Evaluation of Image Segmentation Quality by Adaptive Ground Truth Composition

Bo Peng[1,2] and Lei Zhang[2] ⋆

[1] Dept. of Software Engineering, Southwest Jiaotong University, China,
boopeng@gmail.com,
[2] Dept. of Computing, The Hong Kong Polytechnic University, Hong Kong,
cslzhang@comp.polyu.edu.hk

Abstract. Segmenting an image is an important step in many computer vision applications. However, image segmentation evaluation is far from being well-studied in contrast to the extensive studies on image segmentation algorithms. In this paper, we propose a framework to quantitatively evaluate the quality of a given segmentation with multiple ground truth segmentations. Instead of comparing directly the given segmentation to the ground truths, we assume that if a segmentation is "good", it can be constructed by pieces of the ground truth segmentations. Then for a given segmentation, we construct adaptively a new ground truth which can be locally matched to the segmentation as much as possible and preserve the structural consistency in the ground truths. The quality of the segmentation can then be evaluated by measuring its distance to the adaptively composite ground truth. To the best of our knowledge, this is the first work that provides a framework to adaptively combine multiple ground truths for quantitative segmentation evaluation. Experiments are conducted on the benchmark Berkeley segmentation database, and the results show that the proposed method can faithfully reflect the perceptual qualities of segmentations.

Keywords: Image segmentation evaluation, ground truths, image segmentation

1 Introduction

Image segmentation is a fundamental problem in computer vision. Over the past decades, a large number of segmentation algorithms have been proposed with the hope that a reasonable segmentation could approach the human-level interpretation of an image. With the emergence and development of various segmentation algorithms, the evaluation of perceptual correctness on the segmentation output becomes a demanding task. Despite the fact that image segmentation algorithms have been, and are still being, widely studied, quantitative evaluation of image segmentation quality is a much harder problem and the research outputs are much less mature.

⋆ Corresponding author

Human beings play an essential role in evaluating the quality of image segmentation. The subjective evaluation has been long regarded as the most reliable assessment of the segmentation quality. However, it is expensive, time consuming and often impractical in real-world applications. In addition, even for segmentations which are visually close, different human observers may give inconsistent evaluations. As an alternative, the objective evaluation methods, which aim to predict the segmentation quality accurately and automatically, are much more expected.

The existing objective evaluation methods can be classified as ground-truth based ones and non-ground-truth based ones. In non-ground-truth based methods, the empirical goodness measures are proposed to meet the heuristic criteria in the desirable segmentations. Then the score is calculated based on these criteria to predict the quality of a segmentation. There have been examples of empirical measures on the uniformity of colors [1, 2] or luminance [3] and even the shape of object regions [4]. Since the criteria are mainly summarized from the common characteristics or semantic information of the objects (e.g., homogeneous regions, smooth boundaries, etc.), they are not accurate enough to describe the complex objects in the images.

The ground-truth based methods measure the difference between the segmentation result and the human-labeled ground truths. They are more intuitive than the empirical based measures, since the ground truths can well represent the human-level interpretation of an image. Some measures in this category aim to count the degree of overlapping between regions with strategies of being intolerant [5] or tolerant [6] to region refinement. In contrast to working on regions, there are also measures [7, 8] matching the boundaries between segmentations. Considering only the region boundaries, these measures are more sensitive to the dissimilarity between the segmentation and the ground truths than the region based measures. Some other measures use non-parametric tests to count the pairs of pixels that belong to the same region in different segmentations. The well-known Rand index [9] and its variants [10, 11] are of this kind. So far, there is no standard procedure for segmentation quality evaluation due to the ill-defined nature of image segmentation, i.e., there might be multiple acceptable segmentations which are consistent to the human interpretation of an image. In addition, there exists a large diversity in the perceptually meaningful segmentations for different images. The above factors make the evaluation task very complex.

In this work, we focus on evaluating segmentation results with multiple ground truths. The existing methods [5–11] of this kind prefer matching the given whole segmentation with ground truths for evaluation. However, the available human-labeled ground truths are only a small fraction of all the possible interpretations of an image. The available dataset of ground truths might not contain the desired ground truth which is suitable to match the input segmentation. Hence such kind of comparison often leads to a certain bias on the result or is far from the goal of objective evaluation.

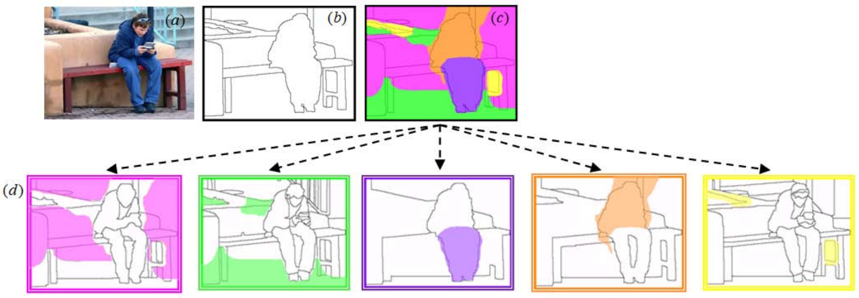

Fig. 1. A composite interpretation between the segmentation and the ground truths - the basic idea. (a) is a sample image and (b) is a segmentation of it. Different parts (shown in different colors in (c)) of the segmentation can be found to be very similar to parts of the different human-labeled ground truths, as illustrated in (d).

We propose a new framework to solve the problem. The basic idea is illustrated in Fig.1. Fig. 1(b) shows a possible segmentation of the image in Fig. 1(a), and it is not directly identical to any of the ground truths listed in Fig. 1(d). However, one would agree that Fig. 1(b) is a good segmentation, and it is similar to these ground truths in the sense that it is composed of similar local structures to them. Inspired by this observation, we propose to design a segmentation quality measure which could generalize the configurations of the segmentation and preserve the structural consistency across the ground truths. We assume that if a segmentation is "good", it can be constructed by pieces of the ground truths. To measure the quality of the segmentation, a composite ground truth can be adaptively produced to locally match the segmentation as much as possible. Note that in Fig. 1(c), the shared regions between the segmentation and the ground truths are typically irregularly shaped; therefore the composition is data driven and cannot be predefined. Also, the confidence of selected pieces from the ground truths will be examined during the process of comparison. Less reliance should be given on the ambiguous structures, even if they are very similar to the segmentation. In the proposed measure, we will integrate all these factors for a robust evaluation. Fig. 2 illustrates the flowchart of the proposed framework. Firstly a new composite ground truth is adaptively constructed from the ground truths in the database, and then the quantitative evaluation score is produced by comparing the input segmentation and the new ground truth.

Researchers have found that human visual system (HVS) is highly adapted to extract structural information from natural scenes [14]. As a consequence, a perceptually meaningful measure should be error-sensitive to the structures in the segmentations. It is also known that human observers may pay different attentions to different parts of the images [6, 8]. The different ground truths segmentations of an image therefore present different levels of details of the objects in the image. This fact makes them rarely identical in the global view, while more consistent in the local structures. For this reason, the evaluation of

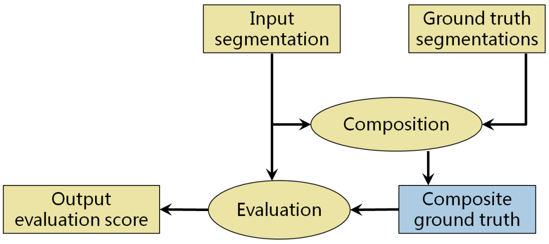

Fig. 2. Flowchart of the proposed segmentation evaluation framework.

image segmentation should be based on the local structures, rather than only the global views. The standard similarity measures, such as Mutual Information [13], Mean Square Error (MES) [14], probabilistic Rand index [9] and Precision-Recall curves [8], compare the segmentation with the whole ground truths, producing a matching result based on the best or the average score. Since the HVS tends to capture the local structures of images, these measures cannot provide an HVS-like evaluation on the segmentation. The idea of matching images with composite pieces has been explored in [15–17], where there is no globally similar image in the given resource to exactly match the image in query. These methods pursue a composition of given images which is meaningful in the perception of vision. In contrast, in this paper we propose to create a new composition that is not only visually similar to the given segmentation, but also keeps the consistency among the ground truths. The construction is automatically performed on the boundary maps instead of the original images. To the best of our knowledge, the proposed work is the first one that generalizes and infers the ground truths for segmentation evaluation.

The rest of paper is organized as follows. Section 2 provides the theoretical framework of constructing the ground truth that is adaptive to the given segmentation. A segmentation evaluation measure is consequently presented in Section 3. Section 4 presents experimental results on 500 benchmark images. Finally the conclusion is made in Section 5.

2 The adaptive composition of ground truth

2.1 Theoretical framework

Many problems in computer vision, such as image segmentation, can be taken as a labeling problem. Consider a set of ground truths $\boldsymbol{G} = \{G_1, G_2, \ldots, G_K\}$ of an image $X = \{x_1, x_2, \ldots, x_N\}$, where $G_i = \{g_1^i, g_2^i, \ldots, g_N^i\}$ denotes a labeling set of X, $i = 1, \ldots, K$, and N is the number of elements in the image (e.g., pixels, regions). Let $S = \{s_1, s_2, \ldots, s_N\}$ be a given segmentation of X, where s_j is the label of x_j (e.g., boundary or non-boundary), $j = 1, \ldots, N$. To examine the similarity between S and $\boldsymbol{G}$, we compute the similarity between

S and a new ground truth G^*, which is to be generated from $\boldsymbol{G}$ based on S. Denote by $G^* = \{g_1^*, g_2^*, \dots, g_N^*\}$ the composite ground truth. We construct G^* by putting together pieces from $\boldsymbol{G}$, i.e., each piece $g_j^* \in \{g_j^1, g_j^2, \dots, g_j^K\}$. Clearly, one primary challenge is how to reduce the artifact in the process of selecting and fusing image pieces. The pieces of the composite ground truth should be integrated seamlessly to keep the consistency of image content. Thus, our principle to choose these pieces can be summarized as: each one of g_j^* should be most similar to its counterpart in S with the constraint of structural consistency across the ground truths. Once G^* is constructed, the quality of segmentation S can be evaluated by calculating the similarity between S and G^*.

G^* is a geometric ensemble of local pieces from $\boldsymbol{G}$. We adopt an optimistic strategy to choose the elements of G^*, by which S will match $\boldsymbol{G}$ as much as possible. G^* can then be taken as a new segmentation of X by assigning a label to each pixel. Meanwhile, g_j^* contains the information of the corresponding location in the K ground truths. To construct such a G^*, we introduce a label l_{g_j} $(l = 1, \dots, K)$ to each g_j^* in G^*. Fig. 3 uses an example to illustrate how to construct the new ground truth G^*. We can see that, given two ground truth images G_1 and G_2, G^* is found by firstly computing the optimal labeling set for the ground truths. Then elements of G^* which are labeled as 1 (or 2) will take their values from G_1 (or G_2). This leads to a maximum-similarity-matching between S and G^*.

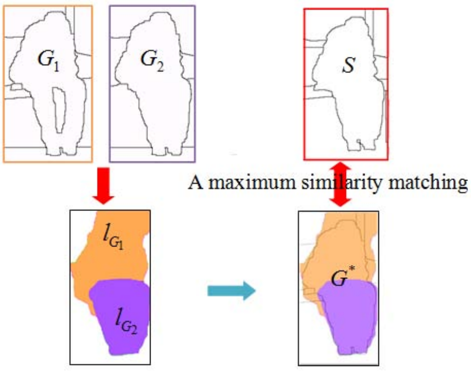

Fig. 3. An example of adaptive ground truth composition for the given segmentation S. G_1 and G_2 are two ground truths by human observers. The optimal labeling $\{l_{G_1}, l_{G_2}\}$ of G_1 and G_2 produces a composite ground truth G^*, which matches the S as much as possible.

Given a set of labels L and a set of elements in G, to construct the segmentation G^* a label $l_g \in L$ needs to be assigned to each of the elements $g \in G$. The label set could be an arbitrary finite set such as $L = \{1, 2, \dots, K\}$. Let $l = \{l_g | l_g \in L\}$ stand for a labeling, i.e., label assignments to all elements in G. We could formulate the labeling problem in terms of energy minimization, and then seek for the labeling l that minimizes the energy. In this work, we propose an energy function that follows the Potts model [18]:

$$E(l) = \sum_j D(l_{g_j}) + \lambda \cdot \sum_{\{g_j, g_{j'}\} \in M} u_{\{g_j, g_{j'}\}} \cdot T(l_{g_j} \neq l_{g_{j'}}) \tag{1}$$

There are two terms in the energy function. The first term $D(l_{g_j})$ is called the data term. It penalizes the decision of assigning l_{g_j} to the element g_j, and thus can be taken as the measure of difference. Suppose that the normalized distances between the ground truths and the segmentation S is $\Delta d(s_j, g_j)$, we can define:

$$D(l_{g_j}) = \Delta d(s_j, g_j) \tag{2}$$

The second term $u_{\{g_j, g_{j'}\}} \cdot T(l_{g_j} \neq l_{g_{j'}})$ indicates the cost of assigning different labels to the pair of elements $\{g_j, g_{j'}\}$ in G^*. M is a neighborhood system, and T is an indicator function:

$$T(l_{g_j} \neq l_{g_{j'}}) = \begin{cases} 1 \text{ if } l_{g_j} \neq l_{g_{j'}} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

We call $u_{\{g_j, g_{j'}\}} \cdot T(l_{g_j} \neq l_{g_{j'}})$ the smoothness term in that it encourages the labeling in the same region have the same labels. In this way, the consistency of neighboring structures can be preserved. It is expected that the separation of regions should pay higher cost on the elements whose label is agreeable by few ground truths, while lower cost on the reverse. Thus we can define $u_{\{g_j, g_{j'}\}}$ in the expression:

$$u_{\{g_j, g_{j'}\}} = min\{\overline{\Delta d_j}, \overline{\Delta d_{j'}}\} \tag{4}$$

where $\overline{\Delta d_j}$ is the average distance between g_j^* and $\{g_j^1, g_j^2, \ldots, g_j^K\}$.

In Eq. (1), the parameter λ is used to control the relative importance of the data term versus the smoothness term. If λ is very small, only the data term matters. In this case, the label of each element is independent of the other elements. If λ is very large, all the elements will have the same label. In the implementation, λ is set manually and its range for different images does not vary a lot (from 20 to 100), which is similar to many graph cut based works.

Minimization of Eq.(1) is NP-hard. We use the effective expansion-moves/swap-moves algorithm described in [19] to solve it. The algorithm aims to compute the minimum cost multi-way cuts on a defined graph. Nodes in the graph are connecting to their neighbors by n-links. Each n-link is assigned a weight $u_{\{g_j, g_{j'}\}}$ defined in the energy function Eq.(1). Suppose we have K ground truths, then there will be K virtual nodes in the graph, representing the K labels of these ground truths. Each graph node connects to the K virtual nodes by t-links. We weight the t-links as $D(l_{g_j})$ to measure the similarity between the graph nodes and the virtual nodes. The K-way cuts will divide the graph into K parts, and bring a one-to-one correspondence to the labeling of the graph. Fig.4 shows an example, where red lines are the graph cuts computed by the expansion-moves/swap-moves algorithm. Labeling of the graph is accordingly obtained (shown in different colors). Finally, we copy the segmented pieces of regions from each ground truth to form the new ground truth G^* that is driven by the given segmentation S.

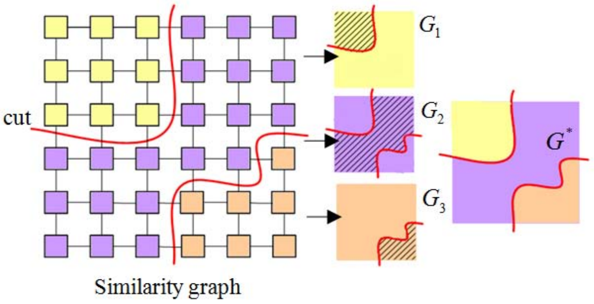

Fig. 4. An example of the construction of G^*. It is produced by the copies of selected pieces (shadow areas) in the ground truths.

2.2 The definition of distance

The distance Δd , which is used in Eq. (2) and Eq. (4), needs to be defined to optimize the labeling energy function Eq. (1). Although many distance measures have been proposed in the existing literature, it is not a trivial work to perform the matching between the machine's segmentation and the human-labeled ground truths. For ground truth data, due to the location errors produced in drawing process, boundaries of the same object might not be fully overlapped. This is an inherent problem for human-labeled ground truths. In Fig.5, we show an example of boundary distortions in different ground truths. If simply matching pixels between the segmentation S and ground truths $\boldsymbol{G}$, S will probably be over-penalized by the unstable and slightly mis-localized boundaries in $\boldsymbol{G}$. Moreover, different segmentation algorithms may produce the object boundaries in different widths. For example, if we take the border pixels in both sides of the adjacent regions, the boundaries will appear in a two-pixel width. To match the segmentation appropriately, the measure should tolerate some acceptable deformations between different segmentations. The work in [20] solves this problem by matching the boundaries under a defined threshold. However, since their method involves a minimum-cost perfect matching of the bipartite graph, it might be limited to performing on the boundaries only.

As an alternative, we consider the structural similarity index *CW-SSIM* proposed by Sampat et al. [21]. It is a general purpose image similarity index which benefits from the fact that the relative phase patterns of complex wavelet coefficients can well preserve the structural information of local image features, and rigid translation of image structures will lead to constant phase shift. *CW-SSIM* overcomes the drawback of its previous version *SSIM* [15] in that it does not

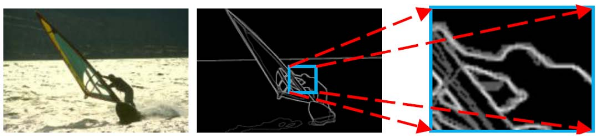

Fig. 5. An example of distorted boundaries among different human-labeled ground truths. Ground truths are drawn into the same image, where the whiter boundaries indicate that more subjects marked as a boundary. (Taken from the Berkeley Dataset [6].)

require the precise correspondences between pixels, and thus becomes robust to the small geometric deformations during the matching process. Therefore, we adopt the principle of *CW-SSIM* and slightly modify it into a new one called *G-SSIM*, which uses the complex Gabor filtering coefficients of an image instead of the steerable complex wavelet transform coefficients used in [21]. Specifically, the Gabor filtering coefficients are obtained by convolving the segmentation with 24 Gabor kernels, which are on 3 different scales and along 8 different directions, respectively. As a result, the *G-SSIM* on each Gabor kernel is defined as:

$$H(\mathbf{c}_s, \mathbf{c}'_s) = \frac{2\sum_{i=1}^{N} |c_{s,i}||c^*_{s',i}| + \alpha}{\sum_{i=1}^{N} |c_{s,i}|^2 + \sum_{i=1}^{N} |c_{s',i}|^2 + \alpha} \cdot \frac{2|\sum_{i=1}^{N} c_{s,i}c^*_{s',i}| + \beta}{2\sum_{i=1}^{N} |c_{s,i}c^*_{s',i}| + \beta} \tag{5}$$

where $\mathbf{c}_s$ and $\mathbf{c}'_s$ are the complex Gabor coefficients of two segmentations s and s', respectively. $|c_{s,i}|$ is the magnitude of a complex Gabor coefficient, and c^* is the conjugate of c. α and β are small positive constants to make the calculation stable.

It is easy to see that the maximum value of *G-SSIM* is 1 if $\mathbf{c}_s$ and $\mathbf{c}'_s$ are identical. Since the human labeling variations cause some technical problems in handling the multiple ground truths. The wavelet measure we used can better solve this problem. It can well preserve the image local structural information without requiring the precise correspondences between pixels and is robust to the small geometric deformations during the matching process. Now we define the distance Δd as:

$$\Delta d(\mathbf{c}_s, \mathbf{c}'_s) = 1 - \overline{H}(\mathbf{c}_s, \mathbf{c}'_s) \tag{6}$$

where $\overline{H}(\mathbf{c}_s, \mathbf{c}'_s)$ is the average value of *G-SSIM* by 24 Gabor kernels. Notice that although the value of Δd is defined on each pixel, it is inherently decided by the textural and structural properties of localized regions of image pairs. With the distance Δd defined in Eq. (6), we can optimize Eq. (1) so that the composite ground truth G^* for the given segmentation S can be obtained. Fig. 6 shows three examples of the adaptive ground truth composition. It should be noted that the composite ground truth does not have to contain the closed form of boundary, since it is designed to work in conjunction with the benchmark criterion (in Sec. 3).

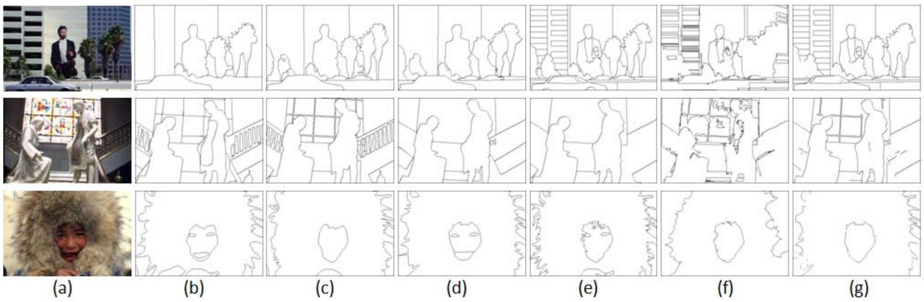

Fig. 6. Examples of the composite ground truth G^*. (a) Original images. (b)-(e) Human labeled ground truths. (f) Input segmentations of images by the mean-shift algorithm [22]. (g) The composite ground truth G^*.

3 The measure of segmentation quality

Once the composite ground truth G^* that is adaptive to S is computed, the problem of image segmentation quality evaluation becomes a general image similarity measure problem: given a segmentation S, how to calculate its similarity (or distance) to the ground truth G^*. Various similarity measures can be employed, and in this section we present a simple but effective one.

In the process of constructing the ground truth, we have allowed for some reasonable deformations of the local structures of S. When the distance $\Delta d(s_j, g_j^*)$ between s_j and g_j^* is obtained, the distance for the whole segmentation can be calculated as the average of all $\Delta d(s_j, g_j^*)$. However, the confidence of such evaluation between s_j and g_j^* also needs to be considered, due to the fact that less reliance should be given on the ambiguous structures, even if they are very similar to the segmentation. For this purpose, we introduce R_{s_j} as the empirical global confidence of g_j^* w.r.t. $\boldsymbol{G}$. For example, we can estimate R_{s_j} as the similarity between g_j^* and $\{g_j^1, g_j^2, \ldots, g_j^K\}$, and there is:

$$R_{s_j} = 1 - \overline{\Delta d_j} \tag{7}$$

where $\overline{\Delta d_j}$ is the average distance between between g_j^* and $\{g_j^1, g_j^2, \ldots, g_j^K\}$. In Eq.(7), R_{s_j} achieves the highest value 1 when the distance between g_j^* and $\{g_j^1, g_j^2, \ldots, g_j^K\}$ is zero and achieves the lowest value zero when the situation is reversed. Since R_{s_j} is a positive factor for describing the confidence, the similarity between s_j and g_j^* should be normalized to [-1,1] such that the high confidence works reasonably for both of the good and bad segmentations. If there are K instances in $\boldsymbol{G}$ and all of them contribute to the construction of G^*, we can decompose S into K disjointed sets $\{S_0, S_1, \ldots, S_K\}$. Based on the above considerations, the measure of segmentation quality is defined as:

$$M(S, \boldsymbol{G}) = \frac{1}{N} \sum_{i=1}^{K} \sum_{s_j \in S_i} (1 - 2\Delta d(s_j, g_j^*)) \cdot R_{s_j} \tag{8}$$

We can see that the proposed quality measure ranges from -1 to 1, which is an accumulated sum of similarity provided by the individual elements of S. The minimum value of $\Delta d(s_j, g_j^*)$ is zero when s_j is completely identical to the ground truths g_j^*; in the meanwhile, if all the ground truths in $\boldsymbol{G}$ are identical, R_{s_j} will be 1 and then $M(S, \boldsymbol{G})$ will achieve its maximum value 1. Note that the measure is decided by both the distances $\Delta d(s_j, g_j^*)$ and R_{s_j} . If S is only similar/dissimilar to G^* without a high consistency among the ground truths data $\{g_j^1, g_j^2, \ldots, g_j^K\}$, $M(S, \boldsymbol{G})$ will be close to zero. This might be a common case for images with complex contents, where perceptual interpretation of the image contents is diverse. In other words, R_{s_j} enhances the confident decisions on the similarity/dissimilarity and therefore preserves the structural consistency in the ground truths.

4 Experiments

In this section, we conduct experiments to validate the proposed measure in comparison with the commonly used F-measure [20] and the Probabilistic Rand (PR) index [11] on a large amount of segmentation results. The code of our method is available at `http://www4.comp.polyu.edu.hk/~cslzhang/SQE/SQE.htm`. It should be stressed that the proposed method is to evaluate the quality of an input segmentation, yet it is possible to adopt the proposed method into a system for segmentation algorithm evaluation.

The F-measure is mainly used in the boundary-based evaluation [20]. Specifically, a precision-recall framework is introduced for computing this measure. Precision is the fraction of detections that are true positives rather than false positives, while recall is the fraction of true positives that are detected rather than missed. In the context of probability, precision corresponds to the probability that the detection is valid, and recall is the probability that the ground truth is detected. In the presence of multiple ground truths, a reasonable combination of the values should be considered. In [20] only the boundaries which match no ground truth boundary are counted as false ones. The precision value is averaged among all the related ground truths. A combination of the precision and recall leads to the F-measure as below:

$$F = \frac{PR}{\tau R + (1-\tau)P} \tag{9}$$

where τ is a relative cost between precision (P) and recall (R). We set it to be 0.5 in the experiments.

The PR index [11] examines the pair-wise relationships in the segmentation. If the label of pixels x_j and $x_{j'}$ are the same in the segmentation image, it is expected that their labels to be the same in the ground truth image for a "good" segmentation and vice versa. Denote by l_j^s the label of pixel x_j in the segmentation S and by l_j^G the corresponding label in the ground truth G. The

PR index for comparing S with multiple ground truths $\boldsymbol{G}$ is defined as:

$$PR(S,\boldsymbol{G}) = \frac{1}{\binom{N}{2}} \sum_{\substack{j,j' \\ j \prec j'}} [I(l_j^S = l_{j'}^S)p_{j,j'} + I(l_j^S \neq l_{j'}^S)(1 - p_{j,j'})] \qquad (10)$$

where $p_{j,j'}$ is the ground truth probability that $I(l_j^S = l_{j'}^S)$. In practice, $p_{j,j'}$ is defined as the mean pixel-pair relationship among the ground truths:

$$p_{j,j'} = \frac{1}{K} \sum I(l_j^{G_i} = l_{j'}^{G_i}) \qquad (11)$$

According to the above definitions, the PR index ranges from 0 to 1, where a score of zero indicates that the segmentation and the ground truths have opposite pair-wise relationships, while a score of 1 indicates that the two have the same pair-wise relationships.

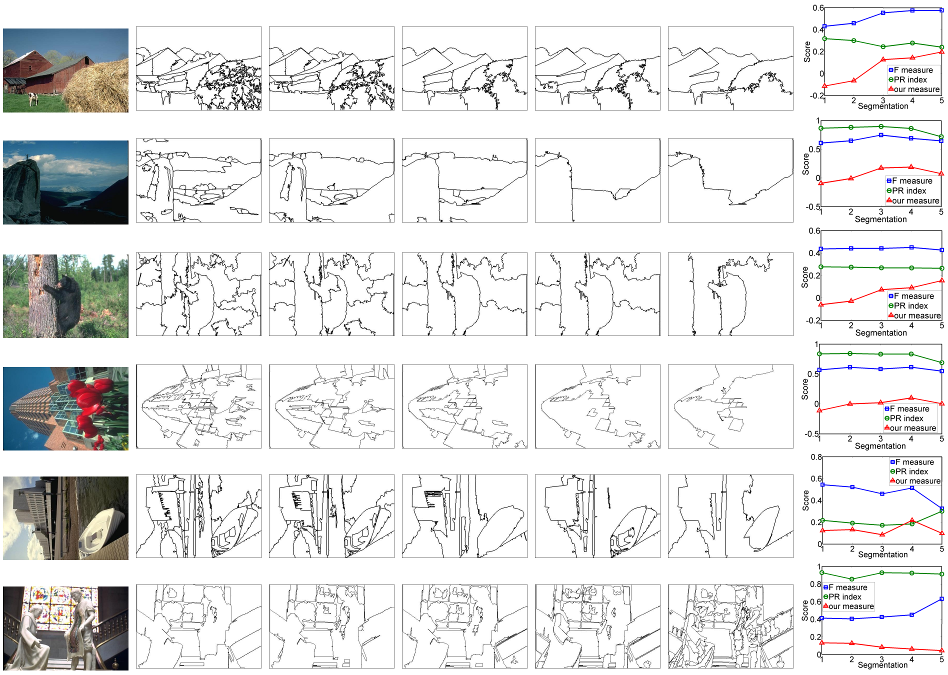

Fig. 7. Example of measure scores for different segmentations. For each original image, 5 segmentations are obtained by the mean-shift algorithm [22].The rightmost column shows the plots of scores achieved by F-measure (in blue), PR index (in green) and our method (in red).

For segmentation algorithm, different parameter settings will lead to segmentations of different granularities. In Fig.7, different segmentations of the given images are produced by the mean-shift method [22]. From the plots of scores (in the rightmost column) we can see that the proposed measure produces the closest results to the human perception. In contrast, both of F-measure and PR index wrongly find the best segmentation results (in 2^{nd}, 3^{rd}, 5^{th} and 6^{th} rows) or reduce the range of scores (in 1^{st} and 4^{th} rows). The success of the proposed measure comes from the adaptive evaluation of the meaningful structures in different levels. However, from the definition of F-measure [20] we can see that the highest value of recall is achieved when a segmentation contains all the boundaries in the ground truths. This strategy is not very intuitive in practical applications. The PR index [11] suffers from issues such as reducing the dynamic range and favoring of large regions. These drawbacks can be observed in the examples in Fig. 7.

Next, we examine the overall performance of the proposed measure on 500 natural images from the Berkeley segmentation database [6]. We apply three segmentation techniques, mean-shift (MS) [22], normalized cut (NC) [23] and efficient-graph based algorithm (EG) [12], to produce segmentations on these images. For each image, we tune the parameters of the used segmentation method such that two different segmentations are produced. It was ensured that both the two produced segmentations are visually meaningful to the observers so that they can make meaningful judgement. Thus 500 pairs of segmentations are obtained. Scores of these segmentations are computed by the proposed measure, the F-measure and the PR index. Then we arranged 10 observers to evaluate these 500 pairs of segmentations by pointing out which segmentation is better than the other one, or if there is a tie between them. We avoid giving the specific instructions so the task becomes generalized. We took the answers from the majority, and if the two segmentations are in a tie, any judgment from the measure will be taken as correct. In the subjective evaluation results, we have 85 pairs of them being classified as ties.

Table 1 shows the comparison results of the three measures. Each of them gives 500 results on the pair of segmentations, where our measure has 379 results which are consistent to the human judgment, while the F-measure and the PR index have only 333 and 266, respectively. Also we count the number of results which are only correctly produced by one measure (we call them "winning" cases). Our measure obtains 114 winning cases, while F-measure and PR index only obtain 4 and 19, respectively. And there are 16 results which are wrongly classified by all of the three measures. The proposed measure outperforms the other two in both of the "consistent" and the "winning" cases. In the meantime, we can have an interesting observation. Since the F-measure and the PR index are based on the boundary and the region relationship, respectively, they have different preferences in the segmentation quality evaluation. The PR index works better than F-measure in terms of "winning" case but it works the worst in terms of "consistent" case. This is mainly because the PR index is a region-based measure, which might compensate for the failure of boundary-based measures. In the PR index, the parameter p_{ij} is based on the ground truths of the image; however, having a sufficiently large number of valid segmentations of an image is often infeasible in practice. As a result, it is hard to obtain a good estimation of this parameter. Moreover, one can conclude that this limitation of ground truth segmentations in existing databases makes the proposed adaptive ground truth composition based segmentation evaluation method a more sensible choice.

Table 1. The number of results which are consistent to the human judgment, as well as the numbers of “winning” cases and failure cases by competing methods.

	Consistent	Winning	Failure
F-measure	333	4	16
PR index	266	19	
Ours	379	114	

5 Conclusions

In this paper, we presented a novel framework for quantitatively evaluating the quality of image segmentations. Instead of directly comparing the input segmentation with the ground truth segmentations, the proposed method adaptively constructs a new ground truth from the available ground truths according to the given segmentation. In this way, it becomes more effective and reasonable to measure the local structures of the segmentation. The quality of the given segmentation can then be measured by comparing it with the adaptively composite ground truth. From the comparison between the proposed measure and two other popular measures on the benchmark Berkeley database of natural images, we can see that the proposed method better reflect the perceptual interpretation on the segmentations. Finally, we compared the measures on 1000 segmentations of 500 natural images. By counting the number of segmentations that are consistent to the human judgment, it was shown that the proposed measure performs better than the F-measure and PR index. The proposed method can be extended to get closed contours for composition, and this will be our future work.

References

1. Liu, J., Yang, Y.: Multiresolution color image segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence. vol. 16, pp. 689-700 (1994)
2. M. Borsotti, P.C., Schettini, R.: Quantitative evaluation of color image segmentation results. Pattern Recognition Letter. vol. 19, pp. 741–747 (1998)
3. Zhang, H., Fritts, J., Goldman, S.: An entropy-based objective segmentation evaluation method for image segmentation. SPIE Electronic Imaging Storage and Retrieval Methods and Applications for Multimedia. pp.38–49 (2004)
4. Ren, X., Jitendra, M.: Learning a classification model for segmentation. IEEE Conference on Computer Vision. pp. 10–17 (2003)

5. Christensen, H., Phillips, P.: Empirical evaluation methods in computer vision. World Scientific Publishing Company (2002)
6. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. International Conference on Computer Vision. pp. 416–423 (2001)
7. Freixenet, J., Munoz, X., Raba, D., Marti, J., Cuff, X.: Yet another survey on image segmentation: Region and boundary information integration. European Conference on Computer Vision. pp. 21–25 (2002)
8. Martin, D.: An empirical approach to grouping and segmentation. Ph.D. dissertation U. C. Berkeley (2002)
9. Rand, W.: Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association. vol. 66, pp. 846–850 (1971)
10. Fowlkes, E., Mallows, C.: A method for comparing two hierarchical clusterings. Journal of the American Statistical Association. vol. 78, pp. 553–569 (1983)
11. Unnikrishnan, R., Hebert, M.: Measures of similarity. IEEE Workshop on Applications of Computer Vision. pp.394–400 (2005)
12. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. International Journal on Computer Vision. vol. 59, pp.167–181 (2004)
13. Viola, P., Wells, W.: Alignment by maximization of mutual information. vol. 3, pp. 16–23 (1995)
14. Wang, Z., Bovik, A.: Modern image quality assessment. New York: Morgan and Claypool Publishing Company (2006)
15. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Process. vol. 13, pp. 600–612 (2004)
16. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M.: Interactive digital photomontage. vol. 23, pp. 294–302 (2004)
17. Russell, B., Efros, A., Sivic, J., Freeman, W., Zisserman, A.: Segmenting scenes by matching image composites. Advances in Neural Information Processing Systems, pp. 1580–1588 (2009)
18. Potts, R.: Some generalized order-disorder transformation. Proceedings of the Cambridge Philosophical Society. vol. 48, pp. 106–109 (1952)
19. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 23, pp. 1222–1239 (2001)
20. Martin, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries using local brightness, color and texture cues. IEEE Trans. on Pattern Analysis and Machine Intelligence. vol. 26, pp. 530–549 (2004)
21. Sampat, M., Wang, Z., Gupta, S., Bovik, A., Markey, M.: Complex wavelet structural similarity: A new image similarity index. IEEE Transactions on Image Processing. vol. 18, pp. 2385–2401 (2009)
22. Comanicu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence. vol. 24, pp. 603–619 (2002)
23. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence. vol. 22, pp. 888–905 (1997)