JACCARD INDEX COMPENSATION FOR OBJECT SEGMENTATION EVALUATION

Ran SHI, King Ngi NGAN, Fellow, IEEE, and Songnan LI

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

ABSTRACT

In this paper, we propose an objective metric for quality evaluation of individual object segmentation in images. Using the Jaccard Index as a base, additional compensation terms are integrated into our metric. These terms not only allow our metric to combine region-based and boundary-based methods, they also describe human visual tolerance and saturation. This metric can also can adaptively adjust "perception function" under different image sizes. The experiment on our subjective object segmentation quality assessment database demonstrates that the proposed metric performs well in matching subjective ratings.

Index Terms— Objective metric, Object segmentation, Human perception

1. INTRODUCTION

Object segmentation is an important and challenging preprocessing step for a variety of applications. It aims at assigning a unique label ("object" or "background") to each pixel. However, different applications have their own requirements for segmentation quality. For some applications, such as image editing, image retargeting and 2D to 3D conversion etc., object segmentation quality directly influences their final performance in terms of visual quality. They require segmentation quality to be as close as possible to the manual extraction. Therefore, it is necessary to design methods to evaluate whether an object segmentation algorithm can achieve this requirement. It is well acknowledged that the most reliable method used to evaluate object segmentation quality is subjective quality evaluation. However, it is infeasible since subjective evaluation is time-consuming and cumbersome. Therefore, there is great demand for designing objective metrics that can automatically evaluate segmentation quality and correlate well with human judgment.

Objective object segmentation metrics are generally classified into two categories: region-based metrics and boundary-based metrics. For region-based metrics, manually labeled ground truth and segmentation result are overlapped. Then, pixels can be distinguished as matching pixels which are labeled as "object" in both ground truth and segmentation result, or as mismatching pixels which are labeled differently. Jaccard Index [1] and F1-score [2] are two typical region-based metrics. Jaccard Index is a ratio of the number of matching pixels to the total number of both matching pixels and mismatching pixels. For F1-score, the mismatching pixels are further divided into false positive and false negative ones. Two indices, namely precision and recall, are adopted to measure these two types of mismatching. Then, F1-score combines them with equal weight to evaluate the overall segmentation quality. In spite of their simplicity, region-based metrics cannot reflect human perception when errors occur at different positions [3].

Boundary-based metrics focus on how to evaluate the distortion of segment result's boundary compared to that of ground truth's. For example in [4], Hausdorff distance was adopted to indicates the maximum of the shortest distance between the segment boundary and the ground truth boundary. It can only reflect local distortion rather than the global one. Some metrics [5, 6] take human perception into account. For [5], each pixel was assigned two likelihood probabilities associated with two boundary sets using fuzzy set theory. Then, the boundary distortion was determined by measuring the difference between the probabilities. Csurka et al. [6] introduced a tolerance band around the ground truth's boundary, whose width was adaptive to the image size. The boundary can be adjusted since it assumed that errors can be tolerated within the band. Then, boundary-based F1-score was used to evaluate the quality. Although boundary-based metrics can reflect human perception, they cannot well describe area distortion.

In [6], Csurka et al. argued that a possible combination between region-based and boundary-based metrics can make evaluation more accurate. However, they only suggested a way to select suitable metrics for certain segmentation algorithms without proposing a concrete metric. Movahedi et al. [7] proposed a mixed measure which calculated the average distance from mismatching pixels to the corresponding boundary. However, average value is not a good index to distinguish different segmentation quality. In this paper, we propose additional terms to extend the Jaccard Index. These terms assign compensation values to mismatching pixels around the boundary. Our metric not only evaluates area mismatches, but also considers human visual tolerance and saturation for describing human perception. The rest of this paper is organized as follows. Section 2 describes our metric in details. Experimental results are presented in Section 3.

(a) Original Image [8] (b) Ground Truth [9] (c) Segmentation Result [10] (d) Four types of errors

Fig. 1: Example of original image, its ground truth, segmentation result and segmentation errors. In (d), the green regions are added regions , the blue regions are added background, the orange regions are border holes and the red region is an inside hole.

Finally, conclusion is drawn in Section 4.

2. PROPOSED METRIC

For evaluating object segmentation quality, our metric aims at measuring the similarity between the ground truth and the segmentation result as shown in Fig. 1. The segmentation result and corresponding ground truth are represented as S and G respectively. Jaccard Index is defined as

$$E = \frac{A(G \cap S)}{A(G \cup S)} \tag{1}$$

where $A(\cdot)$ is the operation of counting amount. In Eq. (1), the numerator corresponds to the number of matching pixels or in other words, true positive (TP). The denominator counts the total number of matching and mismatching pixels. Actually, it is an absolute "0" or "1" assignment and counting problem in the Jaccard Index. In [3], mismatching pixels are classified into four types: added region (AR), added background (AB), border holes (BH) and inside holes (IH). Fig. 1(d) shows an example of the four types of errors. It is easy to understand that added region and added background's pixels are wrongly labeled as "object" in the segmentation result. On the other hand, those pixels mistakenly labeled as "background" are defined as either border holes or inside holes. Added background and border holes are adjacent to the true boundary, while added region and inside holes are not. For our metric, we design compensation terms for the original Jaccard Index as

$$E = \frac{A(G \cap S) + \sum\limits_{i \in AB} COM(i)}{A(G \cup S) - \sum\limits_{j \in BH} COM(j)} \tag{2}$$

In [3, 5, 6], two human perception properties for object segmentation evaluation are discussed. Based on their conclusions, these two properties can be interpreted as: one that humans can put up with added background and border holes to some extent according to the distortion of the boundary, i.e. human visual tolerance; the other that for human beings, it is hard to quantify similarity when errors become large, i.e. human visual saturation. In our metric, these two properties are embodied in that mismatching pixels might be assigned with compensation values according to their tolerance degrees. This means that for pixel i in added background, if it can be tolerated, we can treat it like true positive pixel. So we add its compensation value $COM(i)$ onto the numerator of the Jaccard Index. Similarly, if human eyes can tolerate pixel j in border holes, it can be regarded as a true background pixel. So its compensation value $COM(j)$ is subtracted from the denominator of the Jaccard Index. Since added regions and inside holes do not distort the object's boundary, there are no compensation values for them in our metric. In other words, since added background and border holes are along the boundary, some of their pixels carry both region and boundary attributes in our metric. The boundary attribute can compensate the loss of region attribute to some extent according to human visual tolerance and saturation. Therefore, We seek a perception function which can properly describe these two properties. Logistic function as shown in Fig. 2(a) is one type of psychometric function which establishes a relationship between a physical stimulus and human response [11]. The following formula is the definition of Logistic function.

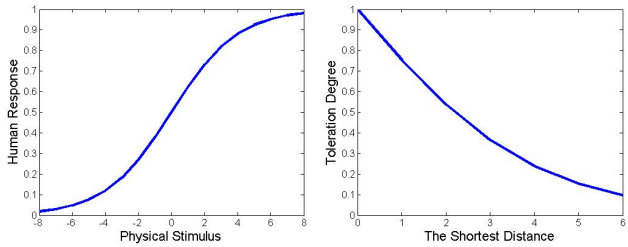
$$y = \frac{1}{1 + \exp(-x/\sigma)} \tag{3}$$

where σ is the bandwidth of the exponent function. x and y can be treated as physical stimulus and human response respectively. In order to be consistent with human visual tolerance and saturation, we adjust the logistic function to obtain the perception function and define $COM(i)$ as follows:

$$COM(i) = \begin{cases} 2 - \dfrac{2}{1 + \exp(-\min\limits_{p \in \Omega_i}(D(i,p))/\sigma)} & \Omega_i \neq \emptyset \\ 0 & o.w. \end{cases} \tag{4}$$

$$\Omega_i = TP \cap SC(i)$$

where $D(i,p)$ is the distance between i and p, and $SC(i)$ is a search circle whose centre point is i and search radius is



(a) Logistic function curve (b) Perception function curve

Fig. 2: The comparison of two curves with $\sigma = 2$.

R. $COM(j)$ can be similarly defined expect $\Omega_j = TN \cap SC(j)$, where TN represents those pixels which are correctly labeled as "background" in the segmentation result. Here, Eq. (4) is treated as the perception function. The shortest distance from a mismatching pixel to the true positive pixels is treated as the physical stimulus, and the output simulates the human tolerance degree which is selected as compensation value. Obviously, if mismatching pixels are far away from true positive pixels, human absolutely cannot tolerate these errors. So it makes no sense to assign compensation values to these pixels. In [6], the search range was made to adpat to the image size. We follow this approach so that R is obtained by $round(\alpha \cdot D_{length})$ where α is a constant and D_{length} is the diagonal length of the image. Since the search range is fixed, the bandwidth σ can be estimated by the following equation.

$$2 - \frac{2}{1 + \exp(-R/\sigma)} = \tau \tag{5}$$

Then equivalently,

$$\sigma = -R \left(\ln \frac{\tau}{2 - \tau} \right)^{-1} \tag{6}$$

where τ is a small constant. When the true positive pixels are only at the boundary of the search circle, the corresponding tolerance degree should be very low. Furthermore, a mismatching pixel's tolerance degree will be 0, when there are no true positive pixels within the search circle. The perception function is drawn in Fig. 2(b) with $\sigma = 2$. From this figure, we can see that the tolerance degree is high when the distance is short. However, when the distance becomes longer, the tolerance degree is lower and the variation of tolerance degree also becomes smaller. These observations demonstrate that our perception function properly conforms to both human visual tolerance and saturation. Different from the an absolute "0" or "1" assignment problem in the Jaccard Index , the compensation value given by Eq. (4) can be between 0 and 1, which also makes our metric more consistent with human perception.

Fig. 3: The interface of our subjective assessment.

3. EXPERIMENTAL RESULTS

In order to evaluate our metric's performance, we develop a subjective object segmentation quality assessment database. The original images are selected from four popular object segmentation databases, such as Microsoft research Asia salient object database (Image Set B) [8], Weizmann [12], VOC2012 [13] and Microsoft research Cambridges grabcut database [14] (it selected some images from the Berkeley Segmentation Database [15]). For individual object segmentation assessment, each image includes only one object to be segmented. The segmentation results are generated by different types of object segmentation algorithms, such as McGuinness interactive segmentation tool [5], Li's Distance Regularized Level Set Evolution (DRLSE) method [16], Mohit Gupta and Krishnan Ramnaths grabcut tool-box [17] and Achantas and Rathus automatic object segmentation methods [9, 10]. In total, 180 segmentation results are selected as our test images, which include various types and intensities of segmentation errors. We refer to the simultaneous double stimulus for discrete evaluation (SDSDE) method [18] to conduct our subjective assessment. The subjective assessment interface is shown in Fig. 3. The original images play an auxiliary role to help viewers understand image content. By referencing the corresponding ground truth, they can select the subject ratings. In order to relieve viewer fatigue, our tests are divided into two sessions. 33 and 31 viewers participated in session 1 and session 2, respectively. The subjective ratings are firstly converted to Z-scores in order to reduce negative influence induced by viewers' rating habits [19]. Then, we conduct a standard screening procedure to reject unreliable viewers. After the screening procedure, 3 out of 33 subjects and 4 out of 31 subjects are rejected in session 1 and session 2, respectively. Finally, Mean Option Score (MOS) is calculated, which indicates the subjective segmentation quality.

Following the work of Video Quality Expert Group [20], each objective score should be mapped to $Q(x)$ by fitting the following function in order to obtain a linear relationship with MOS:

$$Q(x) = \beta_1 \times (0.5 - \frac{1}{1 + \exp(\beta_2 \times (x - \beta_3))}) + \beta_4 \times x + \beta_5 \tag{7}$$

Table 1: Performance of six segmentation quality metrics

	SROCC	LCC	RMSE	OR
MM	0.57	0.32	13.29	0.16
F1	0.85	0.85	7.38	0.03
JI	0.85	0.85	7.37	0.03
FC	0.84	0.84	7.71	0.02
BF	0.86	0.86	7.26	0.01
OUR	0.88	0.88	6.61	0.01

In our experiment, we empirically set $\alpha = 0.02$ and $\tau = 0.1$ respectively. To evaluate its performance, we use four common performance evaluation criterions, i.e., the Linear Correlation Coefficient (LCC), the Spearman Rank-Order Correlation Coefficients (SROCC), the root mean squared error (RMSE), and the outlier ratio (OR), which use $Q(x)$ and MOS as their inputs [21]. Higher values for the first two criterions indicate better performance, while the last two are vice versa. Our metric is compared against five object segmentation evaluation metrics which are the Jaccard Index (JI) [1], F1score (F1) [2], Fuzzy Contour (FC) [5], Boundary F1score (BF) [6] and Mixed Measure (MM) [7]. The comparison results are shown in Table 1.

Since Mixed Measure does not adopt a suitable normalization method, its performance is the worst of all. In fact, its objective scores are all very low, which leads to mapping failure. That is the reason its LCC is much lower and RMSE is much higher. Although Fuzzy Contour considered human visual properties, it treated all pixels in the image as members of fuzzy sets, which made the metric insensitive to the distortion of boundary. Its performance is worse than that of the other three metrics. Two region-based metrics, Jaccard Index and F1score without perceptual factor have similar performance, but both perform worse than the Boundary F1score. Boundary F1score not only integrats human perception, but also adjusts the tolerance band according to different image sizes. So, its performance is better than those of the other three metrics whose parameters are fixed. Our metric achieves the best performance in terms of all four criterions. It indicates our "compensation" strategy reasonably integrates the advantages of region-based and boundary-based metrics, allowing our objective metric to more closely approximate human judgement.

4. CONCLUSION

For object segmentation evaluation, region and boundary information are two main factors in designing objective metrics. Our novel metric takes advantage of psychometric functions to describe human visual tolerance and saturation. It treats the maximum tolerance degree as compensation for the Jaccard Index. The experimental results demonstrate how well our metric matches subjective ratings on our object segmentation assessment database. In future work, we will integrate semantic information into our metric, another important factor in the subjective evaluation of object segmentation quality.

5. REFERENCES

[1] Ge Feng, Song Wang, and Tiecheng Liu, "New benchmark for image segmentation evaluation," *Journal of Electronic Imaging*, vol. 16, no. 3, pp. 033011–033011, 2007.

[2] DMW Powers, "Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[3] Elisa Drelie Gelasca and Touradj Ebrahimi, "On evaluating video object segmentation quality: a perceptually driven objective metric," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 3, no. 2, pp. 319–335, 2009.

[4] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[5] Kevin McGuinness and Noel E OConnor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognition*, vol. 43, no. 2, pp. 434–444, 2010.

[6] Diane Larlus Gabriela Csurka and Florent Perronnin, "What is a good evaluation measure for semantic segmentation?," in *2013 24th British Machine Vision Conference*, 2013.

[7] Vida Movahedi and James H Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2010, pp. 49–56.

[8] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum, "Learning to detect a salient object," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 353–367, 2011.

[9] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, "Frequency-tuned salient region detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 1597–1604.

[10] Esa Rahtu, Juho Kannala, Mikko Salo, and Janne Heikkilä, "Segmenting salient objects from images and videos," in *Computer Vision–ECCV 2010*, pp. 366–379. 2010.

[11] Felix A Wichmann and N Jeremy Hill, "The psychometric function: I. fitting, sampling, and goodness of fit," *Perception & psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.

[12] Sharon Alpert, Meirav Galun, Ronen Basri, and Achi Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration.," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2007.

[13] Everingham M., Van Gool L., Williams C.K.I., Winn J., and Zisserman A., "The pascal visual object classes challenge 2012 (voc2012) results," http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.htm.

[14] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM Transactions on Graphics (TOG)*, 2004, vol. 23, pp. 309–314.

[15] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, 2001, vol. 2, pp. 416–423.

[16] Chunming Li, Chenyang Xu, Changfeng Gui, and Martin D Fox, "Distance regularized level set evolution and its application to image segmentation," *Image Processing, IEEE Transactions on*, vol. 19, no. 12, pp. 3243–3254, 2010.

[17] Mohit Gupta and Krishnan Ramnath., "Interactive segmentation tool-box," http://www.cs.cmu.edu/ mohit-g/segmentation.htm.

[18] Lin Ma, Weisi Lin, Chenwei Deng, and K Ngan, "Image retargeting quality assessment: A study of subjective scores and objective metrics," 2012.

[19] Andre M van Dijk, Jean-Bernard Martens, and Andrew B Watson, "Quality asessment of coded images using numerical category scaling," in *Advanced Networks and Services*, 1995, pp. 90–101.

[20] Video Quality Expert Group (VQEG), "Final report from the video quality experts group on the validation of objective models of video quality assessment i," 2010.

[21] Songnan Li, Fan Zhang, Lin Ma, and King Ngi Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, pp. 935–949, 2011.