OBJECTIVE EVALUATION OF SEGMENTATION QUALITY USING SPATIO-TEMPORAL CONTEXT

Andrea Cavallaro, Elisa Drelie Gelasca, and Touradj Ebrahimi

Swiss Federal Institute of Technology, CH-1015 Lausanne, Switzerland

ABSTRACT

In this paper, we propose an automatic method for the objective evaluation of segmentation results. The method is based on computing the deviation of the segmentation results from a reference segmentation. The discrepancy between two results is weighted based on spatial and temporal contextual information, by taking into account the way humans perceive visual information. The metric is useful for applications where the final judge of the quality is a human observer or the results of segmentation are otherwise processed in a human-like fashion. The proposed evaluation has been applied both to automatically provide a ranking among different segmentation algorithms and to optimally set the parameters of a given algorithm.

1. INTRODUCTION

Segmentation represents the preliminary step of a wide variety of applications, ranging from video coding to video surveillance, and from virtual reality to video editing. Although segmentation techniques have been widely studied over the past decades, very little has been proposed by way of methodology for assessing segmentation results. This is a consequence of the difficulty in universally defining a good segmentation.

Segmentation results depend on the task at hand and on the visual content. No single segmentation technique is useful for all applications. Moreover different techniques are not equally suited for a particular application. An effective evaluation of segmentation is important in selecting the most appropriate technique for a specific application, and furthermore for optimally setting its parameters. However, except for well constrained situations, the design of an evaluation metric is a difficult task, and there is no standard method for objective evaluation of segmentation quality.

Common practices for evaluating segmentation results are based on human intuition or judgment (*subjective evaluation*) and consist in *ad hoc* subjective assessment by a representative group of observers. A significant number of observers is required to produce statistically relevant results, thus making subjective evaluation a time-consuming and expensive process.

To avoid systematic subjective evaluation, an automatic procedure is desired. This procedure is referred to as *objective evaluation*. Quality metrics for objective evaluation of segmentation may judge either segmentation algorithms or segmentation results. The metrics are referred to as analytical methods or empirical methods, respectively. *Analytical methods* evaluate segmentation algorithms by considering their principles, their requirements and their complexity [1]. The advantage of these methods is that an evaluation is obtained without implementing the algorithms. However, because of the lack of a general theory for image segmentation, and because segmentation algorithms may be complex systems composed of several components, not all properties (and therefore strengths) of segmentation algorithms may be easily evaluated. *Empirical methods*, on the other hand, do not evaluate segmentation algorithms directly, but indirectly through their results. To choose a segmentation algorithm based on empirical evaluation, several algorithms are applied on a set of test data that are relevant to a given application. The algorithm producing the best results is then selected for use in that application.

The paper is organized as follows. Empirical methods are discussed in Section 2. Section 3 introduces objective and perceptual elements contributing to the quality of a segmentation result. Evaluation results are presented in Section 4. Finally, in Section 5 we draw the conclusions and we comment on future directions.

2. STATE OF THE ART

Empirical methods evaluate a segmentation algorithm based on the quality of its results. Empirical methods are referred to as *discrepancy methods* when the segmentation result is compared to a reference segmentation. This reference segmentation represents the *ground truth*, or the ideal segmentation, and can be generated either manually or via a reliable procedure. The result of the evaluation is a disparity between the reference segmentation and the actual segmentation result. Discrepancy methods are based on directly measuring the deviation between two partitions. The deviation is evaluated through *discrepancy parameters*, which characterize each method. Discrepancy parameters are based on the spatial and temporal deviations. These deviations may be appropriately weighted to take visually desirable properties of a segmentation mask into account. For example, pixel errors are separated into two groups, those that belong to the result but not to the reference (false positive) and those that belong to the reference but not to the result (false negative) [2]. Furthermore the temporal stability of the segmentation mask shape may be considered. The discrepancy parameters may also be formulated as misclassification penalties regarding shape and motion errors [4]. The discrepancy parameter spatial accuracy is defined in [5] by shape fidelity, geometrical similarity, edge content similarity, and statistical data similarity. Discrepancy parameters are combined over the time interval, to assess the quality of the entire spatio-temporal segmentation. The parameters are weighted so as to combine them in correct proportions and to match evaluation results produced by human viewers.

In applications where the final judge of quality is a human being, it is also important to consider the human visual system to design a quality evaluation procedure, in addition to the objective discrepancy parameters. Traditional evaluation methods do not consider this aspect and just consider objective criteria, such as discrepancy between two results. One distinctive feature of the method in [5] is the evaluation of object relevance for judging the quality of segmentation. The overall segmentation quality depends on the estimated importance of segmented objects in the scene.

3. PROPOSED METHOD

The discrepancy evaluation is defined based on two kinds of errors, namely objective errors and perceptual errors. *Objective errors* quantify the deviation of the results under test from the ground truth. *Perceptual errors* weight these deviations according to human perception and possible human visual attractors.

3.1. Objective errors

A direct comparison of the results of the change detector under test with the reference segmentation allows us to identify two types of errors: false positive pixels, and false negative pixels. A *false positive* is a pixel erroneously detected as belonging to the semantic partition. Let us denote by $C(n)$ the set of pixels detected as changed at frame n, and with $C_r(n)$ the pixels belonging to the reference segmentation. The ensemble of false positive errors, $\epsilon_p(n)$, can be expressed as

$$\epsilon_p(n) = card\big(C(n) \cap \bar{C}_r(n)\big) \tag{1}$$

where the function $card(\cdot)$ represents the cardinality of a set, and $\bar{C}_r(n)$ is the complement of $C_r(n)$, that is, $\bar{C}_r(n) = I \setminus C_r(n)$, where $\setminus$ is the set difference operator. A *false negative* is a pixel erroneously detected as not belonging to the semantic partition. False negatives, $\epsilon_n(n)$, are pixels appearing in the reference segmentation $C_r(n)$, but not in the result under analysis, $C(n)$. They can be expressed as

$$\epsilon_n(n) = card\big(\bar{C}(n) \cap C_r(n)\big) \tag{2}$$

Using Eq. 1 and Eq. 2, we can derive a measure of the *absolute spatial accuracy* at frame n

$$\epsilon(n) = \epsilon_p(n) + \epsilon_n(n) \tag{3}$$

corresponding to the amount of false detections for each time instant n. The larger ϵ, the lower the spatial accuracy.

Using the definitions of false positive and false negative pixels we can now introduce some figures of merit which we will use to measure the accuracy of the change detection results. The significance of the same value of $\epsilon(n)$ depends on the size of the image at hand, and on the amount of change detected, that is, $card\big(C_r(n)\big)$. The larger the image size, the less important is $\epsilon(n)$. Similarly, the larger $card\big(C_r(n)\big)$, the less important is $\epsilon(n)$. For this reason, we introduce a relative measure of the total amount of false detection, referred to here as *relative spatial accuracy*. The relative spatial accuracy can be computed by normalizing the total amount of false detection by the dimension of the frame $N = card(I)$, where I is the set of all pixels in the image, and by the number of elements in the reference mask, $card\big(C_r(n)\big)$:

$$\epsilon'(n) = \begin{cases} \frac{1}{N}\epsilon(n) & \text{if } card\big(C_r(n)\big) = 0, \\ \frac{1}{N \times card\left(C_r(n)\right)}\epsilon(n) & \text{otherwise.} \end{cases} \tag{4}$$

The value of $\epsilon'(n)$ is proportional to the amount of errors. The quality of the results is *inversely* proportional to the amount of deviation between the two results at hand. We define the measure of spatial accuracy, $\nu(n)$, as follows:

$$\nu(n) = 1 - \epsilon'(n), \tag{5}$$

$\nu(n) \in [0, 1]$. The value $\nu(n) = 1$ indicates perfect spatial accuracy in frame n, that is, a perfect match between segmentation results and reference segmentation.

3.2. Perceptual errors

The measure of spatial accuracy proposed in Eq.(5) is an objective discrepancy parameter that quantifies the deviation of the segmentation result at hand without taking human perception into account. Considering human perception is fundamental since the different kinds of errors are not visually significant to the same degree. To accommodate human perception, the different error contributions are weighted according to their *visual relevance*. Visual relevance is defined based on spatial and temporal contextual information.

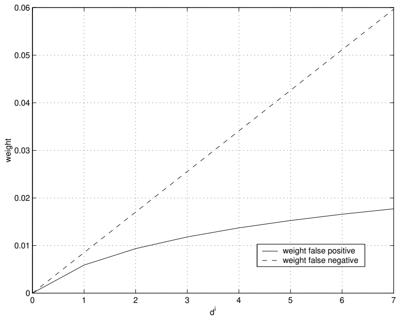

Fig. 1. Spatial context: weights for false positive and false negative pixels

3.2.1. Spatial context

The spatial context of an error pixel is characterized by the distance from the nearest correctly detected object. These properties are related to the focus of attention. An observer looks at points that attract the attention. In addition to this, a false positive contribute differently to the quality than a false negative. False negatives are more significant, and the larger the distance, the more significant the error. The weights w_p and and w_n increase with distance (differently for false positive and false negative), and are normalized by the maximum distance in the frame, that is its diagonal, D. Let d_p^i be the distance of the i^{th} false positive pixel from the contour of the reference mask, and d_n^j the distance of the j^{th} false negative. Then the weighting factor for the false positive pixel, w_p^i, is defined as

$$w_p^i = \frac{\alpha_p \log(d_p^i + 1)}{D}, \tag{6}$$

and that for the false negative pixel, w_n^j, as

$$w_n^j = \frac{\alpha_n d_n^j}{D} \tag{7}$$

The weights for false negative pixels increase linearly and they are larger than those for false positive pixels at the same distance from the border of the object. As we move away from the border, in fact, missing parts of objects are more important than added background. The weights for false positive and false negative pixels are depicted in Figure 1. By considering the spatial context, the measure of the spatial accuracy, $\epsilon_w(n)$, becomes

$$\epsilon_w(n) = \sum_{i=1}^{\epsilon_p(n)} w_p^i + \sum_{j=1}^{\epsilon_n(n)} w_n^j. \tag{8}$$

The weighted spatial accuracy can be derived using Eq.(5), as

$$\nu_w(n) = 1 - \epsilon_w(n). \tag{9}$$

3.2.2. Temporal context

Observers get used to a certain quality and are sensitive to differences in quality over time. A given error may be perceived differently, depending on its *temporal context*. We can identify two phenomena related to the temporal context, namely the surprise effect and the fatigue effect. The *surprise effect* amplifies the changes in the objective spatial accuracy. The *fatigue effect* is related to the fact that we get used to a certain quality thus judging it acceptable if it persists long enough.

To take these phenomena into account in the discrepancy metric, we introduce weights inversely proportional to the temporal duration of the appearance of an error. We combine Eq.(5) and the variation of spatial accuracy in time, $|\frac{d}{dn}\nu(n)|$,[1] to construct a *perceptual spatio-temporal quality measure*

$$\zeta(n) = \frac{\nu(n)}{2}\left[1 + \alpha_t \frac{d}{dn}\nu(n)\right], \tag{10}$$

where $\alpha_t \in [0, 1]$. This takes into account not only the quality but also the stability of the results, by modeling surprise and fatigue effects. Incorporating the spatial context into Eq.(10), we obtain

$$\zeta_w(n) = \frac{\nu_w(n)}{2}\left[1 + \alpha_t \frac{d}{dn}\nu_w(n)\right]. \tag{11}$$

that represents the perceptual quality measure for evaluating the performance of segmentation algorithms.

4. RESULTS

In this section the proposed objective evaluation metric is used to compare and rank different segmentation results. The values of parameters for the evaluation metric have been set according to the qualitative criteria discussed in Section 3.2 and by comparison with informal experiments with human observers: $\alpha_p = 3$ in Eq.(6), $\alpha_n = 2,5$ in Eq.(7), and $\alpha_t = 0.5$ in Eq.(10). The objective evaluation is given with respect to the hand-segmented sequence *Hall Monitor*, provided by the European project COST 211.[2] The quality metric is used to (a) select the best parameter values for a give segmentation algorithm, and (b) rank a group of segmentation algorithms. In particular, we consider here segmentation results obtained by change detection algorithms. First we analyze the influence of the window size on the change detection algorithm described in [7]. Then, we compare change detection results of the methods proposed in [8] (SM), [9] (EDGE) and [10] (CECD). The objective evaluation between different window sizes for the sequence *Hall Monitor* is shown in Figure 2. The dimensions of the win-

[1]The larger this variation, the smaller the temporal coherence.

[2]http://www.tele.ucl.ac.be/EXCHANGE/

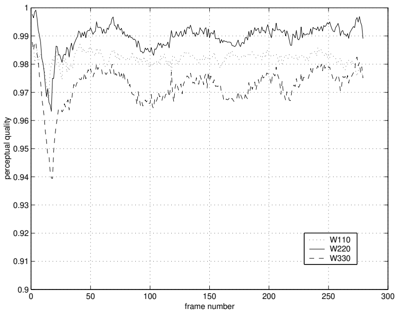

Fig. 2. Objective comparison for perceptual spatio-temporal quality metrics between different window sizes in [7] for the sequence *Hall Monitor*. The symbols of the legend refer to the dimensions of the window. W110: 9 pixels, W220: 25 pixels, W330: 49 pixels.

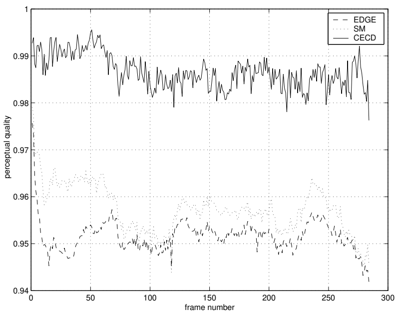

Fig. 3. Objective comparison for perceptual spatio-temporal quality metrics among change detection techniques for the sequence *Hall Monitor*. The symbols of the legend refer to the change detection technique: SM [8], EDGE [9], CECD [10].

dows under test are 9, 25, and 49 pixels. Among the different window sizes under analysis, the choice corresponding to a 5×5 window gives the best results. The objective evaluation of the illumination invariant methods is shown in Figure 3. The three methods under test are compared first in case of illumination variation. The method based on color edges (CECD) provides better results than the other two and can be therefore be selected for use in the change detection application at hand.

5. CONCLUSIONS

An objective metric evaluation of segmentation quality has been presented in this paper. The metric judges the spatial accuracy and temporal coherence of a partition based on direct comparison of results. The deviation of the results from a reference segmentation is weighted according to the visual importance. The visual importance has been determined by considering spatial and temporal contextual information. The metric has been used to compare change detection results obtained with the state-of-the-art methods. The weights of the quality metric have been set according to qualitative criteria and by comparison with informal experiments with human observers. Our future research direction is to automatically determine the most appropriate weights, at least for specific applications. To this end, it would be interesting to define a general formalism for the set up so as to compare subjective and objective results.

6. REFERENCES

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

[2] X. Marichal and P.Villegas, "Objective evaluation of segmentation masks in video sequences," in *Proc. of X European Signal Processing Conference*, Tampere, Finland, pp. 2193–2196, 2000.

[3] C. Erdem, A. M. Tekalp, and B. Sankur, "Metrics for performance evaluation of video object segmentation and tracking without ground-truth," in *Proc. Int. Conference on Image Processing*, Thessaloniki, Greece, 2001.

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

[5] P. Correia and F. Pereira, "Objective evaluation of relative segmentation quality," in *Proc. Int. Conference on Image Processing*, vol. 2, Vancouver, Canada, pp. 308–311, September 2000.

[6] L. Shapiro and R. Haralick, *Computer and Robot Vision*. Reading: Addison-Wesley, 1992.

[7] A. Cavallaro and T. Ebrahimi, "Video object extraction based on adaptive background and statistical change detection," in *Proceedings of SPIE Electronic Imaging - Visual Communications and Image Processing*, San Jose, California, USA, pp. 465–475, 2001.

[8] K. Skifstad and R. Jain, "Illumination independent change detection for real world image sequences," *Computer Vision, Graphics, and Image Processing*, vol. 46, pp. 387–399, june 1989.

[9] D. Aubert, "Passengers queue measurement," in *Proc. of 10th International Conference on Image Analysis and Processing*, Venice, Italy, pp. 1132–1135, 1999.

[10] A. Cavallaro and T. Ebrahimi, "Change detection based on color edges," in *Proceedings of IEEE International Symposium on Circuits and Systems* , Sydney, Australia, 2001.