Robust Object Detection for Video Surveillance Using Stereo Vision and Gaussian Mixture Model

Lars Klitzke, Carsten Koch
Hochschule Emden/Leer
University of Applied Sciences
Department of Informatics and Electronics
Constantiaplatz 4
26723 Emden, Germany
{Lars.Klitzke | Carsten.Koch}@hs-emden-leer.de

ABSTRACT

In this paper, a novel approach is presented for intrusion detection in the field of wide-area outdoor surveillance such as construction site monitoring, using a rotatable stereo camera system combined with a multi-pose object segmentation process.

In many current surveillance applications, monocular cameras are used which are sensitive to illumination changes or shadow casts. Additionally, the object classification, spatial measurement and localization using the 2D projection of a 3D world is ambiguous. Hence, a stereo camera is used to calculate a 3D point cloud of the scenery which is nearly unaffected by illumination changes, therefore enabling robust object detection and localization in the 3D space. The limited viewing range of the stereo camera is expanded by mounting it onto a rotatable tripod. To detect objects in different poses of the camera, pose specific Gaussian Mixture Models (GMM) are used. However, changing illumination outside the current field of view of the camera or spontaneously changing lighting conditions caused by e.g. lights controlled by motion sensors, would lead to false-positives in the segmentation process if using the brightness values. Hence, segmentation is performed using the calculated point cloud which is demonstrated to be robust against changing illumination and shadow casts by comparing the results of the proposed method with other state of the art segmentation methods using a database of self-captured images of a public outdoor area.

Keywords

Video Surveillance, Gaussian Mixture Model, Stereo Vision, Object Detection

1 INTRODUCTION

The usage of video systems for surveillance applications in public, industrial and private domains has been increasingly popular in the last years. Ongoing technological development led to smarter systems which enable automatic identification of critical situations or suspicious objects. The aim of those so called Intelligent Video-System (IVS) is to relieve the strain on human security guards of these systems, because they are typically monitoring multiple screens simultaneously and need to reliably detect salient behaviour. However, this is a challenging task even if they just need to monitor two screens at the same time due to the effects of fatigue and hence, inattentiveness [LCK13]. IVS can reduce the cost of video surveillance systems and increase the productivity at the same time.

The vast majority of current outdoor video surveillance systems uses monocular cameras in which the process of image segmentation is challenging. The cameras are sensitive to shadow casts of objects and (spontaneously) changing illumination conditions. Typical environments of outdoor surveillance are influenced by effects of artificial lights sources which may be (de)activated spontaneously. Additionally, the real world position and true size of detected objects cannot be determined unambiguously by using the 2D projection of the 3D world. Nevertheless, these information may be relevant for surveillance systems e.g. to locate objects in an area accurately or to track their movement in the real world. Additionally, objects can be classified using these informations as e.g. humans, animals or trees to decrease the false-positive object detection rate.

In this paper, we present an approach for intrusion detection in the field of wide area outdoor surveillance, e.g. construction site monitoring by using a multi-pose object segmentation process combined with a rotating stereo camera.

The calculated 3D point cloud of the stereo camera system is used to detect objects. The generation of this point cloud is nearly unaffected by changing lighting conditions. For segmentation, a system pose specific Gaussian Mixture Models (GMM) is used to detect objects in a wider area without false-positives caused by changing illumination outside the current field of view.

It is shown that the 3D-segmentation process by analysis of distance information instead of the brightness values, which are very sensitive to varying illumination conditions, is more robust in cases of overlapping objects, changing lighting conditions and shadow casts. The outdoor scene image datasets generated for this work have been made available online[1].

This paper is structured as following: Section 2 gives an overview of related work on camera surveillance applications. The proposed method for object detection using the calculated 3D point cloud is described in section 3. In section 4 the rotating stereo system is shown and the segmentation process is compared to other state of the art methods. A conclusion of this paper and an overview about further developments is given in section 5.

2 RELATED WORK

The field of camera based surveillance systems has been broadly addressed in the last decades. For instance Haritaoglu *et al.* [HHD98] presented a system called W^4S which uses a stereo camera combined with an intensity based model. They show that the segmentation process is more robust in case of overlapping objects using the distance information instead of the intensity values.

Another approach was presented by Douret and Benosman [DB04]. They used a network of cameras for an intelligent traffic control system to retrieve the height of objects by assuming a plane ground model. However, Kumar *et al.* [KMP09] showed that a missing link between object and ground can lead to inaccurate position information. Therefore, they compared the result of their proposed stereo localization procedure using two pan-tilt-zoom (PTZ) cameras with a monocular approach. With the PTZ cameras they are able to determine the position of objects even when using different zoom levels and in case of overlapping objects. This is realized by using a neural network based interpolation method with an offline calculated look-up table to rectify the images online. However, compared to the rectification process of static stereo camera systems, this procedure is more computational expensive.

Nevertheless, because of the flexibility of PTZ camera and due to the need of observing even greater areas than actual static and monocular cameras can cover, much work has been done on surveillance systems using (dual) PTZ cameras [KMP09][ZWW10][ZOS13]. However, all of those systems perform the image segmentation in the 2D space and localize the detected objects in the world afterwards. The following work will present a more robust segmentation method based on the calculated distance information of the stereo camera. A detailed overview of the current state of the art of intelligent video systems is given by Liu *et al.* in [LCK13].

The main task of a video surveillance system is intrusion detection. A robust segmentation of the image into foreground and background is vital for this task. The problem of image segmentation is widely discussed in the field of image processing. Due to this, there are several methods which can be used. Sen-Ching and Kamath [SCK04] evaluate the result of Frame Differencing, Kalman Filter, Median Filter and Mixture of Gaussian (MoG) using an urban video sequence. They showed that the result of MoG (Gaussian Mixture Model (GMM) respectively), which was proposed by Friedman and Russel [FR97] and extended by Stauffer and Grimson [SG99], outperforms the other methods in many cases, e.g. outdoor surveillance.

Due to this and because of the insensitivity to local movement in the scene, e.g. swaying branches and adaptation to changing illumination conditions, the GMM is used in section 3.2, combined with a 3D point cloud. The latter is provided by a rotating stereo camera realising an even more robust segmentation, while at the same time covering a greater area than a static monocular approach without the computational complexity of handling PTZ cameras.

3 INTRUSION DETECTION

3.1 Online calibration

Typically, cameras for surveillance applications are mounted at elevated positions on walls or poles. This leads to a typical constellation as shown in Fig. 1. To directly measure the true size of detected objects, the calculated point cloud using the stereo

[1] `https://mux.hs-emden-leer.de/lkl`

camera needs to be transformed from the camera's coordinate space C to the world coordinate space W. This transformation is used to rotate and translate the point cloud, so that the camera is virtually located on the ground level and points straight ahead. An online calibration method is used in which the user needs to select the ground of the scenery in the image.

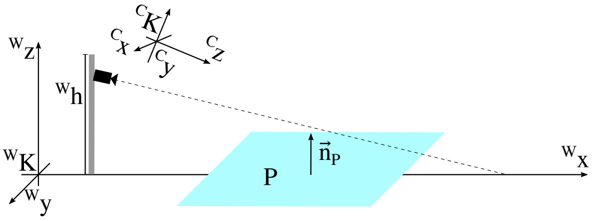


Figure 1: Model of a surveillance camera mounted at a wall or pole looking downwards.

At first, the normal ${}^{C}\vec{n}_P$ of the plane (the selected ground) in the camera's coordinate space is estimated using a method from Kovesi [Kov00] based on RANSAC. Then, the rotation ${}^{C}\mathbf{\Omega}_W$ between ${}^{C}\vec{n}_P$ and the vector ${}^{C}\vec{e} = [0\ 1\ 0]^T$ is calculated with Rodigues' rotation formula [Cor11]. Afterwards, the translation vector ${}^{C}\boldsymbol{\tau}_W$ of the camera relative to the ground can be estimated. Therefore, a median point $\bar{x}$ of the point cloud is calculated and rotated using (1).

$$\bar{x}' = {}^{C}\mathbf{\Omega}_W \bar{x} \tag{1}$$

At least, the translation vector only consists of the third part of the rotated median point (see (2)), because it represents the height of the camera in the world.

$$ {}^{C}\boldsymbol{\tau}_W = \begin{bmatrix} 0 \\ 0 \\ -\bar{x}'_2 \end{bmatrix} \tag{2}$$

As already stated in the beginning, the system is able to pan and tilt. Due to this, the calculated parameters become invalid if the system moves. To overcome this, the user can define specific poses $\xi_1, ..., \xi_n$ of the camera system for each of which the described calibration method needs to be performed once. Hence, a system pose ξ_i is defined by (3).

$$\xi_i = \left[{}^{C}\mathbf{\Omega}_W\ {}^{C}\boldsymbol{\tau}_W \right] \tag{3}$$

3.2 Object detection

In the following, the processing chain (see Fig. 2) for object detection is described using the previously estimated parameters for the system poses. We assume that the stereo camera is calibrated, so we start with the rectified images of the left and right camera. First, the disparity map has to be calculated. The Semi-Global Matching (SGM) algorithm proposed by Hirschmüller is an established method for this task. Based on the disparity map the 3D point cloud of the scenery is reconstructed.

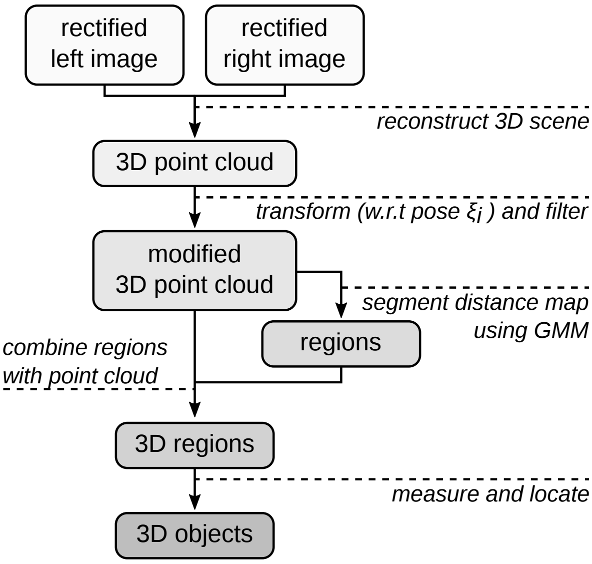


Figure 2: Overview of the processing chain for detecting objects based on the rectified stereo camera images.

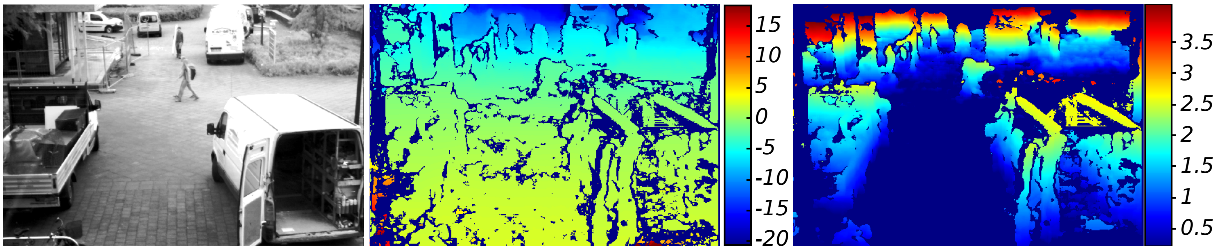
In the next step, this raw point cloud ${}^{C}\mathbf{P}$ is transformed with respect to the current system pose ξ_i from (3) using (4) and the homogeneous coordinates ${}^{C}\tilde{\mathbf{P}}$ of the point cloud ${}^{C}\mathbf{P}$.

$$ {}^{W}\mathbf{P} = \xi_i \cdot {}^{C}\tilde{\mathbf{P}} \tag{4}$$

The component of the point cloud representing the height over the ground with respect to the input image shown in Fig. 3a can be seen in Fig. 3b. Then, a pre-segmentation step is applied using knowledge about the application environment in order to remove points outside the application specific and user defined ranges, e.g. points belonging to the ground or too far away and hence error-prone points. In the case shown in Fig. 3a, the codomain for each axis $\mathbb{W}_{x,y,z}$ is

$$\begin{aligned} \mathbb{W}_z &= \{i | (i \in \mathbb{R}) \wedge (0.3 \leq i \leq 4)\}, \\ \mathbb{W}_x &= \{i | (i \in \mathbb{R}) \wedge (8 \leq i \leq 60)\}, \\ \mathbb{W}_y &= \{i | (i \in \mathbb{R}) \wedge (-20 \leq i \leq 20)\} \end{aligned} \tag{5}$$

to remove points which height over the ground is smaller than 0.3 m, higher than 4 m or with a distance to the camera less than 8 m or greater than 60 m. The horizontal interval $\mathbb{W}_y$ has been selected to contain the entire field of view. The estimation of the ranges for i from (5) is shown in section 4.2.



(a) (b) (c)

Figure 3: Presentation of the transformation and filtering process using the portion of the point cloud, which represents the height over the ground in meter. (a) Rectified input image of the left camera, (b) Raw point cloud, (c) Transformed point cloud using the calculated rotation and translation vector with points outside the user defined range removed (marked as dark blue).

The result of the point cloud transformation followed by the point cloud filtering is shown in Fig. 3c where the points representing the ground are removed (marked as dark blue).

This filtered point cloud is further segmented using a Gaussian Mixture Model [SG99], which is an established method for this task. Therefore, the state of a pixel in respect to foreground or background is modelled by several Gaussian distributions which is demonstrated in section 4.3.

As already stated in the introduction, instead of performing the segmentation in 2D space, the component of the point cloud representing the distance to the camera is used as input for the GMM. The reason for this approach is that the calculated point cloud is nearly unaffected by (spontaneous) illumination changes which may occur on construction sites through lights controlled by motion sensors. Additionally, due to the ability of the system to move between specific poses, the illumination of a scene which is currently not in the camera's field of view, can change. This has great impact on GMM training. Since the GMM would not be able to adapt the background model to this change if intensity values were used, this would lead to pixel misclassification. However, this does not occur when using the point cloud for the previously mentioned reason.

The result of the GMM is a binary mask $\mathbf{M}$ with foreground pixels marked as 1 and background pixels as 0. By applying a morphological opening, the noise in the resulting mask is reduced and foreground regions $\mathbf{R}$ can be selected using blob colouring. Nevertheless, there is still a chance that regions are selected falsely, due to the noisy mask. Therefore, small regions are discarded using (6) with $N(a)$ returning the number of pixels and ρ as region size criterion. Using the empirically determined value of $\rho = 0.005$ gives reasonable results. For each region in $\mathbf{R}'$ the corresponding data of the 3D point cloud $\mathbf{P}$ is aggregated, so that we have a set of 3D regions $\mathbf{U}$, see (7).

$$\mathbf{R}' = \{x|\ x \in \mathbf{R}\ \wedge N(x) > \rho \cdot N(\mathbf{M})\} \quad (6)$$

$$\mathbf{U} = \left\{\mathbf{P}(r)\ |\ r \in \mathbf{R}'\right\} \quad (7)$$

However, it turns out that pixels around the marked regions are occasionally selected in error. Hence, those pixels need to be removed to enhance the subsequent estimation of position and size. A common statistical method for outlier elimination is used in which a sorted set of values is divided into four equal sized groups by determining the quartiles (Q_1, Q_2, Q_3) of the set, with Q_1 as the median value of the total set, Q_2 as the median value of the lower and Q_3 of the higher subset. Then, the interquartile distance $I = Q_3 - Q_1$ is calculated which is used to define the so called lower fence $F_L = Q_2 - \eta I$ and upper fence $F_U = Q_3 + \eta I$ with η as spreading factor. Finally, values outside the range of $[F_L; F_U]$ are considered as outliers and removed from the dataset.

This method is used to create a set of 3D objects $\mathbf{O}$ by estimating the position $L(\mathbf{U}_i)$ and size $S(\mathbf{U}_i)$ of an object $\mathbf{U}_i$.

$$\mathbf{O} = \{(\mathbf{L}(u), \mathbf{S}(u))\,|\ u \in \mathbf{U}\} \quad (8)$$

The position is then defined by the distances to the optical axis of the virtual left camera in each direction with $L(\mathbf{U}_i) = [d_x, d_y, d_z]$ and the size describes the width and height of the object $S(\mathbf{U}_i) = [w, h]$ whereas $d_z = h$. In case of the distance in the x axis the previously described method is applied to the subset $\mathbf{U}_{i,x}$ which mean value corresponds to d_x. All of those values in $\mathbf{U}_{i,x}$ which are marked as outliers are also removed from the set $\mathbf{U}_i$, so that these values are ignored in the following calculations.

The width of the object $\mathbf{U}_i$ is defined as the difference between the left and right edge of the object. These edges are estimated by calculating the mean of the subsets $\mathbf{M}_l$ and $\mathbf{M}_r$ of $\mathbf{U}_{i,y}$ with $N(\mathbf{M}_{l,r}) \geq 0.1\eta N(\mathbf{U}_{i,y})$, whereas $\mathbf{M}_l$ represents the subset for the left and $\mathbf{M}_r$ for the right edge. Additionally outliers of those subsets are removed using the previously described approach. This could on the one hand lead to inaccurate width estimations in case of high value changes, but on the other hand ensures that the edges are not defined by a single value which might be inaccurate. This procedure is also used to estimate the height h of an object whereas the subset $\mathbf{U}_{i,z}$ is used. Finally, d_y is defined as the left edge of the object increased by half the object's width.

The last step in the processing chain is object tracking. Currently, only a naive method is used to compare the currently detected 3D objects with already known ones. Two objects are considered equal if the size of the newly detected object is in the range of two standard deviations from the already known object's size measurements and the location difference of the detected object to the known one is in the range of two standard deviation from the known object location changes. For that reason, each 3D object contains a list of timestamped size and location measurements.

4 EXPERIMENTS

4.1 Stereo camera system architecture

To develop and evaluate the method presented in the last section, a dataset of images is required and this in turn requires a stereo camera system. As already mentioned in the introduction, the camera system is motorized to perform pan and tilt movement. This is accomplished by mounting the stereo camera, which consists of two monochrome GigE cameras from TheImagingSource (model DMK-23GM021) with a resolution of 1280 × 920 pixel to a pan/tilt system of Invescience LC as illustrated in Fig. 4. The pan/tilt actuators are controlled with a custom application running on a connected PC.

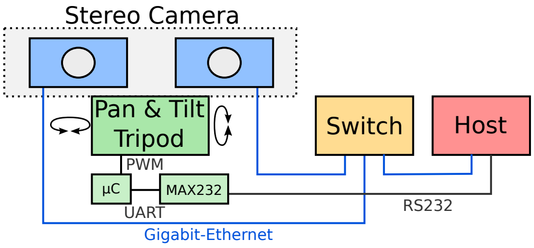


Figure 4: Architecture of the stereo camera surveillance system.

4.2 Application model

Designing a stereo camera system for a specific application or environment is a challenging task. This is due to the fact that the system is influenced by various factors, e.g. the baseline width, the image resolution and the elevation of the camera's planned location [LSP+10]. Changing any parameter of the system directly impacts the field of view of the stereo system or the precision of the calculated distance at a pixel [LSP+10]. Furthermore, the range of disparity values also depends on the target application. Additionally, determination of minimum and maximum values of disparity reduces the search space and processing times. Due to this, a model was created which is used to simulate a stereo camera system attached to a wall or pole for a specific surveillance application (see Fig. 5).

With respect to the given application parameters, the theoretically calculated disparity D of an object at a specific distance Z (blue box in Fig. 5) is calculated by (9) and the absolute depth estimation error $|\delta Z|$ is calculated using (10) as stated by Chang and Chatterjee [CC92].

$$D = \frac{bf_x}{Z} \tag{9}$$

$$|\delta Z| = \frac{bf_x}{D^2}\delta D \tag{10}$$

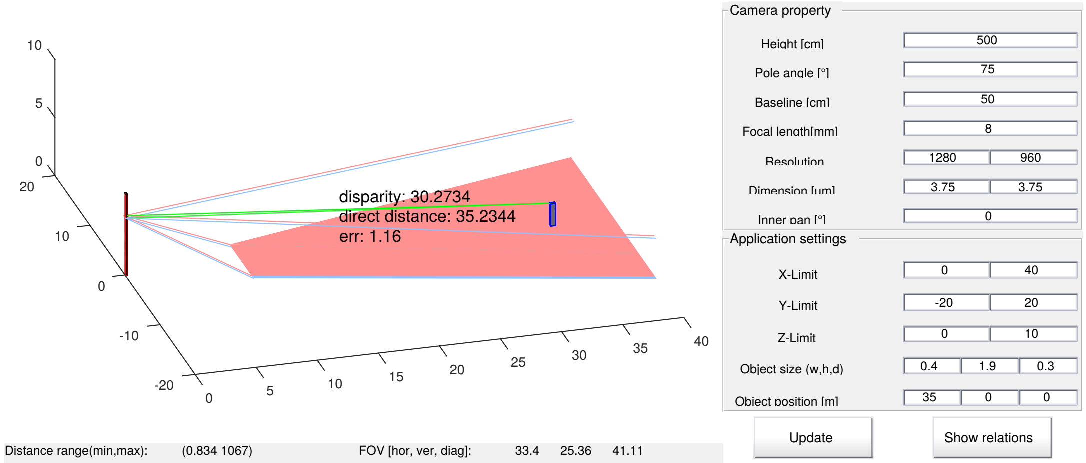
Here, f_x is the focal length in pixels, b is the baseline in meters, D is the disparity and δD is the uncertainty of the estimated disparity in pixels, which is assumed to be 1 in the model as a worst case value. By changing the position of the object, the relevant disparity range and the theoretical distance error can be estimated with respect to the stereo system parameters. Additionally, this model is used to estimate the application-specific ranges used in section 3.2 for the pre-segmentation of the point cloud. This model is made publicly available on the Mathworks file exchange platform[2].

4.3 Demonstration

Using the estimated parameters of the stereo camera, stereo images for evaluation of the proposed object detection method were acquired. In Fig. 6 the stereo system is shown with a baseline width of 15 cm. The datasets for evaluation were recorded using a baseline width of 55 cm in order to decrease the depth error.

The distance value of a specific pixel is representatively monitored over 592 images (see Fig. 7a and 7b). The pixel's state is modelled using three Gaussian distributions, see Fig. 7c. In Fig. 7b three distinctive situations are shown which correspond to Fig. 7d-7f showing objects crossing that pixel.

[2] `http://mathworks.com/matlabcentral/fileexchange/55420-stereo-camera-application-model`



Camera property

Height [cm]	500		
Pole angle [°]	75		
Baseline [cm]	50		
Focal length[mm]	8		
Resolution	1280	960	
Dimension [um]	3.75	3.75	
Inner pan [°]	0		

Application settings

X-Limit	0	40	
Y-Limit	-20	20	
Z-Limit	0	10	
Object size (w,h,d)	0.4	1.9	0.3
Object position [m]	35	0	0

Distance range(min,max): (0.834 1067) FOV [hor, ver, diag]: 33.4 25.36 41.11

Figure 5: Model of a stereo camera attached to a wall or pole. Areas in blue and red represents the field of view of the left and right cameras. Parameters of the system can be changed with the text fields on the right side. Parameter-dependend informations are shown in the figure.

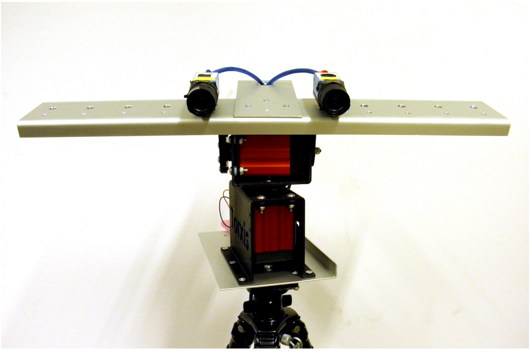


Figure 6: Stereo system with a baseline width of 15 cm and a motorized tripod head for pan and tilt movement.

Table 1 shows the parameters of three distributions over 592 images, sorted by their fitnesses. The dominant distribution is D_1 with a mean of 15.686, variance of 0.084 and a fitness of 3.378 representing the background state of the pixel. This demonstrates the feasibility of robust foreground-background segmentation using distance information and a GMM, therefore enabling the performance of object detection using this approach. Additionally, it can be seen that D_2 represents the value range of the detected objects. This demonstrates the ability of the GMM to perform multi-modal background modelling.

4.4 Evaluation

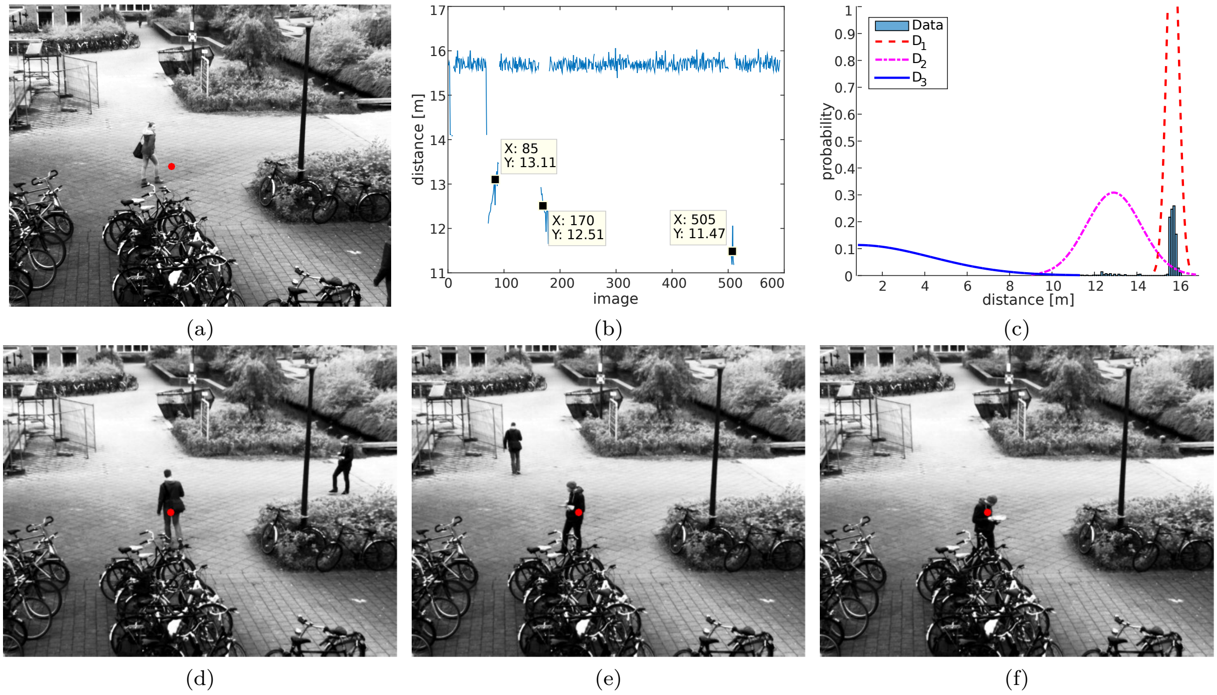
The method proposed in this paper (in the following referred to as `GMMD`) is evaluated for use in the field of outdoor surveillance, e.g. construction site monitoring, which is influenced by changing illumination and characterized by a dynamic background, by comparing it with other state of the art methods.

Parameter	$\mathbf{D_1}$	$\mathbf{D_2}$	$\mathbf{D_3}$
Variance	0.084	1.673	12.352
Mean	15.686	12.904	0.778
Weight	1	0.707	0.367
Fitness	3.378	0.546	0.104

Table 1: Parameters of the three distributions $\mathbf{D_1}$, $\mathbf{D_2}$, $\mathbf{D_3}$ modelling the selected pixel's state after 592 images.

A dataset of 3716 timestamped images with a frame rate of 10 images per seconds is recorded. The stereo camera system was placed at the first floor of the Hochschule Emden/Leer covering the campus as already seen in Fig. 3a. This dataset includes situations with global illumination changes, shadow casts and overlapping objects. The first is caused by manual camera aperture manipulation with varying speed to simulate global intensity changes. From theses images, 15 situations are manually labelled using the *Interactive Segmentation Tools* of McGuiness and O'Connor [MO10] for the ground truth data. Thereby, even foreground objects are marked which are already a part of the scene from the beginning of the sequence and are more or less static.

The result of the `GMMD` is compared with the results of the GMM using the greyscale image of the left camera (`GMM`), Frame Differencing (`FD`) and Median Filter (`MD`). The learning rate of the `GMMD` and `GMM` is set to 0.005 because of the image capturing frame rate. Three distributions are used. The initial variances of the `GMMD` and the `GMM` are set to 0.5 and 0.2 respectively. This is due to the fact that the distance estimation is assumed to be more noisy than the grayscale image with values in the range of $[0, 1]$. Additionally, due to the frame rate



(a) (b) (c)

(d) (e) (f)

Figure 7: Results of the background modelling process for one pixel over 592 images. (a) Image showing the selected pixel. (b) Plot of the distance values, (c) Plot showing the probability of a distance to occur overlaid with the three distributions of the GMM, (d)-(f) Images of the three distinctive situations.

the `FD` compares only each fourth image to ensure movement in the image.

For each of the situations and methods, the recall and precision value [SCK04] is calculated to quantify the results of the methods in respect with their resulting foreground masks $M_{\text{GMMD}}, M_{\text{GMM}}, M_{\text{FD}}, M_{\text{MD}}$ and the ground truth mask G using (11) and (12) respectively.

$$\text{Recall}(M) = \frac{|M_{\text{correct marked}}|}{|G_{\text{marked}}|} \quad (11)$$

$$\text{Precision}(M) = \frac{|M_{\text{correct marked}}|}{|M_{\text{marked}}|} \quad (12)$$

The results of the methods are shown in Fig. 8 whereas T is a threshold for classifying foreground and background pixel with respect to the pixel value changes. In general, methods performing well have a high recall and precision value. However, it is evident that none of the tested methods reach a recall value greater than 62%, which in some extend depends on the ground truth masks. This is due to the fact that static objects are also labelled as foreground but can not be detected by the tested methods. The results show the `GMMD` performing reasonable well in all situations with an average precision of 72.8% and a relatively low standard deviation of 8.5%. No other method has shown behavior this robust (see table 2). For instance, the precision of the `GMM` has a much higher standard deviation (28.4%). In Fig. 9 five prominent situations are shown which describe the behaviour of the `GMM`. The situation (3) corresponds to the rightmost dataset of Fig. 8 with a precision of 5.24% because of a sudden change in the global intensity which results in an inverted foreground mask [SCK04]. The proposed `GMMD` however is unaffected by these changes and produces a foreground mask with a precision of 77% and a recall of 37%. The situations (2) and (5) in Fig. 9 show the classification

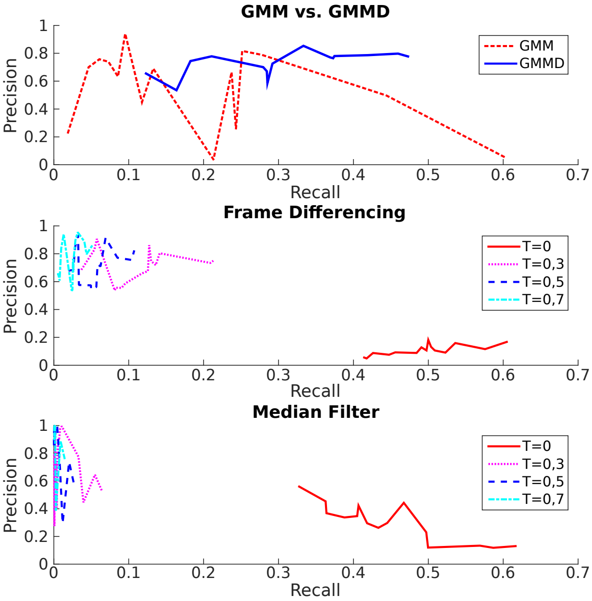


Figure 8: Results for the manually labelled images

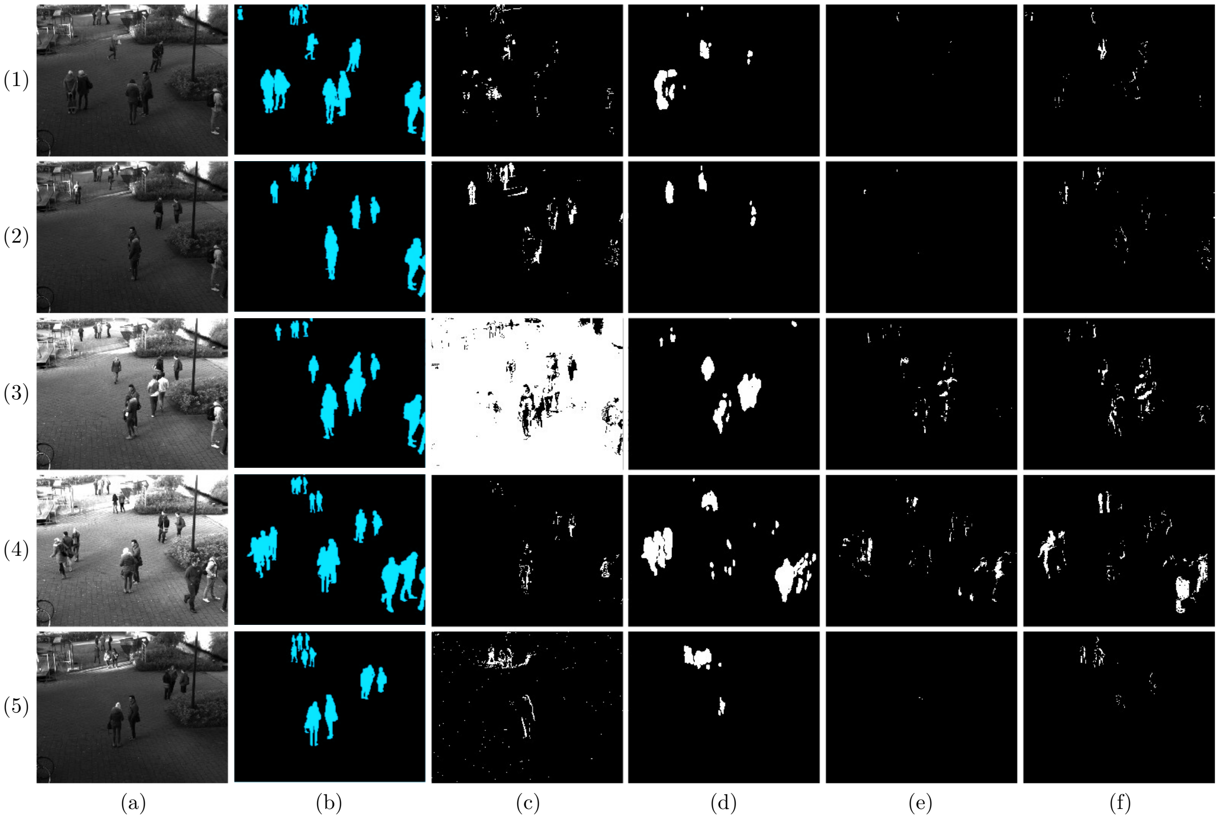


Figure 9: Results of the methods for five prominent situations. (a) Rectified left image, (b) Ground truth foreground mask, (c) Result GMM, (d) Result GMMD, (e) Result MF, (f) Result FD

of shadow casts as foreground pixels in the case of the GMM. However, situation (1) and (2) shows an advantage of the GMM over the GMMD because it even detects far-away objects.

	Recall		Precision	
Dataset	**Mean**	**Std. Dev.**	**Mean**	**Std. Dev.**
GMM	0.194	0.164	0.551	0.284
GMMD	0.308	0.106	0.728	0.085
MF $(T=0)$	0.452	0.086	0.301	0.137
MF $(T>0)$	0.008	0.014	0.798	0.220
FD $(T=0)$	0.492	0.055	0.109	0.039
FD $(T>0)$	0.065	0.050	0.732	0.127

Table 2: Results of the experiment. The results of MF and FD using a threshold greater zero are merged.

5 CONCLUSION

In this paper an approach for object detection in the field of outdoor surveillance for e.g construction site monitoring was presented which combines an actuated stereo camera system with camera pose specific Gaussian Mixture Models (GMM). A novel processing chain for detection of objects based on a calculated 3D point cloud and the current camera pose was described. Additionally, the actuated stereo surveillance system is described and a model is presented for the estimation of application-specific parameters which simplifies the stereo camera system design process.

Furthermore, the presented approach is compared to other state of the art methods using a self captured image database. With an average precision of 72.84% and a recall value of 30.82% it outperforms the other methods. Additionally, it was shown that the proposed method is robust against changing illumination and shadow casts which often occurs in outdoor surveillance applications like construction site monitoring even while moving the stereo camera. However, overexposed pixels cause an incomplete distance map due to the pixel correlation process for the disparity calculation and hence lead to an inaccurate segmentation and the detection range of the method is limited by the stereo camera's distance calculation error.

For a more robust identification of objects, the current naive matching and tracking method need to be extended in future work. Additionally, for 1280 × 960 images, the current disparity map calculation time on a single threaded i7-3770 PC is 0.44 s per image without hardware acceleration. Follow-up works on GPU, FPGA [GEM09] or even on CPU [SLAR14] should lead to a higher number of frames analysed per second. Additional performance gains can be achieved by parallelization of the processing chain presented in section 3.2 using a pipeline architecture which is typically used in the field of processor design.

Currently, the ground is assumed to be plane which is typically not the case in the real world. Hence, the ground selection phase of the online calibration process need to be extended to build a more realistic model of the ground in order to improve the object measurement and localization.

Furthermore, the application of the morphological operation for noise reduction could split up objects into multiple regions and hence multiple 3D objects which degrades the object detection rate. In follow-up works this can be addressed by clustering regions with respect to their 3D representation to enhance the object detection process.

REFERENCES

[CC92] C. Chang and S. Chatterjee. Quantization Error Analysis in Stereo Vision. In *Signals, Systems and Computers, 1992. 1992 Conference Record of The Twenty-Sixth Asilomar Conference on*, pages 1037–1041 vol.2, October 1992.

[Cor11] Peter I. Corke. *Robotics, Vision & Control: Fundamental Algorithms in Matlab*. Springer, 2011.

[DB04] J. Douret and R. Benosman. A multi-cameras 3D volumetric method for outdoor scenes: a road traffic monitoring application. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference*, volume 3, pages 334–337 Vol.3, Aug 2004.

[FR97] Nir Friedman and Stuart Russell. Image Segmentation in Video Sequences: A Probabilistic Approach. In *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pages 175–181. Morgan Kaufmann Publishers Inc., 1997.

[GEM09] Stefan K. Gehrig, Felix Eberli, and Thomas Meyer. A Real-Time Low-Power Stereo Vision Engine Using Semi-Global Matching. In Mario Fritz, Bernt Schiele, and Justus Piater, editors, *Proceedings of the 7th International Conference on Computer Vision Systems: Computer Vision Systems*, volume 5815 of *Lecture Notes in Computer Science*, pages 134–143. Springer, 2009.

[HHD98] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W^4S: A Real-Time System for Detecting and Tracking People in 2 1/2D. In *Computer Vision-ECCV'98*, pages 877–892. Springer, 1998.

[KMP09] Sanjeev Kumar, Christian Micheloni, and Claudio Piciarelli. Stereo Localization Using Dual PTZ Cameras. In *Computer Analysis of Images and Patterns*, volume 5702 of *Lecture Notes in Computer Science*, pages 1061–1069. Springer Berlin Heidelberg, 2009.

[Kov00] P. D. Kovesi. MATLAB and Octave Functions for Computer Vision and Image Processing, 2000. Available from: `http://www.peterkovesi.com/matlabfns/`.

[LCK13] Honghai Liu, Shengyong Chen, and Naoyuki Kubota. Intelligent Video Systems and Analytics: A Survey. *Industrial Informatics, on IEEE Transactions*, 9(3):1222–1233, August 2013.

[LSP$^+$10] David F. Llorca, Miguel A. Sotelo, Ignacio Parra, Manuel Ocaña, and Luis M. Bergasa. Error Analysis in a Stereo Vision-Based Pedestrian Detection Sensor for Collision Avoidance Applications. *Sensors*, 10(4):3741–3758, 2010.

[MO10] Kevin McGuinness and Noel E. O'Connor. A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition*, 43(2):434–444, February 2010.

[SCK04] S. Cheung Sen-Ching and Chandrika Kamath. Robust techniques for background subtraction in urban traffic video. In *Electronic Imaging 2004*, pages 881–892. International Society for Optics and Photonics, 2004.

[SG99] Chris Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, on 1999. IEEE Computer Society Conference*, volume 2. IEEE, 1999.

[SLAR14] R. Spangenberg, T. Langner, S. Adfeldt, and R. Rojas. Large scale semi-global matching on the cpu. In *Intelligent Vehicles Symposium Proceedings, 2014 IEEE*, pages 195–201, June 2014.

[ZOS13] Guang Zheng, Shunichiro Oe, and Zengguo Sun. Moving Object Tracking and 3D Measurement Using Two PTZ Cameras. In *Intelligent Networking and Collaborative Systems (INCoS), on 2013 5th International Conference*, 2013.

[ZWW10] Jie Zhou, Dingrui Wan, and Ying Wu. The Chameleon-Like Vision System. *Signal Processing Magazine, IEEE*, 27(5):91–101, September 2010.