Repositorio Institucional de la Universidad Autónoma de Madrid

https://repositorio.uam.es

Esta es la **versión de autor** de la comunicación de congreso publicada en:
This is an **author produced version** of a paper published in:

2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2010. IEEE 2010. 180-187

DOI: http://dx.doi.org/10.1109/AVSS.2010.21

Copyright: © 2010 IEEE

El acceso a la versión del editor puede requerir la suscripción del recurso
Access to the published version may require subscription

On the Evaluation of Background Subtraction Algorithms without Ground-truth

Juan C. SanMiguel, José M. Martínez
Video Processing and Understanding Lab
Universidad Autónoma de Madrid — E-28049 Madrid, Spain
E-mail: {JuanCarlos.SanMiguel, JoseM.Martinez}@uam.es

Abstract

In video-surveillance systems, the moving object segmentation stage (commonly based on background subtraction) has to deal with several issues like noise, shadows and multimodal backgrounds. Hence, its failure is inevitable and its automatic evaluation is a desirable requirement for online analysis. In this paper, we propose a hierarchy of existing performance measures not-based on ground-truth for video object segmentation. Then, four measures based on color and motion are selected and examined in detail with different segmentation algorithms and standard test sequences for video object segmentation. Experimental results show that color-based measures perform better than motion-based measures and background multimodality heavily reduces the accuracy of all obtained evaluation results.

1. Introduction

In recent years, the design and development of video-surveillance systems have received an enormous attention by the research community motivated by the increase of the demand of the public and private sector. In such kind of systems, the segmentation of moving objects plays a key role because it is the fundamental processing stage for mid or high-level analysis (e.g., object tracking or event detection). Many segmentation algorithms have been proposed based on background subtraction [1] because of their low computational requirements and the typical static camera use. The segmentation algorithm usually operates in different conditions generated by indoor, outdoor or crowded environments.

In this context, the evaluation of the segmentation algorithm is crucial to estimate its accuracy and to tune its parameters for optimal performance. It has been typically approached by comparing the obtained results of the segmentation algorithms with manually annotations of video foreground objects, that is, a Ground-truth (GT) [1][2]. However, manual annotation is a time-consuming, prone to human error and expensive process. Moreover, it only represents a small percentage of data variability (e.g., object models, shadows). This restriction makes difficult to extrapolate the results of the segmentation performance evaluation to new sequences. Furthermore, GT annotations are not available when the segmentation algorithm is operating in an online mode. Conversely, the evaluation not-based on ground-truth (NGT) is a desirable option to overcome these limitations. It has received little attention by the video-surveillance community because of the difficulty on defining a criteria of good segmentation (which is often application-dependent and hard to define explicitly) [3][4]. The evaluation of object segmentation without ground-truth presents several advantages: it does not require manual annotations, it allows to rank algorithms over large datasets and it is suitable for automatic control of online segmentation (self-tuning).

In this paper we propose a hierarchical organization of state-of-art approaches for video object segmentation evaluation without ground-truth (NGT measures) and we present a comparative evaluation of selected approaches in order to understand their advantages and drawbacks on a public available dataset for the evaluation of video object segmentation. The result of this evaluation is a recommendation on which measure performs better under specific characteristics of the sequences that affect the accuracy of the segmentation algorithm.

This paper is organized as follows: Section 2 describes the proposed classification, selected NGT approaches are presented in section 3; Section 4 discusses the evaluation methodology; Section 5 describes the optimum parameter study; Section 6 shows the experimental results and finally Section 7 describes the conclusions and future work.

2. State of art

In the video-surveillance domain, evaluation of foreground segmentation has been mainly approached empirically, although there are some analytical proposals like [5]. Moreover, empirical evaluation can be divided into the use (or not) of annotations of foreground objects.

Evaluation based on ground-truth information (GT), also known as relative evaluation [6] or empirical discrepancy evaluation [7], is based on directly measuring

the deviation between the segmentation results (e.g., binary masks) and the manual annotations of foreground objects. For instance, [1] proposes to analyze the performance of seven background subtraction methods by measuring the number of pixels correctly and erroneously classified as foreground and background under different conditions considered in the test sequences. Moreover, [2] defines more sophisticated statistics using GT information with the aim to detect the split/merge of foreground objects, false alarms and the detection of failures. Additionally, the visual relevance of segmentation errors can be considered to weight them [8].

Evaluation not-based on ground-truth information (NGT), also known as stand-alone evaluation [6] or empirical goodness evaluation [7] is based on inspecting desired properties of the empirical results. Therefore, these methods rely on a priori information about expected segmentation results (e.g., matching of color and object boundaries). Among the existing NGT approaches, we can define three categories: boundary-based, model-based and assisted-based evaluation methods. The following subsections will focus on the three categories. The complete hierarchy of the foreground segmentation evaluation methods is depicted in Fig. 1.

2.1. Boundary-based evaluation

Boundary-based evaluation methods rely on inspecting the properties of the internal and external regions defined by the boundary of the segmented object.

Internal inspection of object boundary regards the study of the homogeneity of object features. For instance, [3] proposes several spatial features like circularity, elongation and compactness of the objects as object homogeneity features. However, these features rely on the temporal correspondence of the video object (tracking). Additionally, [3] proposes motion uniformity as temporal homogeneity features not restricted to use tracking information. Moreover, [9] proposes the difference (in gray-level) between the original and the segmented images as a measure of segmentation quality.

It should be noted that unsupervised evaluation of image segmentation has received more attention than the video object segmentation during last years [7]. However, its applicability is restricted to regions with uniform properties and video objects are usually composed of various color regions (e.g., people, vans) so it is expected that these unsupervised approaches will fail.

Contrast inspection regards the study of the feature differences between the internal and external regions defined by object boundaries. For instance, color contrast is proposed in [4][10]; moreover, motion contrast is also proposed in [4]. Both contrast measures are obtained in the neighborhoods of each boundary pixel. [11] describes
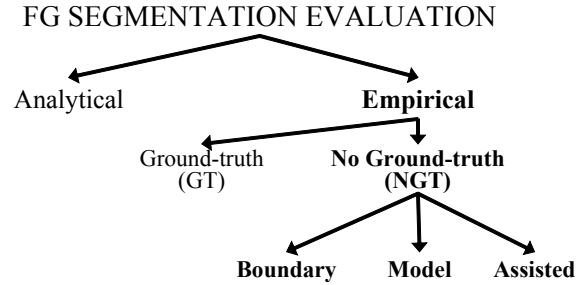
FG SEGMENTATION EVALUATION

Figure 1: Hierarchy of the segmentation evaluation methods

an improved version of the pixel neighborhood used in [4] to address the problem of unreliable and/or unavailable feature estimates. [12] defines the edge profiles to analyze color differences under low contrast conditions.

2.2. Model-based evaluation

Model-based evaluation methods examine the impact of the segmentation results on the following analysis stages (e.g., object classification). They are based on the availability of video object models (e.g., person) or artifacts (e.g., shadows) expected to appear in the video sequence. This evaluation approach is useful to measure if the segmented regions satisfy the system requirements (e.g., human detection).

Within this category we can find several proposals. [13] proposes to use different high-level modules to detect expected foreground objects (people, non-people and illumination changes) and to estimate segmentation accuracy at bounding box level to feedback the segmentation module. A block-based human model is proposed in [14] to evaluate the accuracy of segmentation masks and to correct segmentation errors. Moreover, [15] proposes to validate foreground masks by building a simple moving object model using foreground and background statistics as well as the frame difference.

2.3. Assisted-based evaluation

Assisted-based evaluation methods rely on using complementary algorithms to evaluate foreground segmentation. The key idea is to automatically build an approximation of the GT information to estimate foreground segmentation quality. The expected accuracy is low because they are very dependant on the results of the complementary algorithm.

For example, [16] evaluates a Single Gaussian Background subtraction stage by using a Frame Difference technique. Similarly, [17] proposes to analyze visual and infrared data for object segmentation. [18] builds an estimated GT by using a region-based segmentation algorithm and matches the boundary of segmented video foreground objects and obtained regions.

3. Overview of selected NGT measures

We have chosen the boundary-based contrast measures based on color and motion difference proposed by [4]. We have decided not to use measures from the other two previously described categories (model and assisted) because the constraints introduced (model of foreground regions and accuracy of the additional algorithm) are hard to satisfy. On the contrary, matching of object and color region boundaries is usually satisfied for the video-surveillance domain. Additionally, the segmentation algorithms selected for the experiments do not include any foreground model information.

The first measure selected is the color contrast along the boundary [4]. It is based on defining normal lines of length $2L+1$ for each boundary pixel and comparing the color differences between the initial (P_I) and ending (P_O) points of each normal line. The neighborhood of these pixels is also considered by using a window of size MxM. The scheme is depicted in Fig. 2. It proposes to estimate the segmentation quality of each boundary pixel, *Boundary Spatial Color Contrast*, as shown in Eq. (1).

$$BSCC(t;i) = \frac{\left\| C_O^i(t) - C_I^i(t) \right\|}{\sqrt{3*255^2}} \qquad (1)$$

where C_O and C_I are the average colors calculated in the MxM neighborhood of the pixels P_I and P_O (using the *RGB* color space quantified into 256 levels) for each i boundary pixel of the foreground region at time t. Thus, it proposes to evaluate the foreground segmentation of each object, O_j, and to combine the segmentation of multiple objects as defined in Eq. (2) and (3).

$$DC1_{O_j}(t) = \frac{1}{K_t} \sum_{i=1}^{K_t} BSCC(t;i) \qquad (2)$$

$$DC1(t) = \min_j (DC1_{oj}) \qquad (3)$$

where K_t is the total number of boundary pixels, *BSCC* is the spatial color contrast of boundary pixel i of the foreground region being analyzed. Its value ranges from 0 (lowest segmentation quality) to 1 (highest segmentation quality).

Additionally, [4] proposes to use this measure to detect incorrectly segmented boundary pixels if they are above a certain threshold, T_1. A metric of segmentation quality could be derived by counting the correctly segmented boundary pixels as defined in Eq. (4) and (5).

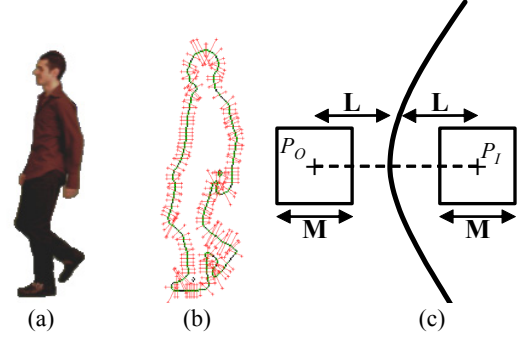$$DC2_{O_j}(t) = \frac{\#(BSCC_j(t) > T_1)}{K_t} \qquad (4)$$

Figure 2: Boundary-based contrast scheme proposed by [4]. (a) Segmented object, (b) its boundary with the normal lines and (c) a zoom on a boundary pixel location

$$DC2(t) = \min_j (DC2_{oj}) \qquad (5)$$

The main advantages of this measure are its low complexity and its possibility to detect failures at finer level (boundary pixel). These aspects make the measure useful for its use to adapt or feedback real-time video segmentation algorithms to improve the segmentation performed. The parameters (of this measure) to study are the normal line length L, the size M of the window around P_O/P_I points and the threshold, T_1, used in the *DC2* metric.

The second measure is based on the motion difference along the object boundary [4]. Similarly to the color-based measure, normal lines of length $2L+1$ are drawn for each boundary pixel and the motion difference between the internal and external parts of object boundaries (supposed, respectively, to be moving and static), *Boundary Motion Contrast*, is computed as shown in Eq. (6).

$$BMC(t;i) = \omega_i \left(1 - \exp\left(-\frac{\left\| v_O^i(t) - v_I^i(t) \right\|}{\sigma^2} \right) \right) \qquad (6)$$

$$\omega_i = R(v_O^i(t)) \cdot R(v_I^i(t))$$

where v_O and v_I are the average motion vectors of the MxM windows centered in the pixels P_I and P_O for each i boundary pixel of the foreground region. R(.) represents the reliability of the motion vectors [4]. The evaluation measures are defined as shown in Eqs. (7) and (8).

$$DM1_{O_j}(t) = \frac{1}{K_t} \sum_{i=1}^{K_t} BMC_i(t) \qquad (7)$$

$$DM1(t) = \min_j (DM1_{oj}) \qquad (8)$$

where K_t is the total number of contour pixels, *BMC* is the boundary motion contrast of pixel i of the foreground region being analyzed, O_j. Its value ranges from 0 (lowest segmentation quality) to 1 (highest segmentation quality).

Similarly, this metric could be used to detect correctly segmented boundary pixels if they are above a certain threshold, T_2, as defined in Eq. (9) and (10).

$$DM2_{o_j}(t) = \frac{\#(BMC_j(t) > T_2)}{K_t} \qquad (9)$$

$$DM2(t) = \min_j(DM2_{oj}) \qquad (10)$$

Similarly to the *BSCC* measure, the main advantages of *BMC* are its low complexity and its possibility to detect failures at finer level (boundary pixel). The only assumption of the measure is that the foreground regions of the sequence have to be moving. For this measure, the parameters to study are the normal line length *L*, the size *M* of the window around P_O/P_I points and the threshold, T_2, used in the *DM2* metric.

4. Evaluation methodology

For the evaluation of the selected NGT measures, we propose to study the variability of their parameters and their application to foreground segmentation algorithms commonly used in video-surveillance. Moreover, their correlation with GT measures is computed to estimate their performance. Thus, it includes three key aspects: use of GT metrics for computing the correlation, selection of the appropriate dataset and the selection of segmentation algorithms. This section describes these aspects whilst section 5 describes the obtained results.

4.1. Ground-truth measures

As GT measures, we propose to use the precision measure for foreground detection (*P1*). Recall of foreground (*R1*) and Precision/Recall of background segmentation (*P0/R0*) have been excluded because the NGT measures are not able to evaluate the accuracy of missed foreground pixel detections (for measure *R1*) or background segmentation results (measures *P0* and *R0*). *P1* measure is defined as Eq. (11).

$$P1 = \frac{TP}{TP + FP} \qquad (11)$$

where *TP* indicates the number of foreground pixels correctly detected, *FP* the number of foreground pixels wrongly detected. Then, the correlation between the *P1* and the NGT measures is computed to study the performance of the evaluation by using the Pearson product-moment correlation coefficient [19]. It ranges from -1 to +1 where a value of +1 indicates perfect positive correlation whilst -1 indicates perfect negative

Table 1: *Description of the selected sequences.*

Seq.	Length (frames)	Background		Foreground	
		Textures	Multimodal	Velocity	Size
ID1	750	High	Low	Low	Medium
ID4	1250	Low	Low	Low	Medium
ID5	752	Medium	High	Low	Medium
ID6	672	Medium	High	Low	Medium
ID7	620	Medium	High	Low	Medium
ID8	794	Medium	Medium	Low	Medium
ID9	1380	Medium	Medium	High	Medium
ID10	307	Medium	Medium	High	Medium
ID11	732	Low	Low	High	Low

correlation between the GT and NGT measures. A value close to 0 indicates no correlation between them.

4.2. Dataset description

The test sequences have been selected from the cVSG dataset [20] that has been designed for the evaluation of moving object segmentation, allowing to combine different foregrounds and backgrounds, and is composed of high-quality uncompressed video sequences of size 720x576 (and associated ground-truth) classified according to several criteria. We propose to use some of these criteria to evaluate the performance of the NGT measures under different complexities of the background and the foreground regions. A description of the selected sequences and their main characteristics are shown in Table 1. The sequences with camera motion have been excluded for the experiments because of the segmentation algorithms do not handle them.

4.3. Selected segmentation algorithms

As segmentation algorithms, we have selected four representative approaches of the background subtraction technique commonly used in video-surveillance: the first two approaches [21][22] independently model each background pixel and the other two [23][24] consider spatial relations of neighborhood pixels. Additionally, a noise filtering stage has been used to remove the salt&pepper noise of the obtained binary foreground segmentation masks by using mathematical morphology [25]. Parameter tuning of each approach has been done according to the results reported in [1].

The first selected approach is the Mixture of Gaussians (MoG) [21] where the movement of each background pixel is represented with a set of weighted Gaussian distributions. The distributions with higher weights are considered to model the background; the remaining to model the foreground. Foreground pixel detection is

Average correlation of DC1 and P1 measures

		L				
		1	3	5	7	9
M	1	65.3±2.6	68.3±3.2	65.3±3.3	62.4±4.3	59.8±12.0
	3	-	72.7±4.1	**73.2±4.3**	69.3±7.3	66.8±8.2
	5	-	-	72.8±3.3	70.0±2.6	67.1±7.5
	7	-	-	-	71.2±4.3	67.1±5.0
	9	-	-	-	-	66.1±10.2

(a)

Average correlation of DM1 and P1 measures

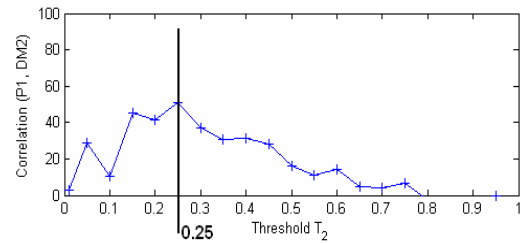|		L				
		1	3	5	7	9
M	1	25±15.4	29±12.8	31.1±8.3	30.0±12.4	29.1±9.8
	3	-	36.3±5.6	**45.1±4.6**	35.2±9.4	34±10.3
	5	-	-	39.3±7.7	36.6±8.3	36.9±6.9
	7	-	-	-	31.5±7.7	38.3±5.0
	9	-	-	-	-	32.5±9.4

(b)

Average correlation of DC2 and P1 measures
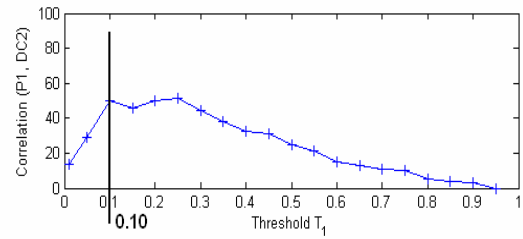
(c)

Average correlation of DM2 and P1 measures

(d)

Figure 3: Optimum parameter estimation for NGT measures. Results are the average of the four segmentation algorithms and they correspond to the correlation (in percentage) of (a) *DC1*, (b) *DM1*, (c) *DC2* and (d) *DM2* with foreground object precision (*P1*) using the Pearson product-moment correlation coefficient. Maximum correlation values are marked in bold.

decided if the pixel does not fall into the deviation around the mean of any of the Gaussians that model the background. This approach is useful to analyze sequences with multimodal background

The second selected approach [22] is the Kernel Density Estimator (KDE) that estimates the probability density function (*pdf*) of each pixel by using the last *N* frames. Foreground/Background pixel detection is decided if its likelihood of belonging to the pixel *pdf* is lower or higher than a predefined threshold. This approach is able to analyze sequences with multimodal backgrounds.

The third selected approach [23] is the Gamma method (GAMMA). It is based on a pixel neighborhood analysis by subtracting a square window between current and background images (around each considered pixel). This subtraction is modeled as a Chi-square distribution considering that the pixel variation follows a Gaussian. The final decision is taken by thresholding the obtained probability of belonging to the Chi-square distribution. This approach eliminates the salt&pepper noise in the foreground binary mask.

The fourth selected approach is the Eigenbackgrounds (EigBG) [24]. It is based on applying principal component analysis to the previous *N* frames in order to capture the spatial relations. A set of basis functions is obtained as a result and each new frame is project into the eigenspace defined by these functions to remove foreground objects. Foreground detection is obtained by comparing each frame with its back projection.

5. Optimum parameter selection

A study of the optimum parameters has been carried out for each NGT measure (*DC1*, *DC2*, *DM1* and *DM2*) for the different segmentation results obtained with the selected algorithms. The optimization process has been divided in two stages using sequences with unimodal backgrounds from the cVSG dataset in order to avoid the adaptation of the parameter values to the high amount of segmentation errors found in multimodal sequences. A summary of the evaluation results is depicted in Fig. 3.

Firstly, we have performed an exhaustive search of the optimum *L* and *M* parameter values by using the measures *DC1* and *DM1* (independent to any thresholding operation). The optimum selection criterion is the maximum correlation between *DC1* and *DM1* with the GT measure *P1*. As it can observed in Fig. 3(a) and 3(b), the optimum values (maximum correlation) for *DC1* measure was 0.732 for the parameters *L=5* and *D=3* whilst for *DM1* measure was 0.451 for the parameters *L=5* and *D=3*. Then, the optimum values of T_1 and T_2 (used in measures *DC2* and *DM2*) are calculated by applying an exhaustive search and considering the optimum values of *L* and *M* for each measure. Results are shown in Fig. 3(c) and 3(d). The optimum selection criterion is the maximum correlation between *DC2* and *DM2* with the GT measure *P1*. Finally, the optimum values were $T_1=0.10$ and $T_2=0.25$ for the *DC2* and *DM2* measures.

Table 2: *Average evaluation results and correlation with P1 measure for background motion test (in percentage)*

Segm. Algorithm	Low Background Motion									Medium-High Background Motion								
	GT	NGT Color				NGT Motion				GT	NGT Color				NGT Motion			
	P1	DC1	Cor.	DC2	Cor.	DM1	Cor.	DM2	Cor.	P1	DC1	Cor.	DC2	Cor.	DM1	Cor.	DM2	Cor.
MoG [21]	96.5	**19.9**	**72.0**	75.7	52.8	15.3	48.0	70.3	45.0	52.5	25.3	20.1	73.4	19.4	**16.3**	**21.2**	77.5	9.4
KDE [22]	88.4	**18.4**	**73.9**	70.2	60.9	10.9	35.6	68.9	34.4	33.4	**18.4**	**18.5**	55.1	16.4	11.5	15.6	69.5	11.0
GAMMA[23]	93.3	**19.7**	**74.0**	74.3	55.4	9.9	41.6	70.4	40.6	50.3	19.7	13.4	**66.2**	**17.6**	8.0	14.9	80.3	5.5
EigBG [24]	84.3	**13.6**	**75.0**	46.6	52.8	7.6	49.3	55.6	48.1	47.1	**13.6**	**18.6**	69.4	11.1	19.6	9.3	99.3	3.4

Table 3: *Average evaluation results and correlation with P1 measure for background texture test (in percentage)*

Segm. Algorithm	Low Background Texture									Medium-High Background Texture								
	GT	NGT Color				NGT Motion				GT	NGT Color				NGT Motion			
	P1	DC1	Cor.	DC2	Cor.	DM1	Cor.	DM2	Cor.	P1	DC1	Cor.	DC2	Cor.	DM1	Cor.	DM2	Cor.
MoG [21]	99.9	**19.0**	**78.3**	78.7	60.3	9.3	41.1	65.3	40.9	96.8	18.8	55.8	74.9	49.8	13.3	**55.9**	67.4	42.9
KDE [22]	92.1	**17.0**	**75.5**	75.4	48.2	8.8	53.5	69.4	40.3	90.6	**16.3**	**63.5**	60.5	35.1	12.8	56.5	73.2	45.1
GAMMA[23]	95.4	**18.2**	**71.3**	77.6	62.4	12.0	51.2	70.9	38.6	94.3	**18.9**	**59.1**	71.9	56.3	15.4	54.9	75.1	39.8
EigBG [24]	80.5	11.0	70.3	**81.3**	**88.1**	9.9	49.3	47.4	41.5	82.9	19.2	61.8	**72.3**	**70.1**	20.4	50.0	59.9	44.5

6. Experimental results

In this section, we study the performance of the NGT evaluation measures under different background and foreground characteristics (previously defined in Table 1). For each one, appropriate test sequences are selected and the segmentation algorithms are applied to obtain the segmentation results. Then, NGT and GT measures are used to evaluate the performance of the obtained segmentations with the objective of understanding in which cases they provide high or low evaluation accuracy. The best obtained results (highest correlation) are marked in bold in each corresponding table.

6.1. Background motion

Background motion (or multimodality) is a key issue in foreground segmentation because it is difficult to handle. The test sequences ID1, ID4 and ID11 were used for computing the results of low background motion and the remaining ones for medium-high background motion. The obtained evaluation results are summarized in Table 2. As we can observe, high background motion reduces the accuracy of the segmentation algorithms increasing the number of wrong detections (see *P1* measure). For the color-based measures, the observed performance decrease (correlation values) is caused by the wrong detected objects (moving background) as they have color-boundaries. For the motion-based measures, the background motion also produces motion boundaries and the measures wrongly evaluate the segmented background as good so their correlation values are heavily reduced.

6.2. Background textures

Different background textures were used to test NGT evaluation performance. Test sequences ID4 and ID1 were used as low and high textured backgrounds. The obtained results are summarized in Table 3. As we can observe,, high-textured backgrounds increase the performance (correlation) of motion-based NGT measures and slightly decrease the performance of color-based NGT measures because of the motion boundaries are less difficult to estimate whilst color boundaries are more difficult to obtain. On the contrary, low textured backgrounds benefit the use of the color-based NGT measures because color boundaries are easier to estimate and decrease the performance of motion-based NGT measures because the extraction of motion boundaries is more difficult.

6.3. Foreground velocity

Different velocities of foreground objects have been tested to study the performance of the NGT measures. The test sequence ID4 was used for low foreground velocity whilst sequence ID11 was used for high foreground velocity. The obtained results are summarized in Table 4. As we can observe, the velocity of the foreground objects does not significantly affect the color-based NGT measures whilst the motion-based NGT measures present a dependence on the object velocity. The computation of the motion boundaries has to be adapted to the specific characteristics of each sequence by setting the optimum parameter for the motion vector calculation (e.g., block size and area search for block matching).

Table 4: *Average evaluation results and correlation with P1 measure for foreground motion test (in percentage)*

Segm. Algorithm	Low Foreground Motion									Medium-High Foreground Motion								
	GT	NGT Color				NGT Motion				GT	NGT Color				NGT Motion			
	P1	DC1	Cor.	DC2	Cor.	DM1	Cor.	DM2	Cor.	P1	DC1	Cor.	DC2	Cor.	DM1	Cor.	DM2	Cor.
MoG [21]	99.9	**19.0**	**78.3**	78.7	60.3	9.3	41.1	65.3	40.9	90.3	**19.0**	**75.3**	74.0	49.1	10.1	45.0	63.1	37.5
KDE [22]	92.1	**17.0**	**75.5**	75.4	48.2	8.8	53.5	69.4	40.3	89.6	**17.3**	**72.0**	71.5	42.3	9.9	36.5	70.1	31.1
GAMMA[23]	95.4	**18.2**	**71.3**	77.6	62.4	12.0	51.2	70.9	38.6	85.9	**15.9**	**68.1**	73.1	59.1	11.5	50.6	66.5	35.4
EigBG [24]	80.5	11.0	70.3	**81.3**	**88.1**	9.9	49.3	47.4	41.5	95.3	20.2	63.9	**70.1**	**69.3**	25.0	39.0	69.6	45.5

Table 5: *Average evaluation results and correlation with P1 measure for foreground size test (in percentage)*

Segm. Algorithm	Low Foreground Size									Medium-High Foreground Size								
	GT	NGT Color				NGT Motion				GT	NGT Color				NGT Motion			
	P1	DC1	Cor.	DC2	Cor.	DM1	Cor.	DM2	Cor.	P1	DC1	Cor.	DC2	Cor.	DM1	Cor.	DM2	Cor.
MoG [21]	70.5	**17.7**	**30.2**	81.1	29.3	17.0	21.0	78.1	19.4	44.4	**21.8**	**26.2**	90.1	11.0	21.0	10.0	88.1	9.9
KDE [22]	60.8	**12.4**	**28.1**	69.9	27.5	12.6	18.8	69.5	12.9	25.1	**14.0**	**19.3**	76.9	12.5	16.6	8.1	77.5	8.9
GAMMA[23]	71.2	**17.4**	**35.4**	81.2	32.7	13.5	20.9	80.3	18.5	49.8	**18.9**	**29.9**	83.0	25.5	18.5	13.1	83.8	11.0
EigBG [24]	57.3	**12.5**	**31.3**	66.7	30.1	10.6	17.0	88.1	11.3	43.3	**15.6**	**21.0**	52.9	13.1	13.6	15.2	89.9	4.3

6.4. Foreground size

Different sizes of foreground objects have been tested to study the performance of the NGT measures. All the test sequences have been used and the obtained results were classified depending if the object had less or more than 200 pixels (considering them as low or high object size). The obtained results are summarized in Table 5. As we can observe in Table 5, the size of the object does not affect the color and motion-based NGT. However, the NGT measures will not work with very thin objects (e.g., walking sticks) because the parameter L could be higher than the dimensions of the object (width or height). The low accuracy (correlation) of NGT measures is because the presence of sequences with multimodal backgrounds produces wrong objects multiple sizes.

7. Conclusions and future work

In this paper we have studied two evaluation measures for foreground segmentation not-based on ground truth information. They rely on comparing the boundaries of the segmented objects against the color and motion boundaries of the video sequence. An evaluation of the measures under different sequence characteristics has been carried out for the video-surveillance domain. Their correlation with ground-truth measures has been used to check their performance. Experimental results showed that the color-based measures are more accurate than the motion-based ones because motion boundaries are more difficult to estimate than color boundaries due to homogeneous or slow-moving object regions. Background multimodality dramatically affects their performance whilst the effect of background textures and foreground velocity/size is less noticeable. Among the color-based measures, *DC1* performs better than *DC2* because it does not need any thresholding operation. Similarly happens with *DM1* and *DM2* motion-based measures.

As future work, we will study the use of these measures to detect segmentation failures and to feedback foreground segmentation algorithms to improve their accuracy.

8. Acknowledgments

9. References

[1] S. Herrero and J. Bescos, "Background subtraction techniques: Systematic evaluation and comparative analysis", in Proc. of ACIVS, pp. 33-42, 2009.

[2] J. Nascimento and J. Marques, "Performance evaluation of object detection algorithms for video surveillance", in IEEE Trans. on Multimedia, 8(4): 761–774, 2006.

[3] P. Correia and F. Pereira, "Stand-alone objective segmentation quality evaluation", in EURASIP Journal on Applied Signal Processing, vol. 4, pp. 389–400, 2002.

[4] C. Erdem, Sankur, and A. Tekalp, "Performance measures for video object segmentation and tracking", in IEEE Trans. on Image Processing, 13(7):937–951, 2004.

| | (a) | (b) | (c) | (d) | (e) | (f) |
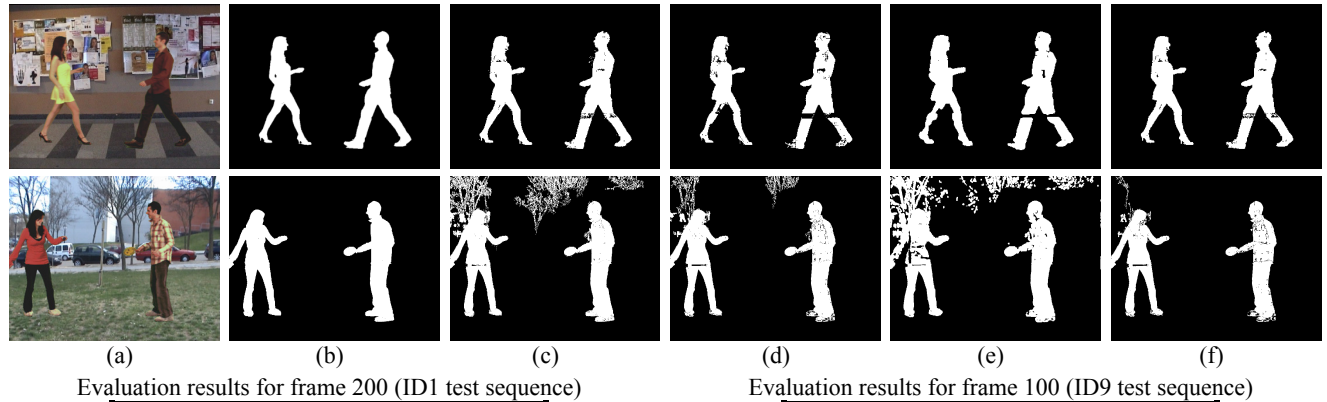
Evaluation results for frame 200 (ID1 test sequence)

Seg. algorithm	GT	NGT			
	P1	DC1	DC2	DM1	DM2
MoG [21]	99.8	17.9	83.5	22.1	85.3
KDE [22]	88.3	15.3	77.3	18.5	80.1
GAMMA[23]	91.5	14.7	68.5	17.0	76.6
EigBG [24]	95.4	19.0	74.5	19.3	79.4

Evaluation results for frame 100 (ID9 test sequence)

Seg. algorithm	GT	NGT			
	P1	DC1	DC2	DM1	DM2
MoG [21]	71.6	20.3	67.5	12.3	66.6
KDE [22]	77.3	23.3	73.4	18.4	77.9
GAMMA[23]	68.9	19.8	61.7	22.3	83.1
EigBG [24]	88.2	26.1	77.0	21.5	81.8

Figure 4: Segmentation masks and their evaluation for frames 200 (first row) and 350 (second row) of ID1 and ID5 test sequences in terms of (a) original frame, (b) ground-truth and segmentation masks of (c) MoG, (d) KDE, (e) GAMMA and (f) EigBG algorithms.

[5] X. Gao, T. Boult, F. Coetzee, and V. Ramesh, "Error analysis of background adaptation", in Proc. of IEEE CVPR, pp. 503-510, 2000.

[6] P. Correia and F. Pereira, "Objective evaluation of video segmentation quality", IEEE Trans. on Image Processing, 12(2):186-200, 2003.

[7] H. Zhang, J. Fritts, and S. Goldman, "Image segmentation evaluation: a survey of unsupervised methods", Computer Vision and Image Understanding, 110:260-280, 2008.

[8] P. Villegas and X. Marichal, "Perceptually-weighted evaluation criteria for segmentation masks in video sequences", IEEE Trans. on Image Processing, 13(8):1092–1103, 2004.

[9] J. Weszka and A. Rosenfeld, "Threshold evaluation techniques", IEEE Trans. on Systems, Man and Cybernetics 8 (8):622–629, 1978.

[10] S. Chabrier, B. Emile, C. Rosenberger, and H. Laurent, "Unsupervised performance evaluation of image segmentation", EURASIP Journal on Applied Signal Processing, pp. 1–12, 2006.

[11] R. Piroddi and T. Vlachos, "A method for single-stimulus quality assessment of segmented video", EURASIP Journal on Applied Signal Processing, pp. 1–22, 2006.

[12] O. Kubassova, M. Boesen, and H. Bliddal, "General framework for unsupervised evaluation of quality segmentation results", in Proc. of IEEE ICIP, pp. 3036–3039, 2008.

[13] M. Harville, "A framework for high-level feedback to adaptive, per pixel, mixture-of-gaussian background models", in Proc. of IEEE ECCV, pp. 543–560, 2002.

[14] M. Rincon, E. Carmona and M. Bachiller, , "Segmentation of moving objects with information feedback between description levels", in Proc. of IWINAC, pp.171–181, 2007.

[15] S. Cheung and C. Kamath, "Robust background subtraction with foreground validation for urban traffic video",

EURASIP Journal on Applied Signal Processing, pp. 2330–2340, 2005.

[16] A. Garcia and J. Bescos, "Video object segmentation based on feedback schemes guided by a low-level scene ontology", in Proc. of ACIVS, pp. 322-333, 2008.

[17] C. Ó Conaire, N. O'Connor, and A. Smeaton, "Detector adaptation by maximising agreement between independent data sources" in Proc. of the IEEE OTCBVS, pp. 1-6, 2007.

[18] L. Goldmann, T. Adamek, P. Vajda, et al, "Towards fully automatic image segmentation evaluation", in Proc of ACIVS, pp. 566–577, 2008.

[19] P. Chen, and P. Popovich, "Correlation: Parametric and nonparametric measures", Thousand Oaks: Sage Publications; 2002.

[20] F. Tiburzi, M. Escudero, J. Bescos, and J. Martinez, "A ground-truth for motion-based video-object segmentation", in Proc. of IEEE ICIP, pp. 17–20, 2008.

[21] G. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking", in Proc. of IEEE CVPR, pp. 246-252 , 1999

[22] A.M. Elgammal,, D. Harwood and L.S Davis, "Non-parametric model for background subtraction", in Proc. of IEEE ECCV, pp.751–767, 2000

[23] A. Cavallaro, O. Steiger and T. Ebrahimi,"Semantic video analysis for adaptive content delivery and automatic description", IEEE Trans. on Circuits and Systems for Video Technology, 15(10):1200–1209, 2005

[24] N. Oliver, B. Rosario, and A. Pentland. "A bayesian computer vision system for modeling human interactions", IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(8):831–843, 2000.

[25] P. Salembier and J. Ruiz. "On filters by reconstruction for size and motion simplification", Proc. of Int. Symposium in Mathematical Morphology, pp.425–434, 2002.