Comparison of Static Background Segmentation Methods

Mustafa Karaman, Lutz Goldmann, Da Yu and Thomas Sikora

Technical University of Berlin, Department of Communication Systems
Einsteinufer 17, Berlin, Germany

ABSTRACT

In the case of a static or motion compensated camera, static background segmentation methods can be applied to segment the interesting foreground objects from the background. Although a lot of methods have been proposed, a general assessment of the state of the art is not available. An important issue is to compare various state of the art methods in terms of quality (accuracy) and computational complexity (time and memory consumption). A representative set of recent techniques is chosen, implemented and compared to each other. An extensive set of videos is used to achieve comprehensive results. Both indoor and outdoor videos with different environmental conditions are used. While visual analysis is used for subjective assessment of the quality, pixel based measures based on available ground truth data are used for the objective assessment. Furthermore the computational complexity is estimated by measuring the elapsed time and memory requirements of each algorithm. The paper summarizes the experiments and considers the assets and drawbacks of the various techniques. Moreover, it will give hints for selecting the optimal approach for a specific environment and directions for further research in this field.

Keywords: change detection, segmentation, background model, color, edges, motion, shadow

1. INTRODUCTION

Video segmentation known as moving object segmentation has attracted the attentions of many researchers in the field of image and video processing, because it plays a very important role in many applications, such as surveillance system, traffic monitoring, object based compression. Although video segmentation has been studied for several decades, it still remains a difficult problem for a computer to automatically and accurately segment moving objects from video sequences. This so called spatio-temporal segmentation combines spatial methods (based on image segmentation) and temporal methods (optical flow, image differencing). The choice of the techniques used depends strongly on the environment conditions and the application.

In some cases (static camera or motion compensated camera), it is assumed that the background is fixed and that differences are solely caused by foreground objects. This assumption leads to the special case of static background segmentation methods, that utilize differencing methods based on the background and the actual frame to obtain the interesting foreground objects.

Although a large number of algorithms exist for this special domain, a general assessment of the current state of the art is not available. This is mainly caused by different application scenarios and the implied environment characteristics. Since segmentation is often used as a preliminary step for further analysis, an objective comparison of the different approaches is needed in order to choose the best suiting method for a specific environment. Thus, our challenge is to compare various state of the art methods in terms of both segmentation quality (accuracy of the obtained object masks) and segmentation complexity (time and memory consumption).

The paper is organized as follows. Section 2 reviews all considered state of the art methods for static background segmentation. In section 3 the experiments and results are summarized. It also includes a description of the used dataset and evaluation criteria.

Further author information: (Send correspondence to M.K.)
M.K.: E-mail: karaman@nue.tu-berlin.de, Telephone: +49 (0)30 314 25451

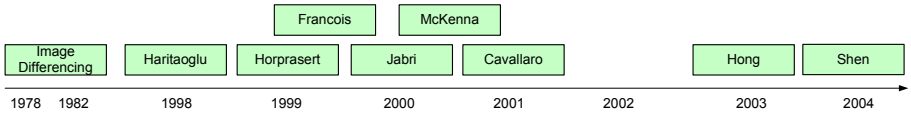

Figure 1. Timeline of the methods.

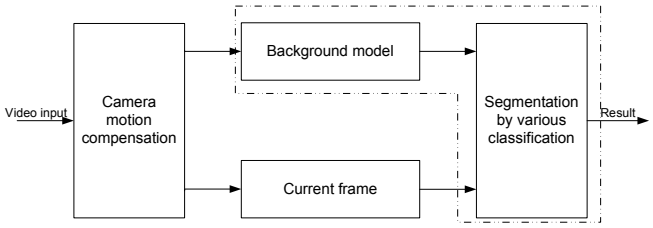

Figure 2. Background segmentation overview.

2. STATE OF THE ART

Figure 1 shows all the methods which are investigated for comparison. It covers the last eight years apart from the simplest method. In this section all methods are reviewed shortly and for more details the appropriate references are mentioned.

2.1. Image Differencing (1982)

The simple image differencing is the first method where the results depends only on the selected threshold value by using different algorithms such as of Otsu,[1] Pun,[2] etc. This method is only mentioned for the evaluation of each method complexity (see Section4).

2.2. Haritaoglu (1998)

Haritaoglu[3] is the one of all investigated methods which uses only grayscale information. The figure 3 describes the overview of the system. First of all, a background model is generated with number of frames in which each pixel is described with three values as the following: minimal intensity (M), maximal intensity (N) and maximal difference value of two successive frames. The difference images are calculated with current image and both the M image and N image. These are used for the classification in which the foreground pixels are given if the difference values are greater than the values of the maximal interframe difference. After the binarization some post processing steps are also needed for elimation of noises. Furthermore, the classified background pixel are then used for updating the background model for considering sudden environmental changes.

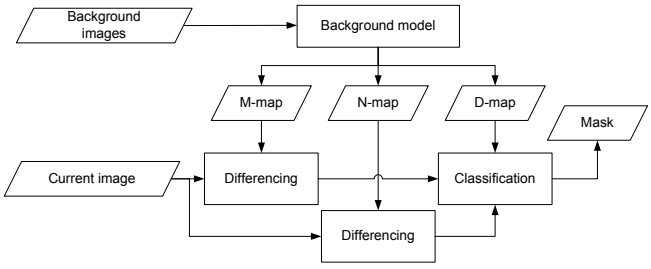

Figure 3. System overview of Haritaoglu.

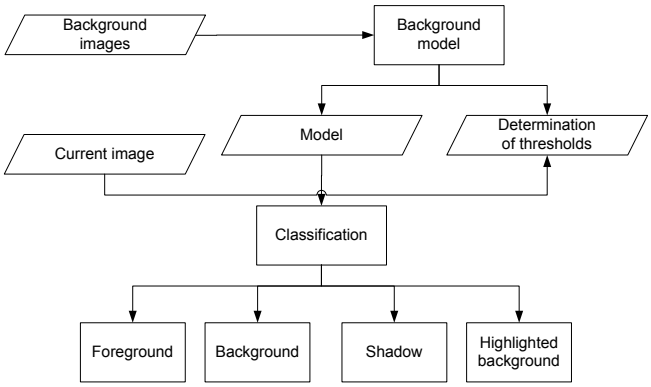

Figure 4. System overview of Horprasert.

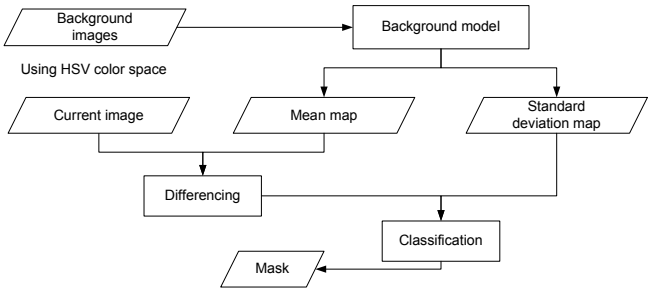

Figure 5. System overview of Francois.

2.3. Francois (1999)

Francois[4] assumes that in the background only very slow global changes can be occured and further the color values of each pixel build a sphere cluster in the RGB color space. With these assumption a background model as a Gaussian distribution is generated by considering the mean value and standard deviation for each pixel. In the system of Francois (Figure 5) the HSV color space is used instead of the RGB. The current image is subtracted from the mean value model and the resulted difference values of each pixel give the information of classifying to either foreground or background regarding to the standard deviation model. Moreover, an update of the background model is also given.

2.4. Horprasert (1999)

Horprasert[5] assumes that the luminance and chrominance has to be separated on the RGB color space by generating a new color model. In that, there is an expected chromaticity line in which the pixel value should be kept. The expected chromaticity is obtained by the arithmetric means of each pixels RGB values calculated over a number of background images. The distortion from this line is given as both chromaticity and brightness distortion being generated by standard deviation. With these distortions several thresholds are determined to classify the pixel to one of the following types: foreground, background, shadow, and highlighted background. Accordingly to this, an incoming image can be classified as one of the last mentioned.

2.5. McKenna (2000)

Similar to Francois's assumption McKenna[6] models the background with mean value and standard deviation, too. However, the sytem considers two information the normalized rgb color space and the edge. For each channel the models are generated. Figure 6 shows the appropriate system where with a number of frames two models are calculated for color and edge separately. For both issues the current image is converted to the adequate form

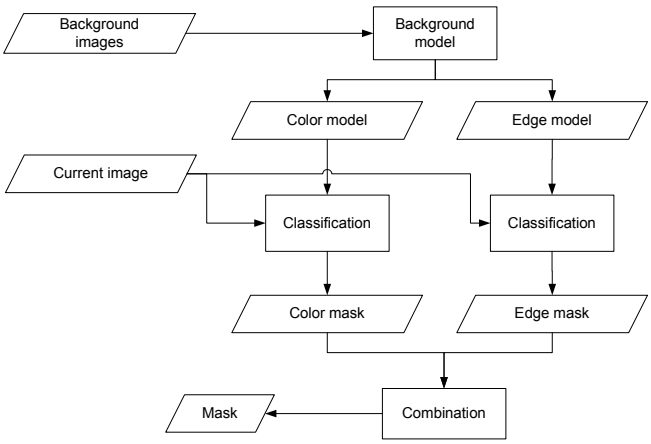

Figure 6. System overview of McKenna.

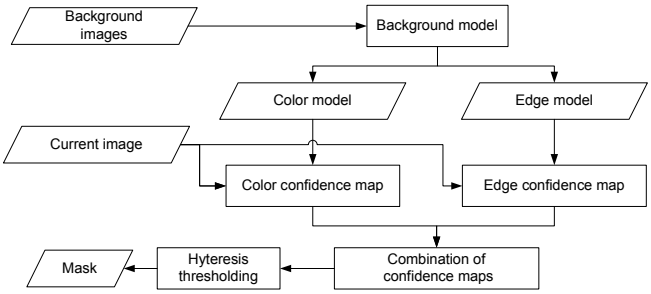

Figure 7. System overview of Jabri.

such as edge image and rgb image which are used apart for the further classification. At least a combination of both classification results gives the final segmentation mask.

2.6. Jabri (2000)

The system of Jabri[7] uses both information the color and the edge similar to the one of McKenna. The background model is trained in both mentioned parts by calculating the mean and standard deviation for each pixel of any color channel. With subtraction of the incoming current image on each channel, confidence maps are generated for both information color and edge. After that a combination of the two maps are utilized by taking its maximum values. At least this output is gone through a hysteris thresholding for binarization.

2.7. Cavallaro (2001)

The system of Cavallaro et al.[8,9] is shown in Figure 8. For each channel of used YCbCr color space an image differencing with background and current images is applied. With each premililary result an edge detection algorithm is utilized using the sobel algorithm. Then all three subresults are fused together which still occurs problems since the detected edges are not really connected as a whole contour. Therefore, a postprocessing step is needed such as the morphological operations to get the resulted mask.

2.8. Hong (2003)

Figure 9 describes the system by Hong.[10] Hong also models the background, but this time both well-known RGB and normalized rgb color are applied. As mentioned in previous methods the mean and standard deviation are used again and these are calculated over each color components. Each color space has its own classification

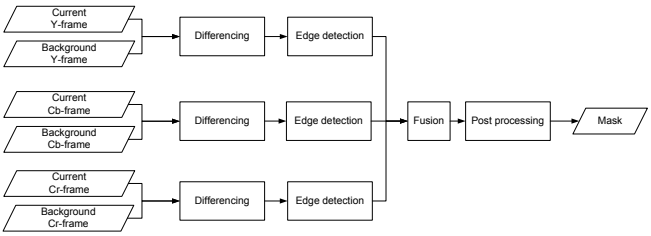

Figure 8. System overview of Cavallaro.

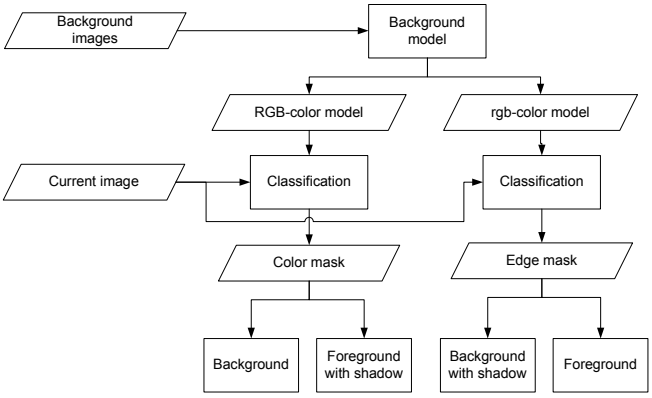

Figure 9. System overview of Hong.

part in which the current image is converted first in each color space. Within each color space the pixel can be classified in four categories as shown in Figure 9 and it is described as following: Until now, Hong is only the one who classifies the pixels into more than two categories namely four.

2.9. Shen (2004)

Shen[11] use the well-known RGB color space and the system can be represented in two section as shown in Figure 10. One of them is the block for generation of fuzzy classificaton and the other one is the block for elimination of falsely detected segmentation regions.

The fuzzy classification is applied to take into account the mobility of pixels precisely instead of the so-called binary classification. Thus, in the fuzzy block a difference image is generated for each RGB color space component. For every channel's result a corresponding threshold is determined by use of unimodal thresholding method for considering the fuzzy set of mobil pixels. Then these thresholds avail to generate fuzzy images which at least are combined to one final fuzzy image. Subsequently, a preliminary mask is achieved by thresholding which describes all detected mobil pixels in all appearences.

To overvome the problems of illumination changes and since there is no sudden adaptive update of the background a combination of temporal imformation and the mentioned above fuzzy color classification is given. The temporal information is achieved by the OR operation of the image differecing of successive frames and the last resulted mask. This output is combined with the preliminary mask of the fuzzy classification block.

3. EXPERIMENTS

3.1. Dataset

We tested all methods with indoor and outdoor video scenes, respectively. These scenes (see 11) can be seen as a representation of environments with various conditions which cover different lightings and background structures. Furthermore, uncompressed and compressed domain (MPEG-2) videos were also considered. All

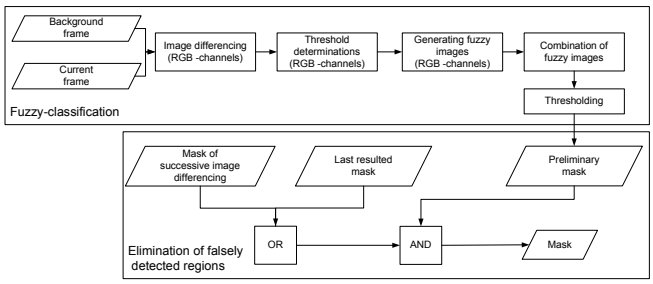

Figure 10. System overview of Shen.

Figure 11. Samples of the used dataset.

videos are in CIF (352x288) format. The experiments have been applied on the MPEG-4 test sequence "Hall", the MPEG-7 test sequence "Highway", the sequence "Group" of the European IST project art.live,[12] as well as on own captured video "House" and "Wall". The ground-truth segmentation data for the test sequence "Hall" is provided by the European project COST 211, the one for the "Group" and "Highway" are provided by EPFL.[13] The two other sequences have been segmented manually by hand to provide the user defined ground-truth.

3.2. Segmentation quality

Objective quality: For the objective evaluation of the segmentation quality ground truth based measures are used. Although some perceptual measures exist[14] we use only well-known pixel based measures. While the true positives (TP) give the number of correctly detected foreground pixels the true negatives (TN) give the number of correctly identified background pixels. In contrast the false negatives (FN) are pixels that are falsely marked as background whereas false positives (FP) are falsely detected as foreground. Figure 12 illustrates the relationships of these parameters and shows the color labels which are used for the subjective evaluation. Based on the above described numbers different evaluation measures can be defined:

The true positive rate (TPR) is given by

$$TPR = TP/(TP + FN) \tag{1}$$

The true negative rate (TNR) is given by

$$TNR = TN/(TN + FP) \tag{2}$$

		Ground-truth		
		Positive	Negative	
Segmentation	Positive	TP	FP	TP+FP
	Negative	FN	TN	FN+TN
		TP+FN	FP+TN	

Figure 12. Statistical parameters for evaluation.

The false positive rate (FPR) is given by

$$FPR = FP/(FP + TN) \tag{3}$$

The false negative rate (FNR) is given by

$$FNR = FN/(TP + FN) \tag{4}$$

While the above mentioned measures can be used for any type of classification, typical measures for two class problems (detection problems) are:

The recall R is the ratio between the number of correctly detected pixels to the number of relevant pixels in the grountd truth data and is defined as $R = TP/(TP + FP)$.

The precision P is the ratio between the number of correctly detected pixels to the total number of pixels and is defined as

$$P = TP/(TP + FN) \tag{5}$$

The F-measure combines these complementary measures with equal weights by calculating

$$F = 2 \cdot P \cdot R/(P + R) \tag{6}$$

Subjective quality: Since the objective quality measures are not very sophisticated subjective quality evaluation by human observers is inevitable. It allows to further analyze the errors of the different segmentation methods. Based on the available ground truth data each pixel is classified into TP, TN, FN or FP and coded by a specific color (see figure 12).

After the decription of the used objective and subjective evaluation methods the experiments with the dataset (see section 3.1) will be summarized. First, it must be noted that the objective as well as the subjective quality differ a lot between different videos. The difference also depends on the used method. Some methods are less restricted and have higher stability than others. Table 1 summarizes our experiments by showing the mean and the variance of the precision, the recall and the f-measure calculated over all videos for each method.

By analyzing the F-measure which is an overall criteria of the segmentation quality 3 groups can be identified: The best performance is shown by Shen, Cavallaro, Horprasert and Jabri with 88 – 83 %. The second group consists of McKenna, Francois and Hong with 79 – 81 %. It is followed by the image differencing and Haritaoglu's method with 69 – 65 %.

The precision can be used to analyze the tendency of an algorithm to oversegment (large number of false positives). The higher the precision, the less likely is oversegmentation. The algorithms with the lowest oversegmentation is Shen with 93 %. It is followed by Hong, Horprasert and McKenna with 80 – 83 %. Image differencing, Cavallaro, Francois and Jabri have a higher tendency to oversegmentation with 70 – 77 %. The highest oversegmentation is reached by Haritoaglu with 55 %.

On the other hand, the recall can be used to estimate the tendency of undersegmentation (large number of false negatives). The higher the recall the less likely is undersegmentation. The amount of undersegmentation of Cavallaro, Jabri and Francois is very small with 92 – 96 %. The second group is formed by McKenna, Horprasert, Shen and Haritaoglu with 87 – 83 %. The highest undersegmentation happens with Hong's and the image differencing method (77 – 73 %).

The subjective quality assessment (see figure 13 for a sample of the video "House") supports the results of the objective quality measures. It can be seen that Shen, McKenna and Horprasert achieve the best segmentation quality with both low false positives and false negatives. They are followed by Jabri and Cavallaro which show a slightly increased number of false positives. Haritaoglu and Francois show a much higher number of FP which corresponds to oversegmentation. On the other hand, the image differencing method shows a high number of FN which corresponds to undersegmentation.

Table 1. Objective measures for all methods without post processing

Algorithm	Mean in% f measure	Recall	Precision	Variance in % f measure	Recall	Precision
Image Differencing	69	73	75	10	17	25
Cavallaro	84	95	77	6	3	10
Shen	88	84	93	7	10	3
Haritaoglu	65	87	55	14	7	19
Francois	79	92	70	6	7	10
McKenna	81	83	80	9	14	6
Hong	79	77	83	12	19	6
Jabri	83	96	73	8	4	10
Horprasert	84	85	83	10	14	5

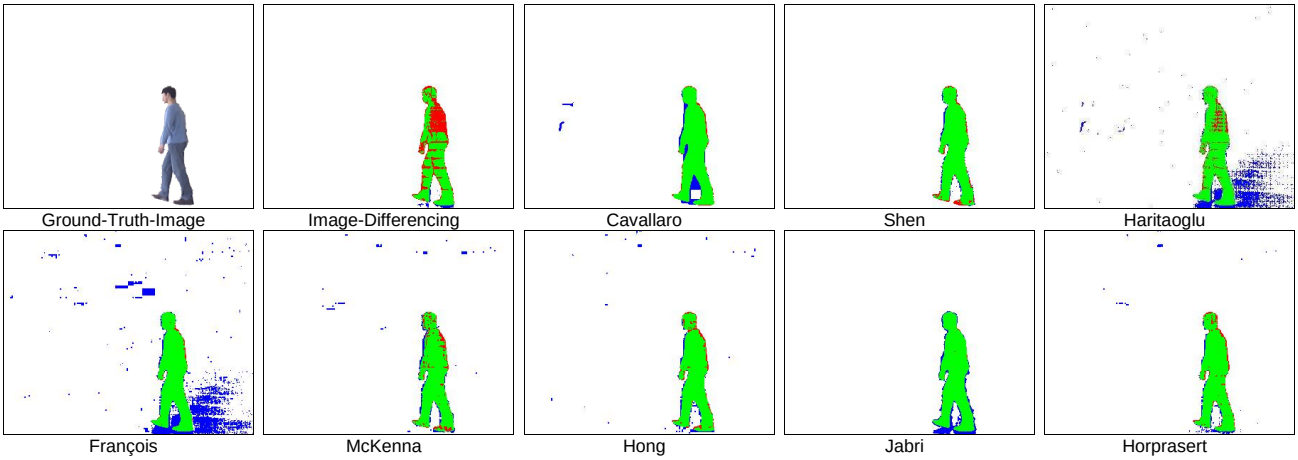

Figure 13. Subjective evaluation without post processing for video "House".

By comparing the objective quality of the methods with or without post processing it was noticed that the TPR increases slightly $(2-4\%)$ since holes in the objects are closed and noise is removed. At the same time the FPR decreases slightly by a similar amount. This is also supported by the subjective comparison of figure 13 and figure 14. As it can be seen small falsely detected objects are removed and falsely discarded holes within the objects are closed. On the other hand it also causes a higher number of false positives at the boundary of the object, which leads to an extension of the object.

For compressed videos the results show small differences to the uncompressed ones. This can be seen in figure 15 which compares the result of an uncompressed sample with a compressed sample. Especially the performance of simpler methods such as image differencing and Haritaoglu's method decreases noticeable. More complex methods and edge-based methods in particular give similar results in comparison to the uncompressed case.

3.3. Computational complexity

Beside the segmentation quality the computational complexity of a segmentation method is another important criteria. This is especially true for real-time applications where a given frame rate must be achieved or special low power hardware is used.

Computational complexity can be split into two parts: time and memory consumption. To a certain extend they are interchangeable. Sophisticated data structures can be used to increase the speed on the cost of higher memory requirements. On the other hand, the memory requirements can be reduced by using elaborate data access functions which usually effects the speed.

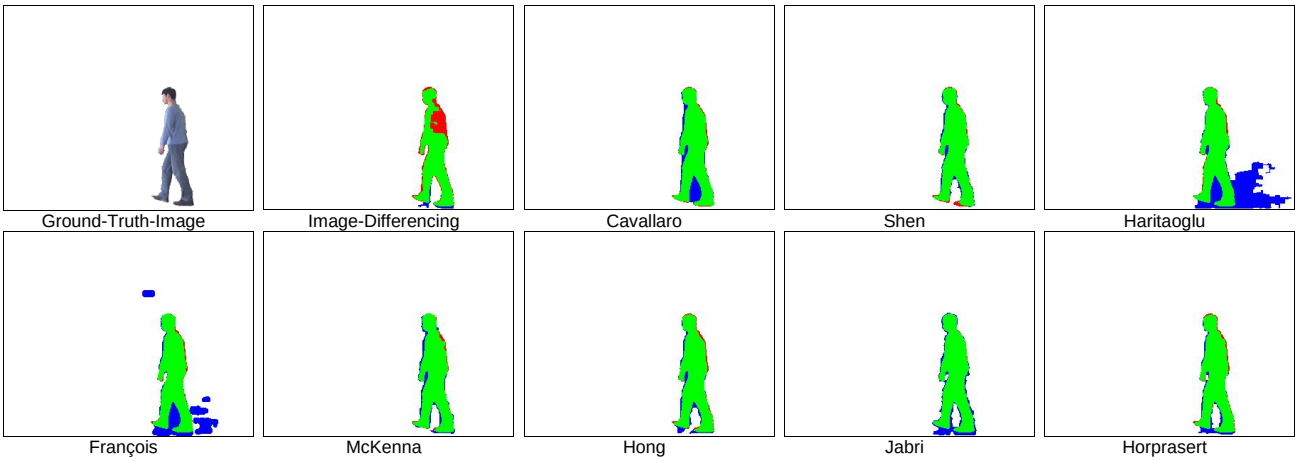

Figure 14. Subjective evaluation with post processing for video "House".

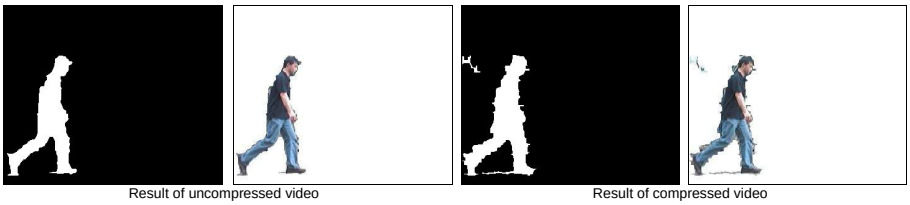

Figure 15. Subjective comparison of uncompressed and compressed video for video "House".

Time consumption is considered by measuring the elapsed time for the segmentation of one frame. The presented results are obtained using a standard personal computer (processor: AMD Athlon XP2600++, memory: 512 MB DDR-RAM) and an implementation in MATLAB (version: 7.0, release: 14). Some preliminary experiments with implementations in C++ (compiler: Visual C++ 7.1) suggest that the overall results are comparable with an increased speed of factor 13.

Memory consumption is only considered in terms of the amount of memory that is used for one background model. It is assumed that all values are stored using double precision (32 bit). Of course, the amount of memory can be decreased by quantizing the float values to integer values and decreasing the number of bits (e.g. 8 bit). But this may also lead to a decrease in performance.

Table 2 shows the time consumption of each segmentation method averaged over all frames of different video sequences. It can be seen that the time consumption is independent from the video sequence since the differences between them are very small. Furthermore, the methods can be divided into 4 groups. The fastest methods are the image differencing, Francois' and Haritaoglu's method with approximately 0.05 s. The second group consists of Hong and Horprasert with a time consumption of about 0.3 s. Third are Cavallaro and McKenna with 0.5 s. The slowest methods are the ones by Shen and Jabri with 0.7 s.

Concerning the memory consumption figure 16 gives the amount of memory that is needed to store one backgorund model for each method. As it can be seen 3 groups exist. The smallest amount of memory and thus the simplest model is used by the image differencing, Haritaoglu's and Cavallaro's method with around 1200 kB. The second group is built by Francois and Horprasert with 2400 – 3200 kB. The methods with the most complicated models are McKenna, Hong and Jabri with 4800 – 7200 kB.

The estimated computational complexity is the same for compressed and uncompressed videos. Since all the methods work with uncompressed videos the time consumption for compressed ones will increase by the decompression time.

Table 2. Mean time consumption (ms) of all segmentation methods

Algorithm	House	Group	Hall	Wall	Highway
Image Differencing	23.5	22.1	21.4	29.6	21.5
Cavallaro	479.6	442.8	221.5	413.9	469.3
Shen	715.3	702.7	700.8	656.2	700.4
Haritaoglu	26.6	25.9	26.0	27.2	29.3
Francois	42.2	45.0	43.8	44.5	46.2
McKenna	432.5	432.5	420.4	389.7	425.7
Hong	343.4	331.2	335.8	298.3	329.7
Jabri	723.8	727.2	710.8	695.9	717.3
Horprasert	255.8	266.1	267.2	244.9	258.1

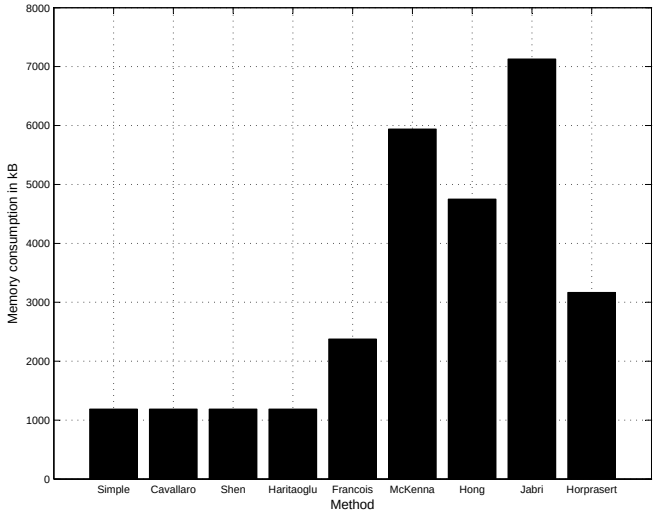

Figure 16. Memory consumption of the background models.

4. CONCLUSION

A comprehensive set of state of the art methods for static background has been presented and compared to each other in terms of the used approach, segmentation quality and computational complexity. An extensive set of video sequence, both indoor and outdoor hase been used obtain general results.Furthermore, uncompressed and compressed domain (MPEG-2) videos were considered. For evaluating the segmentation quality subjective as well as objective measures were used. The subjective evaluation was obtained by visually analyzing the segmentation results. For objective evaluation manually segmented ground truth data was used. Based on this simple pixel-based error measures were applied. Moreover, the segmentation results with and without post-processing (median filtering, morphological operations) were compared to each other. For assessment of the computational complexity, the elapsed time for the techniques was measured. Additionally, the memory requirements of the individual methods were estimated.

The thorough investigation supports multiple conclusions concerning the information, background models, classification and combination strategies used. While color is a powerful clue for segmenting foreground objects from the background, grayscale information is simply not enough for robust detection. Its use should be limited to scenarios were color information is not available, such as night vision or infrared cameras. Concerning the color spaces it can be concluded, that the separation of luminance and chrominance information improves the robustness against shadows. Edge information alone lacks robustness due to falsely detected edges but can improve the performance if used in combination with color. Generally the combination of complementary information (color, edge, motion) leads to higher performance. Statistical background models are better than simple average

background images since they also model the variations of the background due to camera noise. Sophisticated classifiers with multiple classes and fuzzy classification give higher performance than simple classifiers with only two classes and thresholding with the cost of higher computational complexity. This is supported by the fusion in that opinion level fusion (e.g. max rule) is supposed to give better results than decision level fusion (e.g. OR).

Furthermore, it can be concluded that the performance depends largely on the ideal combination of used information, background model, classification and combination strategies. Shen for example uses only one feature (RGB color space) a very simple background model (average image) but a very sophisticated classification step (fuzzy classfication and false detection removal) and obtains the best and most robust results. Jabri uses two features (RGB color and edge information), a statistical background model (Gaussian model) in combination with a sophisticated classification step and obtains good results. On the other hand, Horprasert utilizes only one feature (RGB color) a very simple backgorund model (luminance and chrominance average image) and a medium complicated classification method (multiple classes) to obtain good results as well. Especially, for real-time applications it is very important to note, that the performance does not only depend on a high number of features, very sophisticated models and classifiers but also on an optimal combination of the differetn steps.

While most of the methods perform well under restricted conditions (motivated by specific applications) they lack robustness in the presence of shadows, ghost effects, highlights, reflections, illumination changes, dark environments and if the foreground objects are similar to the background. Although some effects are considered directly or indirectly in the analyzed methods, more research needs to be done in order to develop a robust general purpose segmentation system. One way is to use more sophisticated models for the background to supress noise. The above mentioned problems should be treated independently which leads to a multi class classfication problem instead of a binary classification problem. Multiple complementary features and algorithms should be combined in order to improve the robustness. Very promising is the consideration of temporal information by using previous detection results or combining the segmentation and detection step with the tracking step. Another possibility is the combination with contour following methods such as snakes to improve the accuracy.

Our future research will focus on the above mentioned issues.

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the research initiative SPP Nr. 1041 "Distributed Processing and Delivery of Digital Documents". It was further developed within VISNET, a European Network of Excellence, funded under the European Commission IST FP6 programme.

REFERENCES

1. N. Otsu, "A threshold selection method from gray-level histograms," in *Proc. of IEEE Trans.Systems,Man,and Cybernetics*, pp. 62–66, 1979.
2. T. Pun, "A new method for gray-level picture thresholding using the entropy of the histogram," in *Signal Processing*, 2, ed., pp. 223–237, 1980.
3. D. Haritaoglu, I. Harwood and L. Davis, "W4: Who? when? where? what? a real time system for detecting and tracking people," in *International Conference on Face and Gesture Recognition*, 3, ed., April 1998.
4. A. R. François and G. G. Medioni, "Adaptive color background modeling for real-time segmentation of video streams," in *Proceedings of the International Conference on Imaging Science, Systems, and Technology*, pp. 227–232, (Las Vegas, NA), 1999.
5. T. Horprasert and D. Harwood, "A statistical approach for real-time robust background subtraction and shadow detection," technical report, Computer Vision Laboratory, University of Maryland, USA, 1999.
6. S. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, "Tracking groups of people," *CVIU* **80**, pp. 42–56, October 2000.
7. H. W. S. Jabri, Z. Duric and A. Rosenfeld., "Detection and location of people in video images using adaptive fusion of color and edge information," in *15th International Conference on Pattern Recognition*, **4**, pp. 627–630, (Barcelona, Spain), September 2000.
8. A. Cavallaro and T. Ebrahimi, "Accurate video object segmentation through change detection," in *Proc. of IEEE International Conference on on Multimedia and Expo*, pp. pp. 445–448, August 2002.

9. A. Cavallaro and T. Ebrahimi, "Change detection based on color edges," in *Proc. of IEEE Interna-tional Symposium on Circuits and Systems (ISCAS-2001)*, May 2001.
10. D. Hong and W. Woo, "A background subtraction for a vision-based user interface," in *ICICS-PCM*, 2003.
11. J. Shen, "Motion detection in color image sequence and shadow elimination," in *Visual Communications and Image Processing*, **5308**, pp. 731–740, SPIE, (San Jose, USA), January 2004.
12. "European project IST 10942 art.live." `http://www.tele.ucl.ac.be/PROJECTS/art.live/.`
13. "Ecole polytechnique federale de lausanne (epfl)." `www.epfl.ch`.
14. S. E. Drelie Gelasca E and E. T, "Intuitive strategy for parameter setting in video segmentation," in *Visual Communications and Image Processing 2003*, *Proc. of SPIE* **5150**, pp. 998–1008, SPIE, SPIE, July 2003.
15. C. Kim and J. Hwang, "Fast and automatic video object segmentation and tracking for content-based applications," *CirSysVideo* **12**, pp. 122–129, February 2002.