On Evaluating Video Object Segmentation Quality: A Perceptually Driven Objective Metric

Elisa Drelie Gelasca and Touradj Ebrahimi

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 formal 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 state-of-the-art. Moreover, on the basis of subjective results, weighting strategies are introduced into the proposed objective metric to meet the specificity of different segmentation applications such as video compression and mixed reality. Experimental results confirm the efficiency of the proposed approach.

Index Terms

video object, segmentation, perceptual metric, objective evaluation, psychophysical tests, subjective quality assessment, video object compression, mixed reality.

I. Introduction

Unsupervised segmentation of digital images is a difficult and challenging task [1] with several key-applications in many fields: image classification, object recognition, etc. The performance of algorithms

E. Drelie Gelasca is with Vision Research Laboratory Dept. of Electrical and Computer Engineering University of California Santa Barbara, Ca 93106. Mailto: elisa.drelie@a3.epfl.ch, Phone: +1(805)8935682; T. Ebrahimi is with the Signal Processing Institute, Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland.

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. III), as well as a number of evaluation criteria for segmentation quality assessment reviewed in Sec. II. 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 specifically designed psychophysical experiments has to be performed.

In this paper, a perceptual metric is derived on the basis of the subjective results. 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 specific applications: video compression and mixed reality. First, in this paper, a perceptual metric is built on synthetic artifacts. 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 specific applications: compression and mixed reality. Finally, this paper also provides a comparison of performance of the proposed perceptual metric against state-of-the-art metrics.

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

II. 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 [2], region-based segmentation [3], and video object segmentation [4], [5], [6], [7], [8], [9], [10]. Nevertheless, there is no standardized procedure for subjective tests on any of these segmentation methods, nor any universally adopted objective metrics. In literature (see Sec. II-A), subjective judgments are based on human intuition.

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 [11]. *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 example the intra-region gray-level uniformity goodness metric will cause poor evaluation for any segmentation algorithm which forms regions of uniform texture. For this reason we have chosen to implement a discrepancy method which makes use of the ground-truth. State of the art discrepancy methods are reviewed in the Sec. II-B and summarized in Tab. I.

A. Subjective Evaluation

A set of general guidelines for segmentation quality assessment has been proposed in the COST211/quat European project [4]. These guidelines concern only how the typical display configuration should look like (see [5]), but they do not specify how the test should be carried out (*e.g.* experimental methodology such as type of questions to observers, etc.). This framework proposes to show people four video sequences at the same time and it does not specify how long they should be. Thus, we performed some informal tests and noticed that using this display configuration and for short video sequences (5-10 seconds) four sequences may be too many since subjects can concentrate only on one of them. Moreover, this layout also shows the original sequence without any segmentation which we believe is not essential, since the

subject, once he/she has learned the task, forms his/her own *implicit* segmentation and does not look any more at the original nor at the reference segmentation.

In [6] 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. For each video sequence, the subject can see the original video sequence as many times as necessary. Then, the segmented video is presented only once and the subject has to answer to four evaluation criteria (such as "how well have been important moving objects individually identified?", or "how well are boundaries provided?").

In informal tests, we tried to combine different questions to describe the aspects of segmentation quality. 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. IV, in which only one question is asked after the video is played back and one video sequence is shown during the test.

B. Objective Evaluation

To evaluate a segmented video by discrepancy methods, Erdem and Sankur [10] combined three empirical discrepancy measures into an overall quality segmentation evaluation: *misclassification penalty*, *shape penalty*, and *motion penalty*. In [5], first the individual segmentation quality are 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 [9] 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. We proposed to evaluate the quality of a segmented object through spatial and temporal accuracy joined to yield a combined metric [12]. This work was based on two other discrepancy methods [8], [13]

described below.

During the standardization work of ISO/MPEG-4, within the core experiments on automatic segmentation of moving objects, it became necessary to compare the results of different object segmentation algorithms, not only by subjective evaluation, but also by objective evaluation. The proposal for objective evaluation [8] agreed by the working group uses a ground truth. This metric is adopted by the research community due also to its simplicity. A refinement of this metric has been proposed by Villegas *et al.* [13], [7]. These two metrics and Nascimento's one [14] have been chosen as term of comparison for a new metric proposed in this paper.

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:

$$MPEG = \frac{1}{K} \sum_k \big(Sqm(k) + Tqm_M(k) \big). \tag{2}$$

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.

2) Weighted Evaluation Criteria: Within the project COST 211 [4] the above approach has been further developed by Villegas and Marichal [7], [13]. For the evaluation of the spatial accuracy, as opposed to the previous method, two classes of pixels are distinguished: those which have object-label in the resulting object mask, but not in the reference mask (false positive) and vice versa (false negative), and they are weighted differently. Furthermore, their metric takes into account the impact of these two classes on the spatial accuracy, that is, the evaluation worsens with pixel distance d to the reference object contour. The

spatial accuracy, qms, is normalized by the sum of the areas of reference objects as follows:

$$qms(k) = \frac{qms^{+}(k) + qms^{-}(k)}{\sum_{i=1}^{N_R} R_i(k)}$$
$$= \frac{\sum_{d=1}^{D_M^+} w_+(d) \cdot |\mathcal{P}_d(k)| + \sum_{d=1}^{D_M^-} w_-(d) \cdot |\mathcal{N}_d(k)|}{\sum_{i=1}^{N_R} R_i(k)}, \quad (3)$$

where D_M^+ and D_M^- are the biggest distance d for, respectively, false positives and false negatives; N_R is the total number of objects in the reference R; $\sum_{i=1}^{N^R} R_i(k)$ is the sum of the area of all the objects i in the reference; $\mathcal{P}_d(k)$ and $\mathcal{N}_d(k)$ are positive and negative pixels respectively; $w_+(d)$ and $w_-(d)$ are the weights for positives and negatives respectively, expressed as:

$$w_+(d) = b_1 + \frac{b_2}{d + b_3}, \qquad w_-(d) = f_S \cdot d, \quad (4)$$

where the parameters b_i and f_S are chosen empirically [7]: $b_1 = 20$, $b_2 = -178.125$, $b_3 = 9.375$ and $f_S = 2$. These functions represent the fact that the weights for false negative pixels increase linearly and they are larger than those for false positives at the same distance from the border of the object. In fact, as we move away from the border, missing parts of objects become more important than added background.

Two criteria are used for estimating temporal coherence, the temporal stability $qmt(k)$ and the temporal drift $qmd(k)$ of the mask. First, the variation of spatial accuracy criterion between successive frames is investigated as follows. The temporal stability is equal to the normalized sum of the differences of the spatial accuracy for two consecutive frames for false positive and false negative pixels:

$$qmt(k) = \frac{qms^{+}(k, k-1) + qms^{-}(k, k-1)}{\sum_{i=1}^{N_R} R_i(k)}. \quad (5)$$

where $qms^*(k, k-1) = |qms^*(k) - qms^*(k-1)|$.

Second, the displacement of the gravity center, $\overrightarrow{G}$, of the resulting object and the reference object mask is computed for successive frames to estimate possible *drifts* of the object mask, $\overrightarrow{qmd}(k)$:

$$\overrightarrow{qmd}(k) = [\overrightarrow{G}_E(k) - \overrightarrow{G}_R(k)] - [\overrightarrow{G}_E(k-1) - \overrightarrow{G}_R(k-1)] \quad (6)$$

that is displacement from time $(k-1)$ to time (k) of the centers of gravity $\overrightarrow{G}$, of the estimated E and reference R masks. The value of drift is the norm of the displacement vector divided by the sum of the reference object bounding boxes,

$$qmd(k) = \frac{||\overrightarrow{qmd}(k)||}{\frac{1}{N_R}\sum_{i=1}^{N_R} BB_i^{x,y}(k)}, \quad (7)$$

where $BB_i^{x,y}(k)$ is the bounding box of the object i in the reference mask R at time k. The authors proposed to define a single quality value by linearly combining all the three presented measures as the

weighted quality metric, $wqm(k)$:

$$wqm(k) = w_1 \cdot qms(k) + w_2 \cdot qmt(k) + w_3 \cdot qmd(k),$$
$$wqm = \frac{1}{K} \sum_k wqm(k). \tag{8}$$

The values of the weights w_i are very much application dependent. If no application is specified, all three weights can be assumed equal to $\frac{1}{3}$.

In this method, the perceptual difference between two kinds of errors is taken into account. The drawback is that the weighting functions defined in Eq. (4), that should be 'perceptual' weights of the evaluation criteria, are defined by means of empirical tests. These empirical tests are not generally sufficient. As well in all other proposed evaluation criteria in the literature, the relevance and the corresponding weight of different kinds of errors should be supported by formal subjective experiments performed under clear and well defined specifications.

3) Object Matching Evaluation Criteria: Nascimento and Marques [14] used several simple discrepancy metrics to classify the errors into region splitting, merging or split-merge, detection failures and false alarms. In their 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 the 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 (C_D): the detected region matches one and only one region; False Alarm (F_A): the detected region has no correspondence; Detection Failure (D_F): 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 (S_M): when the conditions M and S simultaneously occur.

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

$$\begin{aligned} mqm(k) &= w_1 \cdot \frac{C_D(k)}{N_R} + w_2 \cdot \frac{F_A(k)}{N_C} + w_3 \cdot \frac{D_F(k)}{N_R} \\ &+ w_4 \cdot \frac{M(k)}{N_R} + w_5 \cdot \frac{S(k)}{N_R} + w_6 \cdot \frac{S_M(k)}{N_R} \end{aligned} \tag{9}$$

where w_i are the weights for the different discrepancy metrics. mqm is the sum of $mqm(k)$ normalized over all frames. 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, while the metrics described in the previous sections are able to monitor intrinsic properties of the segmented objects such as shape irregularities and temporal instability of the mask along time.

TABLE I

OBJECTIVE MEASURES USED IN EVALUATING IMAGE AND VIDEO OBJECT SEGMENTATION SYSTEMS.

Criteria	Objective Metric
Positions of mis-segmented pixels	Cav. [12], Erdem [10], Villegas [7], [13]
Classes of mis-segmented pixels	Cav. [12], Villegas [7], MPEG [8]
Number of objects	Correia [5], Nascimento [14]
Shape changes	Erdem [10], Correia [5], Mech [9]
Temporal stability	Villegas [7], [13], MPEG [8], Erdem [10], Cav. [12]
Temporal drift	Villegas [7], [13]

III. 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 [15] 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.* [16] 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 [17] 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 [18] 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.* [19]'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.* [20] 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.

IV. Subjective Evaluation

The proposed subjective experiment methodology corresponds to the five-step procedure described in detail in [21]: *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

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

Fig. 1. Sample frames of original tested video sequences and segmentation applications (compression and mixed reality) of the tested video sequence ‘Highway’ (frame #95).

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 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 72 sequences were generated: 21 test segmented sequences (3 original $\times$ 7 segmentations plus 3 references $\times$ 3 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. 2) 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.

TABLE II

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

TABLE III

MAV VALUES OBTAINED FOR EACH SEGMENTATION ALGORITHM FOR ALL THE TEST VIDEO SEQUENCES IN GENERIC, COMPRESSION AND MIXED REALITY FRAMEWORKS.

	'Group'			'Hall monitor'			'Highway'			$\overline{MAV}$		
Alg.	Gen.	Cmpr.	Mix.	Gen.	Cmpr.	Mix.	Gen.	Cmpr.	Mix.	Gen.	Cmpr.	Mix.
reference	8.77	11.20	12.54	26.74	15.51	12.60	15.31	14.45	8.03	16.94	12.5	13.20
Jabri	57.46	22.63	51.71	40.37	10.60	44.54	37.94	49.82	39.00	42.25	19.41	48.69
Horprasert	69.94	48.63	72.80	57.57	20.17	57.14	32.06	42.62	34.77	53.19	29.8	57.52
Shen	57.83	33.94	57.70	55.26	60.71	62.22	54.26	53.57	33.60	55.78	50.26	57.79
François	68.57	39.57	76.08	61.43	66.71	77.42	30.20	45.28	40.66	53.40	46.58	66.26
Kim	72.00	40.00	73.00	86.89	52.00	82.54	71.14	45.51	84.51	76.67	45.8	80.01
McKenna	83.36	76.43	84.77	56.86	71.37	74.65	54.26	71.57	68.62	68.82	73.12	76.01
Image D.	99.74	90.00	95.08	60.00	48.40	57.68	67.54	75.34	79.42	75.76	71.24	77.40

A. 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. 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. The setting up of a subjective experiment differs for each application. Therefore, our experiments were focused on two kinds of applications for segmented objects: video compression and mixed reality.

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 (d). 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. IV-B). In fact, the parts of the object that are undetected will be compressed as erroneously considered parts of the background.

Mixed Reality. The goal of video manipulation is to put together video objects from different sources in order to create new video content. In particular, in the *mixed reality* application considered here, video segmentation serves to extract real objects that are then inserted in a virtual background. One of the possible application is to create narrative spaces and interactive games and stories [22]. In order to evaluate different segmentation results in mixed reality scenario, we created a virtual background for each original sequence: we extracted the contour of the background image to recall a virtual background in black and white as in comics scenarios. For the test sequence 'Group' we applied a virtual background created in the context of the European Project art.live [22] processed the same way to extract only the

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

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

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

contours. Figure 1 (e) shows a sample frame for 'Highway'.

B. Subjective Results

Standard methods [23] 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 II shows the subjective ranking during the *interview stage* of the subjective experiment 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 III 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. III). 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. VI for specific scenario analysis).

V. 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. IV).

A. Spatial Artifacts

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:

$$\mathbf{S}_{A_r}(k) = \frac{\sum_{j=1}^{N_{Ar}} |\boldsymbol{A}_r^j(k)|}{|n(k)|}, \tag{10}$$

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:

$$\mathbf{S}_{H_i}(k) = \frac{\sum_{j=1}^{N_{Hi}} |\boldsymbol{H}_i^j(k)|}{|n(k)|}, \tag{11}$$

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 [9] by adding this weighting factor, D^j:

$$D^j = 1 + \frac{\overline{d^j} + \sigma_d^j}{d_{max}^j}, \tag{12}$$

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. (12) and Eq. (10), we obtain, for the border artifacts, the corrected relative spatial error $\mathbf{S}_{A_b}(k)$, for j added backgrounds:

$$\mathbf{S}_{A_b}(k) = \frac{\sum_{j=1}^{N_{Ab}} D_{\mathbf{A}b}^j \cdot |\boldsymbol{A}_b^j(k)|}{|n(k)|} \tag{13}$$

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

$$\mathbf{S}_{H_b}(k) = \frac{\sum_{j=1}^{N_{Hb}} D_{\mathbf{H}b}^j \cdot |\boldsymbol{H}_b^j(k)|}{|n(k)|} \tag{14}$$

B. Temporal Artifacts

The most subjectively disturbing effect is the temporal incoherence of an estimated sequences of object masks. In video segmentation, an artifact often varies its characteristics through time. A non smooth change of any spatial error deteriorates the perceived quality. 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:

$$\mathbf{F}_\Lambda(k) = \frac{|\Lambda(k)| - |\Lambda(k-1)|}{|\Lambda(k)| + |\Lambda(k-1)|}, \tag{15}$$

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 $k-1$. In this equation if the error disappears/appears suddenly it is evenly penalized by the normalization since it causes in the human observer an annoyance due to the unexpected change in the segmentation quality. By doing so, also the *surprise* effect [24] can be taken into account into the metric. This effect is meant to amplify the changes

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

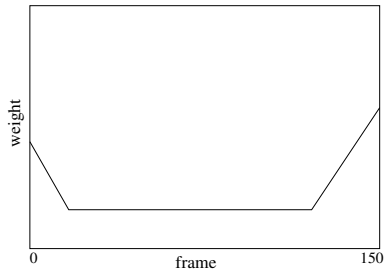

Fig. 3. Weighted function considering human memory in video quality evaluation proposed in [26].

in the spatial accuracy. To model this effect, Eq. (15) is combined to the relative spatial artifact measures to construct an objective spatio-temporal error measure $\mathbf{ST}(k)$ for each artifact. This takes into account not only the quality but also the stability of the results:

$$\mathbf{ST}_\Lambda(k) \quad = \quad \mathbf{S}_\Lambda(k) \cdot \frac{1 + \mathbf{F}_\Lambda(k)}{2}, \tag{16}$$

In modeling the relation between instantaneous and overall quality [25], we can identify two other phenomena related to the temporal context, namely the *fatigue* effect and the *expectation* effect. The fatigue effect is related to the fact that after a while the human gets used to a certain visual quality thus judging it more acceptable if it persists long enough. In subjective experiments on coded video sequences [26] the characteristics of *short-term* human memory have been studied; Fig. 3 shows the characteristics of the weighting functions for the short-term characteristics of human memory. The first gradient is called the beginning effect of human memory (it lasts around 50 frames) and presents higher values at the first frames. With our subjective experiments, we aim at finding the weighting function for 60 frame long video sequences. In fact, our test video were only 5 seconds time long (60 frames) and thus not long enough to cause fatigue effect in the human observers. On the other hand, since they were short video we experienced a different phenomenon: *expectation* effect. By expectation we mean that a good segmentation at the beginning could create a good overall impression on assessing the overall quality of the sequence under test and vice-versa. To model this effect, the overall objective spatio temporal metric, $\mathbf{ST}$ is formulated as follows:

$$\mathbf{ST}_\Lambda \quad = \quad \frac{1}{K} \sum_{k=1}^{K} w_t(k) \mathbf{ST}_\Lambda(k), \tag{17}$$

where the temporal weights $w_t(k)$ that model the *human memory effect* have been empirically defined [21] as:

$$w_t(k) = (a \cdot e^{\frac{k-30}{b}} + c) \tag{18}$$

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

C. Perceptual Objective Metric

In [21] a detailed description of the *synthetic* artifacts used to study and characterize the perception of the spatial and temporal artifacts previously described can be found. 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 [21] 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*: $\mathbf{PST}_\Lambda$. The best fitting function for each artifact was the Weibull function, W. Thus the perceptual artifact metrics are described by:

$$\begin{aligned} W(x,S,k) &= 1 - e^{-(Sx)^k} \;\; where \;\; x = \mathbf{ST}_\Lambda \\ \mathbf{PST}_\Lambda &= W(\mathbf{ST}_\Lambda, S, k) \end{aligned} \tag{19}$$

where the parameters S and k have been obtained in [21] 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 $\mathbf{PST}_{H_b}$. In the following the details of the subjective experiments where the parameters S and k have been obtained are summarized.

The annoyance of the **added region** artifacts was studied by varying its amount on a total of 75 sequences. Moreover, different positions and shapes of added region artifacts were tested to check if they are perceived the same way. To test this hypothesis we used the statistical F test. In this experiment, 28 naive subjects were asked to perform the annoyance task. The subjective experiments showed that the added region annoyance perception is not influenced by the shape or position of the artifact but only by its size (see rows 1 and 2 of Tab. IV). In the **holes** experiment there were two goals. The first goal was to test the two objective metrics, one proposed for inside holes (see Eq. 11) and the second for border holes (see Eq. 14) The second goal was to determine the psychometric annoyance functions for the two kinds of synthetic artifacts. Finally, we studied whether the annoyance caused by a boundary hole could be worse than for an inside hole (for large holes). In this experiment 28 naive subjects were asked to perform the annoyance task on 48 test sequences. This subjective experiment indicated that both

TABLE IV

F VALUES TO TEST IF DIFFERENT FITTING CURVES ARE NEEDED TO DESCRIBE THE PERCEIVED ANNOYANCE FOR DIFFERENT SHAPES AND POSITIONS OF ADDED REGIONS AND HOLES.

Artifact model	$\mathbf{F_c}$ (critical)	$\mathbf{F}$ (value)	$\mathbf{p(F < F_c)}$
added region shape	F(2,68)=3.13	1.43	0.24
added region position	F(4,66)=2.51	0.64	0.63
inside hole position	F(2,28)=3.34	0.13	0.87
hole distinction	F(2,44)=3.21	5.01	0.01

the kind and the size of the hole should be jointly taken into account and not only the distance when an objective metric is proposed. Besides, in the objective metrics proposed in the literature, holes are only considered in terms of uncorrelated set of pixels and of their distances from the reference boundary of the object [7], [12]. With this experiment it was proved that a cluster of error pixels should be distinguished and their characteristics should be thoroughly studied instead of considering each error pixel individually. In other words, in the literature, methods reported in [12], [13], [7] claim that as we move away from the border, holes become more annoying but we proved that this depends on also the kind and the size of the hole, as show by experiments [21]. Furthermore, two positions of inside holes have been also tested: one further than the other to the object borders. Hence, the F-test has been used to investigate whether the perceived annoyance of these two positions could be described with two different fitting curves. As reported in row 3 of Tab. IV the F value shows that the same curve can be used to fit both positions for inside holes. This validates the simple characterization that made about inside holes without considering the distance of the inside hole from the border of the ground truth (see Eq. (14) and (11)). To further confirm the hypothesis that a distinction between inside holes and border holes has to be made applied the F-test on these two sets of data to see if a unique fitting curve can interpolate both kinds of artifacts (see row 4 of Tab. IV). The F-value in this case is equal to 5.01 that is above the threshold of $F(2, 44)$ equal to 3.21. This means that an overall fitting curve is not sufficient to describe both phenomena so two metrics, $\mathbf{PST}_{H_i}$ and $\mathbf{PST}_{H_b}$, were proposed. These results showed that inside hole for small sizes are more annoying than holes on the border, but on the other hand by increasing their size border holes become more annoying than inside holes as the shape of the object becomes less recognizable.

The performance of the proposed objective metric for **added background** (see Eq. 13) was tested on 5 dilated masks plus 16 test sequences with different amount of added background concentrated in some

parts of the object boundaries. For the video *Hall monitor*, five new segmented video sequences were created by varying the number of dilations of correctly segmented video sequences from one dilation to eight dilations. Subjects in this experiment were 8 male students, aged between 23-28. For the second test, with big amounts of added background concentrated in only some parts, 31 subjects judged the 16 test sequences. The experimental results showed that the added background measure of Eq. (13) matches the human annoyance perception both when the artifact is uniformly distributed along the object boundaries and when it is concentrated in some parts of the object boundaries. Finally, the overall perceptual metric is given by the combination of all the four kinds of artifacts. A simple linear combination of artifacts [21] estimates the total annoyance:

$$\mathbf{PST} = a \cdot \mathbf{PST}_{A_r} + b \cdot \mathbf{PST}_{A_b} + c \cdot \mathbf{PST}_{H_i} + d \cdot \mathbf{PST}_{H_b} \tag{20}$$

The perceptual weights were found by means of subjective experiments [21] on combined synthetic artifacts: $a = 2.86$, $b = 4.50$, $c = 4.77$, $d = 5.82$.

VI. 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 state of the art metrics. Second, the parameters of the novel metric are optimized according to specific applications. Moreover, 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 for the segmentation algorithms presented in Sec. III have been plotted versus the subjective annoyance values for the three frameworks. The Pearson and Spearman correlation coefficients are reported in Tab. V. The correlation coefficients for the perceptual metric, **PST** are larger (Pearson= 0.86, Spearman=0.79) compared to the state of the art metrics ($MPEG$ metric, matching quality metric mqm, and weighted quality metric wqm) for all 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 state of the art metrics which do not include perceptual factors. However, MPEG metric outperforms wqm and mqm metrics in mixed reality scenario and no state of the art metric performs well in the case of compression scenario.

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. (20) 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 by maximizing the correlation coefficients (Pearson and Spearman) in the specific scenarios.

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 the mixed reality scenario, the weights obtained for added background ($b = 8.31$), inside ($c = 12.57$) and border holes ($d = 8.74$) are larger than those for added background (6.71). In fact, every artifact that changes the shape or allows to see the virtual background beneath the real objects causes a lot of annoyance in the subjects who are focusing their attention on the virtual story or the interactive game. For both applications, the difference in perception of the four artifacts has been so 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. VI. If the performance of the segmentation algorithms are considered in the general case, the best one in both subjective (Tab. III) and objective (Tab. VI) 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. In the mixed reality case, overall **Jabri** was still the best performing segmentation algorithm. This is due to the fact that almost no shape changes are caused by this segmentation. In fact, only few added regions are present and they do not bother the human viewers since they pay attention to the moving objects. **Francois** and **Image Differencing** shows again the worst case since it produces a lot of inside and border holes (see Fig. 2) that allow to see the virtual background beneath the objects.

TABLE V

CORRELATION COEFFICIENTS BETWEEN THE OBJECTIVE METRICS AND SUBJECTIVE RESULTS (MAV VALUES) FOR ALL THE TEST VIDEO SEQUENCES IN GENERIC AND SPECIFIC APPLICATION FRAMEWORKS. PST METRIC PARAMETERS: $a = 2.86$, $b = 4.50$, $c = 4.77$, $d = 5.82$

	'Generic'		'Compression'		'Mixed Reality'	
Metric	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
wqm	0.69	0.71	0.37	0.32	0.74	0.65
mqm	0.53	0.44	0.50	0.47	0.67	0.55
MPEG	0.73	0.67	0.49	0.41	0.78	0.68
PST	**0.86**	**0.79**	**0.78**	**0.79**	**0.94**	**0.91**

VII. CONCLUSIONS

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

TABLE VI

OBJECTIVE METRIC VALUES OBTAINED FOR EACH SEGMENTATION ALGORITHM FOR ALL THE TEST VIDEO SEQUENCES IN GENERIC, COMPRESSION AND MIXED REALITY FRAMEWORKS.

	'Group'			'Hall monitor'			'Highway'			$\overline{PST}$		
Alg.	Gen.	Cmpr.	Mix.	Gen.	Cmpr.	Mix.	Gen.	Cmpr.	Mix.	Gen.	Cmpr.	Mix.
reference	2.96	6.84	2.95	2.96	6.84	2.95	2.96	6.84	2.95	2.96	6.84	2.95
Jabri	21.59	14.86	21.59	26.31	31.36	26.31	23.84	16.24	23.84	29.31	20.82	23.91
Horprasert	31.76	43.64	31.76	31.35	40.23	31.35	20.48	20.59	20.48	27.36	34.82	27.86
Shen	28.78	40.98	35.82	35.89	63.76	35.91	24.81	37.83	24.81	29.82	47.52	32.18
François	40.84	46.86	40.84	43.87	74.36	43.87	29.19	35.80	29.19	37.96	53.00	37.96
Kim	28.98	41.67	28.98	43.42	54.13	43.42	35.13	44.44	35.13	35.84	46.76	35.84
McKenna	42.73	69.18	42.73	56.86	68.26	39.41	31.12	54.66	31.12	43.57	64.03	37.75
Image D.	46.64	92.33	46.64	60.00	62.84	28.78	36.76	50.40	36.76	47.8	68.52	37.39

performance of such a metric compared against the usually adopted MPEG and the wqm, mqm metrics and its adaptability to take into consideration different segmentation applications. The optimal perceptual parameters have been found for specific segmentation applications: video compression and mixed reality.

ACKNOWLEDGMENT

The authors would like to thank Carli and Arrigoni for running some experiments, Karaman for generating some of the segmentation masks, Marichal and Nascimento for their test code.

REFERENCES

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

[2] 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*, vol. 19, pp. 1338–1359, 1997.

[3] 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*, vol. 2, 2001, pp. 416–425.

[4] "Compare your segmentation algorithm to the cost 211 qam." [Online]. Available: http://www.iva.cs.tut.fi/COST211/Call/Call.htm

[5] P. Correia and F. Pereira, "Objective evaluation of video segmentation quality," *IEEE Transaction on Image Processing*, vol. 12, pp. 186–200, 2003.

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

[7] P. Villegas and X. Marichal, "Perceptually-weighted evaluation criteria for segmentation masks in video sequences," *IEEE Transactions on Image Processing*, vol. 13, no. 8, pp. 1092–1103, August 2004.

[8] 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.

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

[10] C.Erdem and B.Sankur, "Performance evaluation metrics for object-based video segmentation," in *Proc. X European Signal Processing Conference, Tampere, Finland*, vol. 2, 2000, pp. 917–920.

[11] Y. J. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognition*, vol. 29, pp. 1335–1346, 1996.

[12] A. Cavallaro, E. D. Gelasca, and T. Ebrahimi, "Objective evaluation of segmentation quality using spatio-temporal context," in *Proc. IEEE International Conference on Image Processing, Rochester(NY),22-25 September 2002*, 2002, pp. 301–304.

[13] X. Marichal and P. Villegas, "Objective evaluation of segmentation masks in video sequences," in *Proc. Of X European Signal Processing Conference, Tampere, Finland*, 2000, pp. 2139–2196.

[14] J. Nascimento and J. S. Marques, ""new performance evaluation metrics for object detection algorithms"," in *6th International Workshop on Performance Evaluation for tracking and Surveillance (PETS, ECCV), Prague, May 2004*, p. ...

[15] 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, no. 2, February 2002.

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

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

[18] J. Shen, "Motion detection in color image sequence and shadow elimination," January 2004, pp. 731–740.

[19] 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," pp. 627–630, September 2000.

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

[21] E. Drelie, "Full-reference objective quality metrics for video watermarking, video segmentation and 3d model watermarking," Ph.D. dissertation, EPFL, Lausanne, 2005.

[22] X. Marichal, B. Macq, D. Douxchamps, T. Umeda, and art.live consortium, "The ART.LIVE architecture for mixed reality," in *Proc. of Virtual Reality International Conference (VRIC)*, Laval, France, June 2002, pp. 19–21.

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

[24] J. W. Senders, "Distribution of visual attention in static and dynamic displays," vol. 3016, February 1997, pp. 186–194.

[25] R. Hamberg and H. de Ridder, "Time-varying image quality: Modeling the relation between instantaneous and overall quality," *SMPTE Journal*, vol. 108, pp. 802–811, 1999.

[26] Y. Inazumi, Y. Horita, K. Kotani, and T. Murai, "Quality evaluation method considering time transition of coded quality," in *ICIP*, vol. 4, 24-28 October 1999, pp. 338–342.