APPLICATION DEPENDENT VIDEO SEGMENTATION EVALUATION - A CASE STUDY FOR VIDEO SURVEILLANCE -

Elisa Drelie Gelasca[1], *Touradj Ebrahimi*[2]

[1] Department of Electrical and Computer Engineering, University of California, Santa Barbara, California.
[2]ITS-STI, EPFL, 1015-Lausanne, Switzerland, mail: {elisa.drelie@a3., touradj.ebrahimi@}epfl.ch

ABSTRACT

Evaluation of the performance of video segmentation algorithms is important in both theoretical and practical considerations. This paper addresses the problem of video segmentation assessment, through both subjective and objective approaches, for the specific application of video surveillance.

After an overview of the state of the art technique in video segmentation objective evaluation metrics, a general framework is proposed to cope with application dependent evaluation assessment. Finally, the performance of the proposed scheme is compared to state of the art technique and various conclusions are drawn.

1. INTRODUCTION

Segmentation of objects in image sequences is a crucial task for a wide variety of multimedia applications such as video object coding, manipulation and identification. The ideal goal of segmentation is to identify the semantically meaningful components of an image and to group the pixels belonging to such components. While it is very hard to segment static objects in images, it is easier to segment moving objects in video sequences. Once the moving objects are correctly detected and extracted, they can serve for a variety of purposes.

In this paper, we do focus on a specific application of video analysis through segmentation, *surveillance*. Human viewers are asked to assess the quality of segmented objects for a “video surveillance scenario”.

Subjective segmentation evaluation is necessary to study and to characterize the *perception* of different artifacts on the overall quality for different applications, but once this task has been accomplished successfully and an automatic procedure has been devised, systematic subjective evaluation can be avoided. In order to compare the performance of two objective metrics, in this paper, the subjective opinions are compared to their objective results. The first objective measure has been previously proposed in [1] for a general scenario and in this paper, its parameters are specifically tuned for the video surveillance framework. The second objective metric has been proposed by Nascimento *et al.* [2].

In literature, although several quality measures have been developed for still image segmentation, they are not directly applicable in an efficient manner to video object segmentation or tracking.

We distinguish between video object segmentation and tracking *evaluation* since they are two different matters despite being related. Video tracking is the process of locating a moving object (or several ones) in time using a camera. An algorithm analyzes the video frames and outputs the location of objects as a function of time, optionally in real time. It is mainly used in video surveillance systems. The issues involved in video object tracking are different from those of video object segmentation evaluation since the ground truth on which these algorithms compare their performance is different. In fact, video surveillance systems concern algorithms for detecting, indexing and tracking moving objects and this evaluation requires specific considerations as follows. The ideal output (ground truth) of a tracking system can be of two types: bounding box and/or center of gravity. In the former case, regions that contain the detected moving objects of interest are segmented with a set of rectangular areas called bounding boxes as shown in Fig. 1 (a) and (b). Detection and false alarms rates in this case are derived by counting how many times interesting and irrelevant regions are detected. In the latter, the manual ground truth consists in a set of points that define the trajectory of each object in the video sequence (center of gravity) as depicted in Fig. 1 (c). In this case, the motion detection and tracking algorithm is then run on the video sequence and tracking results and ground truth centers of gravity are compared to assess tracking performance.

This paper does not directly evaluate tracking performance but proposes an alternative way to reach the same objectives.

Figure 1 (d) depicts the original frame and (e) shows the result of an ideal video object segmentation - ground truth. As depicted, it does not represent a binary detection problem. Several types of errors (such as shape errors along the boundaries of the object, content similarity, etc.) should be considered (not just mis-detection and false alarms). Thus, proposed tests based on the selection of rectangular regions with and without objects are unrealistic since practical segmentation algorithms have to segment the image into *foreground* –objects of interest– and *background*.

Section 2 presents the state of the art evaluation metrics for video object tracking evaluation [3, 4, 5, 6, 2]. In particular, the details of Nascimento *et al.* [2] metric, *mqm*, that will be used in the comparison with the proposed application dependent metric (see Sec. 3), *PST*, are given. For an overview on the state of the art of video object segmentation evaluation [7, 8, 9, 10, 11, 12, 13, 14] the reader is referred to [1, 17].

2. STATE OF THE ART

Recently, a number of measures have been proposed for video object tracking evaluation. Since we are interested in how the object is segmented for different applications and the evaluation of tracking raises different problems, the reader is introduced to fora such as PETS [15] and CAVIAR [16] for a complete overview on that issue.

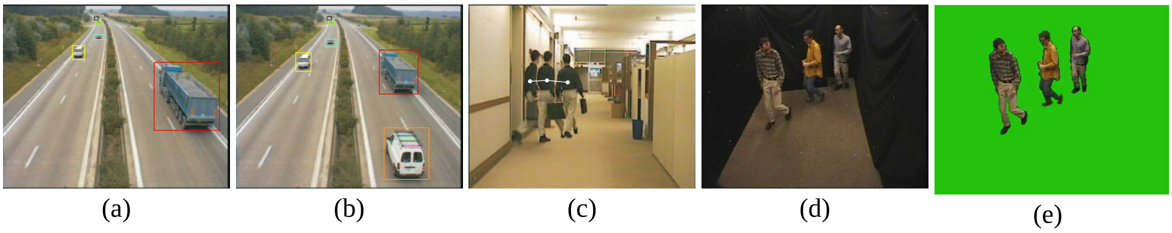


Figure 1: (a) and (b): samples of ground truth for tracking evaluation through bounding box for ‘Highway’ video sequence. (c): sample of ground truth for tracking evaluation through center of gravity for ‘Hall’. Sample of original video sequence ‘Group’ in (d) and the corresponding ideal object segmentation (ground truth) in (e).

Table 1: Objective Measures used in evaluating video tracking systems.

Measure	Source
False Alarm	El. [6], Nasc. [2], Ol. [3], Ob. [4]
Misdetection	El. [6], Nasc. [2], Ol. [3], Ob. [4]
Split and/or Merge	El. [6], Nasc. [2], Ob. [4]
Area Matching	El. [6], Nasc. [2]
Occlusion manag.	El. [6]
Center of gravity	El. [6], Senior [5]

In the following, we will refer to some representative works [3, 4, 5, 6] that can be found in the literature and specifically to Nascimento and Marques’s metric [2] that can be applied also to a more general object segmentation evaluation case. Table 1 shows all the state of the art methods described here and grouped by measure.

Standard measures used in communication theory such as mis-detection rate, false alarm rate and Receiver Operating Characteristics (ROC) are used in [3, 4]. An ROC curve is generated by computing pairs (P_d, P_f), where P_d is the probability of correct signal detection and P_f is the false alarm probability. For example, Oberti *et al.* [4] compute the false-alarm (P_f) and the mis-detection probabilities $(1 - P_d)$ on the basis of discrepancies between the resulting objects and matching area (false alarm) or between the reference area and the matching one (mis-detection). The global performance curve summarizing the curves obtained under different working conditions is obtained by imposing an operating condition $(P_f = 1 - P_d)$ and by plotting the corresponding values against different values of the variable of interest (scene complexity, distance of objects from sensors).

In Oliveira’s metric [3], a specific parameter of the tracking algorithm is varied and the false alarm/detection and split/merge rates are plotted against it. Senior *et al.* [5] employed the trajectories of the centroids of tracked objects and their velocities to evaluate their discrepancy measures.

An interesting framework for tracking performance evaluation which uses pseudo-synthetic video is adopted by Ellis [6]. Isolated ground truth tracks are automatically selected from the PETS2001 dataset, according to three criteria: path, color and shape coherence (in order to remove tracks of poor quality). Pseudo-synthetic video are generated by adding more ground truth tracks and the complex object interactions are controlled by the tuning of perceptual parameters. The metrics used are similar to those in the previously described works: tracker detection rate, false alarm rate, track detection rate, occlusion success rate, etc.

However, these approaches have several limitations. As already mentioned, object detection can not be considered as a simple binary detection problem. Several types of error should be considered and just mis-detection and false alarms are not enough. For example, the proposed test in [5] is based on employing the centroid and areas of rectangular regions but practical algorithms have to segment the image into background and foreground and do not have to classify rectangular regions selected by the user.

To overcome these limitation Nascimento and Marques [2] used several simple discrepancy metrics to classify the errors into region splitting, merging or split-merge, detection failures and false alarms. In this scenario, the most important thing is that all the objects have to be detected and tracked along time. Object matching is performed by computing a binary correspondence matrix between the segmented and the ground truth images. The advantage of this method is that ambiguous segmentations are considered (e.g., it is not always possible to know if two close objects correspond to a single group or a pair of disjoint regions: both interpretations are adopted in such cases). In fact, by analyzing this correspondence matrix, the following measures are computed: Correct Detection (CD): the detected region matches one and only one region; False Alarm (FA): the detected region has no correspondence; Detection Failure (DF): the test region has no correspondence; Merge Region (M): the detected region is associated to several test regions; Split Region (S): the test region is associated to several detected regions; Split-Merge Region (SM): when the conditions M and S simultaneously occur.

The normalized measures are obtained by normalizing the amount of FA by the number of objects in the segmentation, N_C, and all the others by the number of objects in the reference, N_R, and finally by multiplying the obtained numbers by 100. The **object matching quality metric** at frame k, $mqm(k)$, is finally given by:

$$\begin{aligned} mqm(k) &= w_1 \cdot \frac{C_D(k)}{N_R} + w_2 \cdot \frac{F_A(k)}{N_C} + w_3 \cdot \frac{D_F(k)}{N_R} \\ &+ w_4 \cdot \frac{M(k)}{N_R} + w_5 \cdot \frac{S(k)}{N_R} + w_6 \cdot \frac{S_M(k)}{N_R} \qquad (1) \end{aligned}$$

where w_i are the weights for the different discrepancy metrics. The overall metric mqm is the sum of $mqm(k)$ along all the frames k. It is evident that this metric is able to describe quantitatively the correct number of detected objects and their correspondence with the ground truth only, whereas the metric described in the next section is able to monitor intrinsic properties of the segmented objects such as shape irregularities and temporal instability of the mask along time.

3. APPLICATION DEPENDENT METRIC

In [1, 17], *five* different clusters of errors have been recognized as typically produced by the most common segmentation algorithms. **Added region** is the over-segmented part of background disjoint from the correctly segmented objects. **Added background** is the over-segmented part of background attached to the correctly segmented object. **Inside holes** are under-segmented parts completely inside the objects. **Border holes** are under-segmented parts directly on the border of the objects. **Flickering** is the temporal variation of any of the above described artifacts. The proposed objective metric is defined on two kinds of metrics, namely the objective metric and the perceptual metric. First, the *objective metric* classifies and quantifies the deviation of the segmentation result from the reference. Second, segmentation errors are measured through the proposed objective criteria and their perception is studied and characterized by means of subjective experiments. Finally, the perception of segmentation errors is modeled and incorporated in the proposed *perceptual metric*. The novelty of this approach consists in:

- tuning the parameters according the specific application;
- classifying the different clusters of error pixels according to the following characteristics: whether they do or they do not modify the shape of the object and afterward their size.

Border holes, $\mathbf{H}_b$, and added backgrounds, $\mathbf{A}_b$, modify the shape while inside holes, $\mathbf{H}_i$, and added regions, $\mathbf{A}_r$ preserve the segmented object shape (see [17]). The linear combination of the perceptual spatio-temporal metric **PST** for each artifact provides the proposed application dependent objective metric:

$$\mathbf{PST} = a \cdot \mathbf{PST}_{A_r} + b \cdot \mathbf{PST}_{A_b} + c \cdot \mathbf{PST}_{H_i} + d \cdot \mathbf{PST}_{H_b} \quad (2)$$

The perceptual weights a, b, c, and d can be tuned according to the application and have been found by means of subjective experiments [17] for the general scenario case: $a = 2.86$, $b = 4.50$, $c = 4.77$, $d = 5.82$. For further details on this objective metric, the reader is referred [17].

The expected segmentation quality for a given application can often be translated into requirements related to the shape precision and the temporal coherence of the objects to be produced by the segmentation algorithm. Video sequences segmented with high quality should be composed of objects with precisely defined contours, having a perfectly consistent partition along time.

A large number of video segmentation applications can be considered and typically they have different requirements. A full classification of segmentation applications into a set of scenarios, according to different application constraints and goals can be found in [18]. The setting up of a subjective experiment differs for each kind of application. Therefore, we have focused our experiments on video surveillance.

Video surveillance is a typical case where knowledge of the specific application can be used to tune the parameters of the evaluation metric: undetected objects or over segmentation will have a bigger impact on the overall annoyance than changes in the shape of the correctly detected objects.

In order to evaluate different segmentation algorithms in the context of a video surveillance applications, the segmentation results (see next section) and the reference segmentation have been used to produce test video sequences where the object boundaries detected by the segmentation algorithm have been underlined by a colored contour on three original sequences used in this experiment: ’Group’, ‘Hall monitor’ and ‘Highway’ as depicted in Fig. 2.

Section 4 presents how the subjective experiments have been carried out for the specific application. The correlation between the subjective scores and the objective results are analyzed in Sec. 5. In that section, an analysis is carried out to determine how to tune the metric parameters according to the specific application.

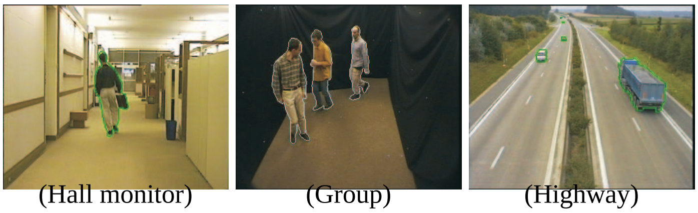
4. SUBJECTIVE EXPERIMENTS

The experimental methodology is composed of a five-step procedure as described in [17]: oral instructions, training, practice trials, experimental trials and interview. After a general introduction on segmentation, the typical artifacts are shown and the original video with the correspondent segmented video are shown as in Fig. 1 (e) and (d). After this introduction the specific application is explained and the corresponding segmentations are shown in the training stage as depicted in Fig. 2.

For a specific application such as video surveillance a *ad hoc* protocol is needed. In the following, a summary of the instructions given to the subjects are described: “[..] *Video-surveillance systems are used in different fields: monitoring the traffic to detect incidents or jams, analysis of the human behavior to identify thefts, brawls or other dangerous situations, security of reserved zones to control the access of a non-authorized person or of abandoned objects. The segmentation can be employed by these systems to identify all the objects in the scene and then detect anomalous situations. For instance, one could introduce a post-processing block for face detection and recognition that activates an alarm if the segmented person is not authorized. Also in less sophisticated systems, where the shots are shown on the monitor and directly controlled by human operators, the segmentation information can be useful to help them in their task through a scene representation as in Fig. 2, with the highlighted objects. Since these systems work in real time, an essential requirement for the algorithms is a low computational cost and what is important is that all the objects are identified and entirely cut out because an only partially detected object could generate an error in the successive phases*[..]” (see [17] for further details).

In the experiments, seven static background segmentation methods were chosen. They are reported in Tab. 2 and reviewed in [1]. The approaches of the selected representative algorithms differ in using various features such as color, luminance, edge, motion and combinations of them. For further details the reader is referred to the corresponding references reported Tab. 2. Table 2 also shows the subjective ranking during the *interview stage* of the subjective experiment. This table reports the tested algorithms from the least to the most annoying and a brief description of the artifacts that are typically introduced.

The test group was composed of 35 subjects aged between 23 and 41 (with 8 females). The subjects were asked one question after each segmented video sequence was presented, “How annoying was the defect relative to the worst case example in the sample video sequences?”. The subject was instructed to enter a numerical value greater than 0. The value 100 was to be assigned to artifacts as annoying as the



(Hall monitor) (Group) (Highway)

Figure 2: Sample frames for video surveillance segmentation application ‘Group’, ‘Hall monitor’ and ‘Highway’.

Table 2: Description of segmentation algorithms artifacts and their perceived strengths gathered in the interview stage.

Algorithm	Artifacts	Strength
Shen [19]	added background	low
	border holes	low
Jabri [20]	added regions	medium
	added background	low
Horprasert [21]	border holes	medium
François [22]	added background	high
McKenna [23]	inside holes	medium
	border holes	medium
	flickering	medium
Image Differencing	inside holes	high
	border holes	high
	flickering	medium
Kim [24]	added regions	high
	added background	high
	flickering	high

most annoying artifacts in the sample video sequences. The subjects were then told that different artifacts would appear combined or alone and they should rate the overall annoyance in both cases. During the experimental trials, subjects were asked to evaluate the segmentation regarding the specific application for the tested segmentation algorithms reported in Tab. 2. The total number of test sequences for this experiment was 24 which included 3 original video sequences (‘Hall monitor’, ‘Highway’,‘Group’) × 8 segmentation algorithms (the reference and the real segmentation algorithms described in Tab. 2).

5. EXPERIMENTAL RESULTS

In this section, three different issues are investigated. First, the performance of the proposed perceptual application dependent metric, **PST**, is analyzed and compared to the state of the art metric, *mqm*. Second, the parameters of the novel metric are optimized according to specific application. Moreover, the results of the metric are used to discuss the performance of the selected state-of-the-art segmentation algorithms according to the scenario.

The performance of the proposed **PST** metric is analyzed in terms of correlation coefficients with the obtained subjective Mean Annoyance Values, *MAV*. The linear correlation coefficient of Pearson and the non-linear (rank) correlation coefficient of Spearman are calculated in order to correlate the subjective and the objective results. The objective results have been plotted versus the subjective annoyance values for the surveillance framework and the Pearson and Spearman correlation coefficients are reported in Tab. 3. The correlation coefficients for the perceptual metric, **PST** are larger (Pearson= 0.86, Spearman=0.77) compared to the state of the art matching quality metric *mqm*, showing a good performance of the proposed application dependent metric. It has to be mentioned that the proposed perceptual metric parameters have been derived on the basis of subjective experiments on *synthetic* artifacts. By testing the metric performance on real segmentation algorithms, it has shown its reliability also in the case of *real* artifacts. The perceptual metric predicts automatically the segmentation quality in a similar way human subjects perceive it (i.e. clusters of errors) and outperforms the state of the art metric which does not include perceptual factors.

Our evaluation metric has been proposed for general purpose segmentation with an ideal segmentation at hand. As mentioned, it is important when evaluating the performance of an algorithm to have *a priori* knowledge on the specific application it is addressing. A novelty in the proposed metric is that the a, b, c, d parameters in Eq. (2) can be easily adjusted depending on application by performing a nonlinear least-squares data fitting using the subjective mean annoyance values (*MAV*). Thus, on the basis of the subjective experiment, the best metric parameters have been also computed ($a = 8.96$, $b = 6.48$ $c = 11.30$ $d = 4.06$) by maximizing the correlation coefficients (Pearson=0.91, Spearman=0.85) for the surveillance scenario. Therefore, in this application, the biggest annoyance weights are given to added regions and inside holes. This can be explained by the fact that human viewers in the surveillance scenario pay attention to mis-detected or over-detected objects that could lead to false alarms (in case of erroneous detection of background parts as moving objects) and missed alarms (in case of mis-detection of moving objects).

If the performance of the segmentation algorithms are considered in details for the surveillance case, the best one in both subjective and objective (see [17]) evaluation is given by **Shen**. This is due to the fact that almost no false alarms or missed alarms are caused by this segmentation. In fact, neither added regions nor missing objects are ever produced. Only few border holes and added backgrounds are present due to the integration of the motion information and a more sophisticated classification part. **Kim** gives the worst results due to under-segmentation and over-segmentation depending on the threshold sensitivity and the incorrect contour filling.

6. CONCLUSIONS

A study on real artifacts produced by typical video object segmentation algorithms has been carried out to test the proposed perceptual metric in the video surveillance scenario. To the best of our knowledge, a comparison among different state of the art video object segmentation systems has received little attention by the image processing community

Table 3: Correlation coefficients between the objective metrics (**PST** and *mqm*) and subjective results (*MAV* values) for all the test video sequences in generic and compression frameworks. PST metric parameters: $a = 2.86$, $b = 4.50$, $c = 4.77$, $d = 5.82$

Metric	'Video Surveillance'	
	Pearson	Spearman
mqm	0.72	0.65
PST	0.86	0.77
PST (opt.)	0.91	0.85

so far, as well as the study of their performances for different applications. Seven state of the art segmentation algorithms have been chosen as typical and analyzed objectively with two metrics. The perceptual objective metric **PST** is able to predict the subjective quality as perceived by human viewers according to specific applications. The results show both the better performance of such a metric when compared against the state of the art, *mqm* metric and its adaptability to take into consideration different segmentation applications. The optimal perceptual parameters have been found for the surveillance segmentation application.

7. ACKNOWLEDGMENTS

The authors would like to thank Mustafa Karaman, Marco Carli and Gaia Arrigoni for collaborating in the experiments and Nascimento for his metric's code.

REFERENCES

[1] E. Drelie Gelasca, T. Ebrahimi, M. Karaman, and T. Sikora, "A Framework for Evaluating Video Object Segmentation Algorithms," in *Proc. CVPR 2006, POCV*, New York, USA, June 16-22. 2006.

[2] J. Nascimento and J. S. Marques, "New Performance Evaluation Metrics for Object Detection Algorithms, " in *6th PETS, ECCV)*, Prague, May 2004.

[3] R. J. Oliveira and P. C. Ribeiro and J. S. Marques and J. M. Lemos, "A Video System for Urban Surveillance: Function Integration and Evaluation, " in *WIAMIS, 2004*

[4] F. Oberti and E. Stringa and G. Vernazza, "Performance Evaluation Criterion for Characterizing Video Surveillance Systems"*Real Time Imaging*, vol. 7, 2001, pp. 457-471.

[5] A. Senior and A. Hampaput and Y. Tian and L. Brown and S. Pankanti and R. Bolle, "Appearance Models for Occlusion Handling, " in *in 2nd IEEE PETS*, 2000.

[6] J. Black and T.J. Ellis and P Rosin, "A Novel Method for Video Tracking Performance Evaluation, " in *The Joint IEEE International Workshop on Visual Surveillance and PETS*, 2003, October, pp. 125-132.

[7] A. Cavallaro and E. Drelie Gelasca and T. Ebrahimi, "Objective Evaluation of Segmentation Quality using Spatio-Temporal Context, " in *Proc. IEEE ICIP, Rochester(NY),22-25 September 2002*, pp. 301-304.

[8] P. Correia and F. Pereira, "Objective Evaluation of Video Segmentation Quality", *IEEE Transaction on Image Processing*, series 2, vol. 12, pp.186-200, 2003.

[9] C. E. Erdem and B. Sankur and A. M. Tekalp, "Performance Measures for Video Object Segmentation and Tracking", *IEEE Transactions on Image Processing*, 2004, vol. 13, num. 7, July, pp. 937-951.

[10] C.Erdem and B.Sankur, "Performance Evaluation Metrics for Object-based Video Segmentation, " in *Proc. EUSIPCO, Tampere, Finland*, vol. 2, pp. 917-920, 2000

[11] R. Mech and F. Marques, "Objective Evaluation Criteria for 2D-Shape Estimation Results of Moving Objects, " in *WIAMIS 2001*, Tampere, Finland, 16-17 May 2001.

[12] M. Wollborn and R. Mech, "Refined Procedure for Objective Evaluation of Video Object Generation Algorithms, " in *ISO/IECJTCI/SC29/WG11 M3448*, 43rd MPEG Meeting, Tokyo, Japan 1998.

[13] P. Villegas and X. Marichal, "Perceptually-Weighted Evaluation Criteria for Segmentation Masks in Video Sequences", *IEEE Transactions on Image Processing*, 2004, vol. 13, num. 8, August 2004, pp. 1092-1103.

[14] R. Piroddi and T. Vlachos, "Perceptually-Weighted Evaluation Criteria for Segmentation Masks in Video Sequences", *IEEE Transactions on Image Processing*, 2006.

[15] IEEE 2005 Winter Vision, "Performance Evaluation of Tracking and Surveillance (PETS)", *http://pets2005.visualsurveillance.org/*.

[16] EU CAVIAR Project, "CAVIAR Test Case Scenarios", *http://homepages.inf.ed.ac.uk/rbf/CAVIAR*.

[17] E. Drelie Gelasca, "Full-Reference Objective Quality Metrics for Video Watermarking, Video Segmentation and 3D Model Watermarking", *PhD thesis, EPFL, Lausanne, December 2005*

[18] P. L. Correia and F. Pereira, "Classification of Video Segmentation Application Scenarios ", *IEEE Trans. on Circuits and Systems for Video Tech.*, series 5, vol. 14, pp. 735-741, 2004.

[19] J. Shen, "Motion Detection in Color Image Sequence and Shadow Elimination, " in *VCIP*, January 2004, pp. 731-740.

[20] S. Jabri and Z. Duric and H. Wechsler and A. Rosenfeld, "Detection and Location of People in Video Images Using Adaptive Fusion of Color and Edge Information, " in *ICPR*, September 2000, pp. 627-630.

[21] T. Horprasert and D. Harwood and L. S. Davis, "A Statistical Approach for Real-time Robust Background Substraction and Shadow Detection, " in *IEEE ICCV Frame Rate Workshop*, 1999.

[22] A. R. J. François and G. G. Medioni, "Adaptive Color Background Modeling for Real-Time Segmentation of Video Streams, " in *International on Imaging Science, System, and Technology*, 1999, pp.227-232.

[23] S. J. McKenna and S. Jabri and Z. Duric and A. Rosenfeld and H. Wechsler, "Tracking Groups of People, " in *Computer Vision and Image Understanding*, vol. 80, 2000.

[24] C. Kim and J. Hwang, "Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, num. 2, February 2002.