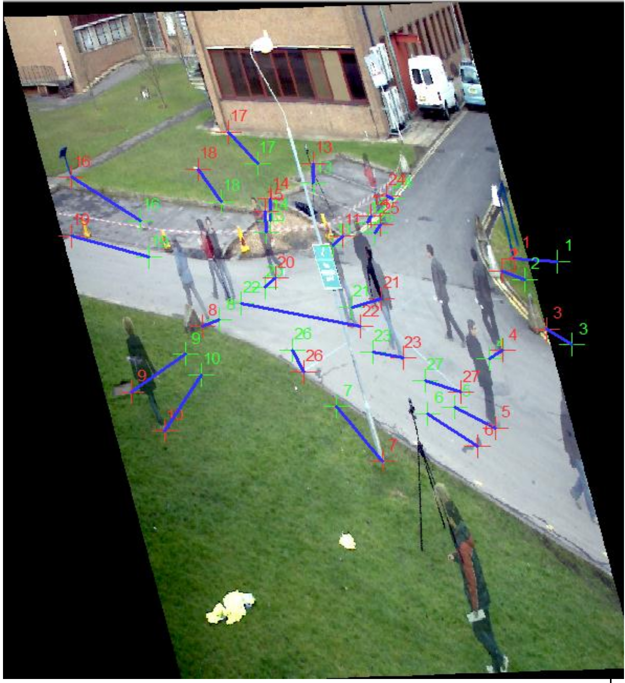
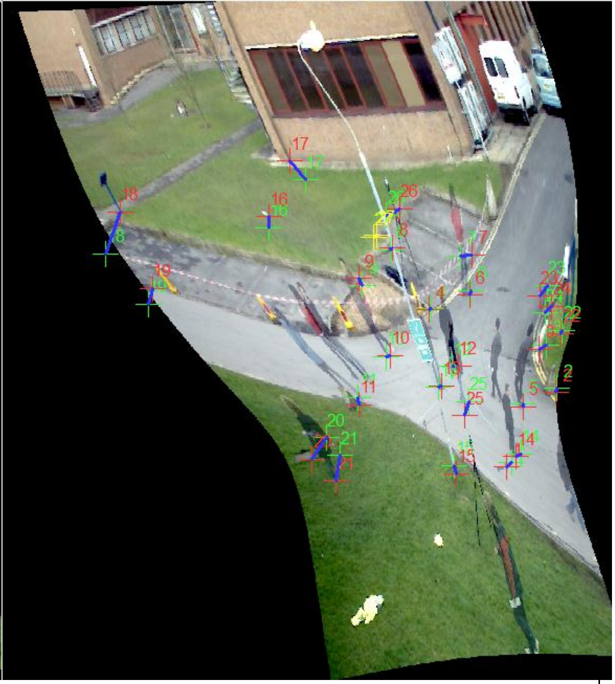
……

……

……

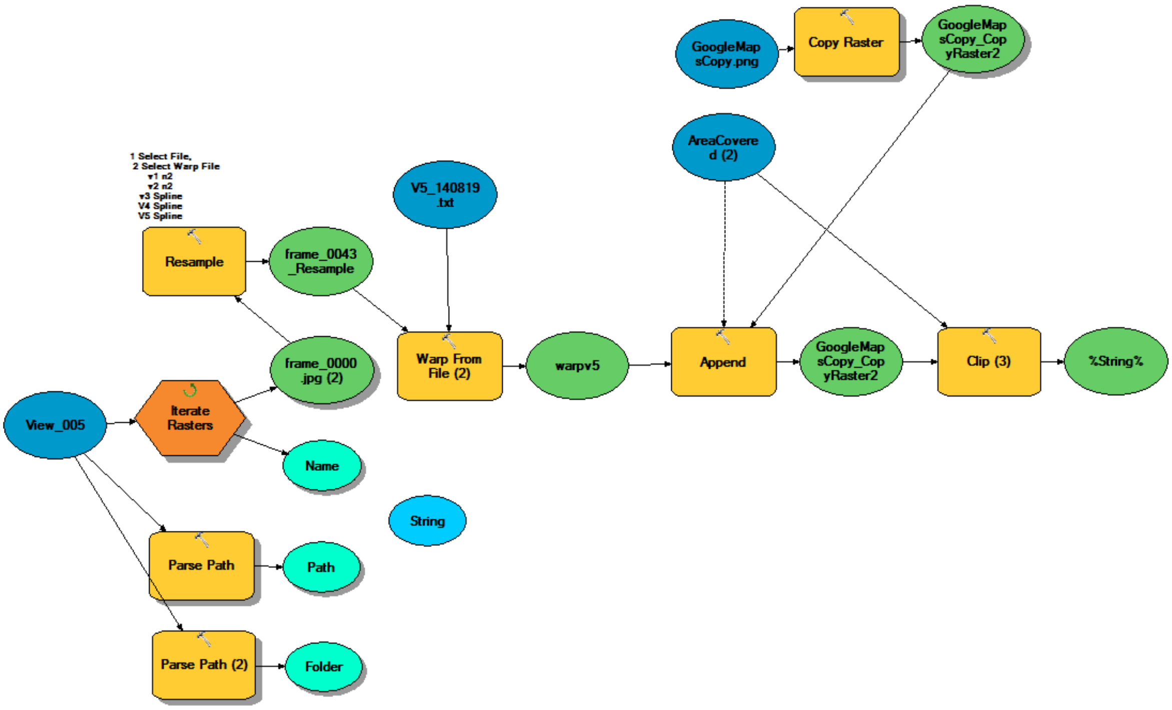
Appendix II Comparison between affine and spline transformation

Comparrison between affine and spline transformation. The bleu lines demonstrate the error at the controll points.the overall RMSE for both transformations is shown below. The RMSE is aproximately three times smaller for the spline in comparisson with the affine transformation. The unit of the RMSE is degrees (based on the WGS 84 projection)

Affine transformation	**Spline transformation**
	
RMSE: 3,32637e-005	RMSE: 1,13493e-005

Appendix III ArcGIS model description

With this model the image sequence from a specific view can be automatically projected onto the reference image (a satellite image from Google maps) based on a predefined projection file. Below an overview of the model is included (for view 5)



Appendix IV Content attached CD

Data

- PETS 2009 dataset scenario S2.L1
- Google maps reference image

Application

- Single camera video tracking application
- Multi-camera video tracking application
- ArcGIS toolbox for automated image projection

Results

- Report (word +pdf)
- Midterm presentation
- Final presentation
- Projected images
- Excel table with performance results
- Demo video single camera video tracking
- Demo video multi-camera video tracking

Relevant literature