5th COST 276 Workshop (2003), pp. 9–14
B. Kovář, J. Přikryl, and M. Vlček (Editors)

Still Image Objective Segmentation Evaluation using Ground Truth

V. Mezaris,[1,2] I. Kompatsiaris[2] and M. G. Strintzis[1,2]

[1] Information Processing Laboratory, Electrical and Computer Engineering Department,
Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
[2] Informatics and Telematics Institute, 1st Km Thermi-Panorama Rd, Thessaloniki 57001, Greece

Abstract

In this paper, an objective segmentation evaluation metric suitable for the evaluation of still image segmentation results is proposed. The proposed metric is based on the spatial accuracy approach, originally proposed for the evaluation of foreground/backgroung segmentation masks generated from video sequences. This approach is extended to still image segmentation evaluation, where both the estimated segmentation masks and the ground truth mask typically contain multiple regions. The proposed method takes into account, using a single metric, not only the accuracy of the boundary localization of the created segments but also the under-segmentation and over-segmentation effects, which can hinder the performance of any segmentation algorithm and decrease the usability of the segmentation results in content-based applications. Several experiments have shown the potential of this approach.

Keywords : *segmentation evaluation; objective evaluation; still image segmentation; spatial accuracy; ground truth.*

1. Introduction

In recent years, the proliferation of digital media has led to the development and deployment of a plethora of multimedia applications supporting the efficient processing, coding, indexing and retrieval of multimedia information and visual information (still images, video) in particular. Although a wide range of radically different approaches on these issues have been reported in the literature, a common characteristic of many recent approaches is the employment of various segmentation tools for enabling a fine-granularity manipulation of visual information [1, 2, 3, 4]. The comparison and selection of suitable segmentation tools becomes therefore an issue of paramount importance to the designer of such fine-granularity content-based applications. This paper proposes a method for the objective evaluation of still image segmentation results, which can be used for the direct comparison of the results of different segmentation algorithms.

Previous work on objective segmentation evaluation includes both *standalone evaluation* methods, which do not make use of a reference segmentation, and *relative evaluation* methods employing ground truth. In [5], methods belonging to both categories are developed for video segmentation evaluation and in particular for individual video object evaluation.

Methods for standalone evaluation of image segmentations have been proposed, among others, in [6], where metrics for intra-object homogeneity and inter-object disparity are proposed. In [7], a combined segmentation and evaluation scheme is developed, to allow for recursive improvement of segmentation accuracy. Although standalone evaluation methods can be very useful in such applications, their results do not necessarily coincide with the human perception of the goodness of segmentation. For this reason, when a reference mask is available or can be generated, relative evaluation methods are preferred.

Recent relative evaluation methods for still image segmentations include [8, 9]. In [8], the use of binary edge masks and scalable discrepancy measures is proposed. In [9], the evaluation is also based on edge pixel discrepancy, but the establishment of a correspondence of regions between the reference mask and the examined one is proposed. In [10], where binary foreground/ background segmentation masks with one foreground object are considered, an area-based rather that edge-based approach is proposed, to estimate a weighted sum of misclassified pixels taking into consideration the visual relevance of segmentation errors.

In this paper, a relative evaluation method using an

area-based approach is proposed for the evaluation of still image segmentation results. The proposed method takes into account the accuracy of the region boundary localization as well as under-segmentation and over-segmentation effects, and is shown to be appropriate for comparing segmentation algorithms on the basis of their performance.

The remainder of this paper is organized as follows: in section 2, the problem of still image segmentation evaluation is formulated. The proposed metric for evaluation is presented in section 3. In section 4, experimental evaluation using natural and synthetic images is discussed, and finally, conclusions are drawn in section 5.

2. Problem Formulation

The problem of still image segmentation evaluation differs considerably from the binary foreground/ background segmentation evaluation problem examined in [10], in that the correctness of the two-class-boundary localization is not the only quantity to be measured. This derives from the presence of an arbitrary number of regions in both the reference mask and the mask to be evaluated. An evaluation metric is therefore desired to take into account the following errors:

- Over-segmentation. A region of the reference mask is represented by two or more regions in the examined segmentation mask.
- Under-segmentation. Two or more regions of the reference mask are represented by a single region in the examined segmentation mask.
- Inaccurate boundary localization, given a correspondence between one region of the reference mask and one of the examined segmentation mask.

The visual relevance of the above segmentation errors should be considered rather than simply their plurality; e.g. over-segmentation by two regions can be more or less important, depending on the properties of the two undesired regions.

3. Evaluation Metric

The proposed evaluation criterion is based on the measure of *spatial accuracy* proposed in [10] for foreground/background masks. In the case of still image segmentation, let $S = s_1,\ s_2, \ldots,\ s_K$ be the segmentation mask to be evaluated, comprising K regions $s_k,\ k = 1, \ldots,\ K$, and let $R = r_1,\ r_2, \ldots,\ r_Q$ be the reference mask, comprising Q reference regions r_q, $q = 1, \ldots,\ Q$. A region is simply defined as a set of pixels $\mathbf{p}$.

For the purpose of evaluating still image segmentation results in accordance with the requirements set in the previous section, a correspondence between the examined segmentation mask and the reference mask has to initially be established, indicating which created region better represents each reference region. This is performed by associating each region r_q of mask R with a different region s_k of mask S on the basis of region overlapping, i.e. s_k is chosen so that $r_q \cap s_k$ is maximized. Let $\mathcal{A} = \{(r_q,\ s_k)\}$ denote the set of region pairs identified using this procedure, and let $\mathcal{N}_R$, $\mathcal{N}_S$ denote the sets of non-associated regions of masks R and S, respectively.

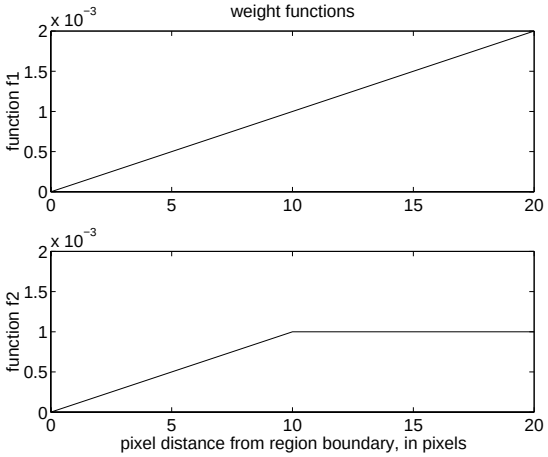

Figure 1: *Weight functions $f_1(.,.)$ and $f_2(.,.)$.*

For every region pair of set $\mathcal{A}$, the criterion of [10] is employed to evaluate the spatial accuracy of the segmentation, as will be explained in the sequel. During this process, the examined reference region r_q is treated as foreground, whereas all other reference regions are treated as background. A weighted sum of misclassified pixels for region pair $(r_q,\ s_k)$, E_q, is the output of this process, indicating the accuracy of region boundary localization:

$$E_q = \sum_{\mathbf{p} \in (r_q - r_q \cap s_k)} f_1(\mathbf{p}, r_q) + \sum_{\mathbf{p} \in (s_k - r_q \cap s_k)} f_2(\mathbf{p}, r_q) \quad \forall (r_q,\ s_k) \in \mathcal{A}$$

Functions $f_1(.,.)$, $f_2(.,.)$ are weight functions, introduced in [10] to deal with the fact that the distance of a misclassified pixel from the boundary of the reference region to which it belongs affects the visual relevance of the error. In particular, function $f_2(.,.)$ is used for false positives (i.e. pixels assigned to s_k but do not belong to r_q), whereas function $f_1(.,.)$ is used for false negatives (i.e. pixels belonging to r_q but not assigned to s_k). The weight functions used in this work are shown in figure 1.

For every reference region r_q not being part of set $\mathcal{A}$, i.e. every region of set $\mathcal{N}_R$, the weighted error E_q is similarly calculated as:

$$E_q = \sum_{\mathbf{p} \in r_q} f_1(\mathbf{p}, r_q), \quad \forall r_q \in \mathcal{N}_R$$

This measure quantifies the error due to under-segmentation, and as can be seen is also a weighted sum of misclassified pixels. Clearly, more visually significant regions (e.g. larger regions) that were missed

in the examined segmentation are assigned a significantly higher error E_q during this procedure than visually less significant ones.

In addition to the above measured errors due to inaccurate boundary localization and under-segmentation, over-segmentation has to be also taken into account. For this reason, a similar error is calculated for every region $s_k \in \mathcal{N}_S$:

$$F_k = \alpha \sum_{\mathbf{p} \in s_k} f_1(\mathbf{p}, r_q)$$

where $\mathbf{p} \in r_q$, i.e. the distance of pixel $\mathbf{p}$ from the boundary of the reference region to which it belongs is employed. Scaling factor α is used to allow the different weighting of this over-segmentation penalty depending on the potential use of the segmentation results and was heuristically set to two in our experiments.

The sum of these error measures, E, for all reference regions and all regions of set $\mathcal{N}_S$, is used for the objective evaluation of segmentation accuracy; values of the sum closer to zero indicate better segmentation.

$$E = \sum_{q=1}^{Q} E_q + \sum_{s_k \in \mathcal{N}_S} F_k$$

An illustrative example of how different segmentation deficiencies are captured by the proposed metric is shown in figure 2 and table 1, where a set of synthetic segmentations of a single synthetic image and corresponding values of the overall error E are presented, respectively.

4. Experimental results

Objective segmentation evaluation experiments were conducted using natural images of the Corel gallery [11] and synthetic images, created using the reference textures of the VisTex database [12]. Reference masks for the former were manually generated.

The segmentation algorithms employed in the evaluation experiments were the K-Means-with-Connectivity-Constraint (KMCC)-based algorithm proposed in [1] and two variants of it. The method of [1] performs segmentation in the combined intensity–texture–position feature space in order to produce connected regions that correspond to the real-life objects shown in the image. A simpler variant of it, that neither uses texture features nor enforces connectivity constraints during pixel classification was used for the purpose of demonstrating over-segmentation. Masks featuring under-segmentation were created using another variant of [1], which performed excessive merging of neighboring regions, i.e. regions were merged even if they exhibited relatively low color similarity.

Several natural and synthetic images used in the experiments are shown in figure 3, along with their reference segmentation masks and the different masks generated using the aforementioned segmentation algorithms. Visual inspection of these masks allows for their subjective comparison and evaluation. In general, the original algorithm of [1] outperforms its variants introduced in this work, as expected. Its variant favoring under-segmentation often misses significant objects (e.g. in the two tiger images), but generally produces acceptable segmentations that could be of use in content-based applications. The variant favoring over-segmentation, on the other hand, typically results to the creation of a plethora of very small, meaningless regions, as a result of the limited use of position features and the no use of texture features for pixel classification. In only a limited number of cases it can be argued that this version performs better (e.g. the first tiger image) or comparably (e.g. the sunset image) to the variant favoring under-segmentation.

Corresponding objective evaluation results for these masks, using the proposed metric, are reported in table 2. On comparing the results of subjective and objective evaluation, it is made evident that objective evaluation results correlate well with the outcome of subjective evaluation. Thus, the proposed evaluation method can be used for facilitating and accelerating the evaluation process by substituting the human evaluator. It also makes possible the accurate and reliable comparison of similar segmentations, which are likely to be too similar for a human evaluator to reach any definite conclusions.

5. Conclusions

A methodology was presented for the objective evaluation of still image segmentations. The proposed metric is based on examining the spatial accuracy of segmentation results using a pre-existing or manually generated reference mask. Its output is a weighted sum of misclassified pixels, effectively indicating how well the examined segmentation mask corresponds to the reference one. The proposed metric was shown to effectively capture deficiencies such as inaccurate boundary localization, over-segmentation and under-segmentation, and its output was shown to correlate well with the outcome of subjective evaluation of segmentation masks by a human observer. The proposed approach is appropriate for comparing segmentation algorithms on the basis of their performance on a set of representative images, as well as for optimizing certain parameters of a segmentation algorithm for a given application scenario.

6. Acknowledgements

This material is based upon work supported by the EU IST project IST-2000-32795 SCHEMA. The assistance of COST276 is also gratefully acknowledged.

References

[1] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis, "A framework for the efficient segmentation of

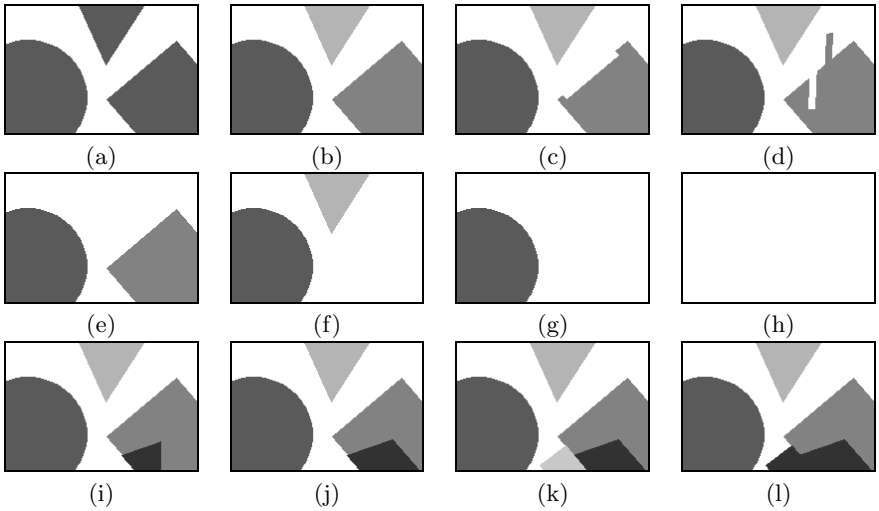

Figure 2: *A manually generated set of segmentations of a synthetic image, used for numerical evaluation experiments. The synthetic image and its ground truth are shown in (a) and (b), respectively. The different segmentations illustrate various errors, namely incorrect boundary localization ((c) and (d)), under-segmentation ((e)-(h)), and over-segmentation ((i)-(l)).*

segmentation mask	description	numerical evaluation output
Figure 2(b)	perfect segmentation (ground truth)	0.0
Figure 2(c)	incorrect boundary localization	2.755405
Figure 2(d)	incorrect boundary localization	3.193779
Figure 2(e)	one region is missed	10.254688
Figure 2(f)	one larger region is missed	34.158249
Figure 2(g)	two regions are missed	44.412938
Figure 2(h)	the image is segmented to a single region	90.919074
Figure 2(i)	a small undesirable region is formed (over-segmentation)	4.141135
Figure 2(j)	a larger undesirable region is formed	11.49188
Figure 2(k)	a second undesirable region is formed	16.314982
Figure 2(l)	the two undesirable regions are merged	16.314982

Table 1: *Numerical evaluation of the segmentations of figure 2*

large-format color images," in *Proc. IEEE Int. Conf. on Image Processing (ICIP02)*, vol. 1, pp. 761–764, Sept. 2002.

[2] C. Carson, S. Belongie, H. Greenspan, and J. Malik, "Blobworld: Image Segmentation Using Expectation-Maximization and Its Application to Image Querying," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, 2002.

[3] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "An Ontology Approach to Object-based Image Retrieval," in *Proc. IEEE Int. Conf. on Image Processing (ICIP03)*, Barcelona, Spain, Sept. 2003.

[4] V. Mezaris, I. Kompatsiaris, and M.G. Strintzis, "Video Object Segmentation using Bayes-based Temporal Tracking and Trajectory-based Region Merging," *IEEE Trans. on Circuits and Systems for Video Technology*, to appear.

[5] P. Correia and F. Pereira, "Objective evaluation of video segmentation quality," *IEEE Trans. on Image Processing*, vol. 12, no. 2, pp. 186–200, Feb. 2003.

[6] M. Levine and A. Nazif, "Dynamic measurement of computer generated image segmentations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 7, pp. 155–164, Mar. 1985.

[7] Y. Ding, G.J. Vachtsevanos, A.J. Yezzi Jr., Y. Zhang, and Y. Wardi, "A recursive segmentation and classification scheme for improving segmentation accuracy and detection rate in real-time machine vision applications," in *Proc. 14th Int. Conf. on Digital Signal Processing (DSP02)*, vol. 2, July 2002.

[8] Q. Huang and B. Dom, "Quantitative methods of evaluating image segmentation," in *Proc. IEEE Int. Conf. on Image Processing (ICIP95)*, vol. 3,

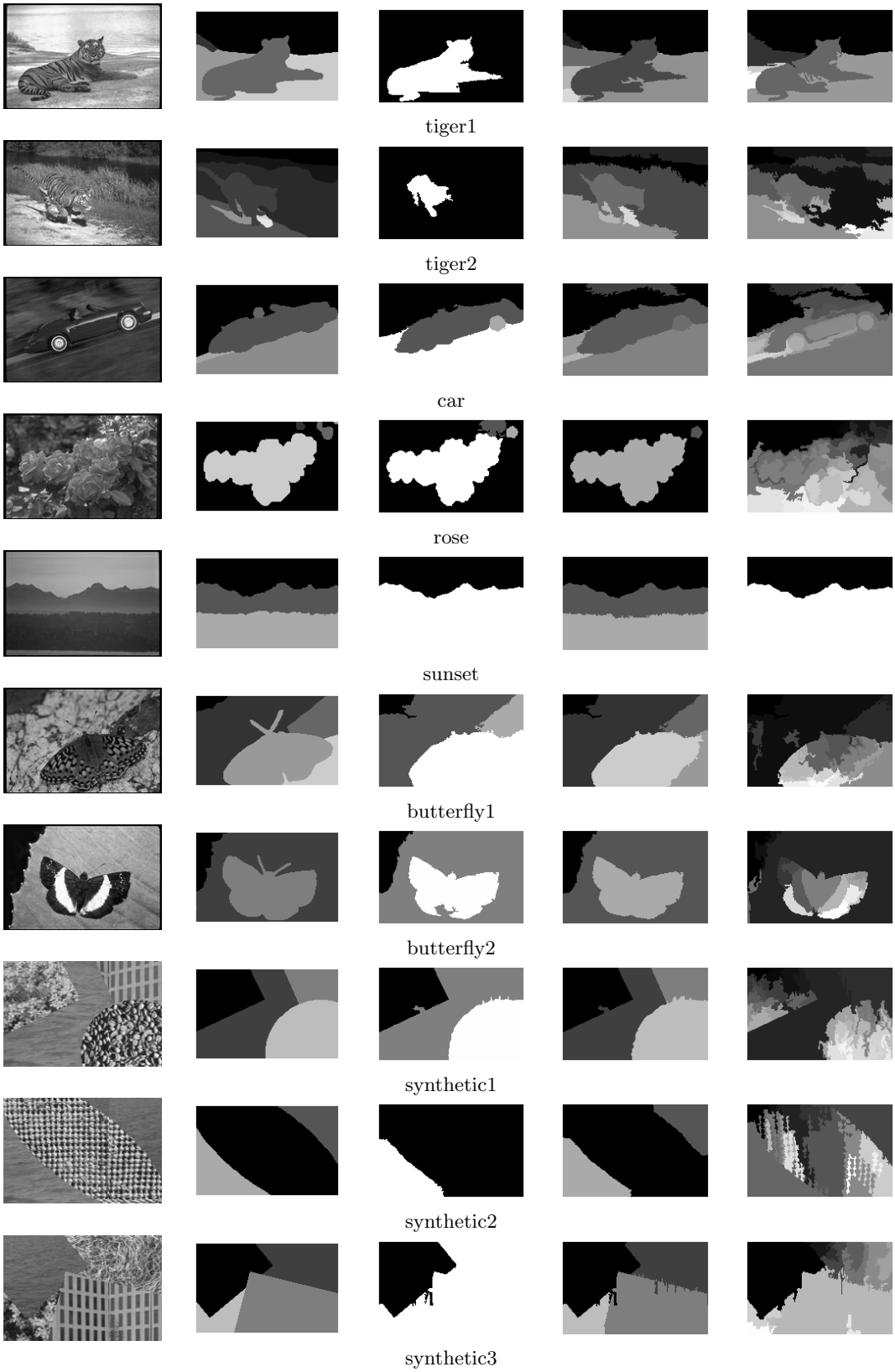

Figure 3: *Images segmented into regions, using the method of [1] (fourth column) and two variants of it, favoring under-segmentation (third column) and over-segmentation (fifth column) respectively. The reference masks (ground truth) are shown in the second column.*

images	third column masks	fourth column masks	fifth column masks
tiger1	55.042781	**12.104017**	13.819027
tiger2	74.53228	**12.979979**	109.187454
car	**18.792529**	54.643714	182.415795
rose	8.119783	**2.853145**	338.944415
sunset	44.384597	**5.722744**	44.383718
butterfly1	17.901865	**9.940959**	85.742792
butterfly2	8.486597	**7.800168**	71.658535
synthetic1	10.410216	**1.260071**	142.923679
synthetic2	16.017338	**1.787774**	205.812701
synthetic3	36.014841	**2.167452**	66.207026

Table 2: *Numerical evaluation of the segmentations of figure 3*

pp. 53–56, Oct. 1995.

[9] C. Odet, B. Belaroussi, and H. Benoit-Cattin, "Scalable discrepancy measures for segmentation evaluation," in *Proc. IEEE Int. Conf. on Image Processing (ICIP02)*, vol. 1, pp. 785–788, Sept. 2002.

[10] P. Villegas, X. Marichal, and A. Salcedo, "Objective Evaluation of Segmentation Masks in Video Sequences," in *Proc. Workshop on Image Analysis For Multimedia Interactive Services*, Berlin, May 1999.

[11] *Corel stock photo library*, Corel Corp., Ontario, Canada.

[12] *MIT Vision Texture (VisTex) database*, `http://www-white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html`.