Visual Quality Evaluation of Image Object Segmentation: Subjective Assessment and Objective Measure

Ran Shi, *Student Member, IEEE*, King Ngi Ngan, *Fellow, IEEE*, Songnan Li, *Member, IEEE*, Raveendran Paramesran, *Senior Member, IEEE*, and Hongliang Li, *Senior Member, IEEE*

***Abstract*— A visual quality evaluation of image object segmentation as one member of the visual quality evaluation family has been studied over the years. Researchers aim at developing the objective measures that can evaluate the visual quality of object segmentation results in agreement with human quality judgments. It is also significant to construct a platform for evaluating the performance of the objective measures in order to analyze their pros and cons. In this paper, first, we present a novel subjective object segmentation visual quality database, in which a total of 255 segmentation results were evaluated by more than thirty human subjects. Then, we propose a novel full-reference objective measure for an object segmentation visual quality evaluation, which involves four human visual properties. Finally, our measure is compared with some state-of-the-art objective measures on our database. The experiment demonstrates that the proposed measure performs better in matching subjective judgments. Moreover, the database is available publicly for other researchers in the field to evaluate their measures.**

***Index Terms*— Object segmentation, visual quality, subjective evaluation, objective measure.**

I. Introduction

OBJECT segmentation as a pre-processing step plays various important roles for different applications. It aims at assigning a unique label ("object" or "background") to each pixel. In this way, the object-level semantic information can be obtained instead of just the pixel-level information, which is more meaningful and easier to analyze and operate on. For some image synthesis applications, such as image editing, image retargeting and 2D to 3D conversion, the subjective quality of the segmentation result is important because the final output is to be evaluated by the end user, i.e., the human viewer. The performance of these applications is directly influenced by the visual quality of the object segmentation process. They require the object segmentation visual quality to be as close as possible to the ground truth (usually generated manually) in order to generate high quality synthesized images. Therefore, it is necessary to evaluate whether an object segmentation algorithm can satisfy this requirement. Although object segmentation algorithms have evolved over the past decades, object segmentation quality assessment is less studied, especially in terms of human visual perception.

Object segmentation visual quality evaluation share many similarities with traditional image quality assessment (IQA) [1], [2]. Firstly, the most reliable evaluation method for object segmentation visual quality is subjective evaluation by human observers. However, subjective evaluation is impractical because it is time consuming involving high labor cost. Therefore, there is a great demand for developing the objective measures which can automatically evaluate the segmentation visual quality that correlates well with human judgment. Secondly, according to the availability of manually generated ground truth, the objective measures in IQA can be classified into three types: full-reference, reduced-reference and no-reference [1]. Moreover, full-reference and no-reference segmentation quality assessment can also be referred to as supervised and unsupervised, respectively [3]. Thirdly, we need subjective segmentation quality databases similar to LIVE [2] and TID2008 [4] to evaluate how well the objective measures correlate with the subjective evaluation. It should contain the source images, the segmentation results and their corresponding ground truths and the associated subjective scores. To the best of our knowledge, there is no such object segmentation visual quality evaluation database available in the public domain to satisfy these requirements.

In this paper, we focus on the full-reference objective measure which can be used to evaluate segmentation algorithms' performance. For single object segmentation, the objective full-reference measures can be classified into three categories: region-based measures, boundary-based measures and hybrid measures [5]. Region-based measures distinguish the pixels as matching pixels or mismatching pixels by overlapping the ground truth with the segmentation result. Matching pixels are labeled as "object" in both the ground truth and the

Manuscript received January 12, 2015; revised June 1, 2015; accepted August 8, 2015. Date of publication August 25, 2015; date of current version September 23, 2015. This work was supported in part by the University of Malaya, Malaysia, under Project UM.C/625/1/HIR/MOHE/ENG/42 and in part by the Research Grants Council, Hong Kong, under Project CUHK 14201115. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lei Zhang.

R. Shi, K. N. Ngan, and S. Li are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: rshi@ee.cuhk.edu.hk; knngan@ee.cuhk.edu.hk; snli@ee.cuhk.edu.hk).

R. Paramesran is with the University of Malaya, Kuala Lumpur 50603, Malaysia (e-mail: ravee58@gmail.com).

H. Li is with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: hlli@uestc.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2473099

segmentation result, while the mismatching pixels are labeled differently. Jaccard Index [6] and F1measure [7] are two typical region-based measures. Jaccard Index is a ratio of the number of matching pixels to the total number of pixels. In F1measure, the mismatching pixels are further divided into false positive and false negative according to their labels in the ground truth. Then, these two types of mismatching pixels are measured by their ratios to the matching pixels, respectively. These two ratios are named as precision and recall. Finally, the F1measure combines the precision and recall with equal weight to evaluate the overall segmentation quality. Since the region-based measures are not sensitive to variation of the object's shape, they can only provide rough evaluation of segmentation quality, but cannot reflect human visual perception.

Different from the region-based measures, boundary-based measures need to extract the ground truth's and the segment's boundaries, then evaluate the segmentation quality by measuring the boundary similarity. In [8], Hausdorff distance was adopted to measure the maximum distortion of the segment's boundary. It can only reflect local distortion rather than the global one. Boundary displacement error as a global measure was proposed in [9], which used the mean of the shortest distances between the pixels on the boundaries of the ground truth and the segment to measure the similarity. It ignored the fact that the shortest distance is just a relative measure regardless of the image size. Some measures [10], [11] take human perception into account. In [10], fuzzy set theory was used to assign each pixel with two likelihood probabilities associated with two boundary sets. Then, the difference between these probabilities indicates the boundary distortion. Csurka and Perronnin [11] and Kohli et al [12] both introduced a concept of tolerance band, so the boundary can be adjusted since they assumed that errors can be tolerated within the band. The difference is that the bandwidth in [11] is adaptive to the image size rather than being fixed as in [12]. Although these boundary-based measures can measure shape variation and reflect the human visual property, they are still sensitive to distortions compared with region-based measures and cannot describe the different influence induced by different error positions.

Some measures can be treated as combinations of region-based and boundary-based measures. For example, Movahedi and Elder [5] proposed a mixed measure based on the average distance from mismatching pixels to their corresponding boundary. However, average value is not a good index for segmentation quality. In [13] and [14], mismatching pixels were assigned with different weights according to their shortest distances to the ground truth boundary and their positions. This strategy merely penalized some errors according to the human visual properties, but did not consider the compensation induced by the human visual properties. So, a better measure should not just combine the region-based and boundary-based measures together as discussed in [11], but also comprehensively integrate the human visual properties into it.

For multiple objects segmentation, Villegas and Marichal [13] proposed two different strategies. If correspondences can be established between objects in the ground truth and the segmentation result, the overall quality can be evaluated by pooling the segmentation quality of each individual object. Otherwise, multiple objects can be treated as a whole. For the first strategy, the pooling method is a linearly weighted addition of each individual object's quality. The weight of the individual object is determined by the area proportion of its corresponding object in the ground truth in order to reduce an excessively negative influence induced by badly evaluated small objects. However, this factor alone is not enough to describe human perception. For the second strategy, all the aforementioned measures for single object evaluation are valid. It is a simple approximation to the first one [13].

In this paper, we present our work from two aspects: Firstly, we construct a novel subjective object segmentation visual quality database [15] which consists of both single object and multiple objects segmentations, with a total of 255 segmentation results, and there are more than thirty viewers involved in each survey session. In addition to a brief introduction of the single object database in our previous conference paper [16], more technical details and analysis about our database are described in this paper. Secondly, we develop a novel objective measure based on four human visual properties for object segmentation visual quality evaluation. It includes a single object segmentation quality assessment method based on [16] and a pooling method for multiple separated objects segmentation.

This paper is organized as follows. Section II gives the details of our subjective database. Section III presents our proposed objective measure for object segmentation visual quality evaluation. The validity of our database and the performance of our measure are discussed in Section IV. We conclude our paper in Section V.

II. Object Segmentation Visual Quality Evaluation Database

In our database, as mentioned above, each segmentation result is assigned a corresponding ground truth and a subjective score. The main database construction stages include source images collection, segmentation results generation, subjective survey and data processing.

A. Source Images Collection

Subjects are sensitive to the distortions in the regions which bear semantic information. However, different objects have different semantic meaning. A comprehensive object segmentation quality evaluation database should include adequate diversity of object categories. The source images with corresponding ground truth in our database are selected from four well known public object segmentation database: Weizmann [17], VOC2012 [18], MSRA (Image B) [19] and Microsoft Research Cambridge's grabcut database [20] (it selects some images from Berkeley Segmentation Database [21]). The single object subset includes 85 source images covering diverse object categories, such as human, animal, car, boat, plant, everyday goods and building. These

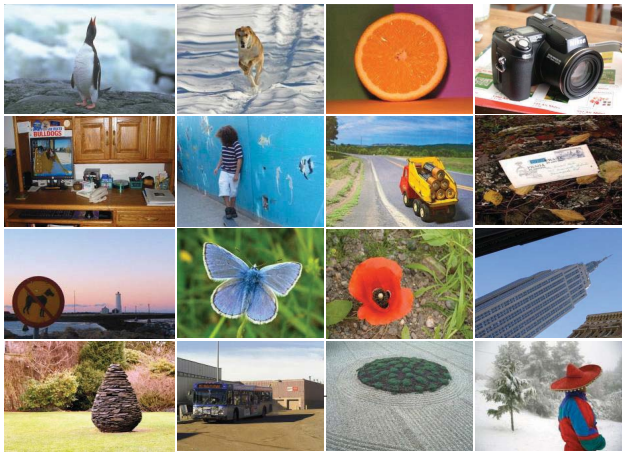


Fig. 1. Samples of the source images in the single object subset.

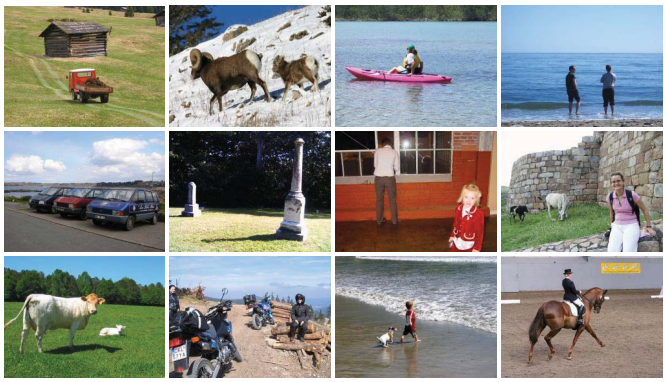


Fig. 2. Samples of the source images in the multiple objects subset.

objects present different sizes and rich texture characteristics. Fig. 1 shows a collection of the source images in this subset.

In the multiple objects subset, there are 35 source images with three possible object locations configurations, such as separateness, overlap and a mixture of both (some objects are separated and the rest are overlapping). Samples of the source images in this subset are shown in Fig. 2.

B. Segmentation Results Generation

A good quality evaluation database should satisfy two requirements: high coverage of possible distortion types and low redundancy of samples. In the single object subset, we consider Gelasca's four basic error types (added region, added background, border hole and inside hole) [14] and their eleven combinations ($C_4^2 + C_4^3 + C_4^4 = 11$), which can cover all possible errors. Examples of the four error types and one of their combinations are shown in Fig. 3. These four error types are also summarized in Table. I. Added region and added background are false positive, and border holes and inside holes are false negative. Added background and border holes are adjacent to the ground truth boundary, while added region and inside holes are not. One real segmentation example of the four error types is shown in Fig. 4. The segmentation quality is subjectively divided into five levels (Excellent, Good, Fair, Poor and Bad) by the authors. In the multiple objects subset, since each individual object has five quality levels as mentioned above, there are 31 possible types of quality level combinations ($C_5^1 + C_5^2 + C_5^3 + C_5^4 + C_5^5 = 31$) in one segmentation result. Note that the subscript "5" represents the five quality levels, and the superscript number indicates how many different levels appear in the result. For example, C_5^1 means that there exist five cases that the segmentation quality of all objects in the result is the same.

We adopt four interactive segmentation methods (namely seeded region growing [22], interactive graph cut [23], simple interactive object extraction [24] and interactive segmentation using binary partition trees [25]) provided by McGuinness' interactive segmentation tool [10], two semi-automatic segmentation methods (i.e., Li's Distance Regularized Level Set Evolution (DRLSE) method [26], and Mohit Gupta and Krishnan Ramnath's grabcut tool-box [27]) and two automatic object segmentation methods (i.e., Achanta's [28] and Rathu's [29]) to generate the segmentation results. The segmentation results generated by different segmentation methods have their own characteristics. For example, Rathu's method may generate large inside holes, while seed region growing will not produce any inside holes and added regions. Since the interactive methods use manual input, they can generate better segmentation quality compared with semi-automatic and automatic methods. This database focuses on evaluating the perceptual quality of the segmentation results rather than the performance of the segmentation methods. Therefore, the above methods satisfied our requirements to generate different types and qualities of segmentation errors.

In real practice, we find there are always some errors perceived on the object boundary more or less. It means that each segmentation result always has the added background or the border hole, and it is unimportant to pursue the results merely having added region, inside hole or their combinations as shown in Fig. 3(a), (d) and (e). So, we exclude these three errors types in the single object part. Although different segmentation methods have their own characteristics, they can still generate similar results in terms of error types and severity. Therefore, it is not necessary to collect all segmentation results into our database. In order to reduce the redundancy, we select 15 segmentation results with diverse object categories for each error type and cover the five perceptual quality levels mentioned above. Finally, there are totally 180 segmentation results in the single object part.

As mentioned above, five quality levels can generate 31 types of quality level combinations. But two cases (four different quality levels and five different quality levels in one result) are quite rare in the real segmentation results, which correspond to six possible combinations. Therefore, the multiple objects subset focuses on the other 25 quality level combinations. We select three segmentation results with different object content for each quality level combination.

C. Subjective Surveys

Compared with traditional subjective image/video quality evaluation, object segmentation subjective quality evaluation

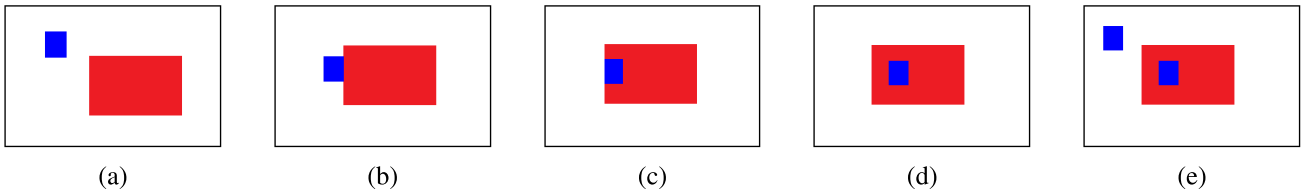


Fig. 3. Illustrations of Gelasca's four basic error types and one combination. The red rectangle and the blue rectangle represent the ground truth and the segmentation error respectively. (a) Added region, (b) Added background, (c) Border hole, (d) Inside hole and (e) The combination of the added region and the inside hole.

TABLE I
SUMMARIZATION OF GELASCA'S FOUR BASIC ERROR TYPES

Type	Abbreviation	Description
Added Region	AR	False positive error, nonadjacent to the ground truth boundary
Added Background	AB	False positive error, adjacent to the ground truth boundary
Border Hole	BH	False negative error, adjacent to the ground truth boundary
Inside Hole	IH	False negative error, nonadjacent to the ground truth boundary

Fig. 4. (a) Original image [19], (b) ground truth [28], (c) segmentation result [29] and (d) segmentation errors [14]. In (d), the purple regions are added regions, the blue regions are added background, the orange regions are border holes and the red region is an inside hole.

lacks standard methodologies. In order to overcome the ambiguity in the segmentation results and make the object being evaluated more apparent, we adopt the simultaneous double stimulus for discrete evaluation (SDSDE) [30] method to conduct our subjective survey. This method is based on the simultaneous double stimulus for continuous evaluation (SDSCE) method, which is a standard subjective evaluation method specified by ITU-R BT.500-13 [31]. The only difference in SDSDE is the use of the absolute category rating (ACR) scale which employs a five-grade discrete segmentation quality scale (5: Excellent, 4: Good, 3: Fair, 2: Poor, 1: Bad) [32], which conforms to the segmentation quality levels in our database. In [33], the experimental data has demonstrated that there are no obvious overall statistical differences between the different rating scales. Therefore, the five-grade discrete scale is employed to reduce the viewer's fatigue and make the subjective rating more distinguishable [30]. The subjective survey interface is shown in Fig. 5. The source image plays an auxiliary role to help the viewers understand the image content. By comparing the segmentation result with the corresponding ground truth, viewers can select the subjective ratings. The subjective ratings are recorded in numerical values according to the 5-grade quality scale during the subjective survey.

In order to analyze the relationship between the individual object quality and the overall quality, an additional rating step is designed for the multiple separated objects case. The procedure is illustrated in Fig. 6. The individual objects are extracted and viewers are asked to rate their quality one by one. Therefore, there are totally 352 segmentation results in the subjective survey. Considering the constraint on survey duration to reduce the effect of the viewers' fatigue, the survey is divided into two sessions (176 images in each session). Since human visual comparison does not satisfy the axiom of symmetry [34], the order of observation of different results could influence human judgment. In order to avoid the contextual and memory effects on the subjective quality ratings, the image triplet (the source image, the ground truth and the segmentation result) are randomly presented to each viewer [31]. Furthermore, the segmentation results generated from the same source image except the multiple separated objects case will not be presented consecutively. Note that multiple separated objects case is treated as one rating group, and the extracted individual object quality rating follows consecutively the overall quality rating.

Before the formal subjective survey, three pre-sessions are arranged. Firstly, a brief introduction of the objective of this survey and how to do the quality evaluation is presented

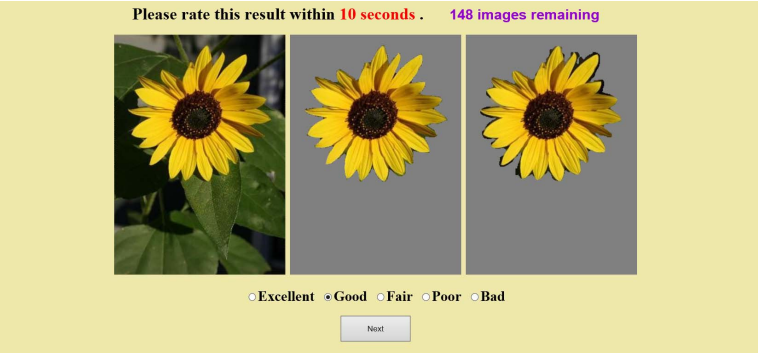


Fig. 5. The interface of our subjective survey [16]. The images from left to right are: the source image, the ground truth and the segmentation result.

Fig. 6. An example of the rating group of one multiple separated objects segmentation result. The number in each ground truth indicates the rating order.

to the viewers. Then, a training session is conducted. There are a total of 10 segmentation results in this session, which includes single object and multiple objects cases with the different perceptual qualities ranging from "Excellent" to "Bad". A quality scale is suggested to the viewers for each training result. It is emphasized that the viewers should rate the segmentation result independently of the suggested quality scale. At the end of this session, viewers should pass a "Ground truth/Ground truth" pair test [31] in order to ensure that the viewers fully understand the survey methodology. Otherwise, the viewers will be trained again until they pass the test. In the next session, a mock evaluation session using 6 segmentation results is conducted to consolidate the viewers' training.

All the viewers in the subjective test are the students from the Chinese University of Hong Kong and Shanghai University. There are 33 and 31 viewers who participated in the Session 1 and Session 2 respectively. Most of them are naive viewers who have little experience on image processing.

D. Data Processing

In order to obtain the final mean opinion score (MOS) and standard deviation value for each segmentation result, the raw data processing as suggested by [30] is adopted. Firstly, the raw score r_{ijk} is converted into $Z\text{-}score_{ijk}$ in order to reduce the negative influence introduced by the viewer's rating habit [35], where ijk indicates the ith viewer rating the jth segmentation result in the session $k = \{1, 2\}$. Then, a standard screening procedure [31] is conducted to reject the unreliable viewers. After this procedure, 3 out of 33 viewers and 4 out of 31 viewers are rejected in Session 1 and Session 2, respectively. Assuming that the $Z\text{-}scores$ assigned by a viewer follow the Gaussian distribution, then 99% of the scores will lie in the range $[-3, +3]$ [36]. The $Z\text{-}scores$ are then linearly mapping to the range of [0, 100] by:

$$\overline{Z}_{ijk} = \frac{100(Z\text{-}score_{ijk} + 3)}{6} \tag{1}$$

Finally, the MOS value and the standard deviation of each segmentation result are computed as follows:

$$MOS_{jk} = \frac{1}{N_k}\sum_{i=1}^{N_k}\overline{Z}_{ijk} \tag{2}$$

$$std_{jk} = \sqrt{\frac{1}{N_k - 1}\sum_{i=1}^{N_k}(\overline{Z}_{ijk} - MOS_{jk})^2} \tag{3}$$

where N_k is the number of remaining viewers of session k after the screening procedure. The MOS value and the standard deviation are treated as the ground truth representing the perceptual quality of the segmentation result. The histogram of the MOS values is shown in Fig. 7. We can see that the MOS values occupy a wild range of perceptual quality from low to high.

E. Analysis of the Subjective Survey

In the single object subset, we find that the subjective evaluation of single object segmentation visual quality can be divided into three levels:

1) In the low level, viewers use low features, such as area and boundary, to measure the similarity between the ground truth and the segmentation result.

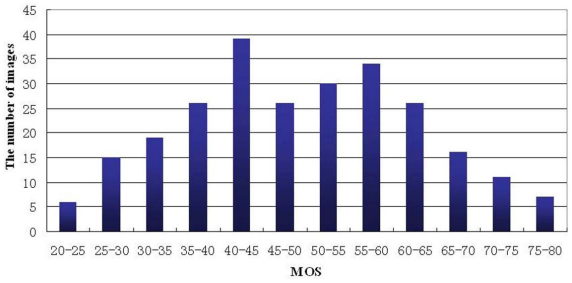


Fig. 7. The histogram of the MOS values.

(a) (b)

Fig. 8. An example of human visual tolerance. (a) Ground Truth, (b) Segmentation Result. The blue-circled part is far worse than the red-circled part in terms of visual quality.

2) In the middle level, error position is an important factor that affects the visual quality. Viewers pay more attention to errors inside the object. When viewing similar areas, the errors which make the object incomplete rather than add redundant background can lead to worse visual quality.
3) In the high level is related to the object semantics. For example, even though the errors all happen at the interior of the object, the visual quality can be diverse since different regions may have different semantic information. Some regions with rich semantic information, such as face and sign, should not be distorted. Otherwise, the visual quality is unacceptable.

For multiple objects part, especially the separateness cases, the overall quality mainly depends on the worst individual object quality, but it is also influenced by area proportion among the individual objects. These observations inspire us to design the objective measure accordingly.

III. Proposed Objective Measure for Object Segmentation Visual Quality Evaluation

Human visual properties are important cues for designing objective measures. In [6], [11], [12], and [37], four human visual properties for object segmentation quality evaluation are discussed. Based on their conclusions, these four properties can be interpreted as:

1) Human can put up with added background and border holes to some extent, i.e., human visual tolerance [10], [11].
2) It is easier for human to quantify small errors, but more difficult for larger ones, i.e. human visual saturation [10].
3) The perceptual importance of false negative and false positive pixels are different [5].
4) The overall quality is mostly determined by the stronger distortions [37].

Our objective measure comprises two parts. One part is single object quality evaluation based on the first three human visual properties. The other part is the pooling method for multiple separated objects based on the fourth property.

A. Single Object Quality Evaluation

The problem formulation is introduced below. Assuming there are n separated objects (overlapped objects are treated as one object in our measure) in the ground truth. It can be represented as $G = \{G_{O_1}, G_{O_2}, \ldots, G_{O_n}\}$. Each object in the ground truth can be assigned a corresponding segment in the segmentation result, so $S = \{S_{O_1}, S_{O_2}, \ldots, S_{O_n}\}$. These correspondences are determined by the shortest distances of pixels in the segmentation result to the objects in the ground truth. Our goal is to measure the similarity $Sim(G, S)$ between the ground truth G and the segmentation result S. There are only G_{O_1} and S_{O_1} in the single object case, so we use G and S for short. Our measure for single object is defined as Eq. (4), as shown at the bottom of this page, where $A(\cdot)$ is the operation of calculating area. Our measure is based on Jaccard Index [5] with two additional terms: compensation term $COM(\cdot)$ and penalty term $PEN(\cdot)$.

1) Compensation Term: The compensation term reflects human visual tolerance. As shown in Fig. 8, we take two local regions as an example. The red-circled added background is along the ground truth boundary. It roughly maintains the shape of the true boundary. However, the blue-circled added background is protruding which destroys the shape of the true boundary. Although these two parts have similar error area, the blue part is far worse in terms of perceptual quality [5]. In other words, we can tolerate the red-circled errors to some extent, thus compensating it. For pixel i in the added background (AB), if it could be tolerated, we can treat it like a part of the true positive rather than an error. Since the numerator of the Jaccard Index $A(G \cap S)$ is the area of the true positive, we add a compensation value $COM(i)$ to the numerator. Meanwhile, if we can tolerate pixel j in border holes (BH), it can be regarded as the true background rather than a missing part. So its compensation value $COM(j)$ is subtracted from the denominator of the Jaccard Index $A(G \cup S)$ which includes the area of missing parts. Since added regions (AR) and inside holes (IH) do not distort object's boundary, we do not assign compensation values to pixels in these regions.

$$Sim(G,S) = \frac{A(G \cap S) + \sum_{i \in AB} COM(i)}{A(G \cup S) - \sum_{j \in BH} COM(j) + \sum_{j \in BH} PEN(j) + \sum_{k \in IH} PEN(k)} \tag{4}$$

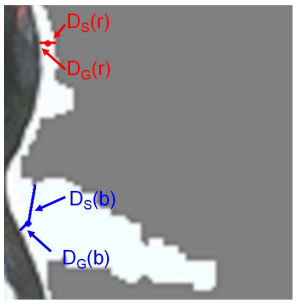
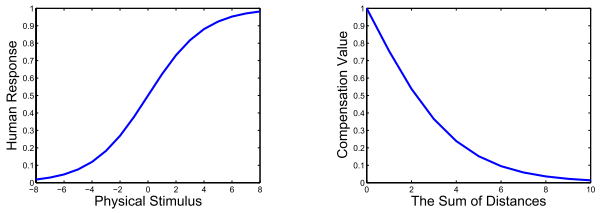


Fig. 9. Zoom in of the segmentation result. "r" and "b" represent the red pixel and blue pixel respectively.



(a) (b)

Fig. 10. The comparison of two curves with $\sigma = 2$. (a) Logistic function curve, (b) Compensation function curve.

The shortest distances of pixel i to the ground truth boundary and the segmentation result boundary are two important factors to measure the tolerance degree [10]. We select one region of the segmentation result and its zoom-in, which is shown in Fig. 9. The red and blue pixels have similar shortest distances to the ground truth boundary. However, the blue pixel has much longer distance to the segmentation result boundary. Since the blue pixel belongs to the protruding error region, it should have less compensation value than the red one. In terms of the sum of the two distances, a smaller sum indicates that this pixel contribute more to maintaining the object's shape well, and vice versa. According to this design, the compensation values of the red pixel and the blue pixel in Fig. 9 can be distinguished. Furthermore, we take the human visual saturation effect into account, i.e., when the object shape is heavily distorted, viewers find it hard to quantify the severity of errors [10]. In other words, when the sum of the two shortest distances is larger, the compensation value should be lower and its variation should become smaller. Logistic function [38] as shown in Fig. 10(a) is one type of psychometric functions which establish a relationship between physical stimulus and the human response. The following formula is the definition of the logistic function.

$$y = \frac{1}{1 + \exp(-x/\sigma)} \tag{5}$$

where σ is the bandwidth of the exponential function. x and y can be treated as the physical stimulus and the human response, respectively. From Fig. 10(a), we can see that the curve becomes approximately horizontal when the physical stimulus is large, which is consistent with the human visual saturation effect mentioned above. According to the above analysis, we adjust the logistic function to define $COM(i)$ as follow:

$$COM(i) = 2 - \frac{2}{1 + \exp(-(D_G(i) + D_S(i))/\sigma)} \tag{6}$$

where $D_G(i)$ and $D_S(i)$ are the shortest distances of the ith pixel to the ground truth boundary and the segmentation result boundary, respectively. $COM(j)$ can be similarly defined. Thus, the sum of these two shortest distances is treated as the physical stimulus, and the output corresponds to the compensation value. If the sum is too large, this pixel severely distorts the boundary's shape, and human can tolerate it no more. So no compensation value should be assigned to this pixel. We control compensation value by setting a bandwidth [16].

$$2 - \frac{2}{1 + \exp(-R/\sigma)} = \tau \tag{7}$$

Or equivalently,

$$\sigma = -R\left(\ln \frac{\tau}{2 - \tau}\right)^{-1} \tag{8}$$

where R is a reference length and τ is the corresponding compensation value of R. In [11], R is adaptive to the image size. We follow this approach making $R = \alpha \cdot D_{length}$ where α is a constant and D_{length} is the diagonal length of the image. τ and α are a pair of parameters. By setting α and τ appropriately, the compensation value could approach 0 when the sum is too large. The compensation function is drawn in Fig. 10(b) with $\sigma = 2$. From this figure, we can see that the compensation value is large when the sum is small. Conversely, when the sum becomes larger, the compensation value is lower and its variation also becomes smaller. These observations demonstrate that our compensation function properly reflects human visual tolerance and saturation.

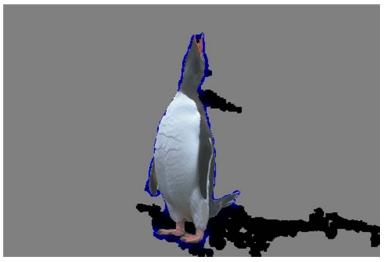


Fig. 11. Compensation value map.

We generate a compensation value map as shown in Fig. 11. The lighter blue pixels indicate larger compensation values. From this map, we can see that some added background regions along the ground truth boundary possess larger compensation values while the severely protruding region have compensation values closely approach to 0 (black), which conforms to the viewers' perception.

2) Penalty Term: The penalty terms mainly describe how the human visual system assigns different weighting to visual quality. Fig. 12 shows a typical example about this property in our database. Although the right image has larger error area, it still obtains a higher MOS. The reason is that false negative pixels (border holes and inside holes) degrade the object itself, which makes it incomplete in the left image; hence, viewers tend to assign higher weights to false negative pixels,

(a) (b)

Fig. 12. An example of different visual quality weights assignment. (a) MOS = 43.551, (b) MOS = 52.301.

(a) (b)

Fig. 13. An example of multiple objects segmentation result. (a) Ground truth, (b) Segmentation result.

and give those results which maintain the completeness of the object better quality ratings. The penalty terms are defined as follows:

$$PEN(k) = \beta \quad k \in IH \tag{9}$$

$$PEN(j) = \beta(1 - COM(j)) \quad j \in BH \tag{10}$$

where β is a constant. For pixels in the inside holes, we assign higher weights. Meanwhile, since pixels in the border holes can be tolerated to some extent, they are given lower weighting.

The formulation of Eq. (4) can also be explained from the perspective of the attributes. Each pixel in the segmentation result is assigned with a basic region attribute by Jaccard Index, i.e., whether it belongs to the true object or not. On the one hand, since added background and border holes are along the boundary, part of their pixels may carry boundary attribute. However, pixels in border holes and inside holes still carry the object completeness attribute. Therefore the boundary attribute may compensate for the loss of region attribute. On the other hand, completeness attribute can aggravate the loss of region attribute. We achieve the combination of region cue, boundary cue and human visual properties by integrating the pixels' multiple attributes.

B. Multiple Separated Objects Quality Evaluation

In this part, we assume that each object's quality has been evaluated by its own single object measure described above. Here, we focus on developing the pooling method to evaluate the overall quality. As mentioned in [13], the perceptual relevance of larger objects is usually higher. Therefore area proportion is an important cue for each object's weight. However, we should not only consider the area proportion of objects in the ground truth but also take into account corresponding segmented objects in the result. One example in our database is shown in Fig. 13. The smaller earthenware on the left in the ground truth is wrongly segmented to be larger in the segmentation result. This area variation makes the left earthenware more visually prominent, thus should be assigned with a higher weight. In addition, we need to consider the human visual property that overall quality is mostly determined by the stronger distortions.

Our pooling method is thus defined as:

$$SimM = \frac{1}{\sum_{p=1}^{n} (w(p) \times \frac{1}{sim(G_{O_p}, S_{O_p})})} \tag{11}$$

$$w(p) = \frac{A(p)}{\sum_{q=1}^{n} A(q)} \tag{12}$$

$$A(p) = A(G_{O_p}) \bigcup A(S_{O_p}) \tag{13}$$

where $A(\cdot)$ represents the area. We use the union area of G_{O_p} and S_{O_p} to measure S_{O_p}'s weight $w(p)$, which covers both cases mentioned above. Our pooling method is a weighted harmonic mean, whose output is mainly determined by the smallest one. We take advantage of this characteristic to describe the human visual property. So, after weighting the overall quality evaluated by our measure strongly depends on the worst case.

IV. Experimental Results

A. Subjective Agreement

Since all segmentation results in the database are subjectively selected by us, it is necessary to firstly test whether viewers can achieve similar opinions on these results before any processing on viewers' ratings. If the ratings are quite different, the MOS as mean value is meaningless and our selection of these results fails the consistency requirement. In [30], each viewer's ratings compose of a vector, and the normalized cross correlation (NCC) and the Euclidean distance (EUD) are used to measure the correlation between two vectors. NCC and EUD are defined as:

$$NCC = \frac{{v_1}^t \cdot v_2}{\| v_1 \| \| v_2 \|} \tag{14}$$

$$EUD = \frac{\| v_1 - v_2 \|}{d} \tag{15}$$

where v_1 and v_2 are the two vectors corresponding to two viewers' ratings and d denotes the dimension of the vector. The higher value of NCC and the lower value of EUD indicate high correlation between two rating vectors. We follow this method to test the subjective agreement of our database. Since there are 33 and 31 viewers in Session 1 and Session 2, 528(C_{33}^2) and 465(C_{31}^2) values of NCC and EUD are obtained from each session, respectively. The average NCC values are 0.95 and 0.94 for Sessions 1 and 2, and the average EUD values are roughly 0.09 for both sessions. The large NCC values indicate that angular difference between every two rating vectors is very small. Meanwhile, the small EUD values reflect the small magnitude difference between the two rating vectors. These observations demonstrate that the viewers have very good agreement on the segmentation results'

TABLE II
PERFORMANCES OF SEVEN QUALITY MEASURES FOR SINGLE OBJECT SEGMENTATION

	SROCC	LCC	RMSE	OR
MM	0.570	0.320	13.290	0.161
F1	0.848	0.850	7.389	0.028
JI	0.848	0.851	7.374	0.028
FC	0.835	0.835	7.712	0.017
BF	0.856	0.857	7.258	0.011
WQM	0.870	0.859	7.187	0.011
OUR	0.912	0.909	5.859	0

visual quality in our database It means that our selections are reasonable and MOS is a reliable index to measure the visual quality.

B. Overall Performance

Since our objective measure consists of the single object measure and the pooling method, we test them respectively using our object segmentation visual quality database. Following the work of Video Quality Expert Group [39], each objective score x is mapped to $Q(x)$ by fitting the following function in order to obtain a linear relationship with MOS:

$$Q(x) = \beta_1 \times (0.5 - \frac{1}{1 + \exp(\beta_2 \times (x - \beta_3))}) + \beta_4 \times x + \beta_5 \tag{16}$$

For the single object measure, we evaluate its performance on the single object subset of our database. We use four common performance evaluation criteria, i.e., the Spearman Rank-Order Correlation Coefficients (SROCC), the Linear Correlation Coefficient (LCC), the root mean squared error (RMSE), and the outlier ratio (OR), which use $Q(x)$ and MOS as their inputs [40]. Higher values of the first two criteria indicate better performance, while the last two are on the reverse. Our measure is compared against six object segmentation evaluation measures which are Jaccard Index (JI) [6], F1measure (F1) [7], Fuzzy Contour (FC) [10], Boundary F1measure (BF) [11], Mixed Measure (MM) [5] and normalized spatial Weighted Quality Measure (WQM) [13]. In this experiment, we empirically set $\alpha = 0.02$, $\tau = 0.1$ and $\beta = 2$. The comparison results are shown in Table II. From the table, we can see that Mixed Measure's performance is the worst. It demonstrates that the average distance has the lowest correlation with the human perception. It also lacks a suitable normalization step which causes mapping failure, leading to much lower LCC and higher RMSE. Two region-based measures, Jaccard Index and F1measure have similar performance. Without involving human visual properties, they cannot achieve higher predictive accuracy. Compared with Fuzzy Contour, Boundary F1measure not only integrates human visual perception, but also adjusts the tolerance band according to the image size. So its performance is better than that of Fuzzy Contour whose parameters are fixed, and even better than those of the two region-based measures. Since Weighted Quality Measure has merits from both region-based and boundary-based measures, its performance is the second best. Our measure achieves the best performance in terms of all four criteria. It indicates our "compensation" and "penalty" strategy not only reasonably organizes region and boundary information, but also suitably takes the human visual properties into account. It allows our objective measure to more closely approximate the human judgment. Examples shown in Fig. 14 exhibit the adjustment of our measure based on Jaccard Index. The Jaccard Index values of the left and middle figures are nearly the same. However, MOS of the middle figure is obviously higher than that of the left one. Since a part of the added background is along the star's boundary which roughly maintains the shape of the star, our measure increases the Jaccard Index value by assigning the compensation value to this region. Meanwhile, although the Jaccard Index value of the right figure is much higher than that of the middle one, their MOSs are similar. Our measure penalizes the border holes which destroy the leaf's shape significantly. These two adjustments make our measure more consistent with the MOS.

Fig. 14. From left to right, their MOSs are 36.453, 46.548 and 45.633, respectively; the Jaccard Index are 0.656, 0.658 and 0.838; the quality scores given by our measure are 0.660, 0.721 and 0.708.

TABLE III
PERFORMANCES OF FOUR POOLING METHODS FOR COMPONENT PERFORMANCE ANALYSIS

	SROCC	LCC	RMSE	OR
GA + LS	0.883	0.907	4.547	0.234
UA + LS	0.886	0.911	4.461	0.234
GA + HM	0.910	0.936	3.804	0
UA + HM	0.916	0.941	3.654	0

We use 42 groups of multiple separated objects to test the performance of the proposed pooling method. Since we have already derived each individual object's MOS in each group by the subject survey, the pooling method becomes the only factor which determines the predictive accuracy of the overall quality. We directly evaluate the overall quality by balancing the individual objects' MOSs. We also compare our pooling method with the typical method adopted in [13]. The typical method can be generalized as "GA (object area in the ground truth) + LS (linear summation)". Correspondingly, our method can be generalized as "UA (union area) + HM (harmonic mean)". In order to evaluate the effectiveness of the processing components, we develop another two combinations "UA + LS" and "GA + HM". The comparison results are shown in Table III. This table clearly shows that each component of our pooling method contributes to the overall performance. Compared with the union area for calculating the weight, the harmonic mean which properly describes the human visual perception plays a more important role.

TABLE IV
PERFORMANCE OF FOUR OBJECT SEGMENTATION VISUAL QUALITY MEASURES

	SROCC	LCC	RMSE	OR
JI + GA + LS	0.797	0.807	7.922	0.031
BF + GA + LS	0.813	0.820	7.680	0.028
WQM + GA + LS	0.845	0.834	7.394	0.016
Our	0.890	0.885	6.231	0

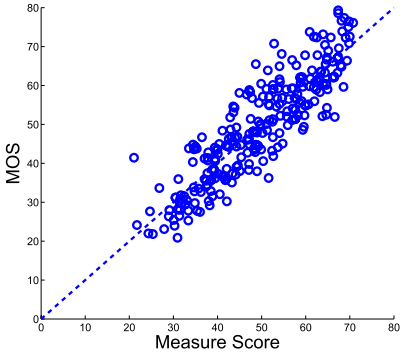


Fig. 15. Scatter plots of our measure on the entire database (after the nonlinear mapping).

As shown in the Table III, the gains provided by HM over LS are much larger than UA over GA. The experimental results demonstrate that our pooling method is better at balancing individual object quality.

Finally, we combine our single object measure and the pooling method to form the complete objective object segmentation visual quality measure. We test it on our database including all 255 segmentation results (180 single object and 75 multiple objects results). Table. IV lists the overall performance. The scatter plot of our measure on the entire database is shown in Fig. 15, where each circle represents one result. The vertical axis denotes the MOS and the horizontal axis denotes the nonlinearly mapped measure output $Q(x)$.

Since Jaccard Index, Boundary F1measure and Weighted Quality Measure have better performance in the previous test, we combine them with the typical pooling method [13] and compare with our measure. The experimental results demonstrate that our measure outperforms the others for object segmentation visual quality evaluation on all possible cases in our database, as shown in Table IV.

C. Parameterization

In the proposed measure, there are three parameters to be determined, i.e., compensation value τ, length proportion α and additional weight β. As we have mentioned, τ and α are a pair of parameters to control the bandwidth in our compensation function. So, it is not necessary to tune these two parameters simultaneously. Thus, we fix $\tau = 0.1$ and seek for the best α.

We simply select SROCC as the index to show our measure's performance under different α and β values on the single object subset. The result is shown in Fig. 16. From this figure, we can see that our measure maintains

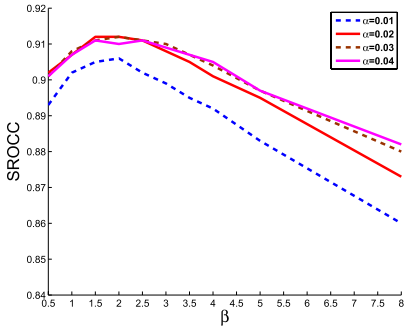


Fig. 16. SROCC of our measure using different α and β values when $\tau = 0.1$.

(a) (b)

Fig. 17. Two compensation value maps using different α. (a) $\alpha = 0.02$, (b) $\alpha = 0.04$.

good performance under different parameter values. The performance of our measure is not very sensitive to these two parameters. It means the design of our measure itself is reasonable which can tolerate certain variation of parameters. According to the SROCC value, $\alpha \in \{0.02, 0.03, 0.04\}$ and $\beta \in [1.0, 3.0]$ are the better choices. However, if we set $\alpha = 0.04$, the compensation range is a little too wide to meet our intuition. One example is shown in Fig. 17. Compared with the compensation value map on the right, the left one better conforms to our intuition that the arms and left legs should not have very high compensation values. Therefore, we set $\alpha = 0.02$, $\tau = 0.1$ and $\beta = 2$ in our measures.

D. Statistical Significance

In order to verify the statistical significance of the predictive accuracy improvement induced by our measure, F-test [41] is conducted on the residuals between the measure outputs (after nonlinear mapping) and MOS. Since the assumption of F-test is that the residual distribution is Gaussian, we need to test whether this assumption is satisfied before we conduct the F-test. A simple criterion proposed in [2] is used to measure the Gaussianity of the residuals: if the residuals have a kurtosis between 2 and 4, they are taken to be Gaussian. After that, we calculate the ratio between the residual variance of our measure and that of compared measure (with the larger variance as the numerator). If the ratio is larger than $F_{critical}$ which is a threshold based on the number of residuals and a given confidence level, then the gain induced by our measure is considered to be significant at the specified confidence level. The F-test results on single object subset and the entire database are given in Table V, where "1" indicates Gaussian, (0) indicating non-Gaussian and the confidence level of

TABLE V
F-TEST RESULTS ON OUR DATABASE

	Single object subset $F_{critical} = 1.279$					All results $F_{critical} = 1.229$		
	F1(1)	JI(1)	FC(1)	BF(1)	WQM(1)	JI+TP(1)	BF+TP(1)	WQM+TP(1)
OUR(1)	1.589	1.581	1.732	1.534	1.505	1.623	1.525	1.414

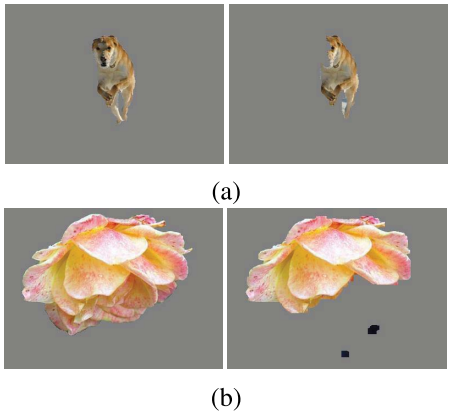


Fig. 18. One example of our failure case. In each pair, the left image is the ground truth and the right image is the segmentation result. The visual quality is indicated by MOS, and evaluated by Jaccard Index and our measure. (a) MOS=32.849, JI=0.744 and OUR=0.575, (b) MOS=46.683, JI=0.744 and OUR=0.512.

$F_{critical}$ is 0.95. As shown in Table IV, all residuals are Gaussian and all ratios are larger than the corresponding $F_{critical}$. It means that the proposed measure outperforms all competitors statistically.

E. Discussion

From the experimental results, we can see that our measure correlates better with the subjective segmentation quality. It can be treated as a candidate benchmark to evaluate an object segmentation method's performance in terms of human perceptual quality. But, there is still about 10% gap between the predictive value and MOS. One reason is that we treat overlapping objects as a whole, which is not so precise. Another reason is about the semantic information. As we have analyzed, subjective evaluation can be divided into three levels from low to high. Our measure which currently stays at low and middle levels which measure the area, the boundary and the region weights. However, it does not describe the semantic information which belongs to the high level. This is the main reason for the performance gap. One failure case of our measure is shown in Fig. 18. Since the face contains important semantic information, distortions on the face should lead to much worse visual quality when compared with lossing a part of the petals. However, our measure predicts that the segmentation result in Fig. 18(a) is better which is not consistent with MOS. Note that our measure provides a good interface to involve the semantic information. The penalty terms can be improved to evaluate the loss of semantic information. More specifically, the penalty value can be varied according to the semantic importance rather than being constant as in the current measure.

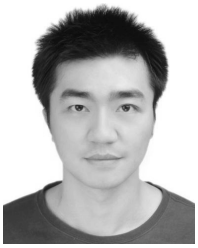
V. CONCLUSION

A subjective segmentation visual quality database is constructed and introduced in this paper. There are totally 255 segmentation results involving single object and multiple objects in the database, and more than thirty viewers participated in each survey session. By analyzing the database, subjective evaluation for single object is divided into three levels and a subjective pooling method for multiple objects is also verified. In the proposed objective measure, the compensation term describes the human visual tolerance and saturation; the penalty term mainly reflects perceptual importance of different error positions; and the harmonic mean is used to approximate the subjective pooling method. Moreover, the compensation and penalty terms are also used in the design of the objective measure in terms of the pixel attribute. In future work, we will integrate semantic information into our measure, which is treated as the high level factor in subjective evaluation of object segmentation visual quality.

REFERENCES

[1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[2] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[3] H. Zhang, J. E. Fritts, and S. A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Comput. Vis. Image Understand.*, vol. 110, no. 2, pp. 260–280, 2008.
[4] N. Ponomarenko, V. V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.
[5] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE Comput. Soc. Conf. Workshops Comput. Vis. Pattern Recognit. (CVPRW)*, Jun. 2010, pp. 49–56.
[6] G. Feng, S. Wang, and T. Liu, "New benchmark for image segmentation evaluation," *J. Electron. Imag.*, vol. 16, no. 3, p. 033011, Jul. 2007.
[7] D. M. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.
[8] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.
[9] J. Freixenet, X. Muñoz, D. Raba, J. Martí, and X. Cufí, "Yet another survey on image segmentation: Region and boundary information integration," in *Proc. 7th Eur. Conf. Comput. Vis. (ECCV)*, 2002, pp. 408–422.
[10] K. McGuinness and N. E. O'Connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognit.*, vol. 43, no. 2, pp. 434–444, Feb. 2010.
[11] G. Csurka, D. Larlus, and F. Perronnin, "What is a good evaluation measure for semantic segmentation?" in *Proc. 24th Brit. Mach. Vis. Conf. (BMVC)*, 2013, pp. 1–11.
[12] P. Kohli, L. Ladický, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *Int. J. Comput. Vis.*, vol. 82, no. 3, pp. 302–324, May 2009.
[13] P. Villegas and X. Marichal, "Perceptually-weighted evaluation criteria for segmentation masks in video sequences," *IEEE Trans. Image Process.*, vol. 13, no. 8, pp. 1092–1103, Aug. 2004.

[14] E. D. Gelasca and T. Ebrahimi, "On evaluating video object segmentation quality: A perceptually driven objective metric," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 319–335, Apr. 2009.
[15] R. Shi, K. N. Ngan, and S. Li. *Image Object Segmentation Visual Quality Evaluation Database*. [Online]. Available: http://www.ee.cuhk.edu.hk/~rshi/, accessed Jan. 2015.
[16] R. Shi, K. N. Ngan, and S. Li, "Jaccard index compensation for object segmentation evaluation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 4457–4461.
[17] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.
[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results*. [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, accessed May 2013.
[19] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[20] C. Rother, V. Kolmogorov, and A. Blake, "'GrabCut': Interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.
[21] 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 *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 2. Jul. 2001, pp. 416–423.
[22] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994.
[23] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1. Jul. 2001, pp. 105–112.
[24] G. Friedland, K. Jantz, and R. Rojas, "SIOX: Simple interactive object extraction in still images," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2005, pp. 1–7.
[25] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 561–576, Apr. 2000.
[26] C. Li, C. Xu, C. Gui, and M. D. Fox, "Distance regularized level set evolution and its application to image segmentation," *IEEE Trans. Image Process.*, vol. 19, no. 12, pp. 3243–3254, Dec. 2010.
[27] M. Gupta and K. Ramnath. *Interactive Segmentation Tool-Box*. [Online]. Available: http://www.cs.cmu.edu/~mohitg/segmentation.htm, accessed Jun. 2013.
[28] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1597–1604.
[29] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proc. 11th Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 366–379.
[30] L. Ma, W. Lin, C. Deng, and K. N. Ngan, "Image retargeting quality assessment: A study of subjective scores and objective metrics," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 626–639, Oct. 2012.
[31] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.500-13, 2012.
[32] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document Rec. ITU-T P.910, 2012.
[33] Q. Huynh-Thu, M.-N. Garcia, F. Speranza, P. Corriveau, and A. Raake, "Study of rating scales for subjective quality assessment of high-definition video," *IEEE Trans. Broadcast.*, vol. 57, no. 1, pp. 1–14, Mar. 2011.
[34] Y. Gavet, M. Fernandes, J. Debayle, and J.-C. Pinoli, "Dissimilarity criteria and their comparison for quantitative evaluation of image segmentation: Application to human retina vessels," *Mach. Vis. Appl.*, vol. 25, no. 8, pp. 1953–1966, 2014.
[35] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," *Proc. SPIE, Adv. Image Video Commun. Storage Technol.*, vol. 2451, pp. 90–101, Feb. 1995.
[36] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[37] S. Li, L. C.-M. Mak, and K. N. Ngan, "Visual quality evaluation for images and videos," in *Multimedia Analysis, Processing and Communications*. Berlin, Germany: Springer-Verlag, 2011, pp. 497–544.
[38] F. A. Wichmann and N. J. Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit," *Perception Psychophys.*, vol. 63, no. 8, pp. 1293–1313, Nov. 2001.
[39] Video Quality Expert Group (VQEG). (2003). *Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment II*. [Online]. Available: http://www.vqeg.org
[40] S. Li, F. Zhang, M. Lin, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.
[41] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. New York, NY, USA: Wiley, 1999.



Ran Shi (S'14) received the B.S. degree in electronic science and technology from the Changshu Institute of Technology, and the M.S. degree in signal and information processing from Shanghai University, in 2009 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, The Chinese University of Hong Kong. His research interests include object segmentation visual quality evaluation, interactive segmentation, and salient object detection.

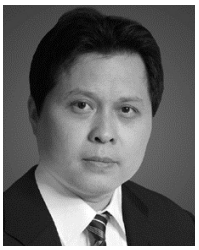
King Ngi Ngan (F'00) received the Ph.D. degree in electrical engineering from Loughborough University, U.K. He was a Full Professor with Nanyang Technological University, Singapore, and the University of Western Australia, Australia. He has been appointed as a Chair Professor with the University of Electronic Science and Technology, Chengdu, China, since 2012, under the National Thousand Talents Program. He is currently a Chair Professor with the Department of Electronic Engineering, The Chinese University of Hong Kong. He holds honorary and visiting professorships of numerous universities in China, Australia, and South East Asia.

Prof. Ngan has published extensively, including three authored books, seven edited volumes, over 370-refereed technical papers, and edited nine special issues in journals. In addition, he holds 15 patents in the areas of image/video coding and communications. He served as an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the *Journal on Visual Communications and Image Representation*, the *EURASIP Journal of Signal Processing: Image Communication*, and the *Journal of Applied Signal Processing*. He was the Chair and Co-Chair of a number of prestigious international conferences on image and video processing, including the 2010 IEEE International Conference on Image Processing, and served on the advisory and technical committees of numerous professional organizations.

Prof. Ngan is a fellow of IET (U.K.), and IEAust (Australia), and an IEEE Distinguished Lecturer in 2006-2007.

Songnan Li (M'13) received the B.Sc. and M.Phil. degrees in computer science and technology from the Harbin Institute of Technology, China, in 2004 and 2006, respectively, and the Ph.D. degree in electronic engineering from The Chinese University of Hong Kong (CUHK), in 2012. He joined CUHK in 2007, as a Research Assistant. From 2012 to 2014, he was appointed as a Post-Doctoral Fellow with the Department of Electronic Engineering, CUHK, where he is currently a Research Assistant Professor with the Department of Electronic Engineering. His research interests include image and video processing, RGB-D computer vision, and visual quality assessment.

Raveendran Paramesran (SM'01) received the B.Sc. and M.Sc. degrees in electrical engineering from South Dakota State University, Brookings, SD, USA, in 1984 and 1985, respectively, and the Ph.D. degree in engineering from the University of Tokushima, in 1994. He was a Systems Designer with Daktronics, USA. He joined the Department of Electrical Engineering, University of Malaya, Kuala Lumpur, in 1986, as a Lecturer. In 1992, he received a Ronpaku Scholarship from Japan to pursue Doctorate degree. He was promoted to Associate Professor and Professor, in 1995 and 2003, respectively. His research areas include image and video analysis, formulation of new image descriptors for image analysis, fast computation of orthogonal moments, analysis of EEG signals, and data modeling of substance concentration acquired from noninvasive methods. His contributions can be seen in the form of journal publications, conference proceedings, authored books, chapters in books, and an international patent to predict blood glucose levels using nonparametric model. He has successfully supervised to completion 14 Ph.D. students and 12 M.EngSc. students (master's by research). He is the President of Malaysia Image Analysis and Machine Intelligence. He is a member of the Signal Processing Society.



Hongliang Li (SM'12) received the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, China, in 2005. From 2005 to 2006, he joined the Visual Signal Processing and Communication Laboratory, The Chinese University of Hong Kong (CUHK), as a Research Associate. From 2006 to 2008, he was a Post-Doctoral Fellow with the Visual Signal Processing and Communication Laboratory, CUHK. He is currently a Professor with the School of Electronic Engineering, University of Electronic Science and Technology of China. His research interests include image segmentation, object detection, image and video coding, visual attention, and multimedia communication system.

Dr. Li has authored or co-authored numerous technical articles in well-known international journals and conferences. He is a co-editor of a Springer book titled *Video Segmentation and its Applications*. He was involved in many professional activities. He is an Editorial Board Member of the *Journal on Visual Communications and Image Representation* (Elsevier Science), and an Area Editor of *Signal Processing: Image Communication* (Elsevier Science). He was a Technical Program Co-Chair in ISPACS2009, a General Co-Chair of the 2010 International Symposium on Intelligent Signal Processing and Communications Systems, a Publicity Chair of the IEEE VCIP 2013, and the Local Co-Chair of the 2014 IEEE International Conference on Multimedia and Expo (ICME). He serves a Technical Program Co-Chair of the IEEE Visual Communications and Image Processing Conference in 2016. He served as the Technical Program Committee Member in a number of international conferences, such as ISCAS2014, ISCAS2013, ICME2014, ICME2013, ICME2012, ISPACS2005, PCM2007, PCM2009, and VCIP2010. He was selected as the New Century Excellent Talents in University from the Ministry of Education, China, in 2008.