Information based multimodal Background Subtraction for Traffic Monitoring Applications

Master's Thesis

Thiemo Alldieck

Aalborg University
School of Information and Communication Technology
Selma Lagerløfsvej 300
DK-9220 Aalborg

AALBORG UNIVERSITY
Master's Thesis

School of Information and Communication Technology
Selma Lagerløfsvej 300
DK-9220 Aalborg
http://sict.aau.dk

Title:
Information based multimodal Background Subtraction for Traffic Monitoring Applications

Theme:
Master's Thesis

Project Period:
Fall 2014 - Summer 2015

Project Group:
15gr1085

Participant(s):
Thiemo Alldieck

Supervisor(s):
Thomas B. Moeslund
Chris Bahnsen

Copies: 3

Page Numbers: 102

Date of Completion:
June 1, 2015

Abstract:

This thesis presents a new approach to background subtraction for multimodal systems. The work is an extension to the Gaussian mixture model background subtraction of Stauffer and Grimson. The background conformity values of two image sources, namely thermal and RGB, are fused in order to enable stable background subtraction for persistent surveillance. Image quality heuristics based on image characteristics and external sources are specified to evaluate the *usefulness* of the modalities and perform the fusion *context aware*. Extensions for the use of the system for the purpose of traffic monitoring are presented. Therefor modulations of a new image representation of the conformity of pixels with the background model are made. The potential of the proposed method has been shown during excessive tests of quantitative and qualitative characteristics.

Preface

This Master's thesis is submitted as part of the Vision, Graphics and Interactive Systems MSc program at Aalborg University. The work has been conducted throughout the 9^{th} and 10^{th} semester from autumn 2014 to summer 2015.

The work originates in the research interests of the Department of Civil Engineering at the Aalborg University. The department uses RGB and thermal video data to analyze the safety of traffic intersections in a semi-automatic process. Therefore a high interest in full automatic and persistent analysis software exists.

The thesis is formed in collaboration with the Visual Analysis of People group at Aalborg University. The work of this thesis aims to lay the foundation for computer vision based persistent multimodal traffic analysis.

I would like to thank Prof. Thomas B. Moeslund and Chris Bahnsen for excellent supervision and support throughout this project. A further thank you goes to Tanja Kidholm Osmann Madsen for proving me with the video material and for conducting an additional recording session for this project.

Aalborg University, June 1, 2015

Thiemo Alldieck
talldi13@student.aau.dk

Contents

Chapter 1

Introduction

The growing congestion of public roads and associated problems lead to a growing need of accurate traffic information. This information can be used for traffic safety analysis, early incident detection, improvement of infrastructure capacity and localization of infrastructural weaknesses.

Particularly road safety is an important subject for traffic researchers. The number of worldwide traffic crashes and injuries is growing and the impact on society is high. Especially vulnerable road users, such as pedestrians, cyclists and motorcyclists, are at high risk of road traffic casualties. Not only costs for medical treatment accrue from an injury, but also physiological complications may result for the victim and his family and friends. Even economic impacts have been measured [Peden et al., 2004]. Therefore, already the purpose of safety analysis justifies the research on new traffic monitoring systems and algorithms. Also does the World Health Organization (WHO) stress the importance of accurate data from different sources in its report on road traffic injury prevention [Peden et al., 2004] and the United Nations [2014] invite its members in its latest resolution on 'Improving global road safety' to 'investments in multisectoral road traffic crash surveillance and analysis'.

Different traffic monitoring systems have been developed and used over the years. Hereby has been shown that the use of cameras offers significant im-

provements over other systems such as inductive loops or microwave detectors. Video surveillance offers a wide band of analysis possibilities as for instance traffic flow, turning movements and vehicle classification [Kastrinaki et al., 2003]. Additionally video cameras are portable, comparatively cheap in acquisition and installation and provide rich information understandable by humans. Image processing techniques play hereby an important role as they provide added value to the raw data, enabling automatic extraction of relevant information [Buch et al., 2011].

The use of cameras for monitoring purposes also introduces a significant drawback. Caused by the functional principle of a camera working in the visual range of light, the quality of the data highly depends on environmental conditions such as rain, fog or day and night cycle. A persistent monitoring of the scene is however often desired. To overcome this problem different detectors have been introduced, working either standalone or in combination with traditional cameras. The potential of these methods has been emphasized by Buch et al. in their review of computer vision techniques for the analysis of urban traffic.

Recently a special interest in thermal or infrared (IR) cameras developed. Thermal cameras capture the radiation emitted by objects that depends on their temperature [Gade and Moeslund, 2014]. Therefore, they are preferably used for surveillance of humans whose body temperature is under normal conditions significantly higher than the air temperature. However, the potential and field of application is much greater as described by Gade and Moeslund. Although still being rarely used, the value of thermal cameras in the field of traffic surveillance has been shown in several works.

To overcome downsides of different sensors, multimodal systems have been developed. The goal of this work is to evaluate, how the usage of two modalities, namely visual and thermal, can assist to solve the problem of persistent traffic analysis. A dynamic fusion method is developed, aiming to enable a situation aware usage of the sensors and therefore compensate for their individual weaknesses. Different image quality heuristics are investigated and a new background

subtraction method based on a trust-based fusion of background conformity values is presented. Concluding an in-depth analysis of the presented method follows. To demonstrate the potential of the proposed method, qualitative and quantitative results are presented, discussed and compared to the state of the art methods.

Chapter 2

Analysis

This chapter provides an introduction and overview over state of the art traffic surveillance and related methods. Hereby methods that work under challenging conditions are of particular interest. A discussion about the term *persistence* and an introduction to image fusion techniques follows. The chapter concludes with the problem statement of this work.

2.1 Methods on Traffic Surveillance

Computer vision based traffic monitoring has been an active field of research in the last decades and is still a subject of high interest. Decreasing hardware cost as well as development of new sensors have opened video analytics for a wide field of applications. Methods have been developed for various operating conditions; however, a standard method for different purposes and conditions is yet to be presented. Particularly research on surveillance at nighttime, difficult light and challenging weather conditions is very limited [Buch et al., 2011].

Computer vision techniques have been developed in the field of traffic surveillance for various purposes. Many of these require the system to work in real-time (RT), which limits the complexity of possible algorithms. Cameras or similar sensors are mounted either stationary, typically on a high pole, or in a vehicle. Both

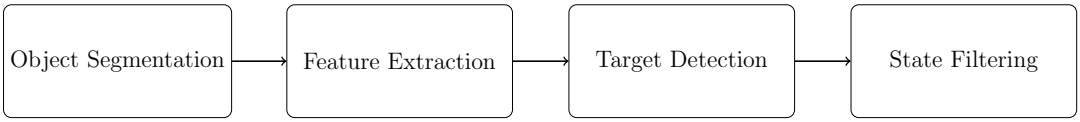


Figure 2.1: Typical steps of a bottom-up tracking approach.

positions introduce additionally limitations and challenges. The camera position on a pole enables a large field of view but causes also a very small level of detail. Moving cameras on the other hand require different image processing techniques as in typical computer vision applications as algorithms can for example not base on fixed points or scene models. However, the scope of this work is limited to methods working with stationary cameras. A further difference to conventional surveillance applications is the fact, that traffic monitoring systems often have to deal with a broad number of different classes, such as cars, trucks, buses, cyclists and pedestrians. Each of these classes have unique properties, which increases the complexity of a *one size fits all* solution.

Traffic monitoring systems fall in general in the category of surveillance systems and can be therefore categorized in one of two processing pipelines. Based on the flow of information the two classes are named *bottom-up* and *top-down* [Al Najjar et al., 2014, 120p]. Top-down techniques require prior knowledge about the objects, as they 'specify a-priori generated hypotheses based on current image data' [Rowe, 2008]. The process relies on the search of learned feature patterns. This might be challenging in the field of traffic monitoring because of the diversity of the road users. Bottom-up techniques on the other hand rely on a segmentation of foreground objects from the background, such as background subtraction or frame differencing. Then feature extraction, target detection and state filtering follows as illustrated in Figure 2.1. No object model is needed and computational costs are generally lower. This makes bottom-up techniques well applicable to traffic surveillance systems. They are however less robust against noise and detection errors and therefore require additional steps to handle resulting difficulties [Al Najjar et al., 2014, 123].

Aside from the tracking of road users, their classification can be of interest in a traffic monitoring system. To improve the working times of monitoring systems, different sensors and sensor fusion (see Section 2.3) are being used. In the following a short overview over crucial steps and state of the art methods in traffic monitoring is given. An in-depth analysis has been provided in the surveys of Kastrinaki et al. [2003] and Buch et al. [2011].

2.1.1 Object Segmentation

The task of object segmentation is a well-researched topic in the field of image processing. However, a golden standard is not found and traffic monitoring applications introduce problems that are ignored by common state of the art techniques. Background model approaches, such as the adaptive Gaussian mixture model by Stauffer and Grimson [1999], assume that foreground objects are constantly in motion and move more or less in the same speed. For traffic, this is obviously not the case, resulting in slow or stationary objects gradually merging into the background. This problem has been addressed by Cheung and Kamath [2005], Vargas et al. [2008] and Yao and Ling [2014]. Cheung and Kamath validate foreground pixels by a moving object model. The latter methods update the background model slower if the pixel was found to be foreground, making it less likely that objects merge into the background. Yao and Ling additionally introduce a prediction step for foreground blobs and texture similarity measures for foreground verification.

2.1.2 Traffic Analysis

Traffic surveillance applications require different levels of understanding of the scene depending on the purpose of the system. Most applications fall in one of the three following categories.

A. Vehicle Counting

Vehicle counting is a common problem of traffic analysis and is solved by the usage of inductive loops. However, the installation causes high costs and interference with the traffic, therefore cameras have been introduced to that field. Only a basic understanding of scene, typically foreground detection, is necessary to perform the task.

State of the art methods like proposed by Bas et al. [2007], Chen et al. [2007b] and Lei et al. [2008] basically rely on background subtraction methods with successive blob analysis. Chen et al. [2007a] provide an adaption for nighttime by detecting the headlights of the vehicle rather than the vehicle itself.

B. Incident Detection

Traffic incident detection requires a basic interpretation of the scene. Incidents can be for example stopped or turning cars, accidents and near accidents as well as actions like passing, tailgating or rule violations. To perform that task, the monitoring system needs to identify and track each vehicle. Occlusion and merging effects of vehicles are hereby the biggest challenges [Kamijo et al., 2000].

Exemplary in this field are the works of Kamijo et al. [2000] and Zou et al. [2009]. Both methods use a Hidden Markov Model (HMM) to classify typical behavior patterns. A slightly different approach is followed by Ki and Lee [2007]. Their methods extract vehicle features such as acceleration, position, area and direction to be able to detect and report accidents. Jackson et al. [2013] present an open source software for tracking and trajectory analysis of generic traffic.

Three different case studies show the potential of the software. Also different commercial systems for incident detection are available.

C. Classification

To be able to conduct arbitrary in depth analyses of traffic and road users, a full understanding of the scene must be present. This includes classification of road users, trajectories and features such as current position, speed and acceleration as well as the dimensions of the vehicles. Some commercial systems for incident detection feature vehicle classification. An all-embracing system is however yet to be presented. In addition, many systems have been developed for highways; systems that can handle the challenging conditions of urban environments are comparatively rare.

Messelodi et al. [2005] present a RT system that is capable of detecting and classifying vehicles including cyclists. Average speed and entering lane of each vehicles are additionally extracted. A significant drawback of the proposed method is that it is not working at nighttime. It is however capable of detecting its working conditions. A method tailored for intersection with heavy pedestrian and bicycle traffic has been presented by Zangenehpour et al. [2014]. Different classifiers have been presented to distinguish road users in the three classes cyclists, pedestrians and motorized vehicle.

2.1.3 Nighttime and bad Weather Surveillance

Surprisingly little work has been done dealing with nighttime and difficult light conditions. In addition, work that covers wide ranges of weather conditions is rare although the interest and amount of workload in that topic is large [Buch et al., 2011].

Some special case methods for tracking during nighttime exist. Robert [2009] presents a method for tracking cars by detection of the two headlights. Dalaff et al. [2003] use image fusion (see Section 2.3) of RGB and thermal images to enable their tracker to work under low light conditions. Other methods, such as presented by Goubet et al. [2006] and Bi et al. [2009], use thermal cameras to detect pedestrians independently of the lighting situation.

Iwasaki et al. [2011] present a pattern-based method that is capable of detecting cars in thermal images under poor visibility conditions such as fog, snow, heavy rain, and nighttime. An update has been presented recently [Iwasaki et al., 2013]. Zhou et al. [2007] built a SVM-based classifier that is able to detect cars by classifying image parts as car or background even under low light conditions.

2.2 Persistence

Video surveillance system characteristics differ by the purpose and setup of the system. One of the most desired qualities is persistence. Persistence is the degree of a system working at all times and under various conditions such as night and day. Different parameters and influences might harm the persistence of a system. Independent of the system setup, these parameters can be separated in two groups. Geometric parameters describe the spatial setup such as camera characteristics and scene geometry. Environmental parameters describe external influences such as weather and lighting situation.

From the setup of a video surveillance system, potential weaknesses can be already derived. Objects far away from the camera position appear smaller and less detailed. Objects behind others might be occluded in the camera view. Problems that might arise are occlusion, self-occlusion, deformation, scaling and mirroring of scene objects. The setup of multiple cameras at different position can lower the problems but introduces identification and registration problems.

The second and larger group of influences on a system are the environmental influences. The degree of how much a parameter harms the system highly depends on the used cameras. In a classical video surveillance situation with a RGB video camera, the lighting situation is of high importance. While indoors the lighting situation might stay more or less the same for a longer period, outdoors it can change dramatically within minutes. Hereby conditions such as presence of streetlights, time of day and presence of clouds or fog can play a role. Depending on the latitude of the location, the amount of sunlight and its angle can be highly dependent on the day of the year, resulting in drastically changing lighting conditions over the year. For outdoor systems, using a thermal camera, the weather condition might influence the quality of the recorded data. Other influencing parameters are possible and will be discussed further in Section 3.2.

The spatial setup of video surveillance systems and resulting weaknesses have been researched with some success already. Research on environmental parame-

ters, that harm the persistence of a system, is still rare. Two parallel strategies can be followed here: The first strategy is the usage of different sensors to minimize the effect on external influences. This strategy called sensor fusion will be further discussed in Section 2.3. Secondly can be investigated on how prior knowledge about influences, such as the lighting situation, can help to improve a system. So have Doshi and Trivedi [2007] shown that satellite images of clouds can help to predict shadows in standard surveillance situation and thus improve adaptive background models, which allows the conclusion that other information sources might be helpful too.

2.3 Multimodal Image Fusion

As discussed earlier in this work, different sensor types produce different types of images with individual strengths and weaknesses. Standard cameras working in the visual range of light capture the reflected colors by a scene. This type of sensing is very similar to the human eye. Poorly illuminated scenes or the presence of visual obstructions such as rain or fog harm the perception. Thermal cameras on the other hand measure the radiation emitted by an object. Images created by this type of sensor are independent of the illumination but are less detailed and provide an unfamiliar visual impression. Image fusion aims to compensate for the individual weaknesses by combining two or more images from different sensors into one. The resulting data ideally contains details from both data sources and can even reveal new features. Generally fusion is not limited to specific types of sensors, this section however focuses on the fusion of RGB and IR images.

Image fusion is used at different stages of a processing pipeline. The three categories are: *pixel-level* fusion, *feature-level* fusion, and *decision-level* fusion [Hall and Llinas, 2001].

Fusion at decision level combines the output from two or more parallel processing pipelines. The results are merged by Boolean operators or weighted average. Serrano-Cuerda et al. [2014] perform parallel segmenting of thermal and RGB data and select the representative output based on confidence heuristics.

Feature-level fusion performs the fusion one step earlier in the processing pipeline. Features from all input images are extracted individually and then fused into a joint feature space. Kwon et al. [2002] present a technique for automatic target recognition (ATR).

Pixel-level fusion is the most common approach. It requires all input images to be spatially and temporally aligned. This alignment, also called registration, is a task for itself. Automatic image registration approaches, as used for example for image stitching, often fail, since there is no correlation between the intensity values of the modalities [Conaire et al., 2006]. This is funded in the fact, that

Figure 2.2: RGB image and result of a naive RGB/IR fusion

the sensors work with different wavebands. A common approach is to manually select corresponding points in both modalities and compute a homography. A homography matrix is a bijective mapping between image points in the different camera views laying on the same 3D plane. This procedure will be further elaborated in Section 4.1.2. However special case automatic methods exist. Since for an algorithm it is hard to tell which points correspond in the images with the lack of correlation, features are used that are most probably present in both modalities. Successful methods have been presented using contours [Heather and Smith, 2005], Harris corners [Hrkać et al., 2007] and Hough lines [Istenic et al., 2007]. However these methods have been tailored for specific datasets and are not generally applicable.

As already discussed, the goal of pixel-level fusion is to enrich the input data. Figure 2.2 shows a scene from the OTCBVS dataset [Davis and Sharma, 2007]. The left image shows the raw RGB image. The right image shows the result of a negative multiplication of the RGB and IR image. Already this naive technique reveals a person standing next to the building. Besides naive fusion though averaging, addition or multiplication of the images, more complex methods have been presented, trying to optimize the information content of the image.

The work of Shah et al. [2010] performs the fusion after different wavelet transforms of the images. This allows a fusion rule based on frequencies rather than

pixels. Details are preserved while simultaneously artifacts can be reduced. A statistical approach is followed by Chen and Leung [2009]. During an expectation-maximization the fusion result is obtained stepwise.

Lallier and Farooq [2000] perform the fusion trough adaptive weight averaging. The weight per pixel is hereby defined by a number of equations that express the interest in the specific pixel. In the context of the work these are the degree of an object being warmer or colder for the thermal domain and the occurrence of contrast differences as well as large spatial and temporal intensity variations for the visual domain.

Instead of fusing the images to a new image that can be represented in RGB, other methods simply combine the inputs and perform the fusion in the subsequent processing. St-Laurent et al. [2007] adapt a state of the art algorithm for moving objects extraction to work with "Red-Green-Blue-Thermal" (RGBT) videos.

2.4 Problem Statement

Following the analysis above, the goal of this work is to create a robust and persistent background subtraction method for traffic monitoring systems using a combination of a RGB and a thermal camera. Object segmentation is the initial and crucial step for successful tracking and qualitative analysis. Therefore this step must be robust against noise and environmental changes. Different speeds of traffic and stopped road users have to be handled by the system.

State of the art camera based traffic monitoring is usually based on video data of the visual domain. Thermal cameras and image fusion are rarely used, although their potential is great especially for low light situations. Under different conditions thermal and RGB cameras vary in data quality. When aiming for a persistent 24 hour surveillance, a fixed fusion based on image characteristics might be challenging, since the surveillance situation can change dramatically. Therefore, an adaptive weighting or even a switch between the modalities might be beneficial. This project investigates in which way prior information can help a system to choose the best modality to work with. Different information sources are investigated and a system design is proposed. In contrast to other works, no situations are excluded to lay the foundation for a system that can work at all times.

Based on this problem statement, the following chapter evaluates system requirements and possible information sources.

Chapter 3

Requirement Specification

In the previous chapter an analysis of state of art traffic monitoring applications and related methods was given. Based on the problem statement given in Section 2.4, requirements are defined in this chapter. This includes an evaluation of possible information sources that may indicate the quality or *usefulness* of the two modalities as well as requirements for the background subtraction algorithm.

3.1 Background Subtraction

As discussed earlier, background subtraction or foreground detection is the first step in many image processing applications. Foreground regions are identified for further processing steps. The foreground is hereby defined as all objects that are not fixed in the scene or all objects of particular interest. In traffic monitoring foreground objects are all road users, such as cars, cyclist and pedestrians. All road users should be marked; false positives should not occur. The coverage of the foreground blobs should be at least on the level of state of the art methods.

Background subtraction for traffic scenes is a comparable hard task. Assumptions made by different background modeling techniques do not hold. In these methods, a statistical background model is build and the current frame is compared to the model. Large differences indicate foreground regions. The

background model is constantly updated to adapt for environmental changes, resulting that slow or stopped objects merge into the background. Other techniques require prior information about foreground objects. The diversity of the traffic in the given setup makes this option impossible.

Background modeling is however the right tool for traffic monitoring because of its flexibility and real-time capability. Adjustments to common methods have been presented, to prevent merging and ghosting effects. The model has to quickly adapt to environmental changes such as illumination changes, shadows or moisture in the RGB modality and automatic camera adjustments and temperature changes in the thermal images and simultaneously preserve foreground objects from merging into the background. The proposal of Yao and Ling [2014] shows great potential and should therefore be consulted.

The background subtraction for the RGB modality should feature a shadow detection. Since shadow detection can also produce false positives, the detection should be triggered only for weather situations where large shadows are likely.

3.2 Quality Heuristics

Knowledge about environmental conditions let us estimate how well a sensor can work in a certain situation. Intuitively we know that a RGB camera is useless in the absence of light. The image quality of thermal cameras is good, when objects are warmer than their environment as it is the case for humans under normal conditions. To adapt this *common sense* to a technical system, the confidence in a sensor must be quantified. Also the quality of data produced by a camera is dependent on a large number of influences. Therefore must be investigated which parameter or set of parameters indicates the quality of recorded data best.

RGB and thermal cameras work in different frequency ranges. Therefore, intensity values have different interpretations and are not correlated. Consequentially influencing parameters are most likely disjoint too. In the following possible quality indicators are evaluated separately for both modalities.

3.2.1 Visible Range

Quantified quality rating of RGB images is a hard task since we are used to this type of sensing. The human eye can adapt to challenging conditions easily and the human brain interprets the visual perception [Gregory, 1997]. Therefore we need an understanding of image processing techniques, to be able to rate the quality of an image.

Starting from the bottom-up technique as described in Section 2.1 an input image is of good quality when the system creates acceptable results using state of the art image processing methods. This means generally all objects of interest are found and false positives are rare. Which results are acceptable is dependent on specific system requirements.

Figure 3.1 shows three situations from the same intersection. While Figure 3.1a shows an ideal image for background subtraction methods, both Figures 3.1b and 3.1c display exemplary challenges. Both images show situations

(a) Good conditions **(b)** Large shadows **(c)** Large reflections

Figure 3.1: RGB images from the same intersection under different conditions.

were ghost objects, namely drop shadows and reflections, are present. In both situations background subtraction techniques based on background modeling would find these areas as foreground objects. Although shadows are handled quite well nowadays [Prati et al., 2003], they still disturb the detection process. An effective method for reflection detection is yet to be presented. In conclusion both images should be rated as of low quality.

Following the argumentation images with low light conditions, such as dusk, dawn and night, should be rated as low quality even though a human might be able to identify the road users. As a heuristic input parameter the elevation angle of the sun be can consulted. As illustrated in Figure 3.2 the solar elevation angle is defined as the angle between the ground plane and the sun's position vector. The sun is visible for angles $\geq 0°$. Between $0°$ and $-18°$ we speak about twilight and below the sun does not contribute to sky illumination, it is night.

On daytime the presence and amount of drop shadows depends on two factors. Firstly only on sunny days drop shadows can appear. A weather database can be accessed to retrieve a description of the weather conditions. Secondly the length of these shadows also depends on the solar elevation angle. Therefore both, weather data and the position of the sun must be consulted to present a heuristic to what extent cast shadows might be present in the scene.

Different weather conditions such as heavy rain, snow and fog also might harm the image quality. The reasons for this are manifold. Examples are reflections

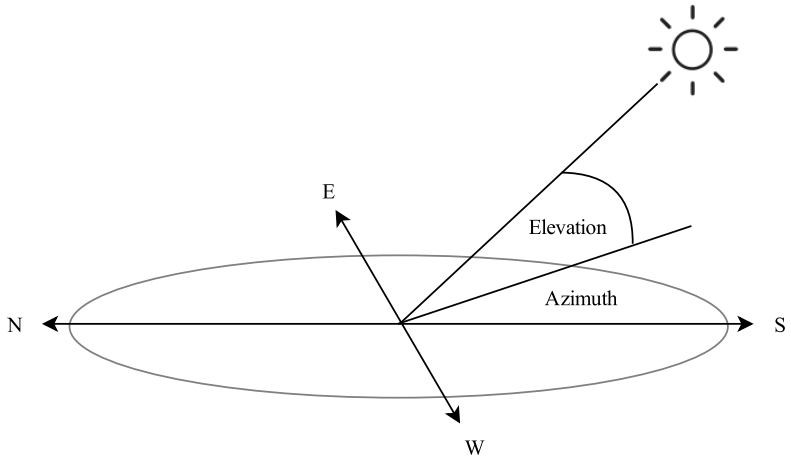


Figure 3.2: Illustration of solar elevation and azimuth angle.

due to moisture, limited visibility due to rain or fog or changing illumination by moving clouds. The quantified quality values of different conditions must be experimentally determined.

3.2.2 Thermal Range

Thermal cameras measure the infrared radiation emitted by all objects. The energy of the radiation is hereby mainly dependent on the temperature of the object. A constant factor called emissivity scales the radiation for different materials [Gade and Moeslund, 2014]. Conversely, this means that objects of different materials emit different level of radiation energy while having the same temperature. With known emissivity the temperature of objects in thermal images can be calculated using the Stefan-Boltzmann law. However many thermal cameras built for surveillance feature Automatic Gain Control (AGC), so that the mapping function between radiation energy and intensity values is unknown. AGC automatically adjusts the image gain to the optimal range. The function comes as a build in feature with state of the art thermal cameras and is necessary to be able to record high quality video data over a longer period.

Figure 3.3: Different road temperatures

The quality of thermal images is high when the contrast between foreground and background objects is high. In traffic surveillance this is the case when road and road users emit a significant different amount of radiation; or simplified there is a noticeable temperature difference. Assuming a more or less constant temperature of road users, the temperature of the road could be a quality indicator. Figure 3.3 shows an example on how different road temperatures affect the contrast between cars and road. Except from direct measurements however, the determination of the road temperature is complicated, since it is influenced by numerous interacting parameters [Chapman and Thornes, 2006]. Therefore, the road temperature cannot be used, as long as this data is not present.

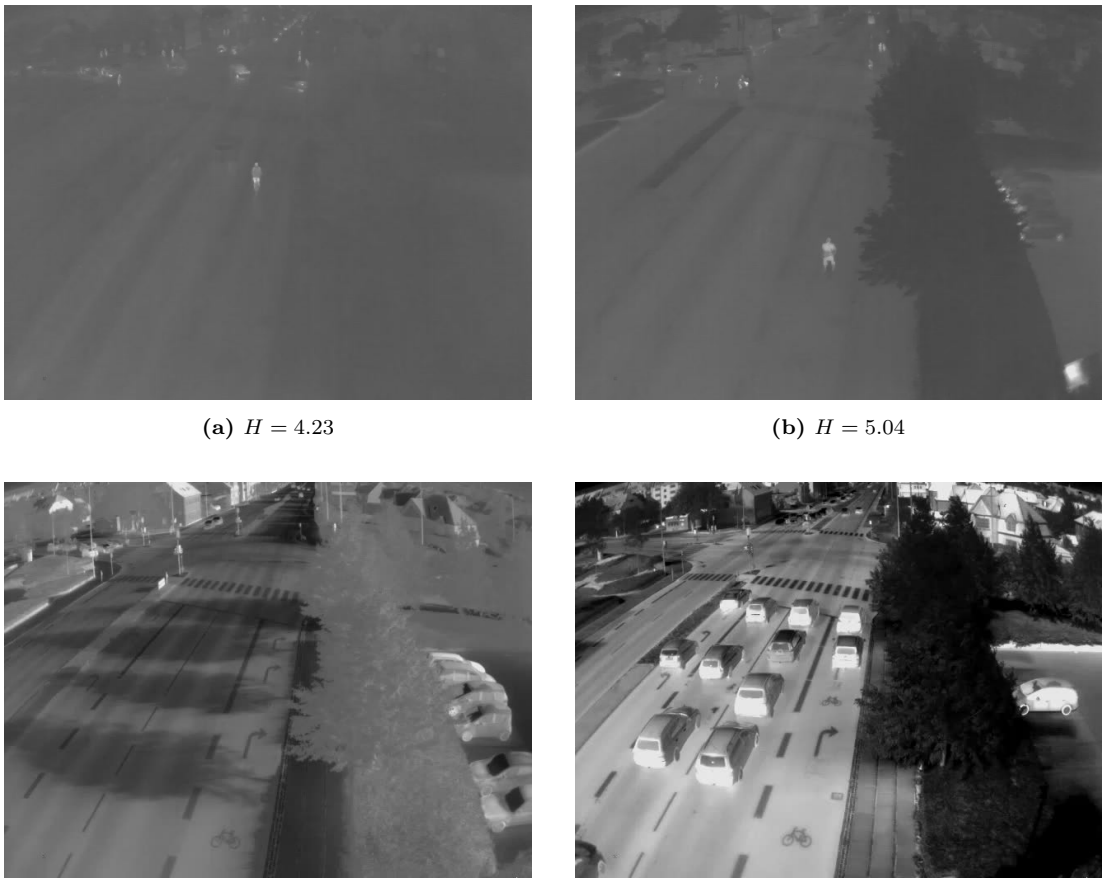
As described earlier, objects of different materials have different intensity values in a thermal image, even when having almost the same temperature. Therefore a certain amount of information is in the image even when showing a scene without foreground objects. In a situation where no objects can be distinguished, the information content is low. Consequentially the image entropy can be used

as a quality heuristic of thermal images. The entropy is the expectation of the *information content* contained in an information source. It is defined as:

$$H = -\sum_{i=0}^{255} p(I_i) \log_2(p(I_i)) \tag{3.1}$$

with p being the probability of a intensity value I in a gray scale image.

Figure 3.4 shows a side-by-side comparison of the same intersection at different times. In the upper images the cars can hardly be seen while in the lower images even details such as road markings stand out clearly. The corresponding entropy values correlate with this impression. Experiments conducted during this work have shown that for thermal images with $H < 4$ the foreground cannot be separated with satisfying results.



(a) $H = 4.23$ **(b)** $H = 5.04$

(c) $H = 6.65$ **(d)** $H = 7.67$

Figure 3.4: Thermal images of the same scene with different entropy values.

Chapter 4

System Design

The following chapter discusses the design of a multimodal background subtraction method that uses environmental information to dynamically fuse the input images as specified in the previous chapters. The basic system design is illustrated in Figure 4.1. The following sections will reference this overview from time to time for better understanding.

The system implementation consists of three major steps that are represented through the following sections. Section 4.2 specifies the image quality heuristics as analyzed in Section 3.2. Section 4.3 elaborates on the heart of the algorithm: the weighted late fusion of the two input streams. Finally Section 4.4 describes special adjustment that have been made in order to optimize the algorithm for the purpose of traffic surveillance.

To begin with preliminary steps are described, that are needed to synchronize and align the input data.

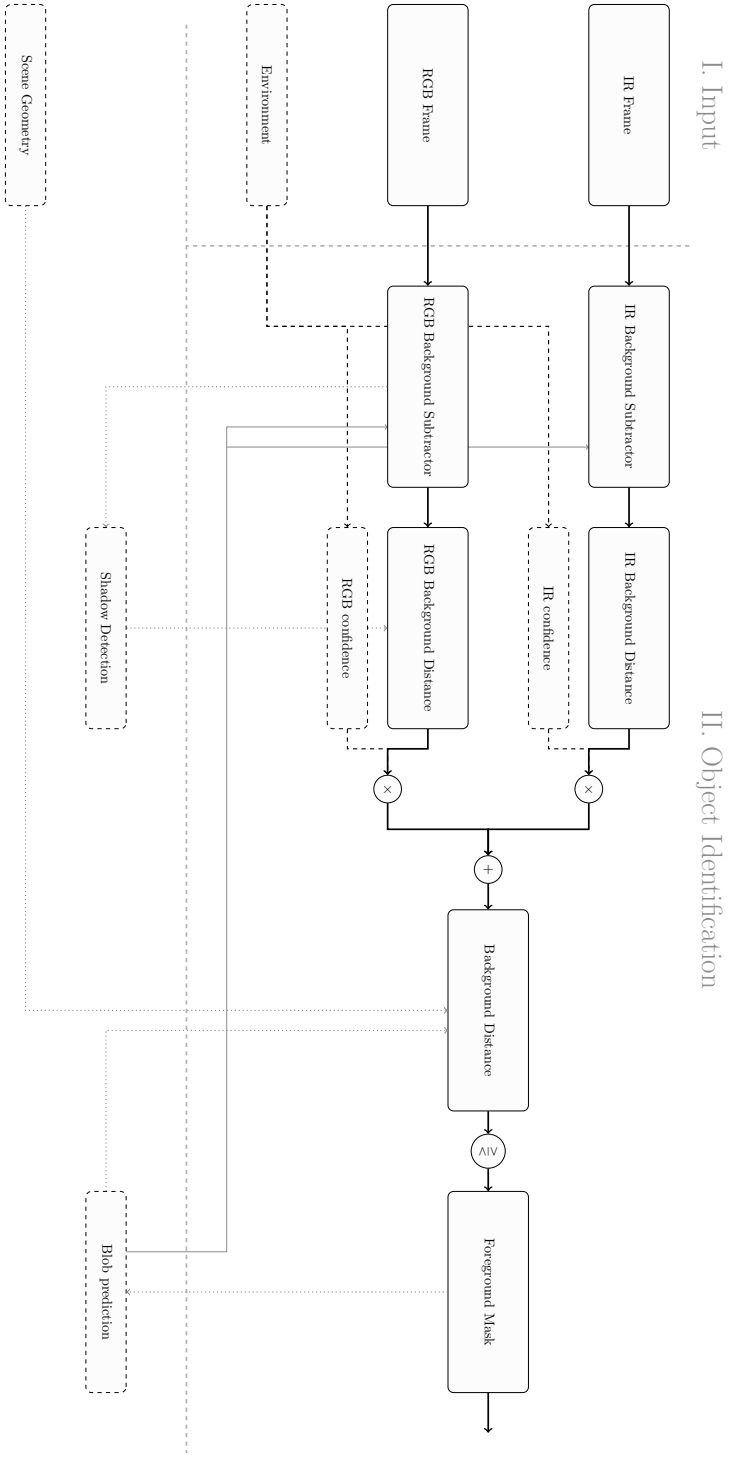


Figure 4.1: System design overview

4.1 Image Preprocessing

In order to provide processable video material, preliminary steps were needed for the main dataset of this project. The steps are not necessarily part of the presented system as they are common methods for computer vision algorithms. They have been however, an important part of the work and are therefore elaborated in the following. The presented solutions are tailored or selected for the given dataset, which will be described in detail in Section 5.1.

4.1.1 Temporal Alignment

In order to be able to fuse information from both cameras, the frames have to be synchronized. Hardware that is able to perform this task at recording time is available but has not been used for the given dataset. As a result the video sequences may have different frame rates and different starting points.

With tools like ffmpeg[1] the frame rate of videos can be adjusted. While doing so frames are doubled or dropped, leading to a slightly temporal inconsistency.

Temporal offsets of the videos have been compensated manually. After synchronization an offset smaller than one frame might be still present. This offset however is negligible, since resulting spatial offsets of moving objects are mostly in sub-pixel level.

The algorithm as presented in the following features a compensation for temporal and spacial inaccuracies.

[1] https://www.ffmpeg.org/ Last downloaded June 1, 2015

4.1.2 Image Registration

The following step is the spatial alignment of the images, also called image registration. During image registration two or more images taken at *different times*, from *different sensors* or from *different viewpoints* are transformed into the same coordinate system [Brown, 1992]. For the given dataset the last two points are the case. In order to perform the registration, the camera geometry is consulted.

The simplest camera model, which describes the projection between the 3D world and the 2D image, is the *pinhole camera model* as displayed in Figure 4.3. With the intercept theorem one can easily see that the point $(X, Y, Z)^T$ is mapped to $(fX/Z, fY/Z, f)^T$ [Hartley and Zisserman, 2003, 153ff]. Using homogeneous vectors, the projection can be expressed as a linear mapping:

$$\begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \mapsto \begin{pmatrix} fX \\ fY \\ Z \end{pmatrix} = \begin{bmatrix} f & & & 0 \\ & f & & 0 \\ & & 1 & 0 \end{bmatrix} \begin{pmatrix} X \\ Y \\ Z \\ 1 \end{pmatrix} \tag{4.1}$$

One important consequence of eq. (4.1) is the *epipolar constraint* as displayed in Figure 4.4 [Hartley and Zisserman, 2003, 325ff]. Given a point in world coordinates X and its image point x of the first camera, its corresponding image point x' of the second camera must lie on the epipolar line l'. For points located on the plane π_i therefore exists a projective transformation from the image plane of the first camera to the image plane of the second camera:

$$x' = \mathbf{H}_{\pi_\mathbf{i}} x \tag{4.2}$$

where $\mathbf{H}$ is a 3×3 matrix and x and x' are homogeneous vectors. The relative position of the cameras is hereby irrelevant. The matrix $\mathbf{H}$ is called homography.

The homography between the ground planes of the two images represents a good approximation of the mapping between all points of the two images.

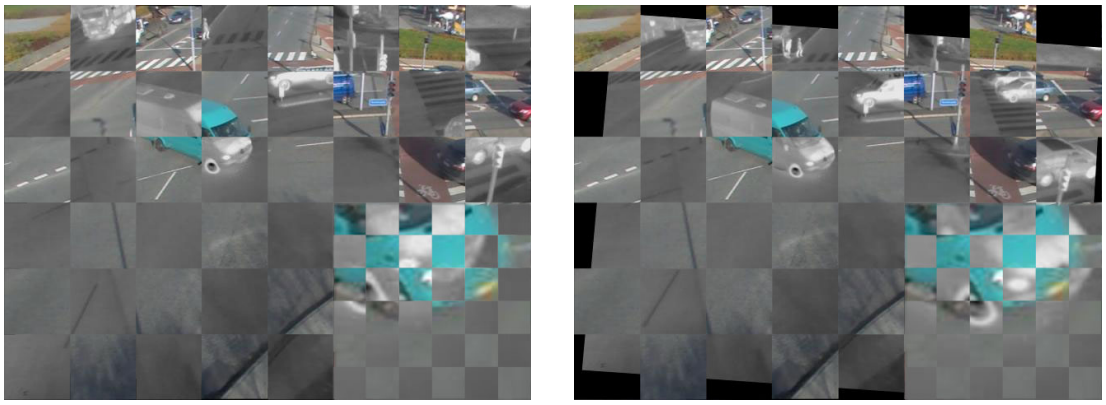


Figure 4.2: Image pair before and after registration.

One image can be registered on the second with mapping all points with the homography. The registration error hereby depends of the distance of a point to the ground plane in world coordinates. For many scenarios, including traffic surveillance, this registration process is sufficient.

To find the homography between the ground planes of the two images, corresponding points are selected manually. Afterwards the homography is calculated through the least squares method. The input and final result of the registration process is displayed in Figure 4.2.

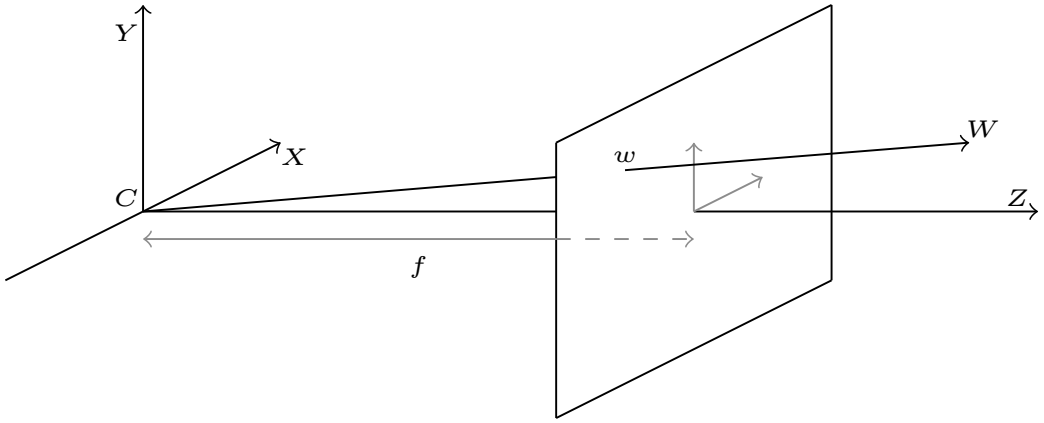


Figure 4.3: Pinhole camera model with camera center C, principle axis Z, focal length f, world point W and corresponding image point w.

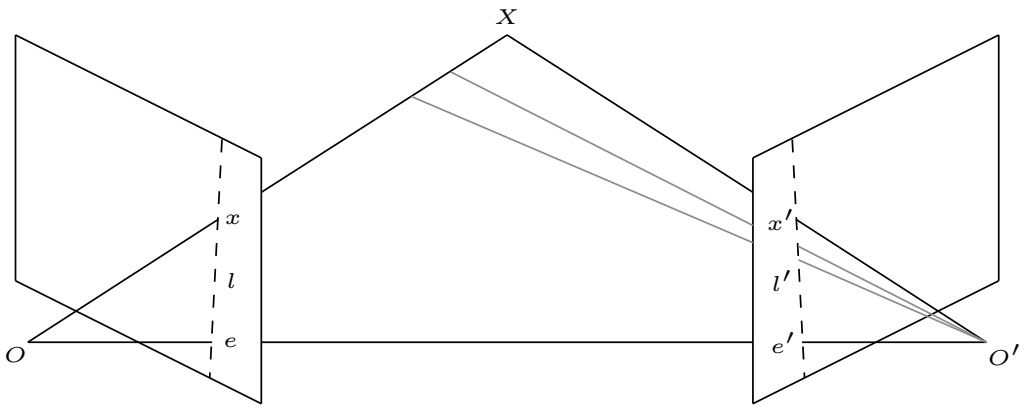


Figure 4.4: Epipolar constraint for point X with its image points x and x', epipolar lines l and l' and epipoles e and e'.

4.1.3 Lens Distortion Correction

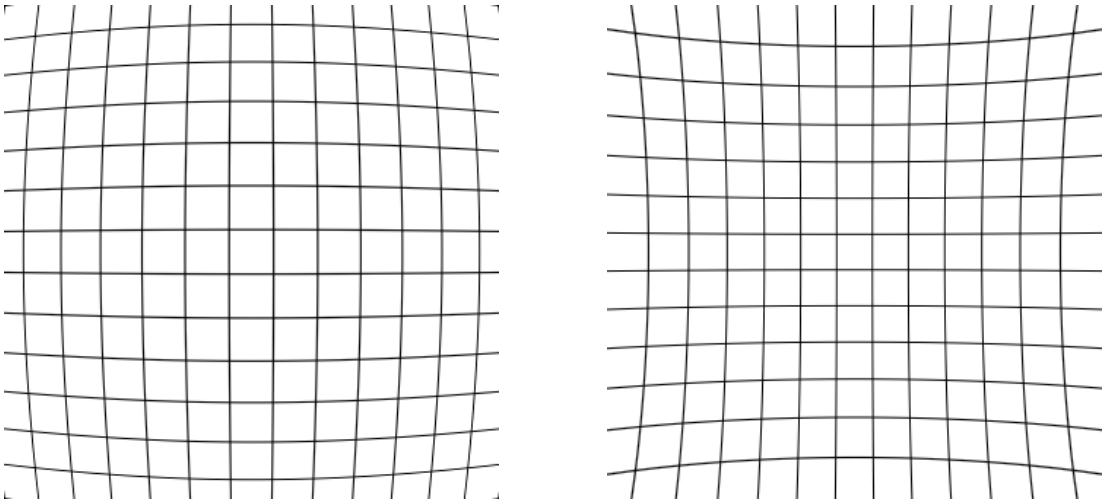
The prior assumption, that a camera follows a linear projection, is not true in many cases. This results that straight lines in world coordinates are not mapped to straight lines in the image [Szeliski, 2010, 52ff]. Image registration based on a homography fails as a consequence. The reason for the nonlinear mapping lies in a *radial distortion* of the lenses. Figure 4.5 displays the two common types of lens distortion. Coordinates in the observed images are displaced away (*barrel distortion*) or towards (*pincushion distortion*) the image center proportional to the radial distance to the center. Fortunately, these types of distortion can be compensated. A simplified model describing the phenomena is the following:

$$\begin{aligned} \hat{x_c} &= x_c(1 + \kappa_1 r^2 + \kappa_2 r^4) \\ \hat{y_c} &= y_c(1 + \kappa_1 r^2 + \kappa_2 r^4) \end{aligned} \tag{4.3}$$

where (x_c, y_c) is the normalized pixel position in reference to the image center and κ_1 and κ_2 are the *radial distortion parameters*. In order to find the original pixel position the parameters must be found.

Since finding the camera parameters for image undistortion is a common task, multiple frameworks exist. Tools like the *Camera Calibration Toolbox for Matlab* [Bouguet, 2008] or the calibration functions of the computer vision library *OpenCV* [Bradski, 2000] are frequently used tools.

The common procedure that is also performed by the mentioned tools, is to first take images of a checkerboard as shown in Figure 4.6. Then the edge points of the squares are found in the resulting image through basic image processing methods or by hand. Afterwards the offset of the points from their corresponding straight line can be measured. From location and offset values, the distortion parameters can be estimated.



(a) Barrel distortion **(b)** Pincushion distortion

Figure 4.5: Different types of lens distortion.

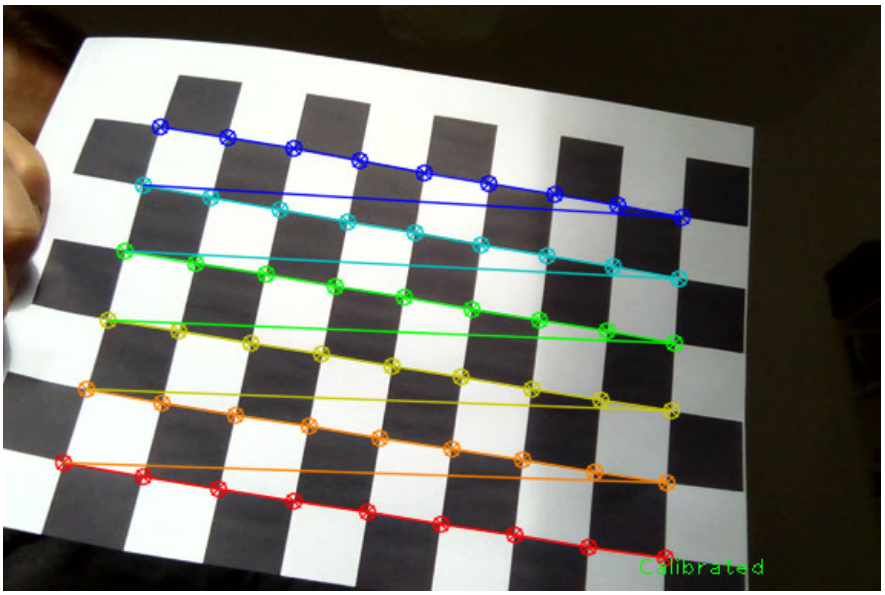


Figure 4.6: Camera calibration with OpenCV.

This method however, requires access to the cameras that have been used for the recording of the dataset. Unfortunately this was not possible during this project. Therefore another solution had to be found.

Alemán-Flores et al. [2014] present a method that is capable of finding points that may have lain on a straight line in world coordinates. In order to do so, a distortion parameter is introduced as a third dimension in Hough line transform. The distortion parameter with the n strongest lines is chosen as the best approximation. Afterwards points that are possible part of these lines are extracted. This way points on straight lines in world coordinates are found, which can then be used for a more precise approximation of the lens distortion parameters similar to the checkerboard method. Found lines by the algorithm in a frame from the dataset used in this work are shown in Figure 4.7.

Once the radial distortion parameters are found through optimization of eq. (4.3), the lens distortion can be corrected. A side-by-side comparison of a distorted image and its corrected correspondence is displayed in Figure 4.8. As expected from the definition, the effect is most noticeable at the corners of the image.

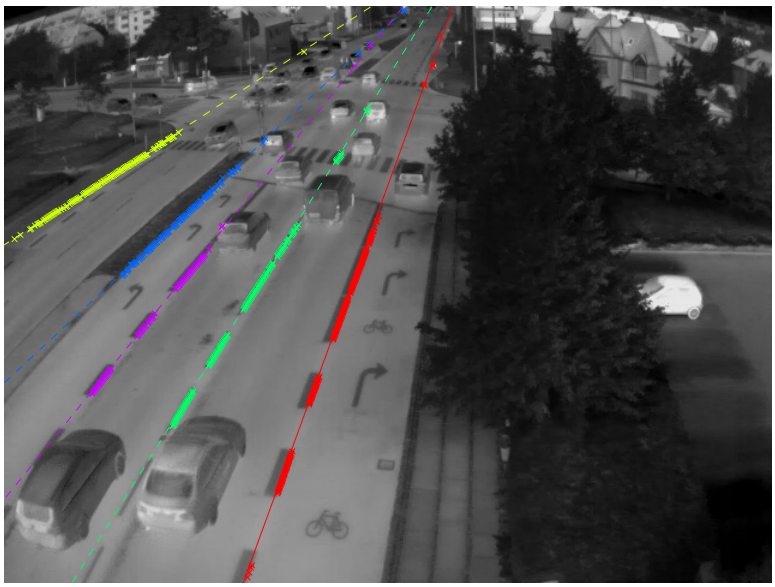


Figure 4.7: Lines extracted through Alemán-Flores et al. [2014].

Figure 4.8: Image before and after lens undistortion.

4.2 Quality Heuristic Specification

In Section 3.2 has been discussed what information sources can be consulted in order to give an estimate how well a background subtraction algorithm will perform on the given modality. This estimate will be specified in the following section in form of heuristic functions that receive the information as an input and calculate a quality value between zero and one as its output.

4.2.1 Thermal Image Quality

The best property that has been found to estimate the thermal image quality is the image entropy as described in Section 3.2.2. The mapping between entropy values and quality is however unknown. From the definition of the entropy images with $H = 0$ contain no information, what specifies the lower boundary. During experiments conducted during this work could be found, that images with $H > 6$ contain enough details for a completely satisfying background subtraction. Images with $H < 4$ show already significant shortcomings.

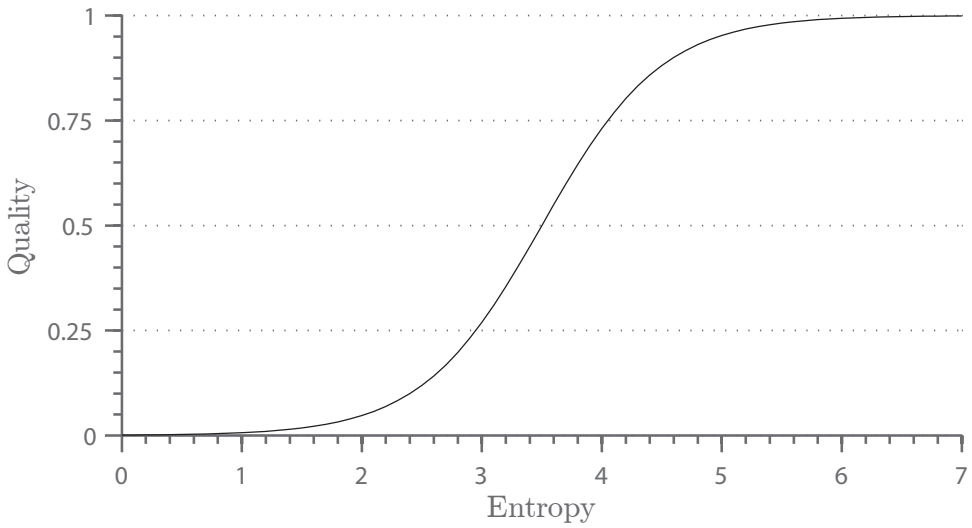


Figure 4.9: Thermal image quality heuristic.

Since the degree of the mapping function between entropy and quality is unknown, it has to be chosen manually. A linear function has been found to be not sufficient. During the experiments could be seen, that the down-rating of low entropy values is too strong. A sigmoid function appeared as a better approximation of the mapping function. The function used during this work is showed in Figure 4.9. However since the true model is unknown, different functions might work better for this or other scenarios. An in-depth analysis is beyond the scope of this work.

4.2.2 RGB Image Quality

For the image quality in the visual domain multiple information sources have been found. In contrast to the thermal domain, all information come from external sources, namely position of the sun and weather conditions. Three functions can be derived from this knowledge base. These are firstly the amount of daylight trough the position of the sun, next the presence of harming influences through weather conditions and last the presence and amount of cast shadows depending on both sources.

A. Illumination

The main influence for the RGB image quality is the illumination of the scene. From the solar elevation model, which is further explained in Appendix A, we can derive information about the illumination. The elevation angle is depended on the hour of the day, day of the year and the location. Figure 4.10 shows the solar elevation angle for summer and winter solstice for the location Aalborg, Denmark.

For angles $> 0°$ the sun is visible, we speak about day. Below $-18°$ the sun has no influence on the illumination of the sky. In between we speak about twilight. Therefore $-18°$ and $0°$ could set the boundaries of a quality heuristic. However the civil twilight begins not before $-6°$. Furthermore is the illumination condition

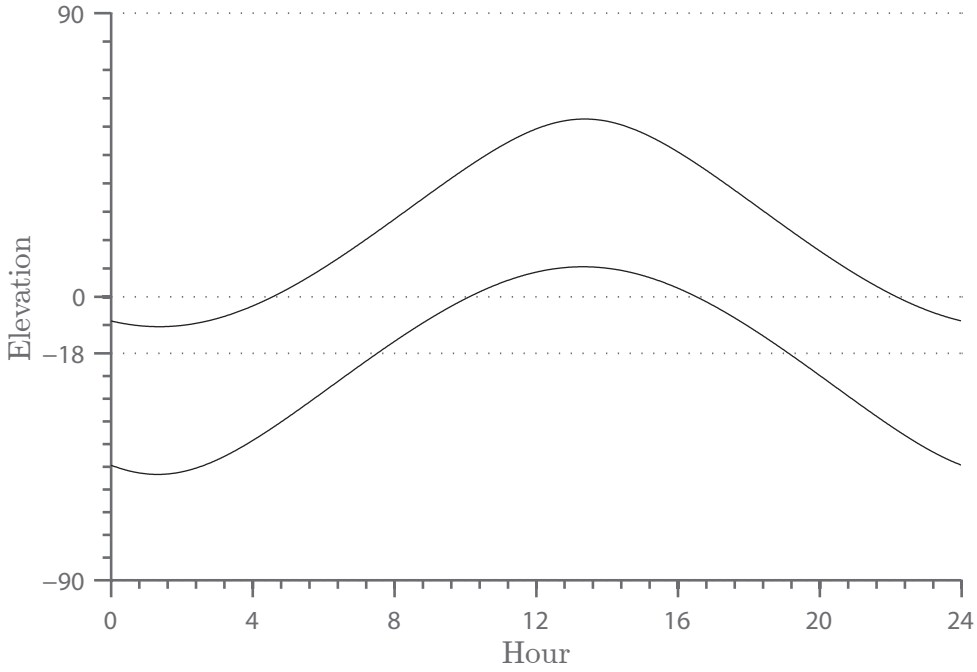


Figure 4.10: Sun's elevation angle for summer and winter solstice for the location Aalborg, Denmark. −18° marks the beginning of the twilight.

is not perfect as soon as the sun is visible. Buildings and the colored sunlight lead to an imperfect illumination. Therefore a perfect illumination has been defined in this project for elevation angles $\geq 6°$. Since the sun and therefore the input function already follows a sine function, the mapping of quality values have been assumed to be linear between $-6°$ and $6°$. However during night streetlights and headlights illuminate the scene. Therefore a minimum quality of 0.2 has been assumed for illumination.

B. Shadows

Although methods for shadow detection exist, the presence of shadows is still a challenge in computer vision. The extent of shadows depends on the solar elevation angle through the formula:

$$L = h/tan(\alpha) \tag{4.4}$$

with h being the object height and α the solar elevation angle. With a unit object height $1 - L$ can serve as a quality function. To prevent negative values a minimum value q_{smin} is introduced. Another modification needs to be made in order to set the quality to 100% for nighttime and bad weather leading to the equation:

$$q_{\text{shadows}} = \max\left(1 - \frac{(\alpha > 0) \wedge w}{\tan(\alpha) * \psi}, q_{\text{smin}}\right) \tag{4.5}$$

where w indicates good weather where shadows are likely and $\psi > 0$ is a scaling factor. The conditions have been hand-picked from the selection the weather database provides. Selected have been situations with clear sky or varying and occasional cloudiness.

The values of q_{smin} and ψ have been set to 0.3 and 50 respectively during this work. Both values are arbitrary and have been hand tuned.

C. Weather

The influence of the weather on computer vision algorithms is manifold. Every weather condition effects to algorithms in a different manner. The weather database Weather Underground by The Weather Channel[2] used during this work differentiates between 133 conditions. An in-depth analysis of all conditions is beyond the scope of the work. Therefore different conditions have been broadly grouped into 5 categories as seen in Table 4.1.

Good conditions	1.0
Low/Varying illumination	0.8
Reflections/moisture	0.6
Particle occlusion/precipitation	0.3
Reduced visibility	0.3

Table 4.1: Different weather categories and corresponding quality rating.

Each of the categories is characterized through a set of quality harming influences, such as varying light, reflections through moisture or particle occlusion. The quality rating of the categories are hand tuned based on experiments. A full list of conditions and their corresponding groups can be found in Appendix B. If a condition falls in multiple categories, the one with the lowest rating is chosen.

[2] http://www.wunderground.com/history/ Last downloaded June 1, 2015

4.3 Background Distance Fusion

This section discusses the main contribution of the work. A new approach to image fusion is presented. Despite other works, not the input data is fused but intermediary results of two parallel background subtractions. In Figure 4.1 (page 26) this part of the system is marked with roman number II.

The system is based on the adaptive Gaussian mixture model background subtraction algorithm presented by Stauffer and Grimson [1999]. The necessary adjustments have been made directly in an OpenCV implementation[3] of the algorithm. The implementation features improvements presented by Zivkovic [2004]. The specific implementation is however interchangeable in the presented method.

4.3.1 Gaussian Mixture Model Background Subtraction

For a better understanding of the proposed method, an introduction into background subtraction with adaptive Gaussian mixture modeling is given in the following. The main idea is to build a statistical model of the scene. Intruding objects can then be detected by identifying parts of the image that do not fit the model.

As the scene model a Gaussian mixture model (GMM) is assumed. Early approaches only assumed a single Gaussian [Wren et al., 1997]. However, it has been shown, that background values may change between different states, e.g. because of shadows or moving leafs. A GMM is able to model this behavior.

The model is built by choosing a time period T. At time t we have a history per pixel x of $X_t = x_t, \ldots, x_{t-T}$. The history is modeled by a GMM with K components.

[3] http://docs.opencv.org/modules/video/doc/motion_analysis_and_object_tracking.html#backgroundsubtractormog2 Last downloaded June 1, 2015

$$P(x|X_t) \approx \sum_{i=1}^{K} \omega_{i,t} \mathcal{N}(x, \mu_{i,t}, \Sigma_{i,t}) \tag{4.6}$$

The mixing weights ω_i are non-negative and add up to one. ω_i describes the *prior probability* of the component i. For computational reasons it is assumed that $\Sigma = I\sigma^2$, i.e. the color channels of the pixels are independent.

Since in X_t are most likely values that belong to foreground objects, we approximate the background model by the first $B \leq K$ larges clusters such as:

$$B = \text{argmin}_b \left(\sum_{k=1}^{b} \omega_k < T \right) \tag{4.7}$$

where T describes the portion of data that should be accounted by the background.

To decide whether or not a pixel of a new image is part of the background, $P(x|X_t)$ can be evaluated and compared to a threshold. In practice however, the components of $P(x|X_t)$ are evaluated individually. A pixel is defined to match the background if it falls within λ standard deviations of the mean of one of the background components:

$$M_{i,t+1} = \begin{cases} 1, & \text{if } |x_{t+1} - \mu_{i,t}| < \lambda\sigma \\ 0, & \text{otherwise} \end{cases} \tag{4.8}$$

When monitoring a scene over a longer time, the scene will most probably change. Reasons can be changing illumination trough moving clouds, moving shadows or even objects that came into the scene and should be treated as being part of the background. Therefore the model has to be updated over time. The update process is done as follows:

$$\omega_{i,t+1} = (1-\alpha)\omega_{i,t+1} + \alpha M_{i,t+1} \tag{4.9}$$

$$\mu_{i,t+1} = (1-\beta)\mu_{i,t} + \beta x_{t+1} \tag{4.10}$$

$$\sigma^2_{i,t+1} = (1-\beta)\sigma^2_{i,t} + \beta(x_{t+1} - \mu_{i,t})^2 \tag{4.11}$$

where α is a constant update rate and β is defined as:

$$\beta = \alpha\mathcal{N}(x_{t+1}, \mu_{i,t}, \sigma^2_{i,t}) \tag{4.12}$$

When more than one match M_i is found, only the one with the most *supporting evidence* and *least variance* (ω_i/σ) is selected and all others are set to 0. If no match is found, a new component is generated and the least probable one is discarded. Finally, the weights ω_i are normalized at each iteration to add up to 1.

An in-depth explanation of the algorithm along with elaboration on possible improvements is given by Power and Schoonees [2002].

4.3.2 Calculating the GMM Distance Map

During the calculation of the foreground mask with the help of the Gaussian mixture model, each pixel is tested against each component of its background model, if it can be accepted as part of the component's Gaussian distribution. The Euclidean distance of the sample value from the mean is hereby the important factor for acceptance. When rewriting eq. (4.8) we have:

$$M_{i,t+1} = \left(\frac{|x_{t+1} - \mu_{i,t}|}{\lambda\sigma_{i,t}} < 1 \right) \tag{4.13}$$

The acceptance distance of the sample as background is now normalized by the specific variance and the threshold value. Large distance values indicate a

high probability of the pixel being foreground whilst small values show a high conformity with the component. With this in mind an approximation of the general conformity of a pixel with the model can be expressed with distance values:

$$D_t \approx \begin{cases} d_{0,t}, & \text{if } M_{0,t} \\ d_{1,t}, & \text{if } M_{1,t} \\ & \dots \\ d_{b,t}, & \text{if } M_{b,t} \\ \min(d_{0,t}, d_{1,t}, \dots, d_{b,t}), & \text{otherwise} \end{cases} \tag{4.14}$$

with

$$d_{i,t} = \frac{|x_t - \mu_{i,t-1}|}{\lambda \sigma_{i,t-1}} \tag{4.15}$$

If a match $M_{i,t}$ is found, the corresponding value of $d_{i,t}$ is used to express the distance. Otherwise the distance to the closest component is used. The resulting values of all pixels form a map expressing the deviation of image regions from the background. The scaling is hereby the same for all pixels, so that a single threshold can be applied. When thresholding the map with the value 1.0, the resulting mask is the same as if calculated through Stauffer and Grimson [1999]. Figure 4.11 shows an image representation of the background distance map of a traffic scene. An arbitrary scaling has been applied for viewing purposes.

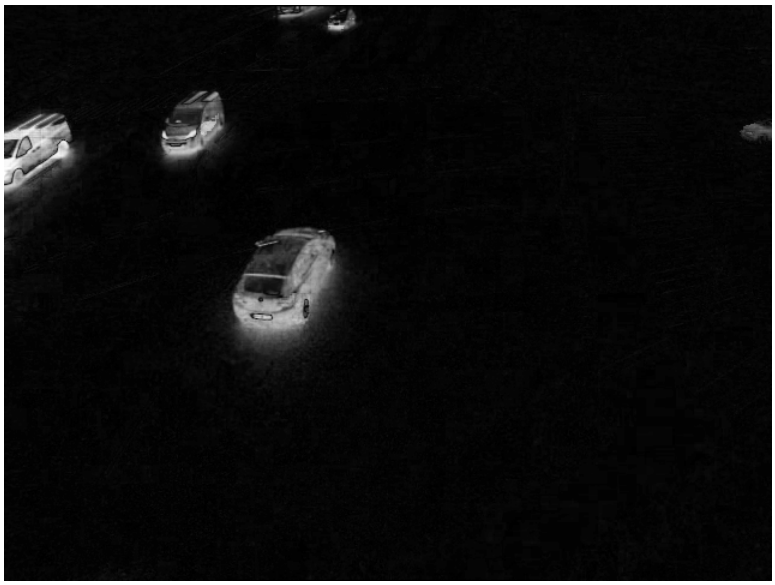


Figure 4.11: Distance map of a traffic scene.

4.3.3 Trust based Fusion and Foreground Identification

At this stage we have heuristics for the image qualities of both modalities as well as maps expressing the background conformity of pixels. Build upon this a trust based fusion is performed.

The trust in a modality can directly be derived from the quality heuristics. The better the quality the higher is the trust. Therefore the different heuristics have to be combined first. Under the assumption, that the different quality heuristics do not interfere each other, the heuristics can simply be multiplied:

$$q_{\mathrm{RGB}} = q_{\mathrm{sun}} \cdot q_{\mathrm{shadows}} \cdot q_{\mathrm{weather}} \tag{4.16}$$

$$q_{\mathrm{IR}} = q_{\mathrm{entropy}} \tag{4.17}$$

However, with the functionality of the background subtraction in mind, these values are not sufficient to describe the trust in each modality. Rapid changes in

the environment can highly disturb the algorithm. The changes can for example be rapid illumination changes for the RGB domain, or a change of the auto gain for the IR domain. These changes are not predictable by the knowledge base of the quality heuristics. Therefore another parameter needs to be introduced. When the background subtraction algorithm fails, a huge number of false positives appear. The resulting number of foreground pixels is much higher as the average of the scene. A quality heuristic based on this phenomena is defined in eq. (4.18), where τ defines the average foreground ratio, γ is an arbitrary scaling factor, $\mathbb{1}$ denotes an *indicator function* and (X, Y) are the image dimensions.

$$\begin{aligned} q_{\mathrm{fg}} &= \max(1 - \gamma(r_{\mathrm{fg}} - \tau), 0) \\ r_{\mathrm{fg}} &= \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} \mathbb{1}\left[D_{x,y} \geq 1\right] \end{aligned} \tag{4.18}$$

Given eq. (4.18) the trust can now be calculated as follows:

$$T_{\mathrm{RGB}} = \min(q_{\mathrm{fg}_{\mathrm{RGB}}}, q_{RGB}) \tag{4.19}$$

$$T_{\mathrm{IR}} = \min(q_{\mathrm{fg}_{\mathrm{IR}}}, q_{IR}) \tag{4.20}$$

To prevent artifacts, the trust in a modality only increases slowly after being rated down. When confronted with dramatic changes, the Gaussian background model needs some time to adapt for the changes. Therefore the trust should only slowly increase while the background is being learned. The trust T at time $t+1$ is calculated:

$$T_{t+1} = \begin{cases} T_{t+1} & \text{if } T_{t+1} \leq T_t \\ \alpha T_{t+1} + (1 - \alpha) T_t & \text{otherwise} \end{cases} \tag{4.21}$$

where α is the update speed of the Gaussian mixture model.

After normalizing the values of T_{RGB} and T_{IR} to add up to 1, the values are used the as weights for the adaptive fusion. Each pixel is fused in the distance map:

$$D_{\mathrm{F}} = w_{\mathrm{RGB}} D_{\mathrm{RGB}} + w_{\mathrm{IR}} D_{\mathrm{IR}} \tag{4.22}$$

with

$$\begin{aligned} w_{\mathrm{RGB}} &= \frac{T_{\mathrm{RGB}}}{T_{\mathrm{RGB}} + T_{\mathrm{IR}}} \\ w_{\mathrm{IR}} &= \frac{T_{\mathrm{IR}}}{T_{\mathrm{RGB}} + T_{\mathrm{IR}}} \end{aligned} \tag{4.23}$$

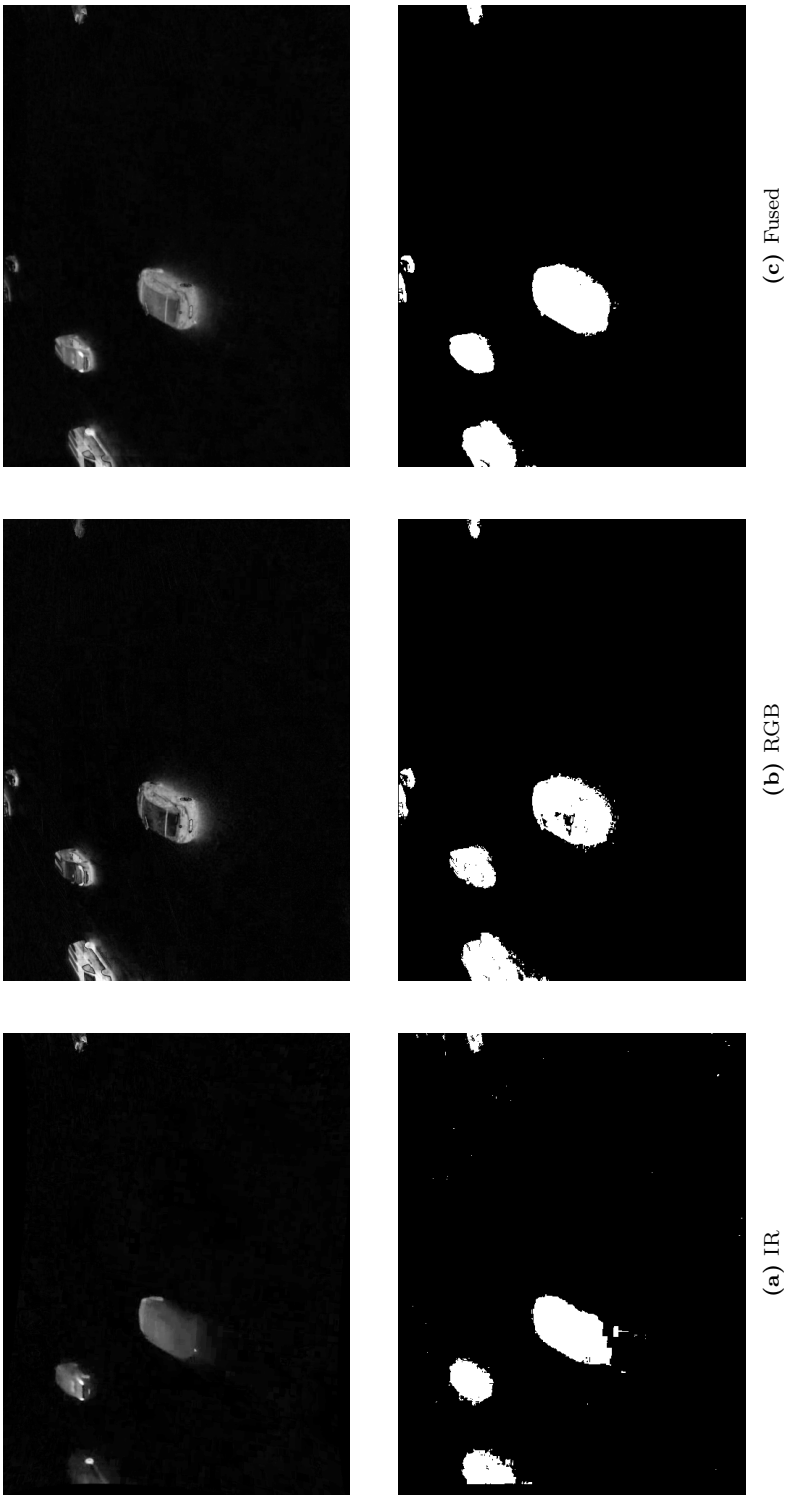
Through the weighting based on the quality respectively trust values, the fusion is adaptive and context aware.

At this stage spatial and temporal registration inaccuracies can be compensated. A simple mean filter applied on the fused distance map dissolves the pixel grid and therefore fuses information of neighboring pixels. Other functions are possible at this stage and have been applied, as further elaborated in Section 4.4.

The final step is the decision whether a pixel is foreground or background. As explained before, all values are scaled to the same level. A simple thresholding per pixel is performed:

$$FG = \begin{cases} 1 & \text{if } D_{\mathrm{F}} \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.24}$$

Figure 4.12 demonstrates the fusion and its effect on the resulting mask.



(a) IR **(b)** RGB **(c)** Fused

Figure 4.12: Distance maps of the different modalities and results after thresholding.

4.4 Application to Traffic Monitoring

The last section described the main contribution of this work. The presented method is an extension to the general adaptive Gaussian mixture model background subtraction method. No constrains have been introduced, so that the method is both applicable for indoor as well as outdoor scenarios. In the following specific extensions for traffic surveillance are presented, showing the modularity of the proposed algorithm. In Figure 4.1 (page 26) these extensions are labeled by roman number III.

4.4.1 Shadow Detection

A common extension to background modeling techniques is shadow detection. Shadows of intruding objects are found as not matching the background model as they appear darker as prior illuminated areas and are therefore declared as foreground. Depending on the purpose of the system, labeling shadow areas as foreground is a false positive. In most surveillance scenarios only the objects and not their shadow are of interest [Prati et al., 2003].

Different shadow detection algorithms have been presented. Prati et al. [2003] distinguish between *deterministic approaches* that use an 'on/off decision process' and *statistical approaches* that 'use probabilistic functions to describe the class membership'. Both methods however can fail and false negatives as well as false positives may occur. Is the task to identify all foreground objects, as it is in traffic surveillance, especially false positives harm the results. Whole objects may be classified as shadow and are therefore lost for further processing. To address this issue, shadow areas have been pruned rather than removed in this work. Since shadow detection is not in the scope of this work, a simple OpenCV implementation of a method presented by Prati et al. [2003] has been used.

As indicated in Figure 4.1 (page 26) the results of the shadow detection are fused into the RGB distance map. In state of the art methods a labeling in the

resulting foreground mask is performed. Instead of making this hard decision, the distance of areas marked as shadows has been scaled down. In this work a fixed scaling has been used. A scaling based on the shadow certainty may be a possible extension. Since the background distance correlates with the certainty of a pixel being foreground, the downscaling can be seen as bringing uncertainty to the decision. Consequentially the decision if a pixel is shadow is only made indirectly, when deciding whether the pixel is foreground or background.

The subsequent fusion of the modalities is the important step for this method to work. Objects that have also been found in the thermal image are most likely found anyway and shadows are voted further down as they are not present in the thermal domain. Especially small areas of false positives can be recovered as being a foreground object using this technique. The mean filter subsequent to the fusion helps the process with removing outliers. Additionally the quality functions allow to predict scenes with drop shadows. Therefore the process can be triggered context aware, only when shadows are most likely present.

4.4.2 Blob Prediction

As discussed in Section 3.1 a successful background subtraction for traffic surveillance must handle the different speeds of the traffic. All objects have to be handled as foreground even when staying in the scene for a longer time. For this purpose the blob prediction method proposed by Yao and Ling [2014] has been integrated in this work. The position of foreground blobs are predicted for each frame and the update rate α is significantly lowered for these areas. Consequentially objects have to stay for a very long time before merging into the background.

To predict blob positions for the current frame $t + 1$ blobs from t and $t - 1$ are matched. Afterwards the displacements between t and $t - 1$ is applied on t. The matching is done with a nearest neighbor search of the blob's centroids. If no neighbor within a range ρ is found, the blob is supposed to be stationary as no prediction about the movement can be made.

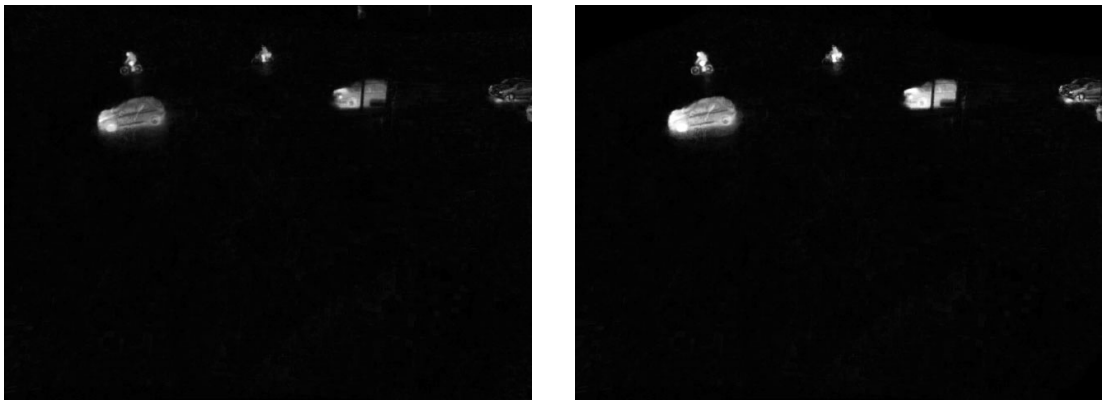


Figure 4.13: Distance map before and after blob prediction based modulation.

In contrast to the method of Yao and Ling [2014] an extension has been made. To prevent artifacts in the background model caused by inaccuracies in the blob prediction, the predicted blobs have been dilated and the edges have been smoothed out. The update rate α has then been calculated as:

$$\alpha = b\alpha_{\mathbf{fg}} + (1 - b)\alpha_{\mathbf{bg}} \tag{4.25}$$

where $0 \leq b \leq 1$ indicates the value in the blob prediction image and $\alpha_{\mathbf{fg}}$ and $\alpha_{\mathbf{bg}}$ are the update rates for foreground respectively background regions.

Another purpose of the blob prediction has been found in this work. Since the boundary of foreground objects changes only gradually, the predicted blobs are a very good estimate of the next frame's foreground. This can help the segmentation as it is more likely to find an object where it is predicted than elsewhere in the scene. Objects follow a trajectory and generally do not appear out of sudden. To express this characteristic another modification of the distance map is done. Analogous to the shadow suppression, predicted areas are scaled up in the distance map. Figure 4.13 demonstrates the effect. One can clearly see how the traffic stands out more clearly in the right image.

By taking the predicted blob areas into account for the decision whether a pixel belongs to the foreground, a spatial-temporal constraint is introduced. The decision is no longer pixel-wise but aware of the history of the whole image.

4.4.3 Scene Geometry based Prior Knowledge

The principle presented in the last sections can be used for another constraint. By looking at the scene geometry one can easily divide the image into three classes. The first class of pixels are areas where no foreground is expected under any circumstances. Examples may be trees or the sky. The second class of pixels describe the areas where objects may move to. A sudden appearance of objects is unlikely or even excluded but objects may move to these areas from other parts of the image. These areas are referenced as neutral zones. The last class describes areas where we expect foreground objects to appear. These areas are called entrance areas in the following. Entrance areas can normally be found at the borders of the image as objects enter the scenery normally from out of the camera's view port. Objects may however also reappear from occlusion or enter from occluded areas. Based on this classification a mask can be manually drawn as seen in Figure 4.14.

The scene classification is *prior knowledge* to the segmentation process. In the adaptive Gaussian mixture model background subtraction by Stauffer and Grimson [1999] all pixels are treated the same. The likelihood of a pixel being foreground is independent to its position. With the introduction of the scene classes this has been changed in this work. Only the entrance areas are treated like by Stauffer and Grimson [1999]. For the other two classes modulations of the distance maps have been made as presented before.

Firstly excluded areas are made impossible to be foreground by setting the corresponding values in the distance map to zero. Secondly the values for neutral zones are scaled down to make it less likely to find foreground pixels in these areas. This is possible because the blob positions have been predicted and uprated

Figure 4.14: Scene area classes. Green: entrance areas, red: excluded areas, rest: neutral.

beforehand. Areas where we expect objects to move to are untouched afterwards or even uprated while unpredicted regions are rated down. This helps to remove noise and find objects more reliable.

Chapter 5

Experiments

To evaluate the performance of the proposed algorithm a series of experiments have been conducted. Hereby both quantitative and qualitative performance have been tested. This chapter begins with an elaboration about the datasets that have been used in this work. A description of the performance metrics and the results of the experiments follows. Concluding an in-depth analysis of the qualitative performance is presented.

5.1 Datasets

The main dataset used in this work contains a large number of multimodal recordings of intersection in Northern Jutland recorded during the year 2013. The recordings have been kindly provided by the Department of Civil Engineering of the Aalborg University. During this work another recording session has been made archiving two goals. Firstly the recordings have been made in winter. Situations with sub-zero temperatures and snow were lacking in the dataset. Secondly the acquisition process can be described better in this work.

In preparation of a recording session two cameras, one thermal and one RGB camera, have been mounted on a high pole, typically a street light. The view ports have been adjusted to overlap as much as possible. The setup is displayed

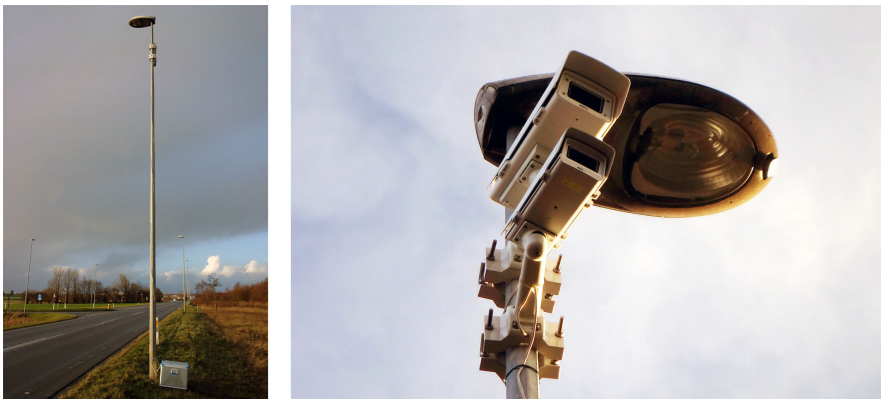


Figure 5.1: Installation for multimodal image acquisition.

in Figure 5.1. A laptop with a power supply for a couple of days has been used to store the captured scene. Through a special program to operate the cameras, one hour videos within a given time frame have been recorded. This way the experiment could run 2-7 days independently. The resulting dataset includes 9 different locations with a total of over 1580 hours recording material in a resolution of 640×480px with varying frame rates.

Since it is impossible to test such a vast amount of data, only roughly 20 scenes have been used during the development of the proposed algorithm. For the experiments a set of 5 different scenes have been selected. The scenes contain different conditions, have been recorded at different locations and therefore represent a good overview over the spectrum of outdoor recording conditions.

To be able to benchmark the proposed algorithm, two commonly used datasets have been additionally used. The *OSU Color-Thermal Database* [Davis and Sharma, 2007] of the *OTCBVS Benchmark Dataset Collection* contains RGB and thermal data of two surveillance scenarios. The videos contain pedestrians recorded on the campus of the Ohio State University. The *INO Video Analytics Dataset*[1] contains a set of multimodal recordings of parking lot situations including cars, cyclists and pedestrians. All scenes that have been tested during the experiments of this work are listed in Tables 5.1 and 5.2.

[1] `http://www.ino.ca/en/video-analytics-dataset/` Last downloaded June 1, 2015

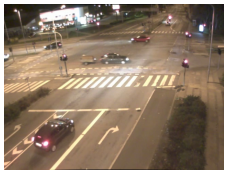
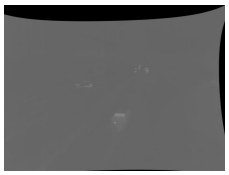
RGB	IR	Name	Description
		Day	Good conditions. High contrast in thermal domain.
		Night	Low light. Reflections and dark objects.
		Auto Gain	AGC of thermal camera during the scene.
		Heavy Rain	Rain and storm. Reflections and reduced view.
		Snow	Reduced view and low contrast in thermal domain.

Table 5.1: Test scenes from the Department of Civil Engineering dataset.

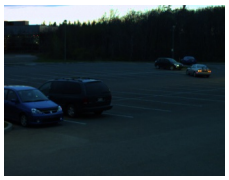
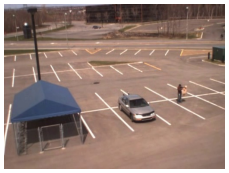
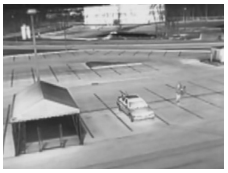
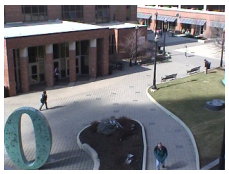
RGB	IR	Name	Description
		INO ParkingEvening	Low light. Dark car appears in the scene.
		INO ParkingSnow	Persons between the cars not visible in RGB.
		INO CoatDeposit	Objects come into the scene and stay.
		INO TreesAndRunner	Twighlight and moving trees.
		OTCBVS 3	Changing illumination due to moving clouds.
		OTCBVS 4	Good conditions. High contrast in thermal domain.

Table 5.2: Test scenes from the benchmark datasets.

5.2 Performance Metrics

The following section explains the quantitative performance metrics that have been used for the evaluation of the experiments. The quality of a segmentation algorithm is commonly determined by two quality measures. *Good detection* means that most foreground pixels of the image have actually been found by the algorithm. *Good discrimination* means a good distinction has been made i.e. not many pixels have been declared as foreground erroneously.

In order to calculate the two quality measures three decision states are identified. *True positives (TP)* are all pixels that have been correctly identified as foreground. Two different error states are distinguished. *False negatives (FN)* are pixels that have not been identified as foreground while actually being part of it, while *false positives (FP)* are pixels that have been misclassified as foreground. With these numbers the two metrics *Detection Rate (DR)* and *False Alarm Rate (FAR)* are defined as follows:

$$DR = \frac{TP}{TP + FN} \tag{5.1}$$

$$FAR = \frac{FP}{TP + FP} \tag{5.2}$$

The Detection Rate is also known under *recall* or *true positive rate* and describes the *sensitivity* of a detector. The False Alarm Rate corresponds to $1 - p$ where p is the detector's *precision* or *specificity*.

In order to evaluate the performance metrics it is required to have access to the true data, commonly referred as *Ground Truth (GT)*. Ground Truth has to be created manually and is a laborious task. Thus only a small sample of the results can be tested. In this work 70 successive frames have been annotated for each test set. Only for the *Auto Gain* set have been annotated 180 frames in order to cover the whole process.

5.3 Quantitative Experiments

In order to evaluate the performance of the proposed method extensive experiments have been performed and evaluated with the described performance metrics. Besides with the algorithm itself, each dataset has been processed with four alternative strategies. Each strategy bases on the Gaussian mixture model background subtraction algorithm presented by Stauffer and Grimson [1999] and improved by Zivkovic [2004]. Firstly both modalities RGB and IR are processed individually. Next a pixel-wise fusion is performed with the creation of RGBT frames. And finally the confidence based selection presented by Serrano-Cuerda et al. [2014] has been implemented.

As this works aspires to create a system that works without the requirement to manually tune its parameters for different conditions, one set of parameters has been defined for all test sequences. In practice however parameters may be tuned to fit the given location and situation. For comparability reasons this fine tuning has not been done in this work. Solely the learning time for each scene has been adjusted to match the specific situation. For example scenes with much traffic need more time to learn a stable background model. For the case of the presented algorithm, both background models first have been learned individually before the described adjustments have been made. This is necessary since predicted foreground regions are learned much slower and false positives are very likely during the learning phase.

The update rate α has been set to be slower for the alternative strategies. Since the state of the art background modeling does not differ between foreground and background in the update step, a quick update rate would result in foreground objects merging into the background. This is also the case in the learning phase of the proposed method. Consequentially the same α has been used here. All important experimental parameters are listed in Table 5.3, where the parameters below the line only apply for the proposed method.

Parameter	Value	Description
α	0.0005	Update rate
K	5	Number of components
λ	4	Number of standard deviations for background acceptance
α_{BG}	0.0033	Background update rate
α_{FG}	0.000033	Foreground update rate
τ	0.1	Foreground ratio
γ	5.0	Foreground deviation weight
ρ	17	Blob match radius (px)
s_{shadow}	0.3	Distance scaling factor for shadow regions
s_{predict}	1.5	Distance scaling factor for predicted regions
s_{neutral}	0.5	Distance scaling factor for neutral regions

Table 5.3: Parameters used in the experiments. The parameters below the line only apply for the proposed method.

For all experiments containing RGB data shadow detection has been performed. The OpenCV implementation described by Prati et al. [2003], the same algorithm as in the proposed method, has been used. Pixels that have been found to be shadow have been classified as being background in the reference methods. Also the region masks have been applied on the resulting data. Excluded areas have been removed from the resulting foreground masks and the resulting masks have been cleaned with morphological operations and hole closing. This way equal conditions have been created for all strategies and differences in the results of the proposed algorithm in contrast to the state of the art methods can be explained by its core contributions.

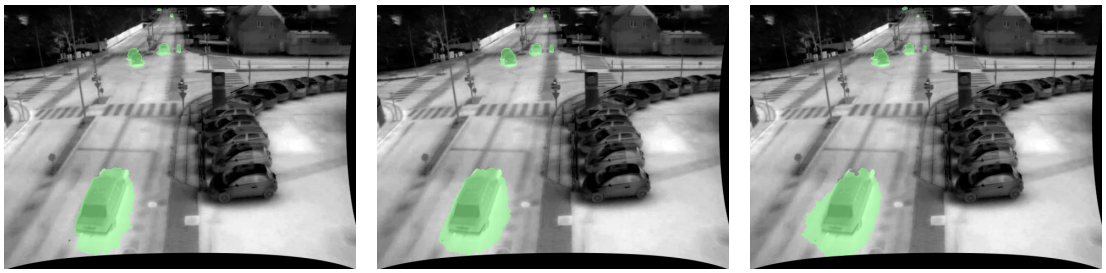
The scenes have been selected as each of them introduces a new scenario with different conditions. State of the art methods often aim to perform best for one special scenario. The proposed method however aims for a good performance *in general*. Therefore a comparison to recent special case methods with best known performance was not desired and even not necessary. The comparison to background modeling on the other hand reveals interesting insights on strengths and weaknesses of the proposed method.

The results of the experiments are displayed in Table 5.4. The general performance of the proposed algorithm can be considered very good due to a minimum Detection Rate of 0.84 and maximum False Alarm Rate of 0.56. It can clearly be seen, that the goal of creating a robust method for a wide bandwidth of condition is achieved. Only the proposed method shows a good performance for every test sequence. The alternative strategies fail at different scenarios, but also better performance as of the proposed method can be seen. The reasons for this are manifold and will be discussed in the following section.

As expected, all fusion approaches show in general a better performance than the single modality methods. The method presented by Serrano-Cuerda et al. [2014] also performs well on the first look. When analyzing the results in detail however, one can easily see, that the results are at most as good as one of the single modalities. This is funded in the design of the algorithm, as it selects one

		Proposed	RGB	IR	RGBT	Select
Day	DR	**0.99**	0.93	0.95	0.97	0.93
	FAR	0.30	**0.09**	0.31	0.29	**0.09**
Night	DR	0.84	0.78	0.48	**0.89**	0.78
	FAR	**0.31**	0.69	0.32	0.66	0.69
Auto Gain	DR	**0.94**	0.86	0.73	0.91	0.81
	FAR	0.25	**0.09**	0.76	0.40	0.58
Heavy Rain	DR	**0.92**	0.46	0.69	0.48	0.69
	FAR	0.22	0.26	**0.11**	0.27	**0.11**
Snow	DR	**0.96**	0.79	0.21	0.92	0.21
	FAR	0.52	0.52	**0.25**	0.55	**0.25**
INO ParkingEvening	DR	0.94	0.93	0.91	**0.95**	0.91
	FAR	0.24	0.27	**0.18**	0.29	0.18
INO ParkingSnow	DR	0.98	0.86	**0.99**	0.96	**0.99**
	FAR	**0.32**	0.78	0.40	0.35	0.40
INO CoatDeposit	DR	**0.97**	0.10	0.10	0.10	0.10
	FAR	0.19	**0.12**	0.30	0.16	0.12
INO TreesAndRunner	DR	**0.94**	0.88	0.84	0.93	0.84
	FAR	0.44	0.65	**0.36**	0.70	**0.36**
OTCBVS 3	DR	**0.95**	0.75	0.94	0.90	0.78
	FAR	**0.56**	0.96	0.74	0.96	0.93
OTCBVS 4	DR	**1.00**	0.94	0.78	0.99	0.78
	FAR	0.55	**0.15**	0.68	0.48	0.68

Table 5.4: Experimental results. The best DR and FAR values of each set are marked bold. Proposed method compared to GMM background subtraction of RGB, IR, and RGBT frames. "Select" indicates result selection based on quality heuristics [Serrano-Cuerda et al., 2014].



(a) IR **(b)** Proposed Method **(c)** RGBT

Figure 5.2: Results of the thermal halo effect and false positive propagation in different approaches.

result of two parallel pipelines. One important characteristic of fusion algorithms is neglected by this design flaw. Fused data or results generally differ from its inputs and therefore contain new features and information. A simple selection obviously makes this impossible. As a result the approach is beaten in 10 out of 11 cases in terms of Detection Rate.

The False Alarm Rate of both, the proposed method and the RGBT approach mirror the weaker modality. The reason is, that a high evidence of foreground in one modality is still present after fusion the data. Only false positives based on weak evidence are successfully smoothed out. In worst case false positives from both modalities are present in the result. The effect is shown in Figure 5.2. It can be seen how the errors of the IR segmentation propagate into the results of the fusion approaches.

A good example for the superiority of fusion approaches in terms of Detection Rate is given by the sequence *INO TreesAndRunner*. Obviously both fusion approaches, the proposed method and the RGBT approach, perform much better than the single modalities. This is the case because both RGB and IR contain frames, that are very hard to segment. The runner passes trees and other objects. The fusion approaches can still rely on the second modality, when the information content of the first is low.

5.4 Special Situation Performance

In the following results of specific test sequences are elaborated in detail. It is shown in which way different details of the design of the proposed algorithm effect the performance. Four different problems that arise during outdoor surveillance are discussed. Emphasis has been put on the adaptive modality weighting of the proposed algorithm and its effect on the segmentation results. To begin with, this context awareness is discussed further.

5.4.1 Context Awareness

One of this work's core contributions is the context awareness of the fusion. It is based on a set of quality heuristics that have been defined in Section 4.2. The goal is to evaluate the *usefulness* of each modality. Instead of using information from the images itself, outside sources have been consulted. Solely the thermal domain has been rated by its own information content. For the tested sequences, the weights calculated with the help of the heuristics are more or less fixed. The time frames are simply too short to see an effect based on for example the altitude heuristic function. The overall concept however has been tested by selecting scenes with various conditions, such as day, night, twilight, heavy rain and snow.

In Figures 5.3 and 5.4 quality functions covering a full day are plotted. Since video data are only present from 5 a.m. to 10 p.m., the IR quality function is only defined for that particular time frame. The tested day was a hot summer day with rather good weather. Because of overcast, no cast shadows have been assumed in the morning. During noon time the temperature was so high, that the thermal camera was overexposed. This problem is represented in the graph through the drop of the entropy based quality heuristic.

The resulting weights for the fusion of the modalities are plotted in Figure 5.5. The foreground quality functions are not considered, due to the necessary computation time. It should be noted that the steep changes in both modalities are

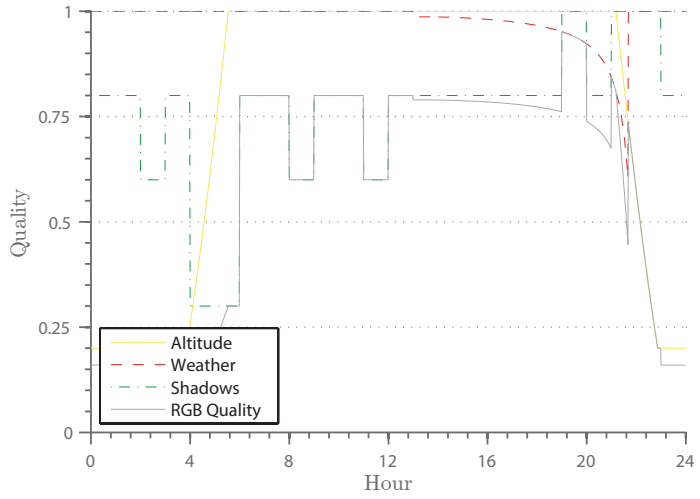


Figure 5.3: RGB quality heuristics and resulting quality function of a full day.

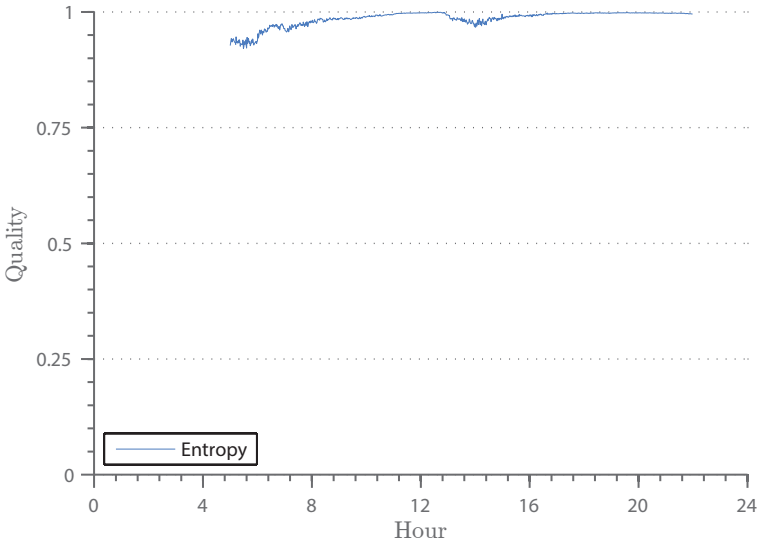


Figure 5.4: Entropy based quality heuristic for thermal domain of a full day.

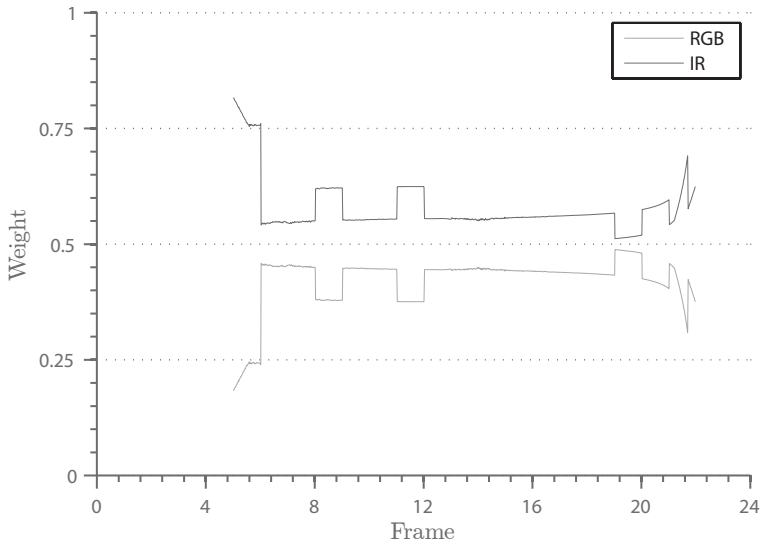
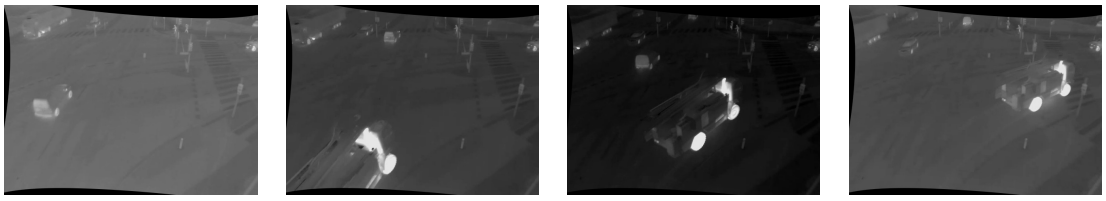


Figure 5.5: Varying weights of the two modalities over a full day.

only present due to the resolution. In practice the weights change rather smooth caused by the auto averaging of the trust functions.

5.4.2 Automatic Gain Control

The first problem that will be discussed further, is the prior mentioned Automatic Gain Control of thermal cameras. When large or hot objects enter the scenery, the camera automatically adjusts its gain in order to preserve a high level of detail. This behavior however highly disturbs the background subtraction. The new background does not longer fit the model, resulting in a high number of false positive foreground pixels. Figure 5.6 displays the described phenomena. The challenge is hereby the short time frame of the adjustment. When the objects leave the scene, the camera adjusts the gain back. Therefore the problem often only persists for 100-200 frames, while highly affecting the segmentation results. Also the background model is affected, since it is updated each frame, which makes it invalid for following frames.

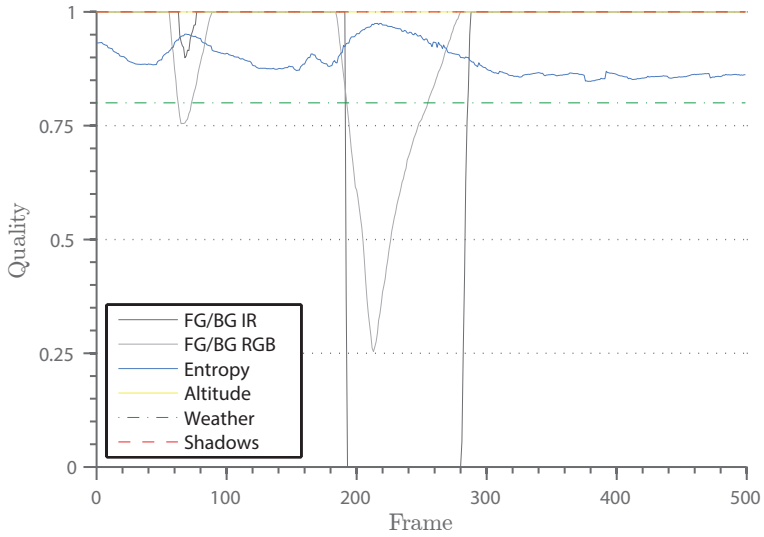


(a) Frame 170 **(b)** Frame 200 **(c)** Frame 230 **(d)** Frame 260

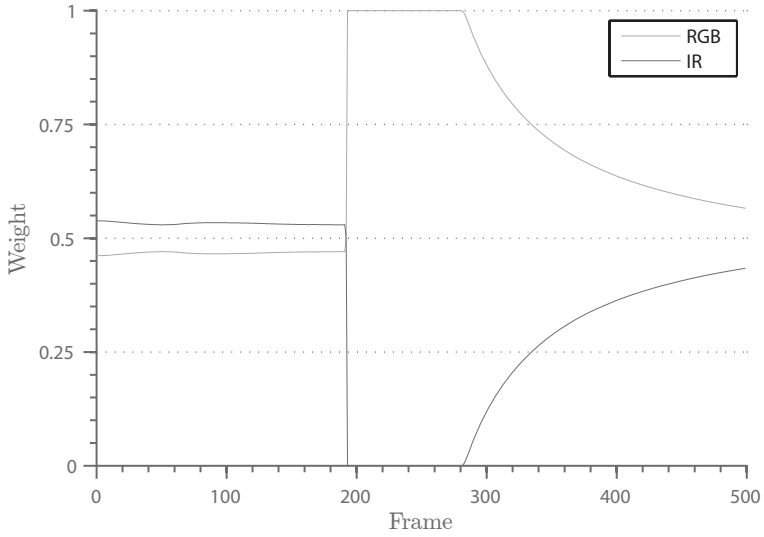
Figure 5.6: AGC of the IR camera triggered by a big truck coming into the scenery.

As seen in Table 5.4 the proposed algorithm handles the described problem well. No segmentation quality reduction can be detected from the raw numbers. The reason for this is the adaptive weighting performed in the fusion step. Through the foreground ratio evaluation described in Section 4.3.3 is detected, that the background model of the thermal domain is invalid. As a result, the IR weight function drops to zero and the segmentation only relies on the RGB domain. Figure 5.7 displays this behavior. It can clearly be seen, how the weight of the IR domain drops parallel to the quality heuristic. After the truck leaves the scene, the camera adjusts back to normal and the quality function instantly rises. The weight however increases only gradually. This is necessary in order to give the background model time, to relearn the background model.

The foreground ratio quality function also decreases while the truck enters the scene. This behavior is normal, since the truck covers a lot of the scene and therefore the number of foreground pixels is unnaturally high. In the end result this however has no effect, since the overall trust in the RGB domain is still higher.

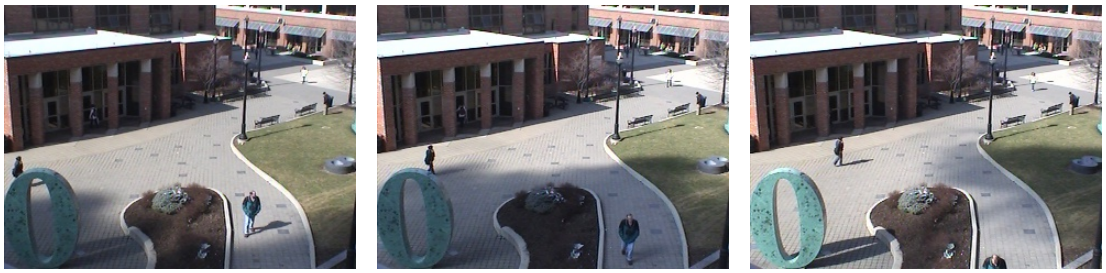


(a) RGB and IR quality heuristics.



(b) Resulting weights.

Figure 5.7: Plot of the effect of the IR ACG on the quality heuristics and modality weights.



(a) Frame 375 **(b)** Frame 425 **(c)** Frame 475

Figure 5.8: Moving cloud casting a large shadow onto the scene.

5.4.3 Changing Illumination

A very similar problem that is common in outdoor surveillance is changing illumination. Although the design of the Gaussian mixture model background subtraction allows variations in the illumination, problems can still arise when abrupt changes happen. This for example is the case when clouds cover or uncover the sun. The algorithm is only designed to adapt for slow changes, e.g. shadows that move over the day. Fast changes of the scenery will cause foreground objects due to the definition of the process.

Figure 5.8 shows an example of such a situation from the *OTCBVS 3* test set. Similar to the problem discussed before, the foreground ratio of the RGB domain is rising, since the background model does not adapt fast enough for these changes. Consequentially a weight shift to the thermal domain is performed by the algorithm, leading to the comparatively low False Alarm Rate of 0.56.

5.4.4 Artifact Reduction

Another contribution of the proposed algorithm can be seen in the results of *OTCBVS 3*. With 0.56 the False Alarm Rate is even lower than the results of the thermal background subtraction. This can be reasoned with the adjustments made to the fused distance map based on the scene geometry. Artifacts are unlikely to appear since unpredicted foreground regions are reduced in the distance

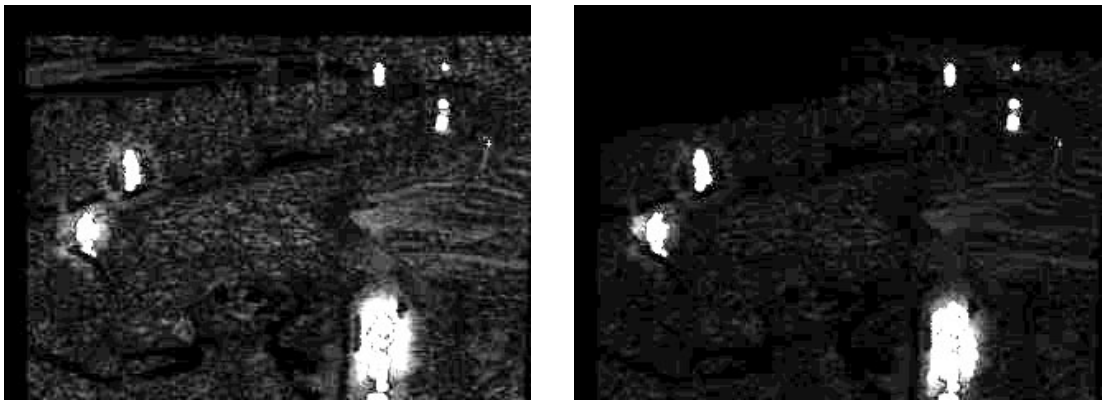
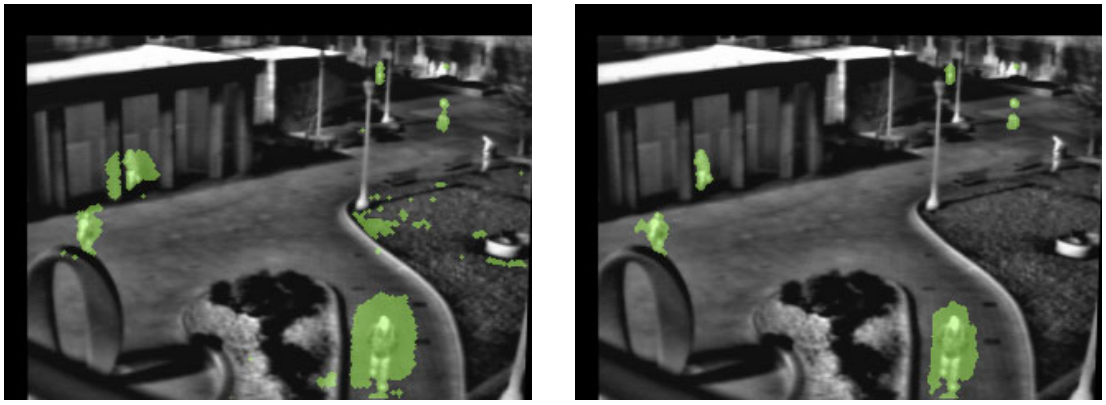


Figure 5.9: Distance map before and after scene geometry based modulation.

map. Figure 5.9 illustrates the principle with a side-by-side comparison of the distance map before and after applying the modulations. The effect on the foreground mask can clearly be seen in Figure 5.10 in comparison to the approach using solely the thermal domain.

5.4.5 Long Staying Objects

The Gaussian mixture model background subtraction presented by Stauffer and Grimson [1999] assumes that foreground objects are constantly in motion. For traffic this is obviously not the case. This has been addressed by Yao and Ling [2014] and the proposed method has been integrated in this work. The original algorithm causes long staying objects to gradually merge into the background. This problem is very much visible in the *INO CoatDeposit* test set. The car coming into the scene merges into the background within a few frames as seen in Figure 5.11. This merging is stopped by predicting foreground regions and lowering their update speed, resulting in a significantly better Detection Rate of the proposed method. The difference is shown in Figure 5.12.



(a) Artifacts in IR mask. **(b)** Reduced artifacts by proposed method.

Figure 5.10: Artifacts in resulting masks of two approaches.

Figure 5.11: Stopped car merges into the background.

Figure 5.12: Background merging prevented through blob prediction and α adjustment.

Chapter 6

Conclusion and Future Work

This work aimed to investigate on how the fusion of thermal and RGB video data can assist camera based traffic monitoring. Related work has been studied and a suitable starting point has been identified. Background subtraction is the first and crucial step of many surveillance algorithms. A stable segmentation lays the base for successful successive analysis. Different environmental conditions are hereby the most challenging problem. Only very few methods have been presented working under extreme situations, although the interest in these methods is high. The goal of this work was therefore to present a method for stable background subtraction as far as possible independent from environmental conditions and thereby laying the base for persistent traffic monitoring.

A new approach to Gaussian mixture model background subtraction using two modalities has been presented. The proposed algorithm fuses an image representation of background model conformity values of each pixel. This allows to include prior knowledge about the modalities in form of quality characteristics. Different image quality heuristics based on image structure and external information sources have been investigated and specified. To match requirements derived from the purpose of traffic monitoring, extensions to the core contribution have been introduced.

The proposed method has been thoroughly tested. The results show a significantly better performance of the proposed method than comparable background subtraction algorithms. Special situation performance suggests that the strategy of including image quality heuristics in the segmentation process has great potential. The modularity of the method allows the implementation of improvements that have been developed originally for the background subtraction by Stauffer and Grimson [1999].

A common problem of image fusion techniques can be seen from the experimental results. Although the algorithm features a suppression of false positives, a propagation to the fused mask can still be noticed. This is especially the case, when the quality rating of the two modalities is similar and therefore information fuses in equal proportions. Based on this observation further development of the proposed method can be derived. Serrano-Cuerda et al. [2014] perform a switch based on image quality heuristics. This work performs an adaptive fusion. The next logical step would be to perform the fusion adaptive per image region. Similar approaches based on image characteristics rather than quality heuristics have been presented. The approaches however do not differentiate between different conditions such as night and day. By specifying quality heuristics for image samples, information about shadows and different lighting conditions within the scene, the main reasons for false positives, could be considered.

The work has been limited to the usage of RGB and thermal imagery. The algorithm however can easily be adapted to work with different imaging sensors. A setup of the proposed system in combination with sensors helping to estimate the image quality would also be an interesting extension. Weather stations and street temperature sensors would enable the heuristics to work much more accurately.

Bibliography

Al Najjar, M., Ghantous, M., and Bayoumi, M. (2014). *Video Surveillance for Sensor Platforms*. Springer.

Alemán-Flores, M., Alvarez, L., Gomez, L., and Santana-Cedrés, D. (2014). Line detection in images showing significant lens distortion and application to distortion correction. *Pattern Recognition Letters*, 36:261–271.

Bas, E., Tekalp, A. M., and Salman, F. S. (2007). Automatic vehicle counting from video for traffic flow analysis. In *Intelligent Vehicles Symposium, 2007 IEEE*, pages 392–397. IEEE.

Bi, L., Tsimhoni, O., and Liu, Y. (2009). Using image-based metrics to model pedestrian detection performance with night-vision systems. *Intelligent Transportation Systems, IEEE Transactions on*, 10(1):155–164.

Bouguet, J. Y. (2008). Camera calibration toolbox for Matlab.

Bradski, G. (2000). *Dr. Dobb's Journal of Software Tools*.

Brown, L. G. (1992). A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376.

Buch, N., Velastin, S. A., and Orwell, J. (2011). A review of computer vision techniques for the analysis of urban traffic. *Intelligent Transportation Systems, IEEE Transactions on*, 12(3):920–939.

Chapman, L. and Thornes, J. (2006). A geomatics-based road surface temperature prediction model. *Science of the Total Environment*, 360(1):68–80.

Chen, S. and Leung, H. (2009). An em-ci based approach to fusion of ir and visual images. In *Information Fusion, 2009. FUSION'09. 12th International Conference on*, pages 1325–1330. IEEE.

Chen, T.-H., Chen, J.-L., Chen, C.-H., and Chang, C.-M. (2007a). Vehicle detection and counting by using headlight information in the dark environment. In *Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on*, volume 2, pages 519–522. IEEE.

Chen, T.-H., Lin, Y.-F., and Chen, T.-Y. (2007b). Intelligent vehicle counting method based on blob analysis in traffic surveillance. In *Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on*, pages 238–238. IEEE.

Cheung, S.-C. S. and Kamath, C. (2005). Robust background subtraction with foreground validation for urban traffic video. *EURASIP Journal on Advances in Signal Processing*, 2005(14):2330–2340.

Conaire, C. O., O'Connor, N. E., Cooke, E., and Smeaton, A. F. (2006). Comparison of fusion methods for thermo-visual surveillance tracking. In *FUSION*, pages 1–7.

Dalaff, C., Reulke, R., Kroen, A., Kahl, T., Ruhe, M., Schischmanow, A., Schlotzhauer, G., and Tuchscheerer, W. (2003). A traffic object detection system for road traffic measurement and management. In *Proc. Image and Vision Computing New Zealand*, pages 78–83. Citeseer.

Davis, J. W. and Sharma, V. (2007). Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding*, 106(2):162–182.

Doshi, A. and Trivedi, M. M. (2007). Satellite imagery based adaptive background models and shadow suppression. *Signal, Image and Video Processing*, 1(2):119–132.

Gade, R. and Moeslund, T. B. (2014). Thermal cameras and applications: a survey. *Machine vision and applications*, 25(1):245–262.

Goubet, E., Katz, J., and Porikli, F. (2006). Pedestrian tracking using thermal infrared imaging. In *Defense and Security Symposium*, pages 62062C–62062C. International Society for Optics and Photonics.

Gregory, R. L. (1997). *Eye and brain: The psychology of seeing* . Princeton university press.

Hall, D. and Llinas, J. (2001). *Multisensor data fusion*. CRC press.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press.

Heather, J. P. and Smith, M. I. (2005). Multimodal image registration with applications to image fusion. In *Information Fusion, 2005 8th International Conference on*, volume 1, pages 8–pp. IEEE.

Hrkać, T., Kalafatić, Z., and Krapac, J. (2007). Infrared-visual image registration based on corners and hausdorff distance. In *Image Analysis*, pages 383–392. Springer.

Istenic, R., Heric, D., Ribaric, S., and Zazula, D. (2007). Thermal and visual image registration in hough parameter space. In *Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on*, pages 106–109. IEEE.

Iwasaki, Y., Kawata, S., and Nakamiya, T. (2011). Robust vehicle detection even in poor visibility conditions using infrared thermal images and its appli-

cation to road traffic flow monitoring. *Measurement Science and Technology*, 22(8):085501.

Iwasaki, Y., Misumi, M., and Nakamiya, T. (2013). Robust vehicle detection under various environmental conditions using an infrared thermal camera and its application to road traffic flow monitoring. *Sensors*, 13(6):7756–7773.

Jackson, S., Miranda-Moreno, L. F., St-Aubin, P., and Saunier, N. (2013). Flexible, mobile video camera system and open source video analysis software for road safety and behavioral analysis. *Transportation Research Record: Journal of the Transportation Research Board*, 2365(1):90–98.

Kamijo, S., Matsushita, Y., Ikeuchi, K., and Sakauchi, M. (2000). Traffic monitoring and accident detection at intersections. *Intelligent Transportation Systems, IEEE Transactions on*, 1(2):108–118.

Kastrinaki, V., Zervakis, M., and Kalaitzakis, K. (2003). A survey of video processing techniques for traffic applications. *Image and vision computing*, 21(4):359–381.

Ki, Y.-K. and Lee, D.-Y. (2007). A traffic accident recording and reporting model at intersections. *Intelligent Transportation Systems, IEEE Transactions on*, 8(2):188–194.

Kwon, H., Der, S. Z., and Nasrabadi, N. M. (2002). Adaptive multisensor target detection using feature-based fusion. *Optical Engineering*, 41(1):69–80.

Lallier, E. and Farooq, M. (2000). A real time pixel-level based image fusion via adaptive weight averaging. In *Information Fusion, 2000. FUSION 2000. Proceedings of the Third International Conference on*, volume 2, pages WEC3–3. IEEE.

Lei, M., Lefloch, D., Gouton, P., and Madani, K. (2008). A video-based real-time vehicle counting system using adaptive background method. In *Signal Image*

Technology and Internet Based Systems, 2008. SITIS'08. IEEE International Conference on, pages 523–528. IEEE.

Messelodi, S., Modena, C. M., and Zanin, M. (2005). A computer vision system for the detection and classification of vehicles at urban road intersections. *Pattern analysis and applications*, 8(1-2):17–31.

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., and Mathers, C. (2004). World report on road traffic injury prevention.

Power, P. W. and Schoonees, J. A. (2002). Understanding background mixture models for foreground segmentation. In *Proceedings image and vision computing New Zealand*, volume 2002, pages 10–11.

Prati, A., Mikic, I., Trivedi, M. M., and Cucchiara, R. (2003). Detecting moving shadows: algorithms and evaluation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(7):918–923.

Robert, K. (2009). Night-time traffic surveillance: A robust framework for multi-vehicle detection, classification and tracking. In *Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 1–6. IEEE.

Rowe, D. (2008). *Towards robust multiple-tracking in unconstrained human-populated environments*. PhD thesis, Universitat Autònoma de Barcelona.

Serrano-Cuerda, J., Fernández-Caballero, A., and López, M. T. (2014). Selection of a visible-light vs. thermal infrared sensor in dynamic environments based on confidence measures. *Applied Sciences*, 4(3):331–350.

Shah, P., Merchant, S., and Desai, U. B. (2010). Fusion of surveillance images in infrared and visible band using curvelet, wavelet and wavelet packet transform. *International Journal of Wavelets, Multiresolution and Information Processing*, 8(02):271–292.

St-Laurent, L., Maldague, X., and Prévost, D. (2007). Combination of colour and thermal sensors for enhanced object detection. In *Information Fusion, 2007 10th International Conference on*, pages 1–8. IEEE.

Stauffer, C. and Grimson, W. E. L. (1999). Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE.

Szeliski, R. (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

United Nations (2014). General assembly resolution 68/269. *Improving global road safety*, A/RES/68/269.

Vargas, M., Toral, S., Barrero, F., and Milla, J. (2008). An enhanced background estimation algorithm for vehicle detection in urban traffic video. In *Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on*, pages 784–790. IEEE.

Wren, C. R., Azarbayejani, A., Darrell, T., and Pentland, A. P. (1997). Pfinder: Real-time tracking of the human body. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):780–785.

Yao, L. and Ling, M. (2014). An improved mixture-of-gaussians background model with frame difference and blob tracking in video stream. *The Scientific World Journal*, 2014.

Zangenehpour, S., Miranda-Moreno, L. F., and Saunier, N. (2014). Automated classification in traffic video at intersections with heavy pedestrian and bicycle traffic. In *Transportation Research Board 93rd Annual Meeting*, number 14-4337.

Zhou, J., Gao, D., and Zhang, D. (2007). Moving vehicle detection for automatic traffic monitoring. *Vehicular Technology, IEEE Transactions on*, 56(1):51–59.

Zivkovic, Z. (2004). Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 28–31. IEEE.

Zou, Y., Shi, G., Shi, H., and Wang, Y. (2009). Image sequences based traffic incident detection for signaled intersections using hmm. In *Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on*, volume 1, pages 257–261. IEEE.

Bibliography

Appendix A

Solar elevation angle

The following formulas describe the stepwise calculation of the solar elevation angle[1].

$$LSTM = 15° \cdot \Delta T_{GMT} \qquad \text{Local Standard Time Meridian}$$

$$B = \frac{360}{365}(d - 81)$$

$$EoT = 9.87 \cdot sin(2B) - 7.52 \cdot cos(B) - 1.5 \cdot sin(B) \qquad \text{Equation of Time}$$

$$TC = 4(\varphi - LSTM) + EoT \qquad \text{Time Correction Factor}$$

$$LST = LT + \frac{TC}{60} \qquad \text{Local Solar Time}$$

$$HRA = 15°(LST - 12) \qquad \text{Hour Angle}$$

[1] http://pveducation.org/pvcdrom/properties-of-sunlight/suns-position
Last downloaded June 1, 2015

Appendix A. Solar elevation angle

$$\delta = 23.45^{\circ} sin(B) \qquad \text{Declination}$$

$$\alpha = sin^{-1}(sin(\delta)sin(\phi) + cos(\delta)sin(\phi)cos(HRA)) \qquad \text{Elevation}$$

ΔT_{GMT} : Time difference to Greenwich Mean Time in hours

d : Days since start of the year

φ : Longitude

Appendix B

Weather Condition Grouping

Good conditions	Clear
Low/Varying lighting conditions	Overcast
	Partly Cloudy
	Mostly Cloudy
	Scattered Clouds
	Squalls
	Funnel Cloud
	Light Mist
	Light Drizzle
Reflections/moisture	[Heavy] Drizzle
	[Light] Rain
	[Light] Rain Showers
	[Light/Heavy] Freezing Drizzle
	[Light] Freezing Rain
	[Heavy] Mist

Particle occlusion/precipitation	[Light/Heavy] Snow
	[Light/Heavy] Snow Grains
	[Light/Heavy] Ice Crystals
	[Light/Heavy] Ice Pellets
	[Light/Heavy] Hail
	[Light/Heavy] Low Drifting Snow
	[Light/Heavy] Blowing Snow
	Heavy Rain Showers
	[Light/Heavy] Snow Showers
	[Light/Heavy] Snow Blowing Snow Mist
	[Light/Heavy] Ice Pellet Showers
	[Light/Heavy] Hail Showers
	[Light/Heavy] Small Hail Showers
	[Light/Heavy] Thunderstorm
	[Light/Heavy] Thunderstorms and Rain
	[Light/Heavy] Thunderstorms and Snow
	[Light/Heavy] Thunderstorms and Ice Pellets
	[Light/Heavy] Thunderstorms with Hail
	[Light/Heavy] Thunderstorms with Small Hail
	Heavy Freezing Rain
	Small Hail

	[Light/Heavy] Fog
	[Light/Heavy] Fog Patches
	[Light/Heavy] Smoke
	[Light/Heavy] Volcanic Ash
	[Light/Heavy] Widespread Dust
	[Light/Heavy] Sand
	[Light/Heavy] Haze
	[Light/Heavy] Spray
	[Light/Heavy] Dust Whirls
Reduced visibility	[Light/Heavy] Sandstorm
	[Light/Heavy] Low Drifting Widespread Dust
	[Light/Heavy] Low Drifting Sand
	[Light/Heavy] Blowing Widespread Dust
	[Light/Heavy] Blowing Sand
	[Light/Heavy] Rain Mist
	[Light/Heavy] Freezing Fog
	Patches of Fog
	Shallow Fog
	Partial Fog

Appendix B. Weather Condition Grouping

Appendix C

Thesis Journal Version

Appendix C. Thesis Journal Version

Machine Vision and Applications (2015) manuscript No.
(will be inserted by the editor)

Information based multimodal Background Subtraction for Traffic Monitoring Applications

Thiemo Alldieck · Chris Bahnsen · Thomas B. Moeslund

Received: date / Accepted: date

Abstract A new approach to background subtraction for multimodal systems is presented. The work is an extension to the Gaussian mixture model background subtraction of Stauffer and Grimson. The background conformity values of two image sources, namely thermal and RGB, are fused in order to enable stable background subtraction for persistent surveillance. Image quality heuristics based on image characteristics and external sources are specified to evaluate the *usefulness* of the modalities and perform the fusion *context aware*. Extensions for the use of the system for the purpose of traffic monitoring are presented. Therefor modulations of a new image representation of the conformity of pixels with the background model are made. The potential of the proposed method has been shown during excessive tests of quantitative and qualitative characteristics.

Keywords Background subtraction · Image Fusion · Traffic Surveillance · Context aware

1 Introduction

The growing congestion of public roads and associated problems lead to a growing need of accurate traffic information. This information can be used for traffic safety analysis, early incident detection, improvement of infrastructure capacity and localization of infrastructural weaknesses. Particularly road safety is an important subject for traffic researchers. The number of worldwide traffic crashes and injuries is growing and the impact on society is high. Not only costs for medical treatment accrue from an injury, but also physiological complications may result for the victim and his family and friends. Especially vulnerable road users, such as pedestrians, cyclists and motorcyclists, are at high risk of road traffic casualties [22]. Therefore, already the purpose of safety analysis justifies the research on new traffic monitoring systems and algorithms.

Different traffic monitoring systems have been developed and used over the years. Hereby has been shown that the use of cameras offers significant improvements over other systems such as inductive loops or microwave detectors. Video surveillance offers a wide band of analysis possibilities as for instance traffic flow, turning movements and vehicle classification, while being comparatively cheap in acquisition and installation [17]. Image processing techniques play hereby an important role as they provide added value to the raw data enabling automatic extraction of relevant information [2].

Computer vision methods have been developed in the field of traffic surveillance for various purposes. Many of these require the system to work in real-time (RT), which limits the complexity of possible algorithms. Further traffic monitoring systems often have to deal with a broad number of different classes, such as cars, trucks, buses, cyclists and pedestrians. Depending on the purpose of the system, applications require different levels of understanding of the scene in terms of objects classification and identification. The task of object segmentation is hereby often the crucial step. Mixture of Gaussian background modeling techniques fail to perform this task, as they assume that foreground objects are constantly in motion and move more or less in the same speed [28]. Different works have addressed this issue by validating foreground pixels by a moving object model [7] or lowering the update speed of the model [29,30]. Build upon object detection algorithms, different monitoring systems have been presented. The range of interpretation goes from vehicle counting [1,6,20] over incident detection [16, 33] to classification [21,31].

Thiemo Alldieck · Chris Bahnsen (✉) · Thomas B. Moeslund
Visual Analysis of People Lab, Aalborg University, Aalborg, Denmark
E-mail: cb@create.aau.dk

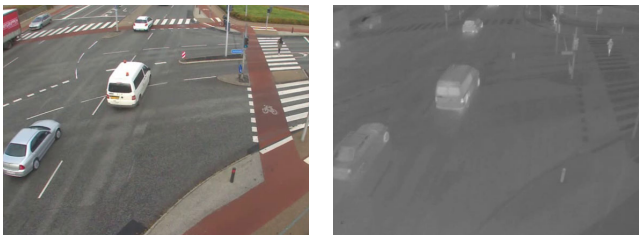


Fig. 1: RGB and thermal image of the same scene

Fig. 2: RGB image and naive fusion of IR and RGB image

The use of cameras for monitoring purposes also introduces a significant drawback. Caused by the functional principle of a camera working in the visual range of light, the quality of the data highly depends on environmental conditions such as rain, fog or day and night cycle. Resulting many applications only work at daytime. A persistent monitoring of the scene however is often desired. Special situation methods have been developed [5,24,33], a standard method for different purposes and conditions is yet to be presented. To overcome this problem different detectors for traffic monitoring have been introduced, working either standalone or in combination with traditional cameras. Hereby special interest in thermal or infrared (IR) cameras developed recently. Thermal cameras capture the radiation emitted by objects that depends on their temperature [10]. This makes the system independent of the lighting situation and visual obstructions caused by fog or rain. On the other hand, as seen in Fig. 1, thermal images are less detailed and provide an unfamiliar visual impression.

In this paper, we present a novel multimodal background subtraction technique. Background subtraction is the first and crucial step of *bottom-up* processing pipelines, as commonly used in RT surveillance systems. To make the system work under different conditions, two different sensors, namely a RGB and a thermal camera, have been used. Two parallel background subtraction algorithms are performed and fused at a new background conformity representation, referred as *GMM distance map*. The methodology of image fusion and related work is discussed in Sect. 2. The presented work integrates image quality heuristic functions for the two modalities as elaborated in Sect. 3. The core contribution of this work is presented in Sect. 4. Subsequently we present extensions for the application of traffic monitoring (Sect. 5). In Sect. 6 we evaluate the work with an own and two commonly used datasets against state of the art background modeling techniques, followed by our drawn conclusions (Sect. 7).

2 Multimodal data fusion

Different sensors have advantages and disadvantages for further processing. To overcome the individual downsides of different sensors, multimodal systems have been developed. These systems use information from multiple sensors and information sources and enrich and combine it. The potential of these methods especially for traffic surveillance has been emphasized by Buch et al. [2]. In this chapter different fusion approaches will be discussed. The focus is hereby on the fusion of video data from IR and RGB cameras.

Different fusion approaches can be classified into three levels: *pixel-level*, *feature-level*, and *decision-level* according to the stage of the data flow where the fusion takes place [12].

Fusion at decision level combines the output from two or more parallel processing pipelines. The results are merged by Boolean operators or weighted average. Serrano et al. [25] perform parallel segmenting of thermal and RGB data and select the representative output based on confidence heuristics.

Feature-level fusion performs the fusion one step earlier in the processing pipeline. Features from all input images are extracted individually and then fused into a joint feature space. In [18] this technique is used for automatic target recognition (ATR).

Pixel-level fusion is the most common approach. Hereby the input images are merged into one. Details that might not be present in one image are hereby added by the other modality. Common examples are structures occluded through dark shadows or smoke in RGB images that are revealed with the help of a thermal image. Pixel-level fusion requires all input images to be spatially and temporally aligned. This alignment, also called registration, is a task for itself. Automatic image registration approaches often fail, since there is no correlation between the intensity values of the modalities [8]. A common approach is to manually select corresponding points in both modalities and compute a homography. However special case automatic methods exist, using features that are most likely present in both modalities e.g. contours [13], Harris corners [14] or Hough lines [15].

Fig. 2 shows a naive pixel-level fusion example with images of the OTCBVS dataset [9]. The right image reveals the person standing next to the building. Beside naive fusion though averaging, addition or multiplication of the images, more complex methods have been presented, trying to optimize the information content of the image.

Shah et al. [26] perform the fusion after different wavelet transforms of the images. This allows a fusion rule based on frequencies rather than pixels preserving details while simultaneously reducing artifacts. A statistical approach is followed in [4]. During an expectation-maximization the fusion result is obtained stepwise.

Lallier and Farooq [19] perform the fusion trough adaptive weight averaging. The weight per pixel is hereby defined by a number of equations that express the interest in the specific pixel. In the context of the work these are the degree of an object being warmer or colder for the thermal domain and the occurrence of contrast differences as well as large spatial and temporal intensity variations for the visual domain.

Instead of fusing the images to a new image that can be represented in RGB, other methods simply combine the inputs to a new format. St-Laurent et al. [27] adapt a state of the art algorithm for moving objects extraction to work with "Red-Green-Blue-Thermal" (RGBT) videos. This way important information are automatically revealed by the object extraction algorithm.

In this paper we present a novel approach to image fusion. Fusion is performed on pixel-level but not in the input data but within two parallel background subtraction pipelines. Therefore an image representation of the conformity of the pixels with the background model is presented, referred as *GMM distance map*. The fusion is performed *context aware* by evaluation of the *usefulness* of the input data. Different heuristics are specified, describing the quality of the modalities. After the fusion, the resulting distance map is manipulated and thresholded in order to create a single foreground mask.

3 Image quality heuristics

The need for image fusion bases on the fact that image processing techniques work only well for data of high quality. The presumption is, that after the fusion process the data is of higher quality than before. The term quality is hereby defined as how well the data fit to the method. Input data that contains noise or unwanted objects that harm the algorithm is of low quality. This is even the case if a human can easily perform the equivalent task.

In the following will be discussed which conditions harm the image quality for surveillance scenarios and which data can be conducted to predict the data quality of a sensor. The quality is hereby rated against state of the art background modeling techniques and may differ for other applications, such as *top-down* tracking approaches. The aim is to build heuristic functions that express the *usefulness* of the modality to be able to perform an adaptive fusion.

3.1 Thermal image quality

As already mentioned thermal cameras measure the infrared radiation emitted by all objects. The energy of the radiation is hereby mainly dependent on the object temperature. A constant factor called emissivity scales the radiation for different materials [10]. With known emissivity the temperature of objects in thermal images can be calculated using the Stefan-Boltzmann law. However many thermal cameras built for surveillance feature Automatic Gain Control (AGC), so the mapping function between radiation energy and intensity values is unknown.

The quality of thermal data is high when the contrast between foreground and background objects is high. In traffic surveillance this is the case when road and road users emit a significant different amount of radiation, resulting well distinguishable intensity values in the images. Assuming a more or less constant temperature of road users, the temperature of the road could be a quality indicator. Except from direct measurements however, the determination of the road temperature is complicated, since it is influenced by numerous interacting parameters [3]. Therefore, the road temperature cannot be used as long as this data is not present.

As described earlier, objects of different materials have different intensity values in a thermal image, even when having almost the same temperature. Therefore a certain amount of information is in the image even when showing a scene without foreground objects. In a situation where no objects can be distinguished, the information content is low. Consequentially the image entropy can be used as a quality heuristic of thermal images. The entropy is defined as:

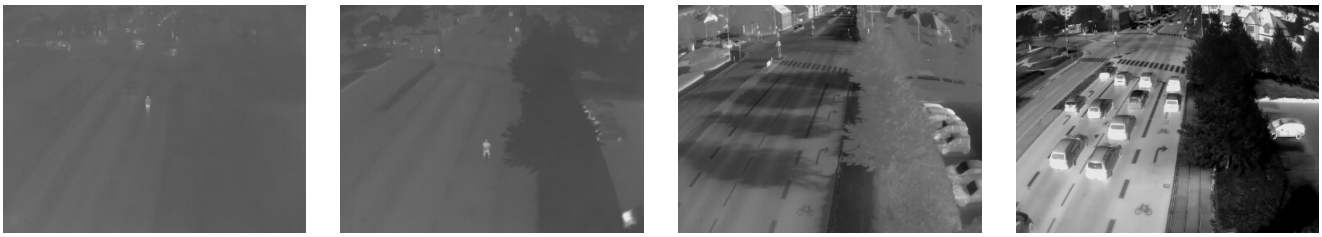
$$H = -\sum_{i=0}^{255} p(I_i)\log_2(p(I_i)) \tag{1}$$

with p being the probability of an intensity value I in a gray scale image. Fig. 3 shows a side-by-side comparison of the same location at different times. The right images appear much more detailed and therefore of higher quality. The corresponding entropy values correlate with this impression.

Since the degree of the mapping function between entropy and quality is unknown, it has to be chosen manually. A linear function has been found to be not sufficient. During the experiments could be seen, that the down-rating of low entropy values is too strong. A sigmoid function appeared as a better approximation of the mapping function. The function used during this work is shown in Fig. 4.

3.2 RGB image quality

Quantified quality rating of RGB images is a hard task since we are used to this type of sensing. The human eye can adapt



(a) $H = 4.23$ (b) $H = 5.04$ (c) $H = 6.65$ (d) $H = 7.67$

Fig. 3: Thermal images of the same scene with different entropy values

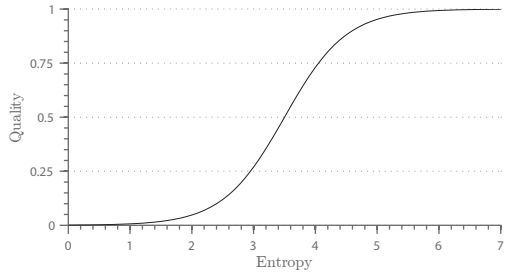


Fig. 4: Thermal image quality heuristic.

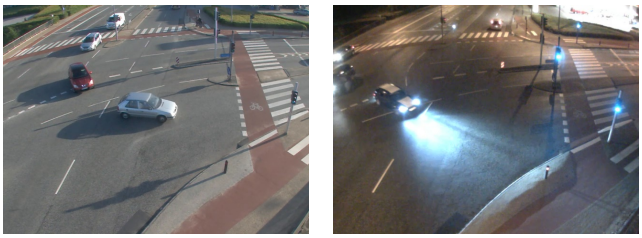


Fig. 5: RGB images with common challenging conditions, such as shadows, reflections and halos

to challenging conditions easily and the human brain interprets the visual perception [11]. An in-depth knowledge of a computer vision method is needed to be able to rate how well input data fits to the selected method. State of the art background modeling tends to produce false positives. Reasons for this are discussed in the following.

3.2.1 Lighting conditions and shadows

Fig. 5 shows a scene at different times of the day. While a human can easily label the cars in the scene, a background subtraction algorithm would be highly disturbed by the large shadows and reflections. Although shadows are handled quite well nowadays, it still disturbs the detection process. An effective method for reflection detection is yet to be presented. In addition in the night cars and background have almost the same color. In conclusion both images should be rated as low quality, even though the reasons are different and so may be the quality rating.

Following the argumentation images with low light conditions, such as twilight and night, should be rated as low quality even though a human might be able to identify the road users. As a heuristic input parameter the elevation angle of the sun can be consulted. The solar elevation angle is defined as the angle between the ground plane and the sun's position vector. The sun is visible for angles $\geq 0°$. Between $0°$ and $-18°$ we speak about twilight and below the sun does not contribute to sky illumination, it is night. In practice however a noticeable illumination is not present before $-6°$, known as the civil twilight. Also the illumination condition is not perfect as soon as the sun is visible. Buildings and the colored sunlight lead to an imperfect illumination. Therefore a perfect illumination has been defined in this work for elevation angles $\geq 6°$

Secondly the image quality depends on the presence and amount of drop shadows. Two factors specify the occurrence. At first only on sunny days drop shadows can appear. A weather database can be accessed to retrieve a description of the weather condition. The length of these shadows is determined by the sun's position. Therefore, both weather data and the solar elevation angle must be consulted to present a heuristic to what extent cast shadows might be present in the scene. The lengths of shadows can be calculated through:

$$L = h/tan(\alpha) \tag{2}$$

with h being the object height and α the solar elevation angle. With a unit object height $1 - \psi L$ can serve as a quality function, where ψ is a scaling factor.

3.2.2 Weather conditions

The weather can disturb the background subtraction process not only through the presence of shadows. Different conditions harm the algorithm through different phenomena such as reflections, particles and reduced visibility just to name a few. A quantitative rating however is not so easily derived. For a reliable image quality heuristic the influence of different conditions has to be experimentally determined. An experienced person might also be able to perform a loose rating manually.

Good conditions	1.0
Low/Varying illumination	0.8
Reflections/moisture	0.6
Particle occlusion/precipitation	0.3
Reduced visibility	0.3

Table 1: Different weather categories and corresponding quality rating.

For this work different conditions have been broadly grouped into 5 categories as seen in Tab. 1. Each of the categories is characterized through a set of quality harming influences, such as varying light, reflections through moisture or particle occlusion. The quality rating of the categories are hand tuned based on experiments.

4 Trust based multimodal background subtraction

The following section discusses the main contribution of the work. A new approach to image fusion is presented. Despite other works, not the input data is fused but intermediary results of two parallel background subtractions. The results are weighted based on the quality heuristics described before, making the system context aware. The algorithm is based on the adaptive Gaussian mixture model background subtraction algorithm presented by Stauffer and Grimson [28]. Fig. 6 illustrates the basic principle of this work.

4.1 Calculating the GMM distance map

During the calculation of the foreground mask with the help of the Gaussian mixture model (GMM), each pixel is tested against each component of its background model, if it can be accepted as part of the component's Gaussian distribution. The Euclidean distance of the sample value from the mean is hereby the important factor for acceptance. A pixel x at time $t+1$ is defined to match the ith component, if it falls within λ standard deviations:

$$M_{i,t+1} = \left(\frac{|x_{t+1} - \mu_{i,t}|}{\lambda \sigma_{i,t}} < 1 \right) \tag{3}$$

The acceptance distance of the sample as background in Eq. (3) is normalized by the specific variance $\sigma_{i,t}$ and the threshold value λ. Large distance values indicate a high probability of the pixel being foreground whilst small values show a high conformity with the component. With this in mind an approximation of the general conformity of a pixel with the model can be expressed with distance values:

$$D_t \approx \begin{cases} d_{0,t}, & \text{if } M_{0,t} \\ d_{1,t}, & \text{if } M_{1,t} \\ \dots & \\ d_{b,t}, & \text{if } M_{b,t} \\ \min(d_{0,t}, d_{1,t}, \dots, d_{b,t}), & \text{otherwise} \end{cases} \tag{4}$$

with

$$d_{i,t} = \frac{|x_t - \mu_{i,t-1}|}{\lambda \sigma_{i,t-1}} \tag{5}$$

If a match $M_{i,t}$ is found, the corresponding value of $d_{i,t}$ is used to express the distance. Otherwise the distance to the closest component is used. The resulting values of all pixels form a map expressing the deviation of image regions from the background. The scaling is hereby the same for all pixels, so that a single threshold can be applied. When thresholding the map with the value 1.0, the resulting mask is the same as if calculated through [28].

4.2 Trust based fusion and foreground identification

At this stage we have heuristics for the image qualities of both modalities as well as maps expressing the background conformity of pixels. Build upon this a trust based fusion is performed.

The trust in a modality can directly be derived from the quality heuristics. The better the quality the higher is the trust. Therefore the different heuristics have to be combined first. Under the assumption, that the different quality heuristics do not interfere each other, the heuristics can simply be multiplied:

$$q_{\text{RGB}} = q_{\text{sun}} \cdot q_{\text{shadows}} \cdot q_{\text{weather}} \tag{6}$$

$$q_{\text{IR}} = q_{\text{entropy}} \tag{7}$$

However, with the functionality of the background subtraction in mind, these values are not sufficient to describe the trust in each modality. Rapid changes in the environment can highly disturb the algorithm. The changes can for example be rapid illumination changes for the RGB domain, or a rapid change of the auto gain for the IR domain. These changes are not predictable by the knowledge base of the quality heuristics. Therefore another parameter needs to be introduced. When the background subtraction algorithm fails, a huge number of false positives appear. The resulting number of foreground pixels is much higher than the average of the scene. A quality heuristic based on this phenomena is defined in Eq. (8), where τ defines the average foreground ratio, γ is an arbitrary scaling factor, $\mathbb{1}$ denotes an *indicator function* and (X,Y) are the image dimensions:

$$q_{\text{fg}} = \max(1 - \gamma(r_{\text{fg}} - \tau), 0) \tag{8}$$

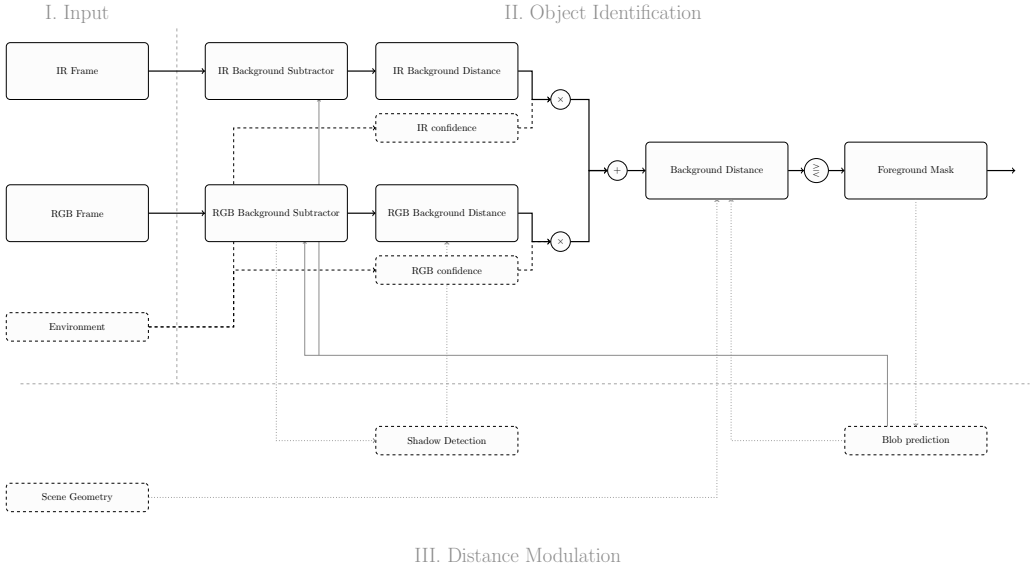


Fig. 6: System design overview.

where

$$r_{\mathrm{fg}} = \frac{1}{XY} \sum_{x=1}^{X} \sum_{y=1}^{Y} \mathbb{1}\left[L_{x,y} \geq 1\right] \tag{9}$$

Given Eq. (8) the trust can now be calculated as follows:

$$T_{\mathrm{RGB}} = \min(q_{\mathrm{fg}_{\mathrm{RGB}}}, q_{RGB}) \tag{10}$$

$$T_{\mathrm{IR}} = \min(q_{\mathrm{fg}_{\mathrm{IR}}}, q_{IR}) \tag{11}$$

To prevent artifacts, the trust in a modality only increases slowly after being rated down. When confronted with dramatic changes, the Gaussian background model needs some time to adapt for the changes. Therefore the trust should only slowly increase while the background is being learned. The trust T at time $t+1$ is calculated:

$$T_{t+1} = \begin{cases} T_{t+1} & \text{if } T_{t+1} \leq T_t \\ \alpha T_{t+1} + (1-\alpha) T_t & \text{otherwise} \end{cases} \tag{12}$$

where α is the update speed of the Gaussian mixture model.

After normalizing the values of T_{RGB} and T_{IR} to add up to 1, the values are used as weights for the adaptive fusion. After image registration, each pixel is fused in the distance map:

$$D_{\mathrm{F}} = w_{\mathrm{RGB}} D_{\mathrm{RGB}} + w_{\mathrm{IR}} D_{\mathrm{IR}} \tag{13}$$

where

$$w_{\mathrm{RGB}} = \frac{T_{\mathrm{RGB}}}{T_{\mathrm{RGB}} + T_{\mathrm{IR}}}$$

$$w_{\mathrm{IR}} = \frac{T_{\mathrm{IR}}}{T_{\mathrm{RGB}} + T_{\mathrm{IR}}} \tag{14}$$

Through the weighting based on the quality respectively trust values, the fusion is adaptive and context aware. This principle is illustrated in part II of Fig. 6.

At this stage spatial and temporal registration inaccuracies can be compensated. A simple mean filter applied on the fused distance map dissolves the pixel grid and therefore fuses information of neighboring pixels. Other functions are possible and have been applied, as further elaborated in the following.

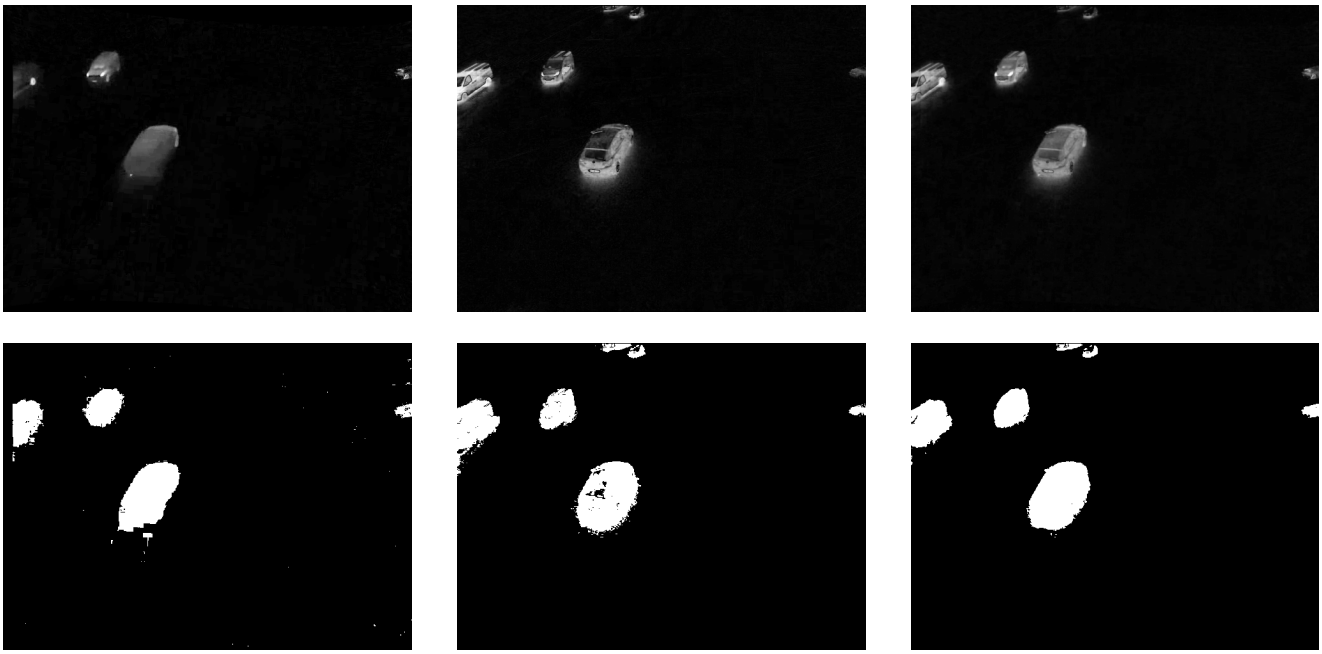
The final step is the decision whether a pixel is foreground or background. As explained before, all values are scaled to the same level. A simple thresholding per pixel is performed:

$$FG = \begin{cases} 1 & \text{if } D_{\mathrm{F}} \geq 1 \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Fig. 8 displays the distance maps, their fusion and its effect on the resulting mask.

5 Application to traffic monitoring

The preceding section described the main contribution of this work. The presented method is an extension to the general adaptive Gaussian mixture model background subtraction method. No constrains have been introduced, so that the method is applicable for both indoor as well as outdoor scenarios. In the following specific extensions for traffic surveillance are presented, showing the modularity of the proposed algorithm. In Fig. 6 these extensions are labeled as number III.



(a) IR (b) RGB (c) Fused

Fig. 8: Distance maps of the different modalities and results after thresholding.

5.1 Shadow Detection

A common extension to background modeling techniques is shadow detection. Shadows of intruding objects are found as not matching the background model as they appear darker as prior illuminated areas and are therefore declared as foreground. Depending on the purpose of the system, labeling shadow areas as foreground is a false positive. In most surveillance scenarios only the objects and not their shadow are of interest [23].

Different shadow detection algorithms have been presented. Prati et al. [23] distinguish between *deterministic approaches* that use an 'on/off decision process' and *statistical approaches* that 'use probabilistic functions to describe the class membership'. Both methods however can fail and false negatives as well as false positives may occur. Is the task to identify all foreground objects, as it is in traffic surveillance, especially false positives harm the results. Whole objects may be classified as shadow. To address this issue, shadow areas have been pruned rather than removed in this work.

In state of the art methods a labeling in the resulting foreground mask is performed. Instead of making this hard decision, the distance of areas marked as shadows are scaled down. In this work a fixed scaling has been used. A scaling based on the shadow certainty may be a possible extension. Since the background distance correlates with the certainty of a pixel being foreground, the downscaling can be seen as bringing uncertainty to the decision. Consequentially the decision whether a pixel is shadow is only made indirectly, when deciding whether the pixel is foreground or background.

The subsequent fusion of the modalities is the important step for this method to work. Objects that have also been found in the thermal image are most likely found anyway and shadows are voted further down as they are not present in the thermal domain. Especially small areas of false positives can be recovered as being a foreground object using this technique. The mean filter subsequent to the fusion helps the process with removing outliers. Additionally the quality functions allow to predict scenes with drop shadows. Therefore the process can be triggered context aware, only when shadows are most likely present.

5.2 Blob prediction

As discussed in Sect. 1 a successful background subtraction for traffic surveillance must handle the different speeds of the traffic. All objects have to be handled as foreground even when staying in the scene for a longer time. For this purpose the blob prediction method proposed by Yao and Ling [30] has been integrated in this work. The position of foreground blobs are predicted for each frame and the update rate α is significantly lowered for these areas. Consequentially objects have to stay for a very long time before merging into the background.

To predict blob positions for the current frame $t+1$ blobs from t and $t-1$ are matched. Afterwards the displacements

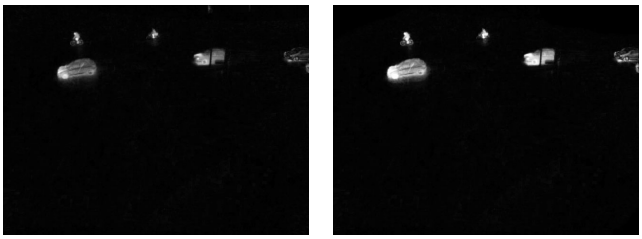


Fig. 9: Distance map before and after blob prediction based modulation.

between t and $t-1$ is applied on t. The matching is done with a nearest neighbor search of the blob's centroids. If no neighbor within a range ρ is found, the blob is supposed to be stationary as no prediction about the movement can be made.

In contrast to [30] an extension has been made. To prevent artifacts in the background model caused by inaccuracies in the blob prediction, the predicted blobs have been dilated and the edges have been smoothed out. The update rate α has then been calculated as:

$$\alpha = b\alpha_{\mathbf{fg}} + (1-b)\alpha_{\mathbf{bg}} \tag{16}$$

where $0 \leq b \leq 1$ indicates the value in the blob prediction image and $\alpha_{\mathbf{fg}}$ and $\alpha_{\mathbf{bg}}$ are the update rates for foreground respectively background regions.

Another purpose of the blob prediction has been found in this work. Since the boundary of foreground objects changes only gradually, the predicted blobs are a very good estimate of the next frame's foreground. This can help the segmentation as it is more likely to find an object where it is predicted than elsewhere in the scene. Objects follow a trajectory and generally do not appear out of sudden. To express this characteristic another modification of the distance map is done. Analogous to the shadow suppression, predicted areas are scaled up in the distance map. Fig. 9 demonstrates the effect. One can clearly see how the traffic stands out more clearly in the right image.

By taking the predicted blob areas into account for the decision whether a pixel belongs to the foreground, a spatial-temporal constraint is introduced. The decision is no longer pixel-wise but aware of the history of the whole image.

5.3 Scene geometry based prior knowledge

The principle presented in the last sections can be used for another constraint. By looking at the scene geometry one can easily divide the image into three classes. The first class of pixels are areas where no foreground is expected under any circumstances. Examples may be trees or the sky. The second class of pixels describe the areas where objects may move to. A sudden appearance of objects is unlikely or even

Fig. 10: Scene area classes. Green: entrance areas, red: excluded areas, rest: neutral.

excluded but objects may move to these areas from other parts of the image. These areas are referenced as neutral zones. The last class describes areas where we expect foreground objects to appear. These areas are called entrance areas in the following. Entrance areas can normally be found at the borders of the image as objects enter the scenery normally from out of the camera's view port. Objects may however also reappear from occlusion or enter from occluded areas. Based on this classification a mask can be drawn as seen in Fig. 10.

The scene classification is *prior knowledge* to the segmentation process. In the adaptive Gaussian mixture model background subtraction by Stauffer and Grimson [28] all pixels are treated the same. The likelihood of a pixel being foreground is independent to its position. With the introduction of the scene classes, this has been changed in this work. Only the entrance areas are treated like in [28]. For the other two classes modulations of the distance maps have been made as presented before.

Firstly excluded areas are made impossible to be foreground by setting the corresponding values in the distance map to zero. Secondly the values for neutral zones are scaled down to make it less likely to find foreground pixels in these areas. This is possible because the blob positions have been predicted and uprated beforehand. Areas where we expect objects to move to are untouched afterwards or even uprated while unpredicted regions are rated down. This helps to remove noise and find objects more reliable.

6 Experiments

To evaluate the performance of the proposed algorithm a series of experiments have been conducted. Hereby both quantitative and qualitative performance have been tested. This section begins with an elaboration about the datasets that have been used in this work. This is followed by a description of the performance metrics and the results of the exper-

iments. Concluding an in-depth analysis of the qualitative performance is presented.

6.1 The datasets

The main dataset used in this work contains a large number of multimodal recordings of intersection in Northern Jutland recorded during the year 2013. To be able to benchmark the proposed algorithm, two commonly used datasets have been additionally used. The *OSU Color-Thermal Database* [9] of the *OTCBVS Benchmark Dataset Collection* contains RGB and thermal data of two surveillance scenarios. The videos contain pedestrians recorded on the campus of the Ohio State University. The *INO Video Analytics Dataset*[1] contains a set of multimodal recordings of parking lot situations including cars, cyclists and pedestrians. All scenes that have been tested during the experiments of this work are listed in Tab. 2 and 3.

6.2 Performance metrics

The following section explains the quantitative performance metrics that have been used for the evaluation of the experiments. The quality of a segmentation algorithm is determined by two quality measures. *Good detection* means that most foreground pixels of the image have actually been found by the algorithm. *Good discrimination* means a good distinction has been made i.e. not many pixels have been declared as foreground erroneously. These metrics are measured by the *Detection Rate (DR)* and *False Alarm Rate (FAR)* defined as follows:

$$DR = \frac{TP}{TP+FN} \tag{17}$$

$$FAR = \frac{FP}{TP+FP} \tag{18}$$

with true positives (TP), false positives (FP) and false negatives (FN). The Detection Rate is also known under *recall* or *true positive rate* and describes the *sensitivity* of a detector. The False Alarm Rate corresponds to $1-p$ where p is the detector's *precision* or *specificity*.

In order to evaluate the performance metrics it is required to have access to the true data, commonly referred as *Ground Truth (GT)*. Ground Truth has to be created manually and is a laborious task. Thus only a small sample of the results can be tested. In this work 70 successive frames have been annotated for each test set. Only for the *Auto Gain* set have been annotated 180 frames in order to cover the whole process.

[1] http://www.ino.ca/en/video-analytics-dataset/

6.3 Quantitative results

In order to evaluate the performance of the proposed method extensive experiments have been performed and evaluated with the described performance metrics. Besides with the algorithm itself, each dataset has been processed with four alternative strategies. Each strategy bases on the Gaussian mixture model background subtraction algorithm presented by Stauffer and Grimson [28] and improved by Zivkovic [32]. Firstly both modalities RGB and IR are processed individually. Next a pixel-wise fusion is performed with the creation of RGBT frames. And last the confidence based selection presented by Serrano-Cuerda et al. [25] is used.

As this works aspires to create a system that works without the requirement to manually tune its parameters for different conditions, one set of parameters has been defined for all test sequences. In practice however parameters may be tuned to fit the given location and situation. For comparability reasons this fine tuning has not been done in this work. Solely the learning time for each scene has been adjusted to match the specific situation. For example scenes with much traffic need more time to learn a stable background model. For the case of the presented algorithm, both background models first have been learned individually before the described adjustments have been made. This is necessary since predicted foreground regions are learned much slower and false positives are very likely during the learning phase.

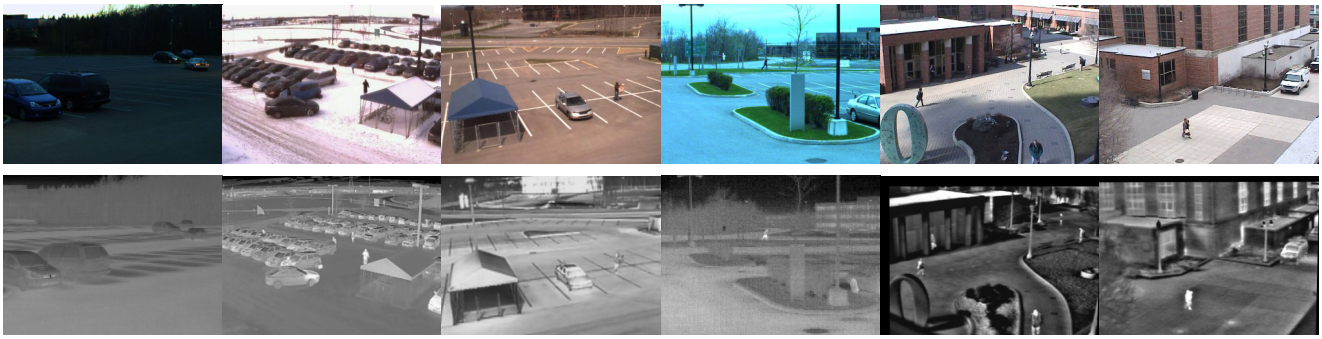
The update rate α has been set to be slower for the alternative strategies. Since the state of the art background modeling does not differ between foreground and background in the update step, a quick update rate would result in foreground objects merging into the background. This is also the case in the learning phase of the proposed method. Consequentially the same α has been used here. All important experimental parameters are listed in Tab. 4, where the parameters below the line only apply for the proposed method.

For all experiments containing RGB data shadow detection has been performed. Pixels that have been found to be shadow have been classified as being background in the reference methods. Furthermore the region masks have been applied on the resulting data. Excluded areas have been removed from the resulting foreground masks and the resulting masks have been cleaned with morphological operations and hole closing. This way equal conditions have been created for all strategies and differences in the results of the proposed algorithm in contrast to the state of the art methods can be explained by its core contributions.

The scenes have been selected, as each of them introduces a new scenario with different conditions. State of the art methods often aim to perform best for one special scenario. The proposed method however aims for a good performance *in general*. Therefore a comparison to recent special case methods with best known performance was not desired

Day	Night	Auto Gain	Heavy Rain	Snow

Table 2: Test scenes from our own dataset.



INO ParkingEvening	INO ParkingSnow	INO CoatDeposit	INO TreesAndRunner	OTCBVS 3	OTCBVS 4

Table 3: Test scenes from the benchmark datasets.

and even not necessary. The comparison to simple background modeling on the other hand reveals interesting insights on strengths and weaknesses of the proposed method.

The results of the experiments are displayed in Tab. 5. The general performance of the proposed algorithm can be considered very good due to a minimum Detection Rate of 0.84 and maximum False Alarm Rate of 0.56. It can clearly be seen, that the goal of creating a robust method for a wide bandwidth of condition is achieved. Only the proposed method shows a good performance for every test sequence. The alternative strategies fail for different scenarios, but also better performance than the proposed method can be seen. The reasons for this are manifold and will be discussed in the following section.

As expected, all fusion approaches show in general a better performance than the single modality methods. The method presented by Serrano-Cuerda et al. [25] also performs well on the first look. When analyzing the results in detail however, one can easily see, that the results are at most as good as one of the single modalities. This is funded in the design of the algorithm, as it selects one result of two parallel pipelines. One important characteristic of fusion algorithms is neglected by this design flaw. Fused data or results generally differ from its inputs and therefore contain new features and information. A simple selection obviously makes this impossible. As a result the approach is beaten in 10 out of 11 cases in terms of Detection Rate.

The False Alarm Rate of both, the proposed method and the RGBT approach mirror the weaker modality. The reason is, that a high evidence of foreground in one modality is still present after fusion the data. Only false positives based on weak evidence are successfully smoothed out. In worst case false positives from both modalities are present in the result.

A good example for the superiority of fusion approaches in terms of Detection Rate is given by the sequence *INO TreesAndRunner*. Obviously both fusion approaches, the proposed method and the RGBT approach, perform much better than the single modalities. This is the case because both RGB and IR contain frames, that are very hard to segment. The runner passes trees and other objects. The fusion approaches can still rely on the second modality, when the information content of the first is low.

Parameter	Value	Description
α	0.0005	Update rate
K	5	Number of components
λ	4	Number of standard deviations for background acceptance
α_{BG}	0.0033	Background update rate
α_{FG}	0.000033	Foreground update rate
τ	0.1	Foreground ratio
γ	5.0	Foreground deviation weight
ρ	17	Blob match radius (px)
s_{shadow}	0.3	Distance scaling factor for shadow regions
$s_{predict}$	1.5	Distance scaling factor for predicted regions
$s_{neutral}$	0.5	Distance scaling factor for neutral regions

Table 4: Parameters used in the experiments. The parameters below the line only apply for the proposed method.

	Proposed	RGB	IR	RGBT	Select
Day	**0.99**	0.93	0.95	0.97	0.93
	0.30	**0.09**	0.31	0.29	**0.09**
Night	0.84	0.78	0.48	**0.89**	0.78
	0.31	0.69	0.32	0.66	0.69
Auto Gain	**0.94**	0.86	0.73	0.91	0.81
	0.25	**0.09**	0.76	0.40	0.58
Heavy Rain	**0.92**	0.46	0.69	0.48	0.69
	0.22	0.26	**0.11**	0.27	**0.11**
Snow	**0.96**	0.79	0.21	0.92	0.21
	0.52	0.52	**0.25**	0.55	**0.25**
INO ParkingEvening	0.94	0.93	0.91	**0.95**	0.91
	0.24	0.27	**0.18**	0.29	0.18
INO ParkingSnow	0.98	0.86	**0.99**	0.96	**0.99**
	0.32	0.78	0.40	0.35	0.40
INO CoatDeposit	**0.97**	0.10	0.10	0.10	0.10
	0.19	**0.12**	0.30	0.16	0.12
INO TreesAndRunner	**0.94**	0.88	0.84	0.93	0.84
	0.44	0.65	**0.36**	0.70	**0.36**
OTCBVS 3	**0.95**	0.75	0.94	0.90	0.78
	0.56	0.96	0.74	0.96	0.93
OTCBVS 4	**1.00**	0.94	0.78	0.99	0.78
	0.55	**0.15**	0.68	0.48	0.68

Table 5: Experimental results. First line DR, second line FAR. The best DR and FAR values of each set are marked bold. Proposed method compared to GMM background subtraction of RGB, IR, and RGBT frames. "Select" indicates result selection based on quality heuristics [25].

6.4 Special situation performance

In the following results of specific test sequences are elaborated in detail. It is shown in which way different details of the design of the proposed algorithm effect the performance. Four different problems that arise during outdoor surveillance are discussed. Emphasis has been put on the adaptive modality weighting of the proposed algorithm and its effect on the segmentation results. To begin with, this context awareness is discussed further.

6.4.1 Context awareness

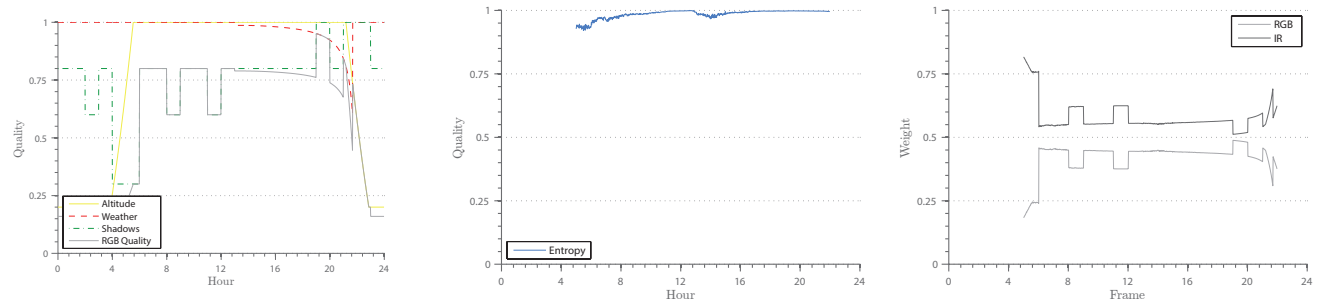
One of this work's core contributions is the context awareness of the fusion. It is based on a set of quality heuristics that have been defined in Sect. 3. The goal is to evaluate the *usefulness* of each modality. Instead of using information from the images itself, outside sources have been consulted. Solely the thermal domain has been rated by its own information content. For the tested sequences, the weights calculated with the help of the heuristics are more or less fixed. The time frames are simply too short to see an effect based on for example the altitude heuristic function. The overall concept however has been tested by selecting scenes with various conditions.

In Fig. 11 quality functions covering a full day are plotted. Since video data are only present from 5 a.m. to 10 p.m., the IR quality function is only defined for that particular time frame. The plotted day was a hot summer day with rather good weather. Because of overcast, no cast shadows have been assumed in the morning. During noon time the temperature was so high, that the thermal camera was overexposed. This problem can be seen in the drop of the quality heuristic.

6.4.2 Automatic Gain Control

The first problem that will be discussed further, is the prior mentioned Automatic Gain Control of thermal cameras. When large or hot objects enter the scenery, the camera automatically adjusts its gain in order to preserve a high level of detail. This behavior however highly disturbs the background subtraction. The new background does not longer fit the model, resulting in a high number of false positive foreground pixels. Fig. 12 displays the described phenomena. The challenge is hereby the short time frame of the adjustment. When the objects leaves the scene, the camera adjusts the gain back. Therefore the problem often only persists for 100-200 frames, while highly affecting the segmentation results. Also the background model is affected, since it is updated each frame, which makes it invalid for following frames.

As seen in 5 the proposed algorithm handles the described problem well. No segmentation quality reduction

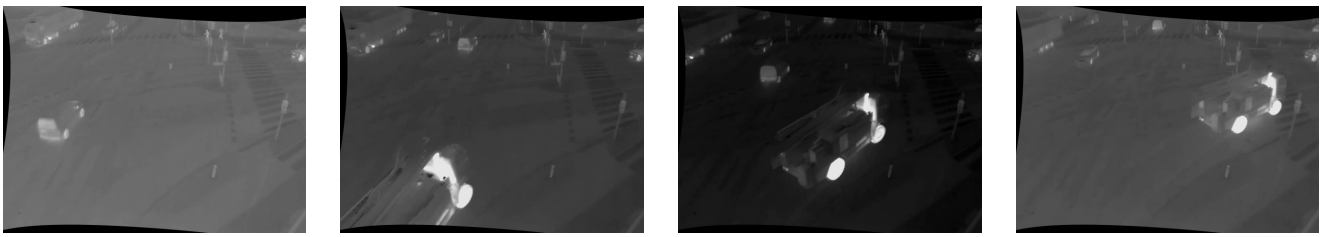


(a) RGB quality heuristics and resulting quality function of a full day.

(b) Entropy based quality heuristic for thermal domain of a full day.

(c) Varying weights of the two modalities over a full day.

Fig. 11: RGB and IR quality heuristics and resulting weights of a full day.



(a) Frame 170

(b) Frame 200

(c) Frame 230

(d) Frame 260

Fig. 12: AGC of the IR camera triggered by a big truck coming into the scenery.

can be detected from the raw numbers. The reason for this is the adaptive weighting performed in the fusion step. Through the foreground ratio evaluation described in 4 is detected, that the background model of the thermal domain is invalid. As a result, the IR weight function drops to zero and the segmentation only relies on the RGB domain. 13 displays this behavior. It can clearly be seen, how the quality heuristic drops parallel to the weight of the IR domain. After the truck leaves the scene, the camera adjusts back to normal and the quality function instantly rises. The weight however increases only gradually. This is necessary in order to give the background model time, to relearn the background model.

The foreground ratio quality function also decreases while the truck enters the scene. This behavior is normal, since the truck covers a lot of the scene and therefore the number of foreground pixels is unnaturally high. In the end result this however has no effect, since the overall trust in the RGB domain is still higher.

6.4.3 Changing illumination

A very similar problem that is common in outdoor surveillance is changing illumination. Although the design of the Gaussian mixture model background subtraction allows variations in the illumination, problems can still arise when abrupt changes happen. This for example is the case when clouds cover or uncover the sun. The algorithm is only designed to adapt for slow changes, e.g. shadows that move over the day. Fast changes of the scenery will cause foreground objects due to the definition of the process.

Similar to the problem discussed before, the foreground ratio of the RGB domain is rising, since the background model does not adapt fast enough for these changes. Consequentially a weight shift to the thermal domain is performed by the algorithm, leading to the comparatively low False Alarm Rate of 0.56.

6.4.4 Artifact reduction

Another contribution of the proposed algorithm can be seen in the results of *OTCBVS 3*. With 0.56 the False Alarm Rate is even lower than the results of the thermal background subtraction. This can be reasoned with the adjustments made to the fused distance map based on the scene geometry. Artifacts are unlikely to appear since unpredicted foreground regions are reduced in the distance map. The effect on the foreground mask can clearly be seen in 14 in comparison to the approach using solely the thermal domain.

6.4.5 Long staying objects

Gaussian mixture model background subtraction presented by Stauffer and Grimson [28] assumes that foreground objects are constantly in motion. For traffic this is obviously

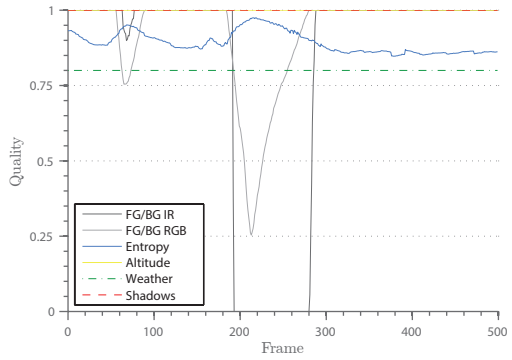


Fig. 13: Quality heuristics for the *Auto Gain* test sequence.

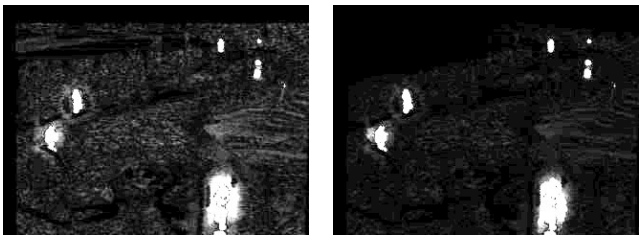


Fig. 14: Distance map before and after scene geometry based modulation.

Fig. 15: Stopped car merges into the background.

not the case. This has been addressed by Yao and Ling [30] and the proposed method has been integrated in this work. The original algorithm causes long staying objects to gradually merge into the background. This problem is very much visible in the *INO CoatDeposit* test set. The car coming into the scene merges into the background within a few frames as seen in Fig. 15. This merging is stopped by predicting foreground regions and lowering their update speed, resulting in a significantly better Detection Rate of the proposed method.

7 Conclusions and future work

This work aimed to investigate on how the fusion of thermal and RGB video data can assist camera based traffic monitoring. Related work has been studied and a suitable starting point has been identified. Background subtraction is the first and crucial step of many surveillance algorithms. A stable segmentation lays the base for successful successive analysis. Different environmental conditions are hereby the most challenging problem. Only very few methods have been presented working under extreme situations, although the interest in these methods is high. The goal of this work was therefore to present a method for stable background subtraction as far as possible independent from environmental conditions and thereby laying the base for persistent traffic monitoring.

A new approach to Gaussian mixture model background subtraction using two modalities has been presented. The proposed algorithm fuses an image representation of background model conformity values of each pixel. This allows to include prior knowledge about the modalities in form of quality characteristics. Different image quality heuristics based on image structure and external information sources have been investigated and specified. To match requirements derived from the purpose of traffic monitoring, extensions to the core contribution have been introduced.

The proposed method has been thoroughly tested. The results show a significantly better performance of the proposed method than comparable background subtraction algorithms. Special situation performance suggests that the strategy of including image quality heuristics in the segmentation process has great potential. The modularity of the method allows the implementation of improvements that have been developed originally for the background subtraction by Stauffer and Grimson [28].

A common problem of image fusion techniques can be seen from the experimental results. Although the algorithm features a suppression of false positives, a propagation to the fused mask can still be noticed. This is especially the case, when the quality rating of the two modalities is similar and therefore information fuses in equal proportions. Based on this observation further development of the proposed method can be derived. Serrano-Cuerda et al. [25] perform a switch based on image quality heuristics. This work performs an adaptive fusion. The next logical step would be to perform the fusion adaptive per image region. Similar approaches based on image characteristics rather than quality heuristics have been presented. The approaches however do not differentiate between different conditions such as night and day. By specifying quality heuristics for image samples, information about shadows and different lighting conditions within the scene, the main reasons for false positives, could be considered.

The work has been limited to the usage of RGB and thermal imagery. The algorithm however can easily be adapted to work with different imaging sensors. A setup of the proposed system in combination with sensors helping to estimate the image quality would also be an interesting extension. Weather stations and street temperature sensors would enable the heuristics to work much more accurately.

References

1. Bas, E., Tekalp, A.M., Salman, F.S.: Automatic vehicle counting from video for traffic flow analysis. In: Intelligent Vehicles Symposium, 2007 IEEE, pp. 392–397. IEEE (2007)
2. Buch, N., Velastin, S.A., Orwell, J.: A review of computer vision techniques for the analysis of urban traffic. Intelligent Transportation Systems, IEEE Transactions on **12**(3), 920–939 (2011)
3. Chapman, L., Thornes, J.: A geomatics-based road surface temperature prediction model. Science of the Total Environment **360**(1), 68–80 (2006)
4. Chen, S., Leung, H.: An em-ci based approach to fusion of ir and visual images. In: Information Fusion, 2009. FUSION'09. 12th International Conference on, pp. 1325–1330. IEEE (2009)
5. Chen, T.H., Chen, J.L., Chen, C.H., Chang, C.M.: Vehicle detection and counting by using headlight information in the dark environment. In: Intelligent Information Hiding and Multimedia Signal Processing, 2007. IIHMSP 2007. Third International Conference on, vol. 2, pp. 519–522. IEEE (2007)
6. Chen, T.H., Lin, Y.F., Chen, T.Y.: Intelligent vehicle counting method based on blob analysis in traffic surveillance. In: Innovative Computing, Information and Control, 2007. ICICIC'07. Second International Conference on, pp. 238–238. IEEE (2007)
7. Cheung, S.C.S., Kamath, C.: Robust background subtraction with foreground validation for urban traffic video. EURASIP Journal on Advances in Signal Processing **2005**(14), 2330–2340 (2005)
8. Conaire, C.O., O'Connor, N.E., Cooke, E., Smeaton, A.F.: Comparison of fusion methods for thermo-visual surveillance tracking. In: FUSION, pp. 1–7 (2006)
9. Davis, J.W., Sharma, V.: Background-subtraction using contour-based fusion of thermal and visible imagery. Computer Vision and Image Understanding **106**(2), 162–182 (2007)
10. Gade, R., Moeslund, T.B.: Thermal cameras and applications: a survey. Machine vision and applications **25**(1), 245–262 (2014)
11. Gregory, R.L.: Eye and brain: The psychology of seeing . Princeton university press (1997)
12. Hall, D., Llinas, J.: Multisensor data fusion. CRC press (2001)
13. Heather, J.P., Smith, M.I.: Multimodal image registration with applications to image fusion. In: Information Fusion, 2005 8th International Conference on, vol. 1, pp. 8–pp. IEEE (2005)
14. Hrkać, T., Kalafatić, Z., Krapac, J.: Infrared-visual image registration based on corners and hausdorff distance. In: Image Analysis, pp. 383–392. Springer (2007)
15. Istenic, R., Heric, D., Ribaric, S., Zazula, D.: Thermal and visual image registration in hough parameter space. In: Systems, Signals and Image Processing, 2007 and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. 14th International Workshop on, pp. 106–109. IEEE (2007)
16. Kamijo, S., Matsushita, Y., Ikeuchi, K., Sakauchi, M.: Traffic monitoring and accident detection at intersections. Intelligent Transportation Systems, IEEE Transactions on **1**(2), 108–118 (2000)
17. Kastrinaki, V., Zervakis, M., Kalaitzakis, K.: A survey of video processing techniques for traffic applications. Image and vision computing **21**(4), 359–381 (2003)
18. Kwon, H., Der, S.Z., Nasrabadi, N.M.: Adaptive multisensor target detection using feature-based fusion. Optical Engineering **41**(1), 69–80 (2002)
19. Lallier, E., Farooq, M.: A real time pixel-level based image fusion via adaptive weight averaging. In: Information Fusion, 2000. FUSION 2000. Proceedings of the Third International Conference on, vol. 2, pp. WEC3–3. IEEE (2000)
20. Lei, M., Lefloch, D., Gouton, P., Madani, K.: A video-based real-time vehicle counting system using adaptive background method. In: Signal Image Technology and Internet Based Systems, 2008. SITIS'08. IEEE International Conference on, pp. 523–528. IEEE (2008)
21. Messelodi, S., Modena, C.M., Zanin, M.: A computer vision system for the detection and classification of vehicles at urban road intersections. Pattern analysis and applications **8**(1-2), 17–31 (2005)
22. Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A.A., Jarawan, E., Mathers, C.: World report on road traffic injury prevention (2004)
23. Prati, A., Mikic, I., Trivedi, M.M., Cucchiara, R.: Detecting moving shadows: algorithms and evaluation. Pattern Analysis and Machine Intelligence, IEEE Transactions on **25**(7), 918–923 (2003)
24. Robert, K.: Night-time traffic surveillance: A robust framework for multi-vehicle detection, classification and tracking. In: Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on, pp. 1–6. IEEE (2009)
25. Serrano-Cuerda, J., Fernández-Caballero, A., López, M.T.: Selection of a visible-light vs. thermal infrared sensor in dynamic environments based on confidence measures. Applied Sciences **4**(3), 331–350 (2014)
26. Shah, P., Merchant, S., Desai, U.B.: Fusion of surveillance images in infrared and visible band using curvelet, wavelet and wavelet packet transform. International Journal of Wavelets, Multiresolution and Information Processing **8**(02), 271–292 (2010)
27. St-Laurent, L., Maldague, X., Prévost, D.: Combination of colour and thermal sensors for enhanced object detection. In: Information Fusion, 2007 10th International Conference on, pp. 1–8. IEEE (2007)
28. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on., vol. 2. IEEE (1999)
29. Vargas, M., Toral, S., Barrero, F., Milla, J.: An enhanced background estimation algorithm for vehicle detection in urban traffic video. In: Intelligent Transportation Systems, 2008. ITSC 2008. 11th International IEEE Conference on, pp. 784–790. IEEE (2008)
30. Yao, L., Ling, M.: An improved mixture-of-gaussians background model with frame difference and blob tracking in video stream. The Scientific World Journal **2014** (2014)
31. Zangenehpour, S., Miranda-Moreno, L.F., Saunier, N.: Automated classification in traffic video at intersections with heavy pedestrian and bicycle traffic. In: Transportation Research Board 93rd Annual Meeting, 14-4337 (2014)
32. Zivkovic, Z.: Improved adaptive gaussian mixture model for background subtraction. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, vol. 2, pp. 28–31. IEEE (2004)
33. Zou, Y., Shi, G., Shi, H., Wang, Y.: Image sequences based traffic incident detection for signaled intersections using hmm. In: Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on, vol. 1, pp. 257–261. IEEE (2009)