ON THE EFFECT OF MOTION SEGMENTATION TECHNIQUES IN DESCRIPTION BASED ADAPTIVE VIDEO TRANSMISSION*

Juan Carlos San Miguel, José M. Martínez

Grupo de Tratamiento de Imágenes
Escuela Politécnica Superior, Universidad Autónoma de Madrid, SPAIN
E-mail: Juancarlos.Sanmiguel@uam.es, JoseM.Martinez@uam.es

Abstract

This paper presents the results of analysing the effect of different motion segmentation techniques in a system that transmits the information captured by a static surveillance camera in an adaptative way based on the on-line generation of descriptions and their descriptions at different levels of detail. The video sequences are analyzed to detect the regions of activity (motion analysis) and to differentiate them from the background, and the corresponding descriptions (mainly MPEG-7 moving regions) are generated together with the textures of the moving regions and the associated background image. Depending on the available bandwidth, different levels of transmission are specified, ranging from just sending the descriptions generated to a transmission with all the associated images corresponding to the moving objects and background. We study the effect of three motion segmentation algorithms in several aspects such as accurate segmentation, size of the descriptions generated, computational efficiency and reconstructed data quality.

1 Introduction

Currently, surveillance systems[1] have more demand, specially for outdoor/indoor security in buildings, and because of the technology evolution the installation of video cameras does not need high economic investments: therefore most of the companies have any surveillance system.

When the surveillance system is big, it consumes a lot of resources, both for transmission and storage. Nowadays, the tasks of high-level interpretation related to the video security (monitoring) are done in their totality by a human who has to process simultaneously a great amount of visual information (coming from the different available cameras) that is presented to him in one or several monitors.

In order to obtain more robust surveillance systems based on video monitoring, the human supervisor could be helped by providing automatic analysis and interpretation tools able to focus his/her attention when a dangerous or strange event takes place in the video captured by a camera, as well as allowing to effectively recover the part of the sequence of video relative to some particular events.

A way for contributing to the above described scenario is the application of technologies allowing the generation and transmission of descriptions of the information captured by surveillance cameras. The descriptions allow the reduction of the data to be transmitted and the interpretation of what is happening in the scene, being possible to trigger events or alarms to focus on some the “active” video signals. If the descriptions are created at different levels of details it would be also possible to transmit the information in an adaptive way (e.g., depending of total bandwidth, giving more bitrate to scenes when something is happening).

In this paper we first overview a system developed for providing the above mentioned functionalities. The descriptions are generated based on motion analysis and object tracking in the sequences. The generated descriptions are created following the MPEG-7 standard for Multimedia Content Description[2]. After presenting the system, we focus in the evaluation of the effects of the different motion segmentation techniques in the performance of the system.

This paper is structured as follows: section 2 gives a brief overview of the state of the art, section 3 overviews the proposed system, and, section 4 explains the different motion segmentation techniques to evaluate. In section 5, experimental results are shown, while section 6 closes the paper with some conclusions.

2 State of the Art

In this paper we focus on two aspects related with the motion detection task: motion-based object segmentation and background generation, both highly related, as it is clear that a good background estimation helps in obtaining a good moving object segmentation.

In this section we provide a short overview of the state of the art with respect to these aspects.

In motion-based object segmentation for surveillance systems, most of the methods in the

* This work is partially supported by Cátedra Infoglobal-UAM para Nuevas Tecnologías de video aplicadas a la seguridad. This work is also supported by the Ministerio de Ciencia y Tecnología of the Spanish Government under project TIN2004-07860 (MEDUSA) and by the Comunidad de Madrid under project P-TIC-0223-0505 (PROMULTIDIS).

literature rely on variations two basic methods: frame difference and background substraction. Some of them make use both, trying to make the advantages of one compesate the drawbacks of the others.

In [3] and [4] these two basic methods are combined with the noise introduced by the camera. It is supposed that noise is always present in images and changes in the same pixels in consecutive images could be by motion or noise. In [3] this decision is taken by thresholding the frame difference between the current frame and a frame representing the background. A locally adaptive threshold is used to model the noise statistics. In [4] a motion detector is presented. It is intended to operate in surveillance applications for long periods of time with time-varying noise level.

In [5], a Bayes decision rule for clasification of background and foreground from a general feature vector is formulated. This rule is combined with a frame difference technique, a background substraction and color analysis.

In [6], moving object shape information is combined with background substraction to update background image in speedway sequences. Finally a frame difference technique is used to extract moving objects.

In [7], temporal change detection is used to prevent errors at background update, althoug it is not used to improve the segmentation made by background substraction, introducing only a little improvement in the whole method.

With respect the background generation/update task, there are a lot of techniques based on pixel analysis, like running average[8], running Gaussian average, mixture of gaussians, kernel density estimators. A brief review of these techniques can be found in [9]. In this review, simple methods (such as the running Gaussian average) offer acceptable accuracy while achieving a high frame rate an having limited memory requirements.

3 System overview

3.1 Introduction

The proposed system is designed to transmit relevant information of sequences of surveillance cameras at adaptive binary rate based on the generation of descriptions.

Sequences are analyzed in order to detect the moving regions, segment them from the background and generate a description of these moving regions (at least, region shape and trajectory). These descriptions can be generated at different levels of detail.

Besides the description of the moving objects, textures can also be extracted and transmitted, both moving regions texture and the estimated background. Depending on the level of quality/detail (closely related with the available bandwidth), textures and background can be transmitted either when there are significant changes, periodically, or on demand. The moving object’s texture does not imply to overcome privacy issues, as if the application requires it, the texture can be not shown in regular use, being only available on demand in special authorized situations.

Two parameters are configurable in the system (their values having major impact in system performance and results):

- Background Update Time (BUT): this parameter indicates how frequently the background scene is updated.
- Object Update Time (OUT): this parameter indicates how frequently the motion detection analysis is done. In other words, when the system examines the motion in the sequence and save motion data to generate descriptions.

3.2 System arquitechture

The system is divided in two applications: the server application that processes the input sequence and the client application that visualizes the sequence synthesized from the received description.

The Server application is composed by different functional modules (see Figure 1). It analyzes a video sequence and extracts motion objects and their trajectory from each group of pictures. Descriptions are generated each BUT time, even if background is not transmitted .

The server is composed by the following modules:

Image Acquisition. This module does the transcoding task adapting different inputs to the system.

Motion Detection. This module does the motion detection task. A description about different techniques that can be used in this module is presented in section 4 and evaluated in section 5.

Background Model. This module updates the

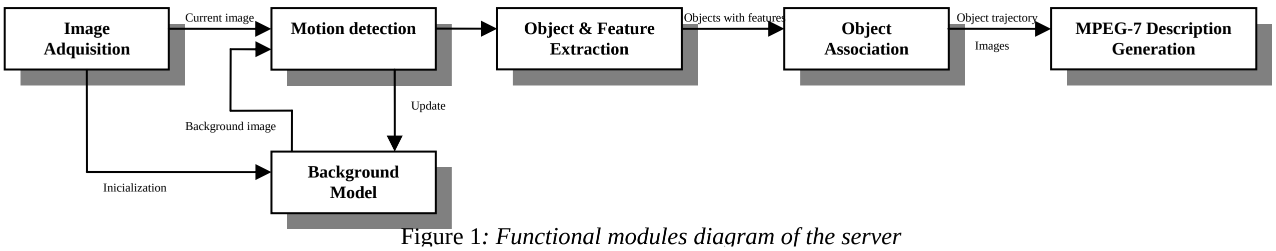

Figure 1*: Functional modules diagram of the server*

background of the sequence periodically.

Object & Feature Extraction. This module extracts different features from the relevant moving object (such as shape description, bounding box, motion trajectory, dominant colour, motion duration,...) to process this information in the next module. Additionally, this data is used in the MPEG-7 description generation

Object Association. This module makes a simple tracking of the relevant objects identified in the previous module. It identifies different objects in the current analyzed frame and tries to associate them with the objects detected in the previous frame. This module does this work following different rules such as distance, colour, and object size. Finally a graph is constructed to follow object trayectory during each group of pictures analyzed. This approach is similar to the one presented in [10].

MPEG-7Description Generation. This module generates the "master" description of all the information obtained from the previous modules. Additionally different types of descriptions can be generated depending on the level of detail selected.

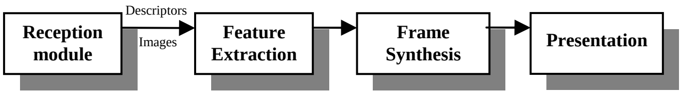

Figure 2*: Functional modules diagram of the client*

The Client application is composed by different functional modules (see Figure 2). It does the synthesis of the original sequence from the information received. The modules are:

Feature Extraction. This module reads the MPEG-7 description received and extracts features needed to synthesize the sequence.

Frames synthesis. This module synthesizes frame to frame the sequence. It uses the trajectory of the objects, its shape description, and any other object description (texture, colour,...), as well as the objects images and the background image, if they are available.

Presentation. This module shows the synthesized sequence frame-by-frame.

3.3 Description's levels of details

From the MPEG-7 master description the system generates different levels of detail, but in a real service implementation, this can be customized in order to avoid unnecessary analysis and adaptation, reducing the computing cost and resources requirements, or increasing performance for the same resources.

One parameter of the level of detail is the granularity of the description, that is, if the moving regions shape and trajectory are more or less detailed,

With respect to regions shape, it can range from a detailed shape description (result of pixel accuracy segmentation) to the centroid of the object. With respect to the motion trajectory, it can range from the trajectory of each pixel in the region's shape to a set of sampled points of the trajectory, including in the range associated interpolation functions.

The second parameter deals with the textures associated to each moving object in the description and the background image.

Based in these two parameters, we have currently implemented four levels:

- ***Level 1***. For each background analysis slot (BUT), the description includes an image representing the background, objects images associated to the object shapes (one each OUT) and the trajectory for each object (considered as rigid objects).
- ***Level 2***. In this level, the moving objects description is reduced to one for the complete background analysis slot. This key-object is selected from the set of available ones.
- ***Level 3***.In this level, images are not included in the description, which consists only of moving regions descriptions. The feature that is used to synthesize the object texture at the client is the dominant colour. The background image is updated each BUT.
- ***Level 4.*** I this level the updated background image is requested on demand.

Object motion information is described by means of the motion of each vertex of the bounding box that fits it, as each OUT the bounding box size can change (this a very rough approximation to a non-rigid object).

Finally, object texture images incorporated to the description depending on the level of description desired. Results for the three first levels are shown in Figure 3.

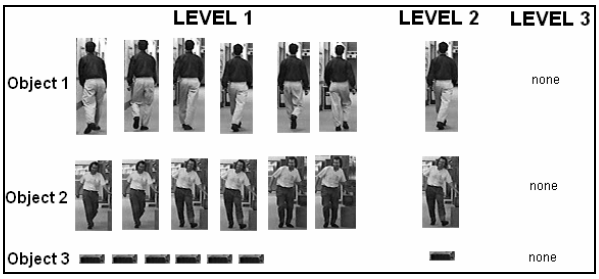

Figure 3: *Object image extraction for the analysis corresponding to a BUT slot (frames 150-180) for first motion detection technique (for Hall Monitor sequence)*

At the receiving side, the client application synthetizes a reconstruction of the original sequence. Depending on the selected level, different quality levels are obtained (see Figure 4).However, levels 2 and 3 have a progressive lost of quality (in level 2 mainly due to the bounding box approach).

3.4 MPEG-7 Description Tools

The proposed system generates one master MPEG-7 compliant description file for each group of pictures analyzed in a BUT.

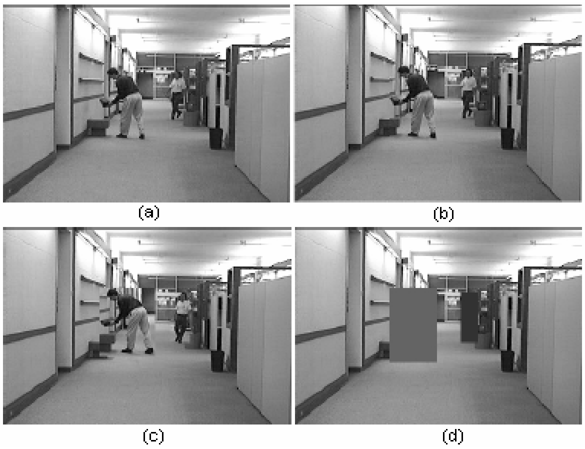

Figure 4: *Example of the three reconstruction levels at the client for Hall Monitor sequence: (a) original frame, (b) level 1, (c) level 2 and (d) level 3.*

The MPEG-7 description tools currently incorporated are:

StillRegion DS. It's used to describe the location of the pictures, format, coding,...

MotionTrajectoryDS. It's used to describe the trajectory for each object. The trajectory is calculated using a "bounding box" that fits to the object to follow.

DominantColor DS. With this description, we have colour information about the tracked object.

4 Motion Detection Techniques

In this section we briefly explain the three different motion detection techniques for which system performance has been evaluated.

- ***Technique 1***. This technique follows an approach similar to the one of the semantic analysis module in [3]. It is based on the noise introduced by the camera (assumed that it is additive and follows a Gaussian distribution) and the change detector decides whether in each pixel position the foreground signal corresponding to an object is present. This decision is taken by thresholding the frame difference between the current frame and the frame representing background. The method is applied in an observation window. We add some new funcionalities: improved processing time and a new background update model (based on a mixture between Average and Running Average methods [8]). The main advantage of this technique is that it can compensate a video signal with a time-varying noise level.
- ***Technique 2.*** This technique[11] is based on frame difference and background subtraction and uses bidirectional temporal change detection to improve a background subtraction method, where the background model uses a simple Gaussian for each point. This technique performs motion detection in different stages. First stage detects motion using a two thresholded frames differences (one between current and previous frame and the other between current and next frame) and a background substraction (between current image and background image in background model). Finally the three masks are combined to obtain a motion mask. This stage is efficient and has little false positives, but fails with homogeneous zones in the foreground and non-moving objects. Mask obtained from this first stage guide the second stage, where current frame luminance values are checked against a background model. Points classified as foreground by the temporal change detector will have greater probability of being classified as motion. This technique is designed to work with different block sizes (at least pixel size). The selection of the size will depend on the requirements of the application. In this paper, we use a 1 pixel block size to compare results with the other techniques. This method assumes a non-complex background, common in many indoor surveillance applications.
- ***Technique 3.*** This technique is based on the statistical approach proposed by [5] (we use the implementation provided in the OpenCV library[12]) that performs a Bayessian classification based on frame difference and background subtraction. This algorithm is based on four parts: change detection, change classification, foreground object segmentation and background learning and maintenance. In the first step, no-change pixels are filtered out by using simple background subtraction and temporal differences. The detected changes are separated as pixels belonging to stationary and moving objects according to inter-frame changes. In the second step, the pixels associated with stationary or moving objects are further classified as background or foreground based on the learned statistics of colours and colour co-occurrences respectively by using the Bayes decision rule. In the third step, foreground objects are segmented by combining the classification results from both stationary and moving parts. In the fourth step, background models are updated.

These techniques use classic methods for motion detection but they have some differences. The first technique uses a window approach and a light statistical model to represent the noise. The second technique uses a pixel approach and a robust background statistical model. The last technique uses complex statistical models for background and change detection.

5 Experimental results

In this section, experimental results of the proposed system are presented, illustrating the subjective effect of motion detection at each different reconstruction level as well as the size of the associated description (data to be transmitted).

The system has been implemented in C++, using OpenCV[12] library for some image processing operations. Tests were executed on a Pentium IV with a CPU frequency of 3.0 GHz and 1GB RAM.

To evaluate performance of the motion detection in the system's results, the three motion detection techniques are tested in terms of processing time, data size generated and sequence reconstruction quality.

Due to space constraints, sample results are shown for a test sequence extracted from AVSS'07 dataset[13]. Additional results can be found at http://www-gti.ii.uam.es/publications/OnTheEffectOfMotionSegmentationTechniquesInDescriptionBasedAdaptiveVideoTransmission

The results shown in Figures 5 and 6 correspond to analysis with 1 sec for BUT and 0.2 sec for OUT. Figure 5 shows the segmentation results for each of the techniques commented in section 4 before and after minimum object size filtering.

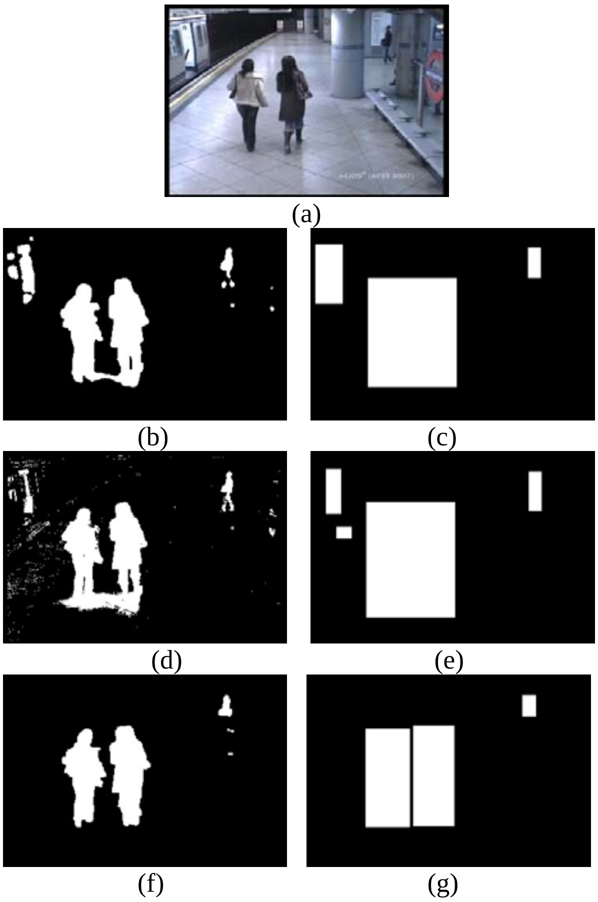
(a)
(b) (c)
(d) (e)
(f) (g)

Figure 5: *Results from AVSS'07 test sequence "AVSS AB Easy"[13]. The results shown correspond to (a) original frame, motion detection and filtered detection for technique 1 (b)(c), technique 2 (d)(e) and technique 3 (f)(g)*

In Figure 5 it can be observed that technique 1 produces good object detection, technique 2 detects more motion that technique 1 but after mask filtering, the bounding box aproximation is the similar. Technique 3 does a detection that losses some object information in the scene. Some of these objects are incorporated to the background model when their movement is stationary (i.e. underground gate) and other objects are not well detected (i.e. man on top-right image).

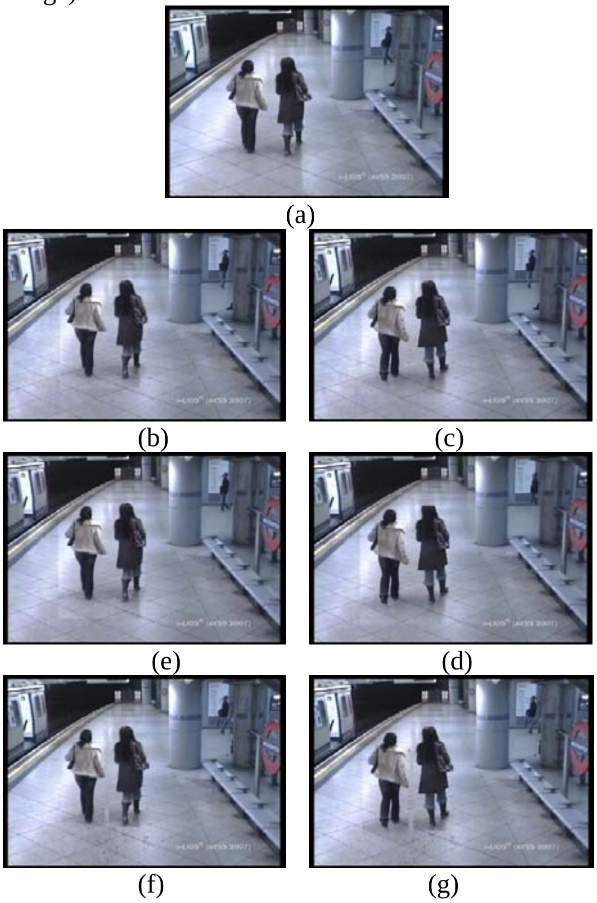
(a)
(b) (c)
(e) (d)
(f) (g)

Figure 6: *Reconstruction levels one and two at the client side from 1735 frame AVSS'07 test sequence"AVSS AB Easy"[13]. Results shown corresponds to (a) original frame, technique1: level 1(b) and level 2 (c), technique 2: level 1(d) and level 2 (e) and technique 3: level 1(f) and level 2(g).*

In Figure 6, we can observe the reconstruction results for the different techniques tested in the system. It can be observed that the reconstruction quality is better for techniques 1 and 2. It is because the first technique does an accurate detection (see Figure 5) and the second one does a similar detection after object filtering. The third technique introduces some noise due to slow object motion, as part of the object is incorporated to the background reference frame and, in consequence, providing a lower quality background estimation, that is finally used in the reconstructed sequence. Also it can be observed that there is a loss of quality between levels 1 and 2 due to the bounding box approach (this effect can be better observed in the sequences available at the above mentioned site).

To compare these segmentation techniques, we evaluate different features showed in the following tables (data for level 4 is not generated because it depends on user interaction).

Table 1: *Data size comparison generated for each analysis level (for AVSS'07 test sequence[13]).*

MOTION DETECTION TECHNIQUE	PARAMS		LEVEL 1 Size	LEVEL 2 Size	LEVEL 3 Size
	BUT	OUT			
Technique 1	1s	0.2s	2,27MB	1,67MB	1,42MB
	3s	0.4s	0,98MB	0.67 MB	0,55 MB
Technique 2	1s	0.2s	2,53 MB	1,87 MB	1,58 MB
	3s	0.4s	1,33 MB	0,91 MB	0,71 MB
Technique 3	1s	0.2s	1,73 MB	1,39 MB	1,27 MB
	3s	0.4s	0,82 MB	0,62 MB	0,55 MB

Table 2: *Comparative of performance speed for the motion detection techniques tested*

ALGORITHM	IMAGE SIZE	
	352x240	640x480
Technique 1	17 fps	4 fps
Technique 2	55 fps	20fps
Technique 3	20 fps	5 fps

Overall results show that fast-detection techniques (such as technique 2) allow visualizing the reconstructed sequence with a short delay. More accurate techniques (such as technique 1), that uses a window approach, reduces the information generated by server application and therefore the communication channel requirements (but increases computational requirements). Moreover, techniques that detect fast motion changes (such as technique 3) allow obtaining good object textures although having the problem mentioned above with respect of the incorporation into the background of the trail of slow motion objects.

At the client side, more accurate techniques give more quality in the reconstructed sequence (at each level) as we can see in Figure 6.

Additionally to the motion detection comparison a data size reduction is reached. Around 90-80% reduction (for level 3 and 1, respectively) of original data size (without incoporating image compression) shows that we can reduce the binary rate in video transmission, as well as supporting the provision of adaptive reduction based on different levels of detail. In level 1, artefacts are barely noticeable, providing an acceptable level of quality.

6 Conclusions

In this paper a study of the effect of motion detection in description based adaptive video transmission is presented.

The contribution of this paper is twofold. On the one hand, a framework for adaptive video transmission based on descriptions is defined and implemented. Transmission rate can be customized to the particular application requirements, always focusing on what is happening in the scene at a semantic level. On the other hand, three algorithms are evaluated for the motion detection task. These algorithms have different properties (based on different approaches) that affects in different ways in system's results, as shown in section 5. Different techniques can be selected depending on system's requirements such as real-time applications, available processing resources, available bandwidth and storage, etc.

References

[1] K.N. Plataniotis, C.S. Regazzoni (eds.), "Special Issue in Visual-centric Surveillance Networks and Services", IEEE Signal Processing Magazine, 22(2), Marzo 2005

[2] B.S. Manjunath, P- Salembier, T. Sikora "Introduction to MPEG-7: Multimedia Content Description Language", John Wiley & Sons, Ltd., 2002.

[3] A. Cavallaro, O. Steiger, T. Ebrahimi, "Semantic Video Analysis for Adaptive Content Delivery and Automatic Description", IEEE Transactions of Circuits and Systems for Video Technology, 15(10):1200-1209, Oct. 2005.

[4] A. Albiol, C. Sandoval, A. Albiol, "Robust Motion Detector for Video Surveillance Applications",. Proc. of the Int. Conference on Image Processing,. ICIP 2003.

[5] L. Li, W. Huang, I.Y.H. Gu, ,Q. Tian, "Foreground object detection from videos containing complex background", Proc. of the ACM Int. Conference on Multimedia, ACM Multimedia'03.

[6] T. Nakanishi, K. Ishiim, "Automatic vehicle image extraction based on spatio-temporal image analysis", Proc. of Int. Conference on Pattern Recognition, IAPR 2006

[7] S. Huwer, H. Niemann, "Adaptive change detection for real-time surveillance applications", Proc. of the IEEE Int. Workshop on Visual Surveillance, IWVS 2000

[8] R. Cucchiara, M. Piccardi, A. Prati, "Detecting moving objects, ghosts and shadows in video streams", IEEE Trans. on Patt. Anal. and Machine Intell., 25(10):1337-1342, Oct. 2003.

[9] M. Piccardi, "Background subtraction techniques: a review", Proc. of IEEE Conference on Systems, Man and Cybernetics, ICSMC 2004.

[10] G. Medioni, "Detecting and tracking moving objects for video surveillance," Proc. of IEEE Int. Conference on Computer Vision and Pattern Recognition, ICCVPR 1999

[11] A. Garcia , J.Bescós, "Real-Time Video Foreground Extraction Based on Context-Aware Background Substraction", Technical Report TR-GTI-UAM-2007-02, 2007. http://www-gti.ii.uam.es/publications/OnTheEffectOfMotionSegmentationTechniquesInDescriptionBasedAdaptiveVideoTransmission/TR-GTI-UAM-2007-02.pdf

[12] OpenCV, open source library for computer vision. http://www.intel.com/technology/computing/opencv/

[13] i-Lids dataset for AVSS 2007