A Framework for Evaluating Video Object Segmentation Algorithms

Elisa Drelie Gelasca, Touradj Ebrahimi
EPFL, CH-1015 Lausanne, Switzerland
elisa.drelie@a3.epfl.ch

Mustafa Karaman, Thomas Sikora
TUB, D-10587 Berlin, Germany
karaman@nue.tu-berlin.de

Abstract

Segmentation of moving objects in image sequences plays an important role in video processing and analysis. Evaluating the quality of segmentation results is necessary to allow the appropriate selection of segmentation algorithms and to tune their parameters for optimal performance. Many segmentation algorithms have been proposed along with a number of evaluation criteria. Nevertheless, no psychophysical experiments evaluating the quality of different video object segmentation results have been conducted. In this paper, a generic framework for segmentation quality evaluation is presented. A perceptually driven automatic method for segmentation evaluation is proposed and compared against an existing approach. Moreover, on the basis of subjective results, perceptual factors are introduced into the novel objective metric to meet the specificity of different segmentation applications such as video compression. Experimental results confirm the efficiency of the proposed evaluation criteria.

1. Introduction

Unsupervised segmentation of digital images is a difficult and challenging task [16] with several key-applications in many fields: image classification, object recognition, etc. The performance of algorithms for subsequent image or video processing, compression and indexing, to mention a few, often depends on a prior efficient image segmentation in which the *a priori* knowledge of the application is also integrated.

Recent multimedia standards and trends in image and video[1] representation have increased the importance of adequately segmenting semantic "objects" in video, in order to ensure efficient coding, manipulation and identification.

Therefore, many segmentation algorithms have been proposed (see Sec. 3), as well as a number of evaluation criteria for segmentation quality assessment reviewed in Sec. 2. The need for a standard quality metric arises from the fact that segmentation is an ill-posed problem: for the same image/video, the optimum segmentation can be different depending on the application.

Many researchers prefer to rely on qualitative human judgment for evaluation. However, subjective evaluation asks for a large panel of human observers, thus resulting in a time-consuming and expensive process. Therefore, there is a need for an automatic objective methodology to allow the appropriate selection of segmentation algorithms as well as to adjust their parameters for optimal performance.

During the last several years, some objective methods for video object segmentation evaluation have been proposed, but no work has been done on studying and characterizing the artifacts typically found in digital video object segmentation to derive a *perceptual* metric. A good understanding of how annoying these artifacts are and how they combine to produce the overall annoyance is an important step in the design of a reliable *perceptual objective quality metric*. To this end, first a series of specially designed psychophysical experiments has to be performed. In this paper, a perceptual metric is derived on the basis of the subjective results. The novelty of the proposed approach consists in studying and characterizing the typical segmentation errors from a perceptual point of view. Different clusters of error pixels are perceptually classified according to the fact if they do or they do not modify the shape of the object.

Second, an objective and subjective study of the annoyance generated by real artifacts introduced by typical video object segmentation algorithms is presented both for an evaluation generic framework and a specific application: video compression. Finally, this paper also provides a comparison of performance of the proposed perceptual metric against a state-of-the-art metric.

2. Overview on Evaluation Methods

The problem of *subjectively* and *objectively* assessing the quality of segmentation has been investigated in different contexts in literature: edge-based segmentation [7], region-based segmentation [11], and video object segmentation [5, 3, 13, 17, 18, 14, 2]. Nevertheless, there is no standardized procedure for subjective tests on any of these seg-

[1] http://www.chiariglione.org/mpeg/

mentation methods, nor any universally adopted objective metrics. In literature, subjective judgments are based on human intuition. A set of general guidelines for segmentation quality assessment has been proposed in the COST211/quat European project [5]. These guidelines concern only how the typical display configuration should look like (see [3]), but they do not specify how the test should be carried out (*e.g*. experimental methodology such as type of questions to observers, etc.) In [13] some criteria related to the computational complexity of the segmentation system are defined together with a number of questions to investigate subjectively the video object segmentation quality for surveillance applications. However, we noticed that in this case subjects had to perform a sort of *memory test* given the large number of questions asked after the video is played back. The capacity of a test subject to reliably assess several elements of a video is limited. The memory of a video fades after time and lends to a tiring and too difficult task to be accomplished. For all the above described reasons, a subjective evaluation methodology is proposed in Sec. 4.

Subjective segmentation evaluation is necessary to study and to characterize the *perception* of different artifacts on the overall quality, but once this task has been accomplished successfully and an automatic procedure has been devised, systematic subjective evaluation can be avoided.

The automatic procedure is referred to as *objective evaluation method*. Quality metrics for objective evaluation of segmentation may judge either the segmentation algorithms or their segmentation results. These are referred to as analytical or empirical methods, respectively [19]. *Empirical methods* do not evaluate the segmentation algorithms directly, but indirectly through their results. Empirical methods are divided into *empirical discrepancy* metrics when the segmentation result is compared to an ideally segmented 'reference' mask (ground truth), and *empirical goodness* metrics when the quality of the result is based on intuitive measures of goodness such as color uniformity. The main disadvantage of such an approach is that the goodness metrics are at best heuristic, and may exhibit strong bias toward a particular algorithm. For this reason, we have chosen to implement a discrepancy method which makes use of the ground-truth. State of the art discrepancy methods and, in particular the MPEG metric chosen as a term of comparison for our metric, are reviewed in Sec. 2.1.

2.1. Objective Evaluation Criteria

To evaluate a segmented video by discrepancy methods, Erdem and Sankur [2] combined three empirical discrepancy measures into an overall quality segmentation evaluation: *misclassification penalty*, *shape penalty*, and *motion penalty*. In [3], first the individual segmentation quality were measured by four spatial accuracy criteria: *shape fidelity*, *geometrical fidelity*, *edge and statistical content similarity* and two temporal criteria: *temporal perceptual information* and *criticality*. Second, the similarity factor between the reference and the resulting segmentation is computed. Furthermore, the multiple-object case was addressed by using the criteria of application-dependent "*object relevance*" to provide the weights for the quality metric of each object. Finally, they combined all these three measures into an overall segmentation quality evaluation.

Another way to approach the problem is to consider it as a particular case of shape similarity as proposed in [14] for video object segmentation. In this method, the evaluation of the spatial accuracy and the temporal coherence is based on the mean and standard deviation of the 2-D shape estimation errors.

During the standardization work of ISO/MPEG-4, within the core experiment on automatic segmentation of moving objects, it became necessary to compare the results of different proposed object segmentation algorithms, not only by subjective evaluation, but also by objective evaluation. The proposal for objective evaluation [18] agreed by the working group uses a ground truth in order to evaluate the segmentation results. This metric is usually adopted by the research community due also to its simplicity. For this reason, the MPEG metric has been chosen as term of comparison for the new metric proposed in this paper. In the following section, the description of this metric is provided in detail.

A refinement of the MPEG metric has been proposed by Villegas *et al*. [17]. For the evaluation of the spatial accuracy, as opposed to the previous method, two classes of pixels are distinguished: false positive and false negative, and they are weighted differently. Furthermore, their metric takes into account the impact of the two classes on the spatial accuracy, *i.e*. the evaluation worsens with pixel distance to the reference object contour. The perceptual difference between two kinds of errors is given by means of 'perceptual' weighting functions. The drawback is that these are defined by means of empirical tests which are not generally sufficient to guarantee the definition of 'perceptual' weights. In this paper, the relevance and the corresponding weight of different kinds of errors is supported by formal subjective experiments performed under clear and well defined specifications.

2.1.1 MPEG Evaluation Criteria

A moving object can be represented by a binary mask, called *object mask*, where a pixel has object-label if it is inside the object and background-label if it is outside the object. The objective evaluation approach used in the ISO/MPEG-4 core-experiment has two objective criteria: the *spatial accuracy* and the *temporal coherence*. Spatial accuracy, Sqm, is estimated through the amount of error pixels in the object mask (both false positive and false negative pixels) in the resulting mask deviating from the ideal mask.

Temporal coherence is estimated by the difference of the spatial accuracy between the mask, M, at the current and previous frame k,

$$Tqm_M(k) = Sqm(k) - Sqm(k-1). \tag{1}$$

The two evaluation criteria can be combined in a single *MPEG error measure*, through the sum:

$$MPEG = \sum_k \big(Sqm(k) + Tqm_M(k)\big). \tag{2}$$

In this metric, the perceptual difference of different classes of errors, false positive and false negative, is not considered and they are all treated the same. In fact, different kinds of errors should be combined in the metric in correct proportions to match evaluation results produced by human observers.

3. Segmentation Algorithms

In the experiments, we chose seven static background segmentation methods. The approaches of the selected representative algorithms differ in using various features such as color, luminance, edge, motion and combinations of them. A quick overview of the principles on which each technique is based is reported. For further details the reader is invited to refer to each appropriate paper. Tuning of parameters has been done on several video sequences and the best parameters for each algorithm were tuned according to visual inspection.

Image Differencing is based on basic background subtraction in which greyscale images are used and an absolute differencing with the background and current frame is applied. The segmentation results depend only on the threshold method used for binarization.

Kim's [10] approach is based on greyscale images and applies the Canny edge operator to the current, background, and successive frames. The motion information obtained by the difference edge map is used for selecting the relevant edges from the current frame. The object mask is achieved by filling the boundaries obtained by the previous edge results with connecting the first and second occurred edge pixels for each vertical and horizontal line, respectively.

Horprasert *et al.* [8] use color and illumination information. This method evaluates for each pixel the brightness and the chromaticity distortions between the background image and the current frame. The background is modeled by four values: the mean and the standard deviation over several background frames and the variation of the brightness and chromaticity distortions. Each pixel of the current frame is classified as *original background*, *shadow*, *highlighted background*, and *foreground*.

François and Medioni's [6] technique operates in the HSV color space and models the background by using the mean and standard deviation. The pixels of the current frame are compared to those of the updated background. For the classification of each pixel the V value is always used and the color information H and S are used in the regions where they are evaluated to be reliable.

Shen [15] uses both RGB and HSI color spaces. The segmentation is executed in two steps. In the first step a fuzzy classification is utilized by considering the mobility of pixels which is generated by combining the results from separately thresholded difference images of each RGB channel. In the second step the falsely detected pixels from the first step are eliminated by using the previous segmentation result and the motion information obtained from successive frames. The HSI color space is used to overcome shadows by considering the basic illumination features of shadow.

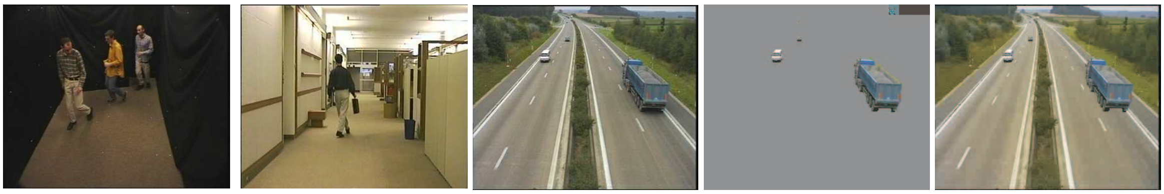
Jabri *et al.* [9]'s system uses both information: RGB pixel values and edges. The background model is trained in both mentioned parts by calculating the mean and standard deviation for each pixel of any color channel. The edge model is built by applying the Sobel edge operator for both horizontal and vertical cases. Confidence maps are generated for color and edge respectively, and a combination of them is utilized by taking its maximum values. Finally, this output goes through a hysteresis thresholding for binarization.

McKenna *et al.* [12] also use color and edge information to model the background. Instead of the RGB color space the normalized RGB space (rgb) is used. The models are generated separately for each channel. The incoming frame is classified separately and a combination of both classification results gives the final segmentation mask.

4. Subjective Evaluation

The proposed subjective experiment methodology corresponds to the five-step procedure described in detail in [4]: *oral instructions* (the subject is made familiar with the task of segmentation), *training* (original and reference sequences are shown), *practice trials* (subjects' responses are collected on a small subset of test sequences), *experimental trials* (the test is performed on the complete set of sequences), *interview* (qualitative descriptions of the perceived artifacts).

The test group was composed of 35 subjects aged between 23-41 (of which 8 females). The subjects were asked one question after each segmented video sequence was presented, “How annoying was the defect relative to the worst 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 most annoying artifacts in the sample video sequences. The subjects were then told that different artifacts would



‘Group’ (a) ‘Hall’ (b) ‘Highway’ (c) Reference (d) Compression (e)

Figure 1. Sample frames for original, reference segmentation and compression segmentation application.

appear combined or alone and they should rate the overall annoyance in both cases. In fact, *five* different clusters of errors were recognized as typically provided 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 textured video objects have been overlapped on a uniform gray background ($Y = 127$, $U = 127$, $V = 127$) and the three original sequences used in this experiment are ’Group’, ‘Hall monitor’ and ‘Highway’ (see Fig. 1 (a), (b), (c)). The seven segmentation algorithms described in the previous section have been applied to each original video sequence. Both *general* and *application* dependent segmentation scenarios were considered in the subjective evaluation. A total number of 48 sequences were generated: 21 test segmented sequences (3 original $\times$ 7 segmentations plus 3 references $\times$ 2 frameworks).

In order to assess if a segmentation is good in a general scenario, viewers were asked to mentally compare the results of the segmentation at hand with the ideal (reference) segmentation (shown in Fig. 1 (d)) and formulate their judgments. Studying how subjective quality scores change in relation to the specific segmentation tasks provides a lot of interesting insights in developing evaluation metrics. In the following, a possible application scenario is described and the subjective results providing general guidelines for the development of segmentation algorithms are presented.

4.1. Application Dependent Evaluation

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. The setting up of a subjective experiment differs for each application.

Table 1. Description of segmentation algorithms artifacts and their perceived strengths gathered in the interview stage.

Algorithm	Artifacts	Strength
Shen	added background	low
	border holes	low
Jabri	added regions	medium
	added background	low
Horprasert	border holes	medium
François	added background	high
McKenna	inside holes	medium
	border holes	medium
	flickering	medium
Image Differencing	inside holes	high
	border holes	high
	flickering	medium
Kim	added regions	high
	added background	high
	flickering	high

In *video compression*, segmentation can improve the coding performance over a low-bandwidth channel. The MPEG-4 coding scheme[2] was adopted to compress the background separately from the objects. Since we only want to study the segmentation artifacts perception, distortions due to compression should not be included in the segmented objects. Thus, the segmented video objects were not actually compressed. In such a way, the compressed background could be transmitted only once and the video objects corresponding to the foreground (moving objects) could be transmitted and added on top of it so as to update the scene. A sample of compressed background test sequence is shown in Fig. 1 (e). Subjects were instructed with the video compression principles and asked to only judge the object segmentation quality in relation to this task. Video compression is a typical case where knowledge of the specific application can be used to tune the parameters of the evaluation metric: undetected object’s parts will have a bigger impact on the overall annoyance than over-segmentation of the detected objects (see Sec. 4.2). In fact, the parts of the object that are undetected will be compressed as erroneously considered parts of the background.

4.2. Subjective Results

Standard methods [1] were used to analyze and to screen the judgments provided by the test subjects. From the data gathered, we calculated the Mean Annoyance Values (MAV) of each test sequence. Table 1 shows the subjective ranking during the *interview stage* of the subjective ex-

[2] Miscrosoft’s MPEG VM software encoder & decoder. Version: FDAMI 2-3-001213, integrator: Simon Winder, Microsoft Corp.

Table 2. MAV values obtained for each segmentation algorithm for all the test video sequences in generic and compression frameworks.

Alg.	‘Group’		‘Hall monitor’		‘Highway’		‘$\overline{MAV}$’	
	Gen.	Cmpr.	Gen.	Cmpr.	Gen.	Cmpr.	Gen.	Cmpr.
reference	8.77	11.20	26.74	15.51	15.31	10.77	16.94	12.5
Jabri	57.46	22.63	40.37	10.60	37.94	25.00	42.25	19.41
Horprasert	69.94	48.63	57.57	20.17	32.06	20.74	53.19	29.8
Shen	57.83	33.94	55.26	60.71	54.26	56.14	55.78	50.26
François	68.57	39.57	61.43	66.71	30.20	33.46	53.40	46.58
McKenna	83.36	76.43	56.86	71.37	54.26	71.57	68.82	73.12
Image D.	99.74	90.00	60.00	48.40	67.54	75.34	75.76	71.24
Kim	72.00	40.00	86.89	52.00	71.14	45.51	76.67	45.8

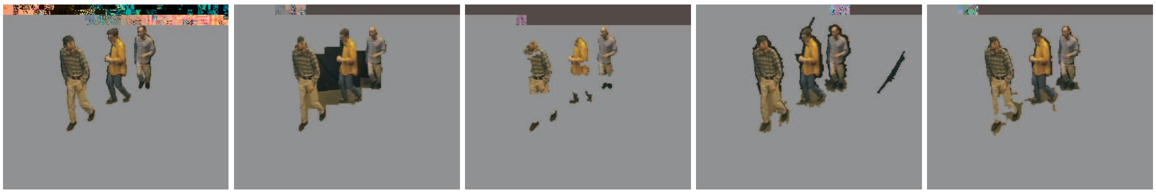

reference **Kim** **Image D.** **Jabri** **Shen**

Figure 2. Sample frames for the reference and some segmentation results of the tested video sequence ’Group’ (frame #100).

periment for the general framework. This table reports the tested algorithms from the least to the most annoying and a brief description of the artifacts that are typically introduced. Table 2 reports the MAV values, gathered in the *experimental trials*, for all video and algorithms, along with the different scenarios considered. The results of the subjective experiments averaged for all the three video sequences are also reported in the last two columns. The averaged Annoyance Values ($\overline{MAV}$) have been computed for each algorithm and the reference in order to provide a general overview on the segmenting performance of the described algorithms. In the general scenario, the subjective results show that the algorithms which on average introduce the most annoying artifacts are the **Kim** and **Image Differencing** algorithms. The least annoying artifacts are generated by **Horprasert**, **Jabri** and **Shen** algorithms (see Fig. 2).

The most annoying artifact is flickering usually due to noise, camera jitter and varying illumination. It produces erroneously segmented regions (different at each frame). A high value of flickering of added regions is generated by **Kim**’s algorithm and it is the most annoying artifact on average for the general scenario (Tab. 2). In fact, no matter what the size of the artifact is, if the segmentation presents temporal instabilities it will annoy the subject a lot more than any other spatial artifact.

In general scenario, the second most annoying artifact according to subjective experiments is that introduced by **Image Differencing** due to the large amount of holes and especially border holes. They are perceived as the most annoying in terms of spatial errors. Holes are usually due to the algorithm’s failures in differentiating the foreground regions from the background when they look very similar in color or texture or other uniformity features that the algorithm exploits to segment. Then the artifacts introduced by **McKenna** are rated as the third most annoying ones. In this case, especially the holes are annoying to human observers, even if they are smaller than those introduced by the **Image Differencing**’s method, but still of considerable amount.

Added background is the fourth annoying artifact and it is generated by **François**’s algorithm. It is mostly caused by erroneously detecting moving shadows as part of the moving foreground objects. Since shadows move along with objects from which they are casted, we observed that this artifact does not annoy too much the human observer and is subjectively rated better than flickering or missing parts of objects in this general scenario.

The least annoying artifacts in average are introduced by **Horprasert**, **Jabri** and **Shen** algorithms. In fact, these algorithms introduce smaller amounts of artifacts compared to others (see Sec. 6 for compression scenario analysis).

5. Proposed Evaluation Criteria

The proposed discrepancy method 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 our approach consists in classifying the different clusters of error pixels according to the following characteristics: if they do or they do not modify the shape of the object and afterward their size. Border holes, $\mathcal{H}_b$, and added backgrounds, $\mathcal{A}_b$, modify the shape while inside holes, $\mathcal{H}_i$, and added regions, $\mathcal{A}_r$ preserve the segmented object shape (see Sec. 4).

The relative spatial error $\mathbf{S}_{A_r}(k)$, for all the j added regions at frame k, $\mathcal{A}_r^j(k)$, is obtained by simply applying:

$$\mathbf{S}_{A_r}(k) = \frac{\sum_{j=1}^{N_{Ar}} |\boldsymbol{A}_r^j(k)|}{|n(k)|}, \tag{3}$$

where $|\cdot|$ is the set cardinality operator; $n(k)$ is the sum of the reference and the result segmentation areas; N_{Ar} is the total number of added regions.

Similarly, for all the j holes inside the segmentation, $\mathcal{H}_i^j(k)$, the relative spatial error, $\mathbf{S}_{H_i}(k)$, is given by:

$$\mathbf{S}_{H_i}(k) = \frac{\sum_{j=1}^{N_{Hi}} |\boldsymbol{H}_i^j(k)|}{|n(k)|}, \tag{4}$$

where N_{Hi} is the total number of holes inside the objects. The spatial error for added background and holes on the border of the object is formulated in a different way. In fact, both kinds of errors are located around the object contours and it has to be distinguished from numerous deviations around the object boundary and a few but larger deviation [14] by adding this weighting factor, D^j:

$$D^j = 1 + \frac{\overline{d^j} + \sigma_d^j}{d_{max}^j}, \tag{5}$$

where d are the distance values[3] of error pixels from the correct object contour. The mean $\overline{d}$ and the standard deviation σ_d of d are calculated and are then normalized by the maximal diameter, d_{max}, of the reference object to which the cluster of errors belongs to. By combining this last Eq. (5) and Eq. (3), we obtain, for the border artifacts, the corrected relative spatial error $\mathbf{S}_{A_b}(k)$, for j added backgrounds:

$$\mathbf{S}_{A_b}(k) = \frac{\sum_{j=1}^{N_{Ab}} D_{\mathbf{A}b}^j \cdot |\boldsymbol{A}_b^j(k)|}{|n(k)|} \tag{6}$$

similarly for j holes on the border, $\mathcal{H}_b^j(k)$, the relative spatial error $\mathbf{S}_{H_b}(k)$ is:

$$\mathbf{S}_{H_b}(k) = \frac{\sum_{j=1}^{N_{Hb}} D_{\mathbf{H}b}^j \cdot |\boldsymbol{H}_b^j(k)|}{|n(k)|} \tag{7}$$

The temporal artifact caused by an abrupt variation of the spatial errors between consecutive frames is called *flickering*. To take this phenomenon into account in the objective metric, a measure of flickering is introduced, $\mathbf{F}(k)$ that can be computed for each kind of artifact Λ=[$\mathcal{A}_r$, $\mathcal{A}_b$, $\mathcal{H}_i$, $\mathcal{H}_b$] as follows:

$$\mathbf{F}_\Lambda(k) = \frac{|\Lambda(k)| - |\Lambda(k-1)|}{|\Lambda(k)| + |\Lambda(k-1)|}, \tag{8}$$

The difference of artifact amounts between two consecutive frames is normalized by the sum of the amount of this artifact in the current frame k and the previous frame $k-1$. To model this effect, Eq. (8) is combined to the relative spatial artifact measures to construct an objective spatio-temporal error measure $\mathbf{ST}(\mathbf{k})$ for each artifact, and finally the artifact is summed along the time axis to obtain the overall objective spatio temporal metric $\mathbf{ST}$ for each artifact Λ:

$$\begin{aligned} \mathbf{ST}_\Lambda(k) &= \mathbf{S}_\Lambda(k) \cdot \frac{1 + \mathbf{F}_\Lambda(k)}{2}, \\ \mathbf{ST}_\Lambda &= \frac{1}{K} \sum_{k=1}^{K} w_t(k) \mathbf{ST}_\Lambda(k), \end{aligned} \tag{9}$$

where the temporal weights $w_t(k)$ that model the *human memory effect* have been empirically defined [4] as:

$$w_t(k) = (a \cdot e^{\frac{k-30}{b}} + c) \tag{10}$$

with $a = 0.02$, $b = 7.8$, $c = 0.0078$, $K = 60$ (total number of frames).

5.1. Perceptual Objective Metric

In [4] *synthetic* artifacts were used to study and characterize the perception of the spatial and temporal artifacts previously described. In the following, a brief description of the parameters obtained for the perceptual metric is given and in the next section, the proposed metric is tested on *real* artifacts. The **ST** values of each artifact metrics were plotted versus the values of MAV and the best fitting psychometric curves were found [4] to describe the human perception of errors. Four psychometric curves were derived through subjective experiments, one for each artifact, to obtain four *perceptual artifact metrics*: **PST**$_\Lambda$. The best fitting function for each artifact was the Weibull function, W. Thus the perceptual artifact metrics are described by:

$$\begin{aligned} W(x, S, k) &= 1 - e^{-(Sx)^k} \quad where \quad x = \mathbf{ST}_\Lambda \\ \mathbf{PST}_\Lambda &= W(\mathbf{ST}_\Lambda, S, k) \end{aligned} \tag{11}$$

where the parameters S and k have been obtained in [4] for the general scenario case with synthetic artifacts: $S = 0.014$, $k = 0.304$ for $\mathbf{PST}_{A_r}$; $S = 0.026$, $k = 0.653$ for $\mathbf{PST}_{A_b}$; $S = 0.331$, $k = 0.2339$ for $\mathbf{PST}_{H_i}$; $S = 0.771$, $k = 0.641$ for $\mathbf{PST}_{H_b}$.

The overall perceptual metric is given by the combination of all the four kinds of artifacts. A simple linear combination of artifacts [4] estimates the total annoyance:

$$\mathbf{PST} = a \cdot \mathbf{PST}_{A_r} + b \cdot \mathbf{PST}_{A_b} + c \cdot \mathbf{PST}_{H_i} + d \cdot \mathbf{PST}_{H_b} \tag{12}$$

[3] For distance computation, 8-connectivity has been used.

The perceptual weights were found by means of subjective experiments [4] on combined synthetic artifacts: $a = 2.86$, $b = 4.50$, $c = 4.77$, $d = 5.82$.

6. Experimental Results

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

The performance of the proposed **PST** metric are analyzed in terms of correlation coefficients with the obtained subjective MAV values. 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 two frameworks 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=0.89) compared to the state of the art MPEG metric (Pearson= 0.73, Spearman=0.67) for both scenarios showing a good performance of the proposed 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 the state of the art 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 MPEG metrics which does not include perceptual factors.

Our evaluation metric has been proposed for general purpose segmentation with an ideal segmentation at hand. 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. (12) can be easily adjusted depending on applications 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 ($a = 2.34$, $b = 0.62$ $c = 8.59$ $d = 13.39$) by maximizing the correlation coefficients (Pearson=0.89, Spearman=0.89).

In the compression scenario, the optimized weights obtained for added regions and background ($a = 2.34$, $b = 0.62$) are really small compared to those for inside and border holes ($c = 8.59$, $d = 13.39$). In fact, in this application we have preserved the quality of the segmented objects and compressed the background. Therefore, the parts of the object that have been erroneously segmented as part of the background have been compressed and annoy the subjects more than having segmentation artifacts like added region or background that have not be compressed. In such a case, the difference in perception of the four artifacts has been numerically quantified.

Since the final goal for an objective metric is to help in choosing the best performing algorithm on a given set of data, the performance of the state of the art segmentation algorithms are discussed on the basis of the **PST** metric results reported in Tab. 4. If the performance of the segmentation algorithms are considered in the general case, the best one in both subjective (Tab. 2) and objective (Tab. 4) evaluation is given by **Jabri** for 'Hall' and 'Group'. In fact, the generated confidence maps and the hysteresis thresholding method which integrates neighbor pixels is more capable than other methods to distinguish homogeneous regions. For the 'Highway', the best performance is achieved by **Horprasert** in which the distortions for brightness and chromaticity obtained from background modeling give a bigger range to classify only the relevant object pixels in the current frame. **ImageDifferencing** and **Kim** give the worst results due to under-segmentation and over-segmentation depending on the threshold sensitivity and the incorrect contour filling of **Kim**.

In the video compression case, overall **Jabri** was estimated as the best performing algorithm as for the general scenario. In fact, even if this algorithm introduces some added background and added regions, they are not much bothering the user in this specific application: they are not compressed as well as the rest of the object and unlike the background. **ImageDifferencing** and **McKenna** shows the worst cases since this last method is not able to deal with similar colors in the background and foreground causing inside and border holes.

Table 3. Correlation coefficients between the objective metrics (**PST** and **MPEG**) 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	'Generic'		'Compression'	
	Pearson	Spearman	Pearson	Spearman
MPEG	0.73	0.67	0.49	0.41
PST	0.86	0.79	0.78	0.79

7. Conclusions

A perceptually driven objective metric for segmentation quality evaluation has been proposed on the basis of psychophysical experiments. A study on real artifacts produced by typical video object segmentation algorithms has been

Table 4. **PST** metric values obtained for each segmentation algorithm in both generic and compression frameworks.

Alg.	‘Group’		‘Hall monitor’		‘Highway’		‘$\overline{MAV}$’	
	Gen.	Cmpr.	Gen.	Cmpr.	Gen.	Cmpr.	Gen.	Cmpr.
reference	2.96	6.84	2.96	6.84	2.96	6.84	2.96	6.84
Jabri	21.59	14.86	26.31	31.36	23.84	16.24	29.31	20.82
Horprasert	31.76	43.64	31.35	40.23	20.48	20.59	27.36	34.82
Shen	28.78	40.98	35.89	63.76	24.81	37.83	29.82	47.52
François	40.84	46.86	43.87	74.36	29.19	35.80	37.96	53.00
Kim	28.98	41.67	43.42	54.13	35.13	44.44	35.84	46.76
McKenna	42.73	69.18	56.86	68.26	31.12	54.66	43.57	64.03
Image D.	46.64	92.33	60.00	62.84	36.76	50.40	47.8	68.52

carried out to test the proposed perceptual metric. 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 so far, as well as the study of their performances for different applications. Seven state of the art segmentation algorithms were chosen as typical and analyzed both objectively and subjectively. First, a classification of the real artifacts introduced is provided according to subjective perception. Second, a perceptual objective metric able to predict the subjective quality as perceived by human viewers has been proposed. The results show both the better performance of such a metric compared against the usually adopted MPEG metric and its adaptability to take into consideration different segmentation applications. The optimal perceptual parameters have been found for a specific segmentation application, the video compression.

Acknowledgment

The authors would like to thank Marco Carli and Gaia Arrigoni for running some experiments.

References

[1] *Subjective Video Quality Assessment Methods for Multimedia Applications Recommendation P.910*. International Telecommunication Union, Geneva, Switzerland, 1996. 4

[2] C.Erdem and B.Sankur. Performance evaluation metrics for object-based video segmentation. In *Proc. X European Signal Processing Conference, Tampere, Finland*, volume 2, pages 917–920, 2000. 1, 2

[3] P. Correia and F. Pereira. Objective evaluation of video segmentation quality. *IEEE Transaction on Image Processing*, 12:186–200, 2003. 1, 2

[4] E. Drelie. *Full-Reference Objective Quality Metrics for Video Watermarking, Video Segmentation and 3D Model Watermarking*. PhD thesis, EPFL, Lausanne, 2005. 3, 6, 7

[5] C. for AM Comparisons. Compare your segmentation algorithm to the cost 211 quat analysis model http://www.iva.cs.tut.fi/cost211/call/call.htm. 1, 2

[6] A. R. J. François and G. G. Medioni. Adaptive color background modeling for real-time segmentation of video streams. pages 227–232, 1999. 3

[7] M. Heath, S. Sarkar, T. Sanocki, and K. Bowyer. A robust visual method for assessing the relative performance of edge detection algorithms. *Transactions on Pattern Analyis and Machine Intelligence*, 19:1338–1359, 1997. 1

[8] T. Horprasert, D. Harwood, and L. S. Davis. A statistical approach for real-time robust background substraction and shadow detection. 1999. 3

[9] S. Jabri, Z. Duric, H. Wechsler, and A. Rosenfeld. Detection and location of people in video images using adaptive fusion of color and edge information, September 2000. 3

[10] 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*, 12(2), February 2002. 3

[11] D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of ICCV-01, July 7-14, 2001, Vancouver*, volume 2, pages 416–425, 2001. 1

[12] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler. Tracking groups of people. *Computer Vision and Image Understanding*, 80:42–56, 2000. 3

[13] K. McKoen, R. Navarro-Prieto, E. Durucan, B. Duc, F. Ziliani, and T. Ebrahimi. Evaluation of segmentation methods for surveillance applications. In *EUSIPCO*, pages pp. 1045–1048, September 2000. 1, 2

[14] R. Mech and F. Marques. Objective evaluation criteria for 2d-shape estimation results of moving objects. In *WIAMIS*, Tampere, Finland, 16-17 May 2001. 1, 2, 6

[15] J. Shen. Motion detection in color image sequence and shadow elimination. pages 731–740, January 2004. 3

[16] M. Sonka, V. Hlavic, and R. Boyle. *Image Processing, Analysis and Machine Vision*. An International Thomson Publishing Company, 2nd edition, 1999. 1

[17] P. Villegas and X. Marichal. Perceptually-weighted evaluation criteria for segmentation masks in video sequences. *IEEE Transactions on Image Processing*, 13(8):1092–1103, August 2004. 1, 2

[18] 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, 1998. 1, 2

[19] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29:1335–1346, 1996. 2