The Objective Evaluation of Image Object Segmentation Quality

Ran SHI[1], King Ngi NGAN[1], Fellow, IEEE and Songnan LI[1], Member, IEEE,

[1] Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
{rshi, knngan, snli}@ ee.cuhk.edu.hk

Abstract. In this paper, a novel objective quality metric is proposed for individual object segmentation in images. We analyze four types of segmentation errors, and verify experimentally that besides quantity, area and contour, the distortion of object content is another useful segmentation quality index. Our metric evaluates the similarity between ideal result and segmentation result by measuring these distortions. The metric has been tested on our subjectively-rated image segmentation database and demonstrated a good performance in matching subjective ratings.

Keywords: Object segmentation, Objective metric, Distortions.

1 Introduction

Object segmentation is an important prior processing step for a variety of applications, such as content-based image retrieval, image retargeting and image compression, etc. The object segmentation quality directly influences their performances. Human subjective quality judgment is a reliable approach to evaluate the object segmentation quality. However, the subjective evaluation is time-costing and cumbersome. Therefore, there is a great demand for designing objective metrics which can automatically evaluate segmentation quality and be in close agreement with human judgments.

F-measure is a classical and popular segmentation metric [1]. It compares segmentation results with manually labeled ground truths to find the mismatching regions. The mismatching regions are then classified as false positive and false negative ones, respectively. Two indexes called precision and recall are adopted to measure these two types of distortions, and they are combined in F-measure to evaluate the overall segmentation quality. In [2], Jaccard index was proposed, which used a region-merging strategy to measure segmentation accuracy. Like F-measure, Villegas and Marichal's metric [3] also classified segmentation errors into false positive and false negative ones. Distance information was introduced to weight these two types of errors. In [4], Erdem and Sankur adopted shape information as an empirical discrepancy measures. Fragmentation was introduced in [5] to measure discrepancy in terms of the quantity of objects. A new fuzzy Jaccard index is proposed in [6] to capture the intrinsic uncertainty in the edge positions in order to

evaluate boundary accuracy. In [7, 18], Gelasca classified all possible segmentation errors into four different types: added region, added background, inside holes and border holes. Each type was assigned a weight which was derived from psychophysical experiments. Although this approach exploited perceptual information, it only used the area and distance information to describe segmentation errors. Furthermore, all these metrics mentioned above do not consider the distortions of the segmented object content.

In this paper, we propose a new objective quality metric to evaluate the subjective quality of the individual object segmentation. The proposed metric measures the similarity between the ground truth and segmentation result in four aspects: quantity, area, external contour and content. Section 2 describes them in detail. Experimental results are presented in Section 3. Finally, conclusion is drawn in Section 4.

2 Proposed Metric

For object segmentation, the image is partitioned into two segments, i.e., object and background. However, the segmented object may have more than one region. Suppose the segmented object consists of $\{R_1, R_2, ...R_n\}$, and $R_i \cap R_j = \varnothing$ for $i \neq j$. We group these regions into two sets:

$$R_{external-region} = \{R_i \mid R_i \cap R_{ground-truth} = \varnothing\} \quad (1)$$

$$R_{object-region} = \{R_j \mid R_j \cap R_{ground-truth} \neq \varnothing\} \quad (2)$$

where $R_{ground-truth}$ represents the ideal segmentation result as exemplified in Fig.1(a). Fig.1 (b) shows these two region sets in the segmentation result. The regions surrounded by red and blue lines are $R_{external-region}$ and $R_{object-region}$, respectively.

(a) (b) (c)

Fig. 1. Different kinds of regions and segmentation errors. (a) ground truth; (b) two kinds of regions; (c) four kinds of segmentation errors.

As aforementioned, in [7, 18] segmentation errors are classified into four types which is illustrated in Fig.1(c). Different types of segmentation errors distort the

object differently. From Fig.1(c), we can observe that added regions (red) increase object's area and quantity; added background (green) increases object's area and changes object's external contour; border holes (yellow) decrease object's area and destroys object's external contour and content; inside hole (pink) decreases object's area and destroy content. A summary of these observations is given in Table 1. Our metric evaluates segmentation quality based on four aspects of the object, i.e., quantity, area, external contour and content. Higher metric score indicates better segmentation quality.

Table 1. The distortions produced by types of segmentation errors.

	Quantity	Area	External Contour	Content
Added region	√	√		
Added background		√	√	
Border hole		√	√	√
Inside hole		√		√

2.1 The quantity of object

Number of objects in the segmentation result should be equal to that in the ground truth. In other words, a substantial disagreement of the object number can be used to indicate a large discrepancy [9]. For individual object segmentation, there is only one object. Since the human visual system (HVS) does not pay attention to very small errors [6, 8], those isolated points should not be treated as external regions. "Opening" and "Closing" operation (the mask is 3×3) is performed first on the segmentation result in order to remove isolated points and fill the very tiny holes. Then, we use the following equation to measure similarity in terms of the object quantity :

$$S_{quantity} = \frac{1}{1+\frac{Area(R_{external-region})}{Area(R_{object-region})}\cdot card(R_{external-region})} \tag{3}$$

where $Area(\cdot)$ and $card(\cdot)$ is operation of computing the region area and quantity, respectively. Different from [5], we introduce relative sizes of added regions in Eq. (3) to replace the scaling parameters used in [5]. With this modification, larger area of added regions will lead to worse segmentation results.

2.2 The area of object

Area is another important index for segmentation quality measure. We use $R_{true-object}$ to denote the portion of object that has been correctly segmented. The measure S_{area} is used to measure the area accuracy :

$$S_{area} = \frac{Area(R_{true-objcet})}{Area(R_{object-region} \cup R_{ground-truth})} \tag{4}$$

The above formula is like Jaccard Index [2]. Due to the area of added regions have been considered in Eq. (3), here we only consider the area variation caused by added background, border holes and inside holes.

2.3 The external contour of object

According to [7], segmentation quality will be quite different when errors are uniformly distributed along the object boundaries from that when errors concentrate in parts of the object boundaries. As shown in Fig.2, although the two pictures have the same area of added background, Fig.2 (a) has a better quality since its contour Fig.2 (b) maintains the shape of the ideal object contour.

(a) (b) (c) (d)

Fig. 2. Different kinds of object's external contour.

According to the fuzzy set theory [6, 15], each pixel *x* is assigned two probabilities values, $f_{\Omega_G}(x)$ and $f_{\Omega_S}(x)$, Ω_G and Ω_S represent fuzzy sets for the ideal external contour and the segmented one, respectively. $f_{\Omega_G}(x)$ is the membership function of Ω_G, and is defined as :

$$f_{\Omega_G}(x) = \frac{refdis_G}{refdis_G + D_G(x)} \tag{7}$$

$$refdis_G = k \cdot length_G \tag{8}$$

where $D_G(x)$ is the shortest distance from *x* to the ideal external contour. *k* is a

scaling value, and $refdis_G$ is a reference distance. $length_G$ is the diagonal of the ideal external contour's bounding box. If $D_G(x)$ is shorter than $refdis_G$, it means x has a high probability belonging to Ω_G. Furthermore, the probability of a pixel belonging to a contour is not only related to the distance between the pixel and the contour, but also related to the scale of the contour, or equivalently, the size of the object. Therefore, we introduce $refdis_G$ as a relative distance rather than using absolute distance as in [6], since $refdis_G$ is adaptive to different sizes of the object

The membership function $f_{\Omega_S}(x)$ of Ω_S can be similarly defined. We compare probability of elements in these two fuzzy sets to describe the similarity between the ideal contour and segmented contour:

$$S_{contour} = \frac{\sum_x \min(f_{\Omega_G}(x), f_{\Omega_S}(x))}{\sum_x \max(f_{\Omega_G}(x), f_{\Omega_S}(x))} \tag{9}$$

where $x \in R_{true-objcet} \oplus R_{ground-truth}$. From Eq. (9), we can see that for a good segmentation which maintains the shape of ideal contour, the two probabilities for each pixel should be similar, which can be satisfied by Fig. 2(b) and violated by Fig. 2(d).

The overall measure for the similarity of object region is defined as follows:

$$S_{object-region} = S_{area} \cdot S_{contour} \tag{10}$$

2.4 The content of object

Based on our observations on a subjective dataset to be introduced in Section 3, it is found that the segmented object without border holes or inside holes usually gets a higher subjective score than those with border holes or inside holes. As shown in Fig.3, the "cake" in the right image obtains a lower subjective score (1.04) than "sun flower" (2.08), even though it has a better external contour and smaller error area.

Fig. 3. The segmentation results of "sun flower" and "cake".

The reason is that the content information of object itself is lost due to border holes and inside holes, which makes the segmented object less recognizable. For example, in "sun flower", object itself is the yellow flower without green leaves. Subjects trend to prefer complete objects which can be easily recognized. We measure the completeness of object content in terms of area and texture :

$$S_{content} = S_{object-area} \cdot S_{object-texture} \tag{11}$$

Different from S_{area}, $S_{object-area}$ does not consider the error area of added background. It only measures the area variation which could lead to loss of object content. A lower value of $S_{object-area}$ indicates that the segmented object losses more content information. $S_{object-area}$ is given by

$$S_{object-area} = \frac{Area(R_{true-object})}{Area(R_{ground-truth})} \tag{12}$$

The texture information is crucial for object recognition [8]. We convert the original color image into a gray image. Then, Sobel operator is applied to this gray image. The magnitude of gradient values is used to approximate the texture information. $S_{object-texture}$ is defined as

$$S_{object-texture} = \frac{\sum Sobel(x)}{\sum Sobel(y)} \tag{13}$$

where pixel $x \in R_{true-object}$ and $y \in R_{ground-truth}$. A higher value of $S_{object-texture}$ indicates the segmented object could be easily recognized.

Finally, the overall objective quality metric by using spatial weighted pooling [10] for individual object segmentation is given as follows:

$$S_{overall} = \frac{\alpha \cdot S_{quantity} + S_{object-region} + \lambda \cdot S_{content}}{1 + \alpha + \lambda} \tag{14}$$

where α and λ are balancing weights among three terms. The values of three parameters α, λ and k are determined by training as discussed in Section 3.

3 Experimental Results

Our dataset consists of a testing set and a training set. The testing set has 76 original images and 152 segmentation images which are generated by using Achanta's [11] and Rahtu's [12] segmentation algorithms. The training set has 18 original images and 18 segmentation images which are also generated by using the same segmentation

algorithms. The original images are selected from Microsoft Research Asia salient object database (Image Set B). Each image includes only one object to be segmented. The ground truth is provided by Achanta [11]. One original image and its corresponding ground truth and segmentation results compose an image group. Since there are no prescribed standards for the subjective evaluation of object segmentation, we incorporate the simultaneous double stimulus for continuous evaluation (SDSCE) method and double-stimulus continuous quality-scale (DSCQS) method [13] to design our subjective assessment. The subjective assessment interface is shown in Fig.4. The original images are provided to help viewers to understand images' content. The viewers can press the "arrow button" to do the switchover among ground truths and segmentation results, and then they can select the subject ratings. The difference of the subjective assessment in this paper is the use of the absolute category rating (ACR) scale [19] which employs a five-grade discrete (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) segmentation quality scale. In [17], the experimental data has demonstrated that there are no obvious overall statistical differences between different rating scales. Therefore, the five-grade discrete is employed to reduce the viewers' fatigue and make the subjective rating more distinguishable. Totally 16 subjects (10 males and 6 females, from 23 to 26 years old) participate in the subjective test to evaluate perceptual quality of each segmented image, where 8 viewers are experts in image processing and the others are not. Each viewer begins with a brief introduction about this subjective study and how to do the quality evaluation. During the subjective assessment, the first five image groups' assessments are treated as a training session which is used to stabilize the viewers' opinion. These images do not belong to testing set or training set, and their ratings are not taken into account in the results of the assessment. The subjective ratings are processed to calculate the Differential Mean Option Score (DMOS) which indicates the quality difference between the ground truth segmentation and the algorithm segmentation result. Outliers are rejected by a standard screening procedure [13]. The three parameters of proposed metric are automatically chosen by maximizing the Spearman Rank-Order Correlation Coefficients (SROCC) using the training set. The training result is $k = 0.03$, $\alpha = 0.5$ and $\lambda = 0.7$.

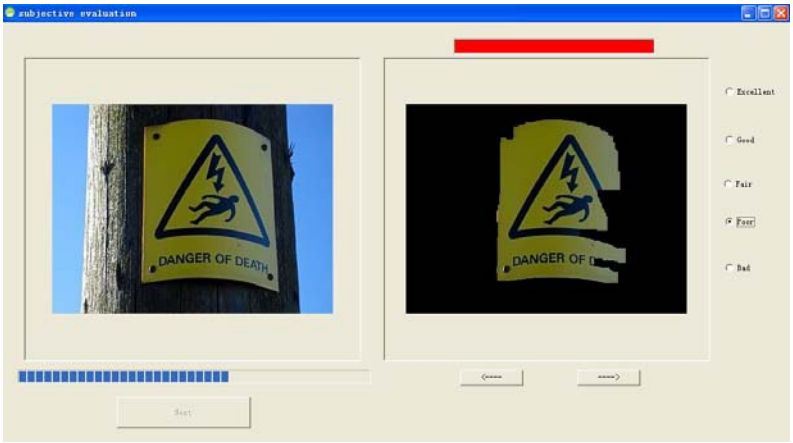

Fig. 4. The subjective evaluation interface.

Following the Video Quality Expert Group's work [14], each metric score is mapped to $Q(x)$ firstly using the following fitting function to obtain a linear relationship between $Q(x)$ and the subjective scores:

$$Q(x) = \beta_1 \times (0.5 - \frac{1}{1 + \exp(\beta_2 \times (x - \beta_3))}) + \beta_4 \times x + \beta_5 \tag{15}$$

To evaluate its performance, we use two common performance evaluation criteria, i.e., the Linear Correlation Coefficient (LCC) and the Spearman Rank-Order Correlation Coefficients (SROCC), which use Q(x) and DMOS as their inputs [16]. Our metric is compared against two popular image object segmentation quality metrics $F_\beta measure$ ($\beta = 0.3$) which is used in [11, 12] and Jaccard Index [2]. We also evaluate our metric's performance without $S_{content}$. The comparison results are shown in the Table 2:

Table 2. Overall performances of four different segmentation metrics.

		$F_\beta measure$	Jaccard Index	Our (without $S_{content}$)	Our
76 images by Achanta	SROCC	0.74	0.80	0.80	**0.82**
	LCC	0.70	0.80	0.82	**0.84**
76 images by Rahtu	SROCC	0.84	0.88	0.88	**0.88**
	LCC	0.85	0.89	0.89	**0.91**
Overall 152 images	SROCC	0.81	0.87	0.87	**0.89**
	LCC	0.82	0.86	0.87	**0.88**

From Table II, we can see that our metric achieves higher SROCC and LCC with $S_{content}$. On the other hand, the experimental results demonstrate that the design of $S_{content}$ is reasonable. The object content is an important measure for segmentation quality evaluation. Fig. 5 shows the scatter plots of the proposed object segmentation quality metric on our databases. In the graph, each circle represents a test image (total 152 images). The vertical axis denotes the DMOS and the horizontal axis denotes the nonlinearly mapped metric outputs.

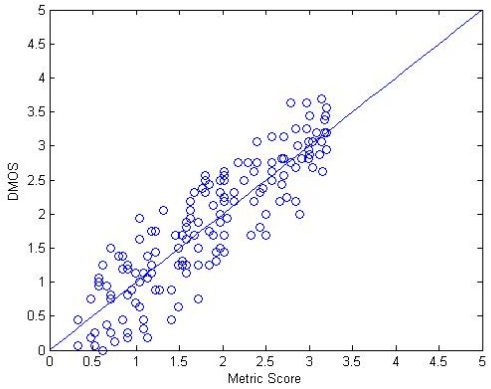

Fig. 5. Scatter plots of the proposed object segmentation quality metric on our dataset (after the nonlinear mapping).

4 Conclusion

In this paper, we propose an objective quality metric for individual object segmentation in images. Our metric is designed based on describing four types of distortions. Relative size of added regions is considered to measure the quantity of object. We use fuzzy set theory to describe the similarity of object's contour, and introduce reference distance to adapt different sizes of the object. Meanwhile, the completeness of object content is treated as an important measure in our metric. The experimental results demonstrate that our metric has a good performance on our individual object segmentation dataset.

References

1. D. Powers, M.S.: Evaluation: From Precision, Recall and F-Factor to ROC, Informedne, Markedness & Correlation. In: Journal of Machine Learning Technologies, vol.2, no.1, pp. 37–63 (2011)
2. F. Ge, S. Wang and T. Liu, M.S.: New benchmark for image segmentation evaluation. In: Journal of Electronic Imaging, vol.16, no.3, 033011 (2007)
3. P. Villegas and X. Marichal, M.S.:Perceptually-weighted evaluation criteria for segmentation masks in video sequences. In: IEEE Trans. Image Process, vol. 13, no. 8, pp. 1092–1103 (2004)
4. C. Erdem and B. Sankur, C.:Performance evaluation metrics for objectbased video segmentation. In: Proc. X Eur. Signal Process Conf, vol. 2, pp. 917–920, Tampere, Finland (2000)

5. K. Strasters and J. Gebrands, M.S.: Three-dimensional image segmentation using a split, merge and group approach. In: Pattern Recognit. Lett., vol. 12, no.5, pp. 307–325 (1991)
6. K. McGuinness, and N. O'Connor, M.S.: A comparative evaluation of interactive segmentation algorithms. In: Pattern Recognition, vol. 43, no.2, pp.434-444 (2010)
7. E. D. Gelasca, M.S.: Full-reference objective quality metrics for video watermarking, video segmentation and 3D model watermarking. In: Ph.D. dissertation, EPFL, Lausanne, Switzerland (2005)
8. P. Correia and F. Pereira, M.S.: Objective evaluation of video segmentation quality. In: IEEE Trans. Image Process, vol. 12, no. 2, pp. 186–200 (2003)
9. Y. J. Zhang, M.S.: A Survey on Evaluation Methods for Image Segmentation In: Pattern Recognition, vol. 29, no.8, pp.1335-1346 (1996)
10. S.N Li, C.M Mak and K.N.Ngan, C.: Visual Quality Evaluation for Images and Videos. In: Multimedia Analysis, Processing and Communication, Springer Berlin/Heidelberg Publisher (2011)
11. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, M.S.: Frequency-tuned salient region detection. In: Proc. IEEE CVPR, pp. 1597-1604, Miami, USA (2009)
12. E. Rahtu, J. Kannala, M. Salo, and J. Heikkila, C.: Segmenting salient objects from images and videos, In: Proc. ECCV, pp. 366-379, Crete, Greece (2010)
13. Internal Telecommunication Union Radio communication Sector, C.: ITU-R Recommendation BT.500-13, Methodology for the Subjective Assessment of the Quality of Television Pictures (2012)
14. Video Quality Expert Group (VQEG) S.: Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment I (2010)
15. L. Zadeh, M.S.: Fuzzy sets and systems. In: Information and Control, vol.8, no.3, pp. 338–353 (1965)
16. S. Li, F. Zhang, L. Ma, and K. N. Ngan, M.S.: Image quality assessment by separately evaluating detail losses and additive impairments. In: IEEE Trans. Multimedia, vol. 13, no. 5, pp. 935–949 (2011)
17. Q. Huynh-Thu, N. N. Garcia, F. Speranza, P. Corriveau, and A. Raake, M.S.: Study of rating scales for subjective quality assessment of high-definition video. In: IEEE Trans. Broadcasting, vol. 57, no. 1, pp. 1–14 (2011)
18. E. D. Gelasca, M. Karaman, T. Ebrahimi and T. Sikora, M.S.:A Framework for Evaluating Video Object Segmentation Algorithms. In: Proc. IEEE CVPR Workshop, pp. 198-198, New York, USA (2006)
19. Internal Telecommunication Union Telecommunication Standardization Sector, C.: ITU-T Recommendation P.910, Subjective video quality assessment methods for multimedia applications (2012)