Evaluation of Segmentation Quality via Adaptive Composition of Reference Segmentations

Bo Peng, Lei Zhang, Xuanqin Mou, and Ming-Hsuan Yang

Abstract—Evaluating image segmentation quality is a critical step for generating desirable segmented output and comparing performance of algorithms, among others. However, automatic evaluation of segmented results is inherently challenging since image segmentation is an ill-posed problem. This paper presents a framework to evaluate segmentation quality using multiple labeled segmentations which are considered as references. For a segmentation to be evaluated, we adaptively compose a reference segmentation using multiple labeled segmentations, which locally matches the input segments while preserving structural consistency. The quality of a given segmentation is then measured by its distance to the composed reference. A new dataset of 200 images, where each one has 6 to 15 labeled segmentations, is developed for performance evaluation of image segmentation. Furthermore, to quantitatively compare the proposed segmentation evaluation algorithm with the state-of-the-art methods, a benchmark segmentation evaluation dataset is proposed. Extensive experiments are carried out to validate the proposed segmentation evaluation framework.

Index Terms—image segmentation evaluation, segmentation quality, image segmentation dataset

1 Introduction

IMAGE segmentation aims to localize object boundaries in accordance to human visual interpretation. It is inherently an ill-posed problem since there exist multiple plausible segmentations for the same input image [46]. As numerous segmentation algorithms have been developed in the past decades, quantitative evaluation of segmentation results has become a crucial problem for performance evaluation. In addition, proper parameter values can be determined based on reliable quantitative evaluation of image segmentation.

Evaluation methods of machine segmentation can be categorized based on whether human labeled segmentations are used as references or not. In the first category, one or more human labeled segmentations of an image are used as references [4], [46], [28], [44], [45], [24] to compute the degree of similarity (or difference) as quality scores. While in the second category, criteria of desired segmentations are defined (without using labeled segmentations) and used to measure the quality of input segmentations [6], [53], [38], [10]. The criteria are usually generalized from common characteristics or semantic information of objects (e.g., homogeneous regions and smooth boundaries) although they may not accurately describe complex objects in natural images. In this work we focus on methods based on reference segmentations labeled by humans.

- *B. Peng is with Department of Software Engineering, Southwest Jiaotong University, Chengdu, China. E-mail: bpeng@swjtu.edu.cn*
- *L. Zhang is with Department of Computing, Hong Kong Polytechnic University, Hong Kong. E-mail: cslzhang@comp.polyu.edu.hk*
- *X. Mou is with Institute of Image Processing and Pattern Recognition, Xian Jiaotong University, Xian, China. E-mail: xqmou@mail.xjtu.edu.cn*
- *M.-H. Yang is with School of Engineering, University of California, Merced, CA 95344 USA. E-mail: mhyang@ucmerced.edu*

In most reference-based segmentation evaluation methods [4], [46], [28], [44], [45], [24], each element (e.g., a pixel, a region) in the given segmentation is equally compared to its counterparts in all the labeled segmentations, and the average is often used as the quantitative score. However, human visual system tends to focus on structural information from natural scenes [50], and a good measure should take visual perception into account. In addition, it is known that different observers may pay attention to different regions in an image [28], [29], and multiple human labeled segmentations reflect different levels of perceived details. Thus, human labeled segmentations of an image are rarely identical on the holistic scale, but more consistent in terms of local structures. Consequently, evaluation of image segmentation should rely more on local structures. On the other hand, using more labeled segmentations as references is likely to facilitate fair evaluation. However, generating reference segmentations is time-consuming, and in practice only a few human-labeled results are available at our disposal. The limited number of labeled results likely leads to certain bias on segmentation evaluation, and likewise the problem with object boundary localization errors is exacerbated [28].

To address the above-mentioned problems, we propose a novel algorithm to evaluate segmentation quality based on multiple segmentations labeled by humans[1]. The main idea of this work is illustrated in Figure 1. Given an input (Figure 1(a)), an image segmentation (Figure 1(b)) is generated by an algorithm, which is not identical to any of the labeled segmentations by humans (Figure 1(d)). While the segments of the good segmentation are different from the labeled segmentations on the holistic scale, they are similar in terms of

1. Preliminary results of this work are presented in [33].

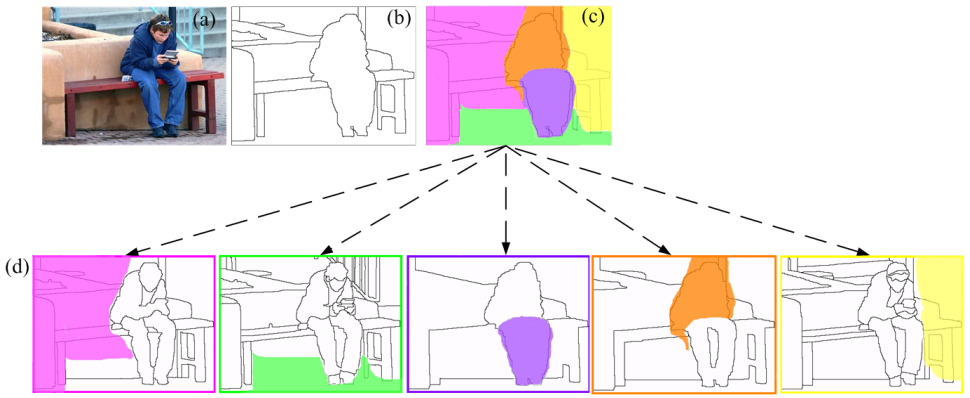


Fig. 1. An illustrative example between a machine segmentation and labeled segmentations by humans. (a) An input image. (b) An image segmentation of (a). (c) Different parts (shown in different colors) of the segmentation in (b) are similar to those segments labeled by humans in (d). In order to better evaluate image segmentation algorithms, it is important to have a good reference segmentation composed from labeled segments by humans.

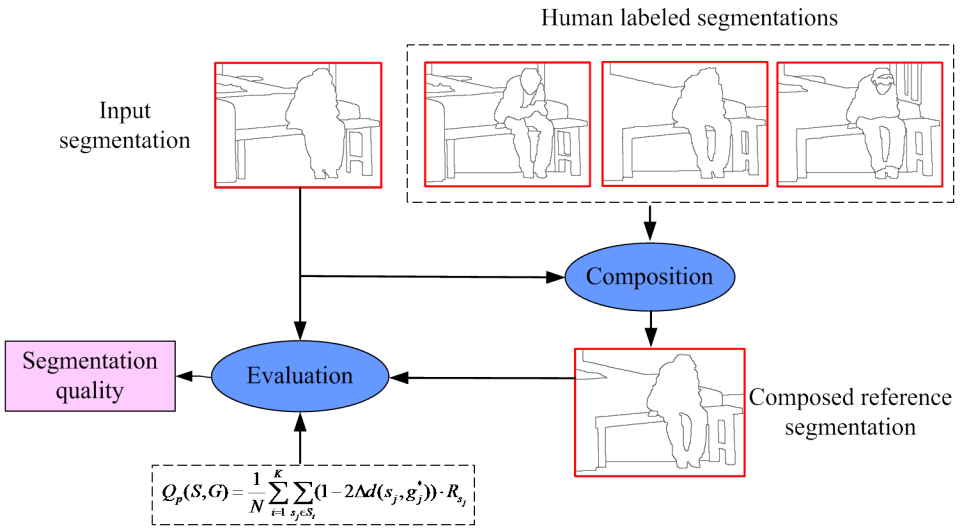


Fig. 2. Proposed evaluation framework based on adaptive composition of the reference segmentation.

local structures (i.e., the parts of Figure 1(b) are shown in Figure 1(c) with different colors, and they are similar to the segments labeled by humans in Figure 1(d)). Motivated by this observation, we propose to construct a reference segmentation which generalizes configurations of the labeled segmentations while preserving structural consistency. The underlying assumption of this work is that if an image segmentation is good, it can be composed by pieces of the labeled results.

Given an image, the composed reference segmentation should locally match the input segmentation as much as possible. Notice that in Figure 1(c), the matched regions between the input segmentation and human labeled results are in irregular shapes and thus the composition process is data-driven. Figure 2 illustrates the main steps of the proposed segmentation evaluation framework. For an input segmentation, a composed reference segmentation is adaptively constructed based on the labeled segmentations in the dataset. The quality score is computed based on the proposed similarity measure between the pair of input segmentation and composed reference segmentation.

The second contribution of this work is that a new benchmark dataset is constructed for evaluating segmentation quality. The developed dataset is motivated by the Berkeley segmentation database (BSDS500) [29] which contains 500 source images and each has 4 to 9 (5.4 on average) segmentations labeled by humans. The proposed dataset consists of 200 source images and most of them have 8 to 13 (10.7 on average) labeled segmentations. In addition, the proposed dataset consists of objects from more diverse categories.

We note that in some existing large-scale databases such as MS-COCO [25] and KITTI [19], the images are labeled based on the given object categories, and only one human labeled segmentation is provided for each image. These datasets are developed for high-level vision tasks, such as object detection, recognition, and tracking. In contrast, the BSDS500 and proposed datasets are not constrained to object categories and they contain multiple pixel-wise labeled segmentations in an image, which can be used for image segmentation and boundary detection tasks considered in this work.

Another important contribution of this work is that a dataset is constructed to quantitatively compare different segmentation evaluation methods. The proposed evaluation dataset is composed of 500 pairs of segmentations generated by state-of-the-art segmentation algorithms such as the efficient graph-based algorithm [16] and mean-shift based method [11]. Each pair of segmentation results is evaluated by human subjects. To the best of our knowledge, this dataset is the largest segmentation evaluation set in the literature in terms of number of images, diversity of objects, and number of human judgment per segmentation.

2 Related Work

In this work, a segmentation describes image regions where pixels have similar properties (e.g., Figure 1 (b)). These segmentations can be either generated by algorithms or annotated by humans. A segmentation consists of several segments (or regions) which are described by their boundaries. We review the representative reference-based segmentation evaluation methods. These methods are designed to measure the similarity or difference between an input segmentation and labeled segmentations based on regions, pixels, or boundaries.

2.1 Region-based methods

An image segmentation can be viewed as a collection of connected but exclusive regions (or segments). Region-based methods compute similarity measures in terms of differences or affinities between two segmentations. For example, the region-based measure in [22] uses the directional Hamming distance to compute discrepancy of two segmentations. The Local Refinement Error (LRE) and Global Consistency Error (GCE) [28] measure to which degree the segmentations S_1 and S_2 agree with each other. Let $R(S, p_i)$ be the set of pixels in segmentation S that are in the same region as pixel p_i, the LRE is defined as:

$$E(S_1, S_2, p_i) = \frac{|R(S_1, p_i) \setminus R(S_2, p_i)|}{|R(S_1, p_i)|}, \tag{1}$$

where $|\cdot|$ is the cardinality of a set, and $\setminus$ denotes set difference. The GCE is defined as:

$$GCE(S_1, S_2) = \frac{1}{N} \min\{\sum_i E(S_1, S_2, p_i), \sum_i E(S_2, S_1, p_i)\}, \tag{2}$$

where N is the total number of pixels in S_1. While the GCE measure accommodates refinements at different granularities, this measure suffers from degenerate cases (e.g., when there are few pixels in a segment) [29].

The Segmentation Covering (SC) [4] measures the similarity between segmentations by weight averaging the overlaps of regions in two segmentations. The covering of a segmentation S_2 by a segmentation S_1 is defined by

$$C(S_1 \to S_2) = \frac{1}{N} \sum_{R \in S_1} |R| \cdot \max_{R' \in S_2} \frac{|R \cap R'|}{|R \cup R'|}, \tag{3}$$

where R and R' are regions in S_1 and S_2, respectively.

Instead of using a single measure, multiple measures can be used to quantify segmentation quality. The approach in [14] performs evaluation in a multi-dimensional fitness (cost) space with multiple measures. In [21], five instances of segmentation are defined, from which the corresponding measures are designed for evaluation.

By considering a region as a cluster of image pixels, comparison of clusters can be used for segmentation evaluation. Meila [30] proposed an information-theoretic distance of clusters. For segmentations, this distance can be interpreted as the average conditional entropy of one segmentation given the other. The Variation of Information (VOI) measure is defined as:

$$VOI(S_1, S_2) = H(S_1|S_2) + H(S_2|S_1), \tag{4}$$

where H and I respectively denote the entropy and mutual information of the given segmentation S_1 and human labeled segmentation S_2. If two segmentations are identical, the VOI value is zero. The upper bound of VOI is finite and depends on the number of elements in the segments. Since clustering has been extensively studied in machine learning, various measures of difference [24] or similarity [47] between clusters can be adopted for segmentation evaluation.

2.2 Pixel-based methods

Significant efforts have been made to design measures for the pair-wise comparisons between a segmented image and multiple human labeled segmentations [36], [17], [44], [45], [46], [24]. The Probabilistic Rand Index (PRI) [46] defines correctness of segmentations by a statistical function. Suppose that $\mathbf{I}(l_i^S = l_j^S)$ is a binary function on the labels of each pair of pixels (x_i, x_j), the PRI is defined as:

$$PRI(S, \{S_k\}) = \frac{1}{\binom{N}{2}} \sum_{i,j,i \neq j} [\mathbf{I}(l_i^S = l_j^S) p_{ij} + \mathbf{I}(l_i^S \neq l_j^S)(1 - p_{ij})], \tag{5}$$

where N is the number of pixels, $\{S_k\}$ is the set of labeled segmentations, and p_{ij} is the probability that the labels of (x_i, x_j) are the same. In practice, the mean pixel pair relationship in all labeled segmentations is used to compute p_{ij} and the range of PRI is within [0, 1]. A score

of zero indicates that the labeling of a test segmentation is completely opposite to that of the labeled segmentation, while a score of 1 indicates that the labels of input segmentation and labeled segmentations are the same on every pixel pair. This measure accommodates region refinements appropriately as it accepts refinements only in regions that human observers find ambiguous. The Normalized Probabilistic Rand (NPR) index [46] extends the PRI measure and allows one to compare segmentations between different images. Specifically, it normalizes PRI with the expected values of input images so that NPR is zero-mean with larger range than PRI.

2.3 Boundary-based methods

Boundary-based quality measures have also been proposed in recent years. The Boundary Displacement Error (BDE) [18] defines the error of one boundary pixel as the distance to the closest pixel in the other boundary image. A near-zero mean and a small standard deviation indicate good quality of the segmentation. The F-measure can be applied to both region-based [3], [34] and boundary-based [28] evaluation. In particular, a precision-recall framework is introduced in [28], where a combination of precision and recall leads to the F-measure as below:

$$F = \frac{PR}{\tau R + (1-\tau)P}, \tag{6}$$

where τ is a relative cost between precision P and recall R. Other discrepancy methods in this category can be found in [7], [31], [13], [42].

2.4 Evaluation with multiple references

While all the measures introduced above can be used for evaluating the segmentation quality, little attention has been paid to how to effectively utilize the multiple human labeled segmentations. Most existing methods compute the quality score by each of the reference segmentations and output the average. For example, the methods based on Mutual Information [49], Mean Square Error [51], Segmentation Covering [4], Probabilistic Rand Index [36] and Recall Curves [28] holistically compare an input segmentation with a collection of references, and compute the average matching result as the final score. On the other hand, the method based on Precision Curves [28] computes the fraction of a segmentation that matches any of the references for evaluation. Since the human visual system (HVS) tends to perceive local structures of an image, these measures may not be appropriate for perceptual evaluation of segmentation quality.

3 Composing Reference Segmentations

Numerous methods have been proposed to process images based on composite pieces [39], [2], [27], [9], [12], [23]. When there is no image in the template set that is holistically similar to an input image, Russell et al. [39] use a composition of templates for scene segmentation. The composition of figure-ground segments of an image [27], [9], [12], [23] can generate plausible segmentation on multiple scales. The approach of combining parts from different photographs into one single composite picture [2], [5] can be applied to many editing tasks, such as relighting, extended depth of field, panoramic stitching, detection of saliency, etc. In this work, we propose to compose a reference segmentation based on labeled segmentations for evaluating a given segmentation. Each composed reference is not only adaptive to a given segmentation, but also structurally consistent to the labeled segmentations. The composition is carried out on the segmentation maps instead of the original image. To the best of our knowledge, the proposed framework is the first one that generalizes and infers the labeled segmentations for evaluation.

3.1 Proposed algorithm

Image segmentation can be considered as a labeling problem. Consider a set of human labeled segmentations $\boldsymbol{G} = \{G_1, G_2, \ldots, G_K\}$ 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 set of labels for each pixel in X, $i = 1, \ldots, K$, and N is the number of pixels in the image. Let $S = \{s_1, s_2, \ldots, s_N\}$ be a given segmentation of X, where s_j is the label of x_j, $j = 1, \ldots, N$. For image segmentation, labels are values that indicate the class a pixel belongs to, e.g., a binary value as the boundary or the non-boundary. Refer to Figure 2, to examine the similarity between S and $\boldsymbol{G}$, we compute the similarity between S and a new reference segmentation G^*, $G^* = \{g_1^*, g_2^*, \ldots, g_N^*\}$, which is generated from $\boldsymbol{G}$ based on S. We construct G^* by putting together pieces from $\boldsymbol{G}$, i.e., each piece $g_j^* \in \{g_j^1, g_j^2, \ldots, g_j^K\}$. Clearly, one primary challenging factor is how to reduce the artifacts in the process of selecting and fusing image pieces. The pieces of the composed reference should be integrated seamlessly to maintain consistency of image contents. Our principle is that each one of g_j^* should be most similar to its counterpart in S with the structural consistency constraints across multiple labeled segmentations. Once G^* is constructed, the quality of segmentation S is evaluated by computing the similarity between S and G^*.

A reference segmentation G^* is a geometric ensemble of local pieces from the set $\boldsymbol{G}$. We use an optimistic strategy to choose the element g_j^*, by which S will match $\boldsymbol{G}$ as much as possible. To construct a reference segmentation G^*, we introduce a labeling set $L = \{l_{g_j} | l_{g_j} \in \{1, \ldots, K\}\}$. For each g_j^*, l_{g_j} indicates the reference index where it is selected from. G^* is generated by firstly computing the labeling set $\{l_{g_1}, l_{g_2}, ..., l_{g_N}\}$, then g_j^* is set to be boundary or non-boundary according to the value of g_i^l. Figure 3 illustrates how to construct a reference segmentation G^* for part of a segmentation (in the red rectangle). Given two human labeled segmentations G_1

and G_2, the labeling set $L = \{l_{G_1} = 1, l_{G_2} = 2\}$ is computed based on S. Then, we assign elements of G^* to the class label at the corresponding location in G_1 or G_2. This leads to the maximum similarity match between S and G^*.

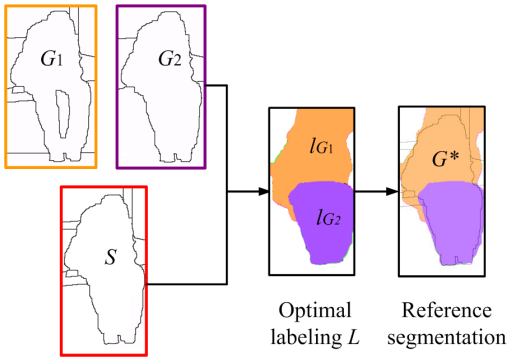


Fig. 3. An example to compose a reference segmentation for segmentation S with labeled segmentations G_1 and G_2. The optimal labelings l_{G_1}, l_{G_2} of G_1 and G_2 generate a reference segmentation G^*, which matches S closely.

For multiple references, the labeling set can be an arbitrary finite set, e.g., $L = \{1, 2, \ldots, K\}$. Let $l = \{l_g | l_g \in L\}$ denote a labeling, i.e., label assignments to all elements in G^*. We formulate the labeling problem in terms of energy minimization, and seek for the labeling l that minimizes the energy. We propose an energy function that follows the Potts model [35]:

$$E(l) = \sum_j D(l_{g_j}) + \lambda \cdot \sum_{\{g_j, g_{j'}\} \in M} u_{\{g_j, g_{j'}\}} \cdot T(l_{g_j} \neq l_{g_{j'}}). \quad (7)$$

The first part $D(l_{g_j})$ of this energy function is the data term, which penalizes the decision of assigning l_{g_j} to element g_j^*, and can be considered as the measure of difference. Suppose that the normalized distance between a reference G and segmentation S is $\Delta d(s_j, g_j)$, we define:

$$D(l_{g_j}) = \Delta d(s_j, g_j). \quad (8)$$

The second part $u_{\{g_j, g_{j'}\}} \cdot T(l_{g_j} \neq l_{g_{j'}})$ of (7) indicates the cost of assigning different labels to the pair of elements $\{g_j, g_{j'}\}$ in G^*. In (7), M is a neighborhood system, and T is an indicator function:

$$T(l_{g_j} \neq l_{g_{j'}}) = \begin{cases} 1 & \text{if } l_{g_j} \neq l_{g_{j'}} \\ 0 & \text{otherwise.} \end{cases} \quad (9)$$

The smoothness term of (7) encourages the elements in the same region to have the same labels, and thus the consistency of neighboring structures can be preserved. It is expected that separation of regions incurs higher cost on the elements which appear in fewer labeled segments and lower cost otherwise. We define $u_{\{g_j, g_{j'}\}}$ as:

$$u_{\{g_j, g_{j'}\}} = \min\{\overline{\Delta d_j}, \overline{\Delta d_{j'}}\}, \quad (10)$$

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

The optimization problem in (7) is NP-hard and we adopt the expansion moves and swap moves algorithm [8] to solve it. The algorithm computes the minimum cost multi-way cuts on a defined graph. Nodes in the graph connect to their neighbors by n-links and each is assigned a weight $u_{\{g_j, g_{j'}\}}$ defined in the energy function (7). Suppose that we have K labeled segmentations, then K virtual nodes are created in the graph. 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 divide the graph into K parts, and generate a one-to-one correspondence to the labeling of the graph.

The parameter λ in (7) controls the relative importance of the data and smoothness terms. If λ is small, only the data term matters and the label of each element is independent of other elements. If λ is large, all the elements have the same label.

In the proposed algorithm, we use L labels, where each label corresponds to one reference segmentation, to compose G^*. While it is NP-hard to compute the exact minimum of the proposed formulation, there are several justifications for using this approach rather than using a binary label (boundary or non-boundary) to compose reference segmentations.

Although an object boundary may be considered as an entity for binary labeling, it is an extremely thin elongated structure. The optimization process for a binary labeling model will have a bias toward shorter boundaries (known as the "shrinking bias"), which makes it difficult to label thin elongated structures [48]. Introducing regularizer terms into energy functions [20], [41] or using connectivity constraints [43], [48], however, will lead to higher-order cost functions or require user guidance for boundary connection. Although the proposed model is non-convex, we experimentally show that it is insensitive to initialization and parameter λ (See Section 6.1), and our approach generates stable evaluation results for image segmentation.

3.2 Distance Δd

The distance Δd in (8) and (10) needs to be defined before we minimize the labeling energy function (7). Although many distance measures have been proposed in the literature, it is not a trivial task to select a suitable measure to compare the machine segmentation with the human-labeled segments. Due to the localization errors from human labeling process, boundaries or regions of the same object may not be fully overlapped in different labeled segmentations. This is an inherent issue for human labeled segmentations. In particular, the boundary based measures are more sensitive to the dissimilarity of segmentations than the region based measures. Figure 4 shows an example of boundary distortions among different human labeled segmentations. If directly comparing the corresponding pixels, the distance will be over-penalized by slightly different boundaries in the

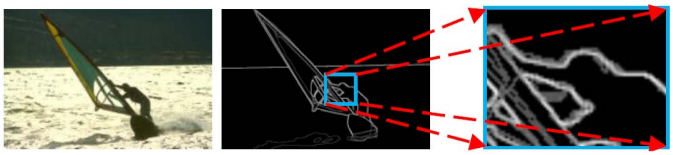


Fig. 4. An example of inconsistent boundaries among different human-labeled segmentations from the Berkeley Segmentation Dataset [29]). The whiter pixels indicate that more human subjects mark them as boundary.

labeled segmentations. Since the HVS is insensitive to such minor inconsistency, to compare the segmentations faithfully, the pixel-based distance measure should be able to accommodate some geometric inconsistency of boundaries.

In [28], this problem is addressed by matching the boundaries with a predefined threshold instead of precise correspondence. In [40], Sampat et al. propose a structural similarity index in the complex wavelet domain. This index is based on the fact that the relative phase of complex wavelet coefficients preserve the structural information of local image patterns well, while rigid translation of image structures leads to constant phase shift. In addition, this index does not require precise correspondence between pixels, and it is robust to small geometric distortions. We use the principle of this index to define a pixel-based distance, which uses the complex Gabor transform coefficients instead of the steerable complex wavelet transform coefficients. The coefficients are computed by convolving a segmentation with 24 Gabor kernels on 3 scales and along 8 directions. With the outputs of these Gabor filters, the similarity index between two segmentations is defined by

$$H(c_x, c_y) = \frac{2\sum_{i=1}^{N} |c_{x,i}||c^*_{y,i}| + \alpha}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + \alpha} \cdot \frac{2|\sum_{i=1}^{N} c_{x,i}c^*_{y,i}| + \beta}{2\sum_{i=1}^{N} |c_{x,i}c^*_{y,i}| + \beta}, \tag{11}$$

where c_x and c_y are the complex Gabor coefficients of two segmentations x and y, respectively; $|c_{x,i}|$ is the magnitude of a complex Gabor coefficient, and c^* is the conjugate of c; and α as well as β are small positive constants for computational stability. It is easy to see that the maximum value of H is 1 if c_x and c_y are identical. Therefore, we define the distance Δd as

$$\Delta d(c_x, c_y) = 1 - \overline{H}(c_x, c_y), \tag{12}$$

where $\overline{H}(c_x, c_y)$ is the average value of $H(c_x, c_y)$ obtained by 24 Gabor filters. With the distance Δd defined in (12), we optimize (7) and obtain the composed reference G^* for the given segmentation S. Figures 5 and 6 show some composed reference segmentations. Given the segmentations (labeled in green), the reference segmentations (labeled in red) are adaptively composed from multiple labeled segmentations (labeled in black).

4 Measuring Segmentation Quality

Once the composed reference G^* for the given segmentation S is obtained, the problem of measuring segmentation quality becomes the problem of computing image similarity. To compute the similarity (or distance) between S and the reference G^*, we propose a measure based on the pixel based distance used in composing the references.

When the pixel-based distance defined in (12) is used to construct a reference G^*, some geometric inconsistency of local boundaries in S has been factored in. 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 computed by the average of all $\Delta d(s_j, g^*_j)$. However, the confidence of g^*_j should also be considered since less weight should be given to those ambiguous structures, even if they are very similar. Thus, we introduce R_{s_j} as the empirical global confidence of g^*_j with respect to $\boldsymbol{G}$. For example, we can estimate R_{s_j} as the similarity between g^*_j and $\{g^1_j, g^2_j, \ldots, g^K_j\}$ and define it as

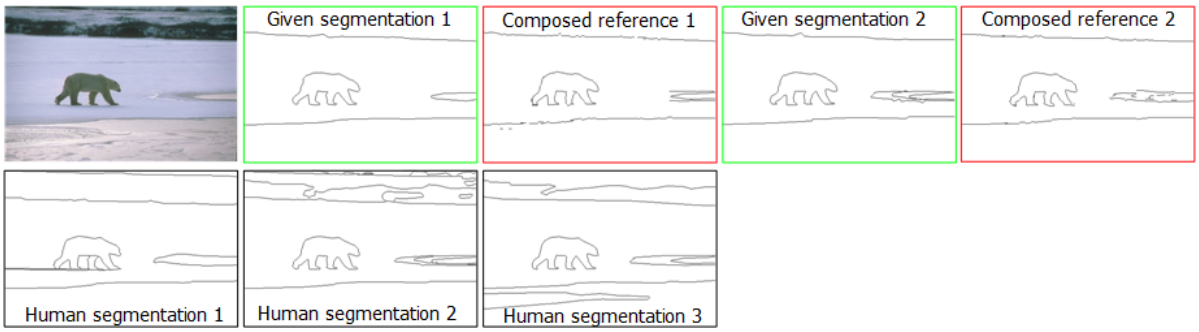
$$R_{s_j} = 1 - \overline{\Delta d_j}, \tag{13}$$

where $\overline{\Delta d_j}$ is the average distance between g^*_j and $\{g^1_j, g^2_j, \ldots, g^K_j\}$. In (13), R_{s_j} achieves the highest value 1 when the distance between g^*_j and $\{g^1_j, g^2_j, \ldots, g^K_j\}$ 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 set $\{S_0, S_1, \ldots, S_K\}$. Finally, we define the pixel-based quality measure as

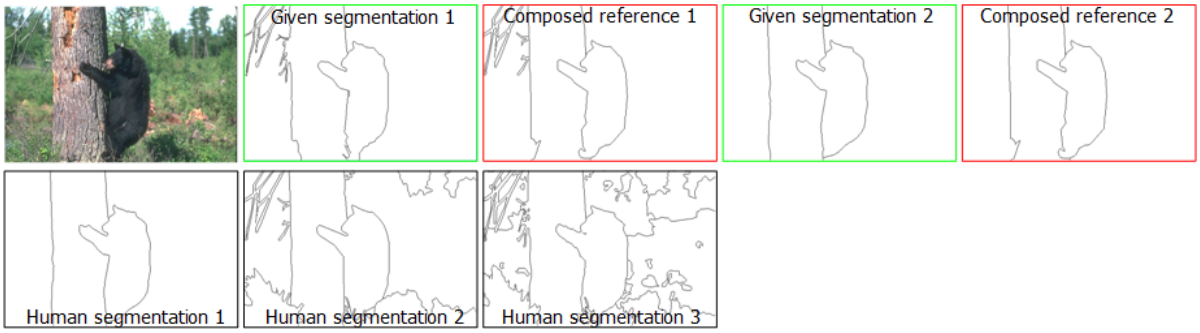
$$Q_p(S, \boldsymbol{G}) = \frac{1}{N}\sum_{i=1}^{K}\sum_{s_j \in S_i} (1 - 2\Delta d(s_j, g^*_j)) \cdot R_{s_j}. \tag{14}$$

The proposed quality measure is related to the accumulated sum of the similarities computed from each element of S. The minimum value of $\Delta d(s_j, g^*_j)$ is 0 when s_j is identical to g^*_j of a reference segmentation. If all the labeled segmentations in $\boldsymbol{G}$ are identical, R_{s_j} is 1 and $Q_p(S, G)$ has the maximum value. Note that the measure is determined by both $\Delta d(s_j, g^*_j)$ and R_{s_j}. If S is only similar to G^* without high consistency among the labels $\{g^1_j, g^2_j, \ldots, g^K_j\}$, the value of R_{s_j} is low and hence the absolute value of $Q_p(S, G)$ is also low. This issue may arise for images with complex contents, where perceptual interpretation of image contents is likely to be different.

The scores of the proposed measure with the composed reference (Q_p) and by averaging over multiple references ($\overline{Q}$) are illustrated in Figures 5 and 6. In Figure 5, there are two input human labeled segmentations for each image, and the remaining three human labeled segmentations are used to compose the reference segmentation. Obviously, these human labeled segmentations

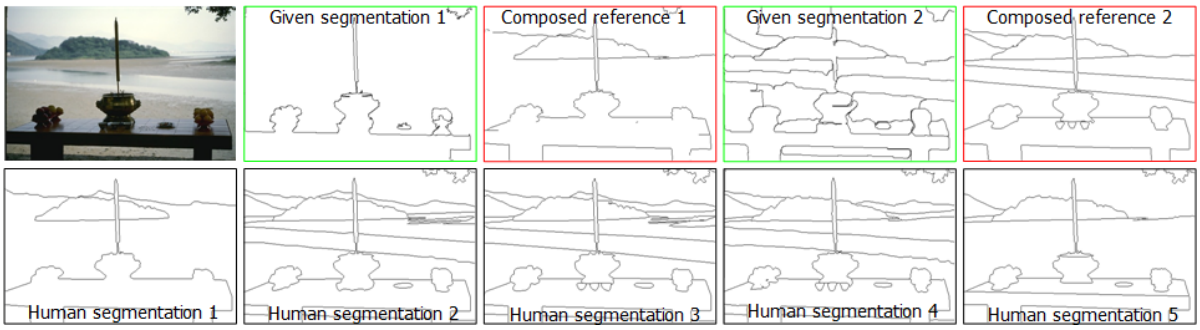


(a) Q_{p1}=0.81, Q_{p_2}=0.80, $\overline{Q_1}$=0.73, $\overline{Q_2}$=0.69

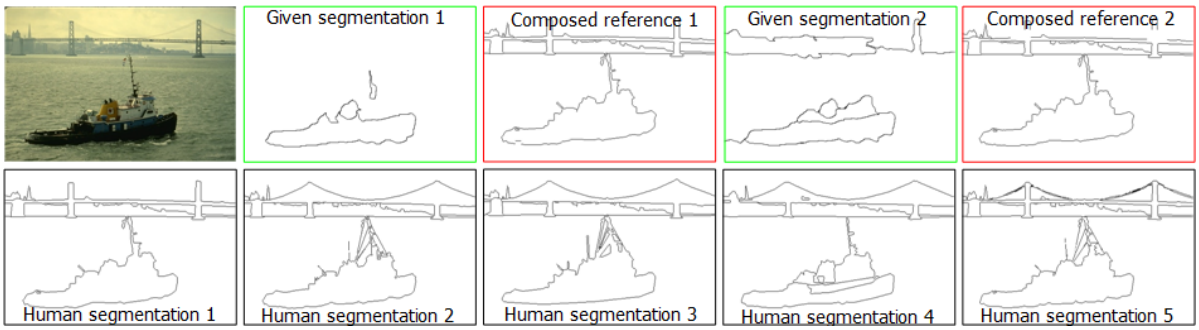


(b) Q_{p1}=0.69, Q_{p2}=0.74, $\overline{Q_{p1}}$=0.56, $\overline{Q_{p2}}$=0.55

Fig. 5. Examples of composed references for human labeled segmentations. For each example, the first row shows the original image, two human labeled segmentations (labeled in green) of it and the corresponding composed reference segmentations G^* (labeled in red). The second row shows the human labeled segmentations used to compose the reference segmentations. The images are from the Berkeley Segmentation Dataset [29].



(a) Q_{p1}=0.50, Q_{p_2}=0.39, $\overline{Q_1}$=0.39, $\overline{Q_2}$=0.36



(b) Q_{p1}=0.52, Q_{p_2}=0.56, $\overline{Q_1}$=0.49, $\overline{Q_2}$ =0.49

Fig. 6. Examples of composed references for machine segmentations. For each example, the first row shows the original image, two machine segmentations (labeled in green) of it and the corresponding composed reference segmentations G^* (labeled in red). The second row shows human labeled segmentations used to compose the reference segmentations. The images are from the Berkeley Segmentation Dataset [29].

have good quality. As expected, the proposed measure rates them with good scores, which are higher than those by the averaging approach (i.e., $Q_p > \overline{Q}$). In Figure 6, there are two input segmentations generated by the Mean Shift algorithm [11] for each image, and all the five human labeled segmentations are used to compose the reference segmentation. One can see that the two input segmentations have visually very different quality, and the proposed measure generates scores with larger discrepancy than the averaging approach (i.e., $|Q_{p1} - Q_{p2}| > |\overline{Q_1} - \overline{Q_2}|$). In Figure 6(b), the first segmentation is worse than the second one and the proposed measure reflects this (i.e., $Q_{p1} < Q_{p2}$). However, the averaging scores of human labeled segmentations suggest otherwise.

5 Datasets

To assess the performance of a segmentation algorithm, it is crucial to develop a dataset with multiple labeled segmentations and human evaluation scores of each image. Although several image segmentation datasets have been developed [29], [3], [1], [26], there are some limitations in terms of number of labels and subject scores. In this work, we develop a new image segmentation dataset and a novel segmentation evaluation dataset with subject scores. These two datasets are available at http://www4.comp.polyu.edu.hk/~cslzhang/ISE/ISE.htm.

5.1 Image segmentation dataset

Several image segmentation datasets have been constructed in past decades. The Weizmann segmentation dataset [3] contains 200 images with labeled foreground (with one or two objects) and background segments. Achanta et al. [1] develop a saliency segmentation dataset with 1000 images. The salient objects in each image are manually segmented based on the salient regions drawn by Liu et al. [26]. However, only one labeled segmentation is generated for each image.

The Berkeley Segmentation Database (BSDS500) [29] is large and representative which has been used in numerous vision problems. It contains 500 source images with 4 to 9 human labeled segmentations per image. Nevertheless, some of the labeled object contours can be delineated more precisely with greater details. Furthermore, the number of labeled results per image can be extended to cover a wide range of visual perception differences.

In order to better evaluate segmentation algorithms, we construct a new segmentation dataset of 200 images where each one is labeled by 6 to 15 persons. We develop a platform that facilitates drawing object boundaries in an image (See Figure 7). Two functions to label image segmentations are provided, either manually or with software aids. As hand tremors often lead to unsmooth segmentation on object boundaries for users with no prior training on digital art, we use the livewire algorithm [15] to facilitate the labeling process. This tool allows a user to initialize a starting point on the boundary and the subsequent point is selected interactively based on the shortest path to best fit the object of interest. For boundaries in the blurred regions or complex objects, users are suggested to use manual segmentations to minimize errors via the interactive process. Compared to the process with all manual labels, these two modules help to generate accurate boundaries with less effort.

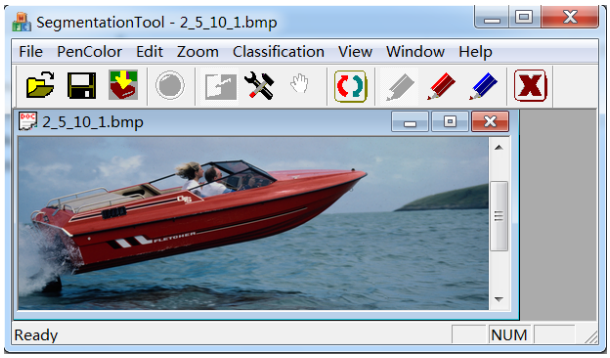


Fig. 7. User interface of the developed image segmentation tool.

To construct the segmentation library, we ask 45 subjects to segment images. Each subject is randomly assigned 50 to 150 images, and is asked to segment each image into 3 to 100 pieces. In order to reduce ambiguous interpretations caused by image contents, subjects are asked to pay more attention to low level features (e.g., color, textures) and pay equal attention to all objects in the scenes. Figure 8 shows some sample segmented images in the developed dataset. Note that the segmentations are labeled with different levels of details that correspond to different visual perception.

Fig. 8. Sample images and the labeled segmentations in the developed dataset.

The statistics of the proposed dataset and the BSDS500 database are summarized in Table 1. Although the number of source images in the proposed dataset is smaller than that of the BSDS500 database, each image in our dataset is labeled by more human subjects. By using the

developed toolkit, the segmentation time in our dataset is much shorter than that with the BSDS500 database. This largely reduces the efforts of human subjects and allows them to focus on drawing boundary details of objects. Figure 9 shows the characteristics of labeled segmentations in the two datasets. In the BSDS500 dataset, most images have 5 to 6 labeled segmentations by humans while most images in our developed dataset are labeled by 8 to 13 subjects.

TABLE 1
Summary of the BSDS500 and proposed datasets.

Dataset	BSDS500	Our Dataset
# images	500	200
# labeled segmentations/image	4-9 (5.4 on average)	6-15 (10.7 on average)
Image type	Natural images	Natural images
Software supported	Yes	Yes
# subjects	30	45
Time/segmentation	5-30 mins	2-4 mins

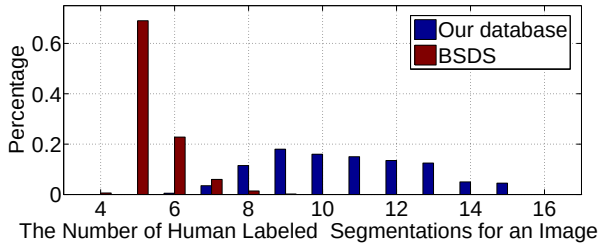


Fig. 9. Distribution of labeled segmentations in the proposed and BSDS500 datasets

5.2 Segmentation evaluation dataset

Several meta-measures [28], [34] have been proposed to compare the performance of a pair of segmentation algorithms based on a segmentation dataset with human labeled results. Nonetheless, these measures are limited for quantitatively evaluating the segmentation algorithms. In [32], 10 different segmentations per image (from a set of 80 images) labeled as "good" or "bad" are used to select parameters for segmentation algorithms. An evaluation dataset consisting of 199 pairs of human and machine segmentations are constructed [54], where the human segmentations are indicated as "good" ones. In addition, results for another set of 249 pairs are presented. However, the number of segmentation pairs or the diversity of machine segmentations is limited. Furthermore, none of these datasets is publicly accessible.

In this work, we design and develop a novel dataset for evaluating segmentation evaluation algorithms. The proposed benchmark dataset contains 500 pairs of segmentations and the corresponding evaluation results by human subjects. The first 200 pairs (Part A) are generated by using 200 images from our segmentation dataset (in Section 5.1), while the other 300 pairs of segmentations (Part B) are generated by 300 images selected from the BSDS500 dataset [29]. As different segmentation algorithms exploit different properties, we generate diverse segmentation results by using the efficient graph-based (EG) algorithm [16], mean-shift (MS) approach [11], compression-based texture merging (CTM) [52] algorithms as well as texture and boundary encoding-based segmentation (TBES) method [37].

As these algorithms are developed based on various criteria, the segmentation results are different and diverse. Table 2 shows the parameter settings of the 4 algorithms used to generate the segmentations (the parameter settings are not the same as those in the original settings). With these parameters, each algorithm generates diverse results ranging from over-segmentation to under-segmentation.

TABLE 2
Parameter settings of the four algorithms for generating segmentations in our evaluation dataset.

Algorithms	Parameter values
EG [16]	$K = \{600, 800, 1000, 1400, 1800\}$
MS [11]	$h_r = \{7, 11, 15, 19, 23\}$, $h_s = 7$, $min_R = 150$.
CTM [52]	$\epsilon = \{0.1, 0.2, 0.3, 0.4, 0.5\}$
TBES [37]	$N_{sp} = 200$. $\epsilon = \{50, 100, 200, 300, 400\}$

After inspecting and removing the results with poor segmentations with little semantic meaning, we obtain 17 segmentations for each image. Next, 10 human subjects are asked to select the best 3 and the worst 3 segmentations from these segmentations. Based on the consensus of human evaluation, a candidate group of good segmentations is constructed (and likewise for the bad segmentations). For each image, we randomly select one segmentation from the group of good results and pair it with a segmentation randomly selected from the group of bad segmentations. We form the segmentation pairs to ensure that the quality difference in each pair is not too small to tell by human subjects. It should be noted that machine segmentations generally have much lower quality than human segmentations, and thus in our dataset there are few pairs in which one segmentation is clearly better than the other. This ensures the level of difficulty to distinguish the quality of different segmentations. In addition, since the segmentations in each pair may be produced by different algorithms, the results are more diverse. Finally, 500 pairs of segmentations are generated in the evaluation dataset. Figure 10 shows some example pairs in our dataset.

In this work, 70 subjects with little or no research experience in image segmentation are asked to evaluate the 500 pairs of segmentations. We note that one may misjudge the segmentation quality when a segmentation contains too many regions. Instruction is given to subjects that all segmented regions should have approximately equal importance in evaluation. We evenly divide

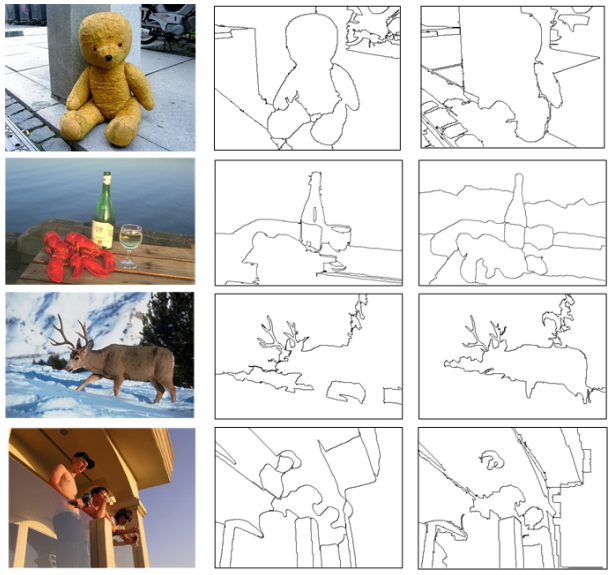


Fig. 10. Sample pairs in the segmentation evaluation dataset.

the 500 segmentations pairs into 10 groups. Each time one subject is only asked to evaluate one group (each one evaluates no more than 4 groups).

The consistency of the evaluation results is measured by the confidence rate, which is defined as the percentage of subjects making the same judgment for the same pair. The distribution of confidence rates is plotted in Figure 11. The confidence rate reflects the level of difficulties in evaluating the pair of segmentations. Figure 11 shows that more than 50% of the evaluations have the confidence rate over 0.8. About 10% of the evaluations have the confidence rate between 0.5 and 0.6, which correspond to the difficult pairs for comparisons.

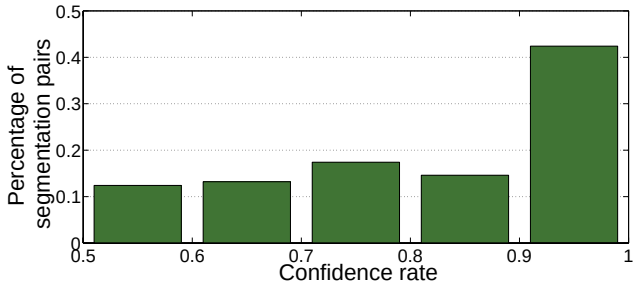


Fig. 11. Distribution of confidence rates on the proposed segmentation evaluation dataset.

6 Experimental Results

We evaluate the proposed segmentation quality measure Q_p in comparison with existing measures based on PRI [46], VOI [30], GCE [28], SC [4], BDE [18] and F-measure [28]. Among the evaluated methods, the PRI and our proposed measure work with multiple labeled segmentations. The other measures operate on each labeled reference individually, and the average results are reported.

All experiments are carried out on a desktop computer with an Intel Core 2 Duo 3.00 GHz CPU and 4GB memory. The run time of the proposed measure consists two parts: 24.6 ± 6.0 seconds for composing the reference G^* and 10.7 ± 1.1 seconds for computing the score Q_p.

6.1 Sensitivity analysis

In Section 3, the energy function (7) is defined to construct the reference segmentation, yet, the value of parameter λ should be chosen before implementing the algorithm. Meanwhile, the initial labeling of graph cut algorithm [8] is randomly decided in the optimization. Experiments in [8] have shown that varying the initial labelings does not significantly change the final result. In this section, we perform extensive experiments to test the effects of λ and initial labeling on the final evaluation score.

In all our experiments in Sections 6.2 and 6.3, we fix λ to be 800. Therefore, we evaluate the effect of λ within a moderate range around 800, more specifically, [500,1200] with an interval of 50. We carry out experiments to analyze the standard deviations of our measure Q_p on each of the 1,000 segmentations in the evaluation database (Part A and Part B). The initial labeling of graph cut is set randomly, then the mean values and standard deviations of Q_p with respect to the 15 different values of λ are computed. Figure 12 shows the results sorted in an ascending order of the mean value of Q_p.

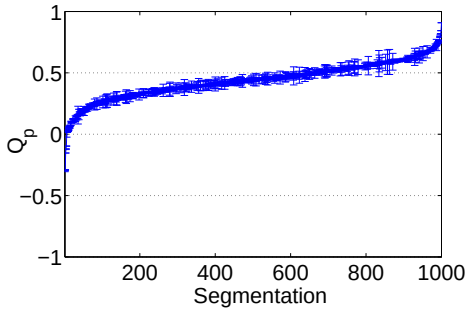


Fig. 12. Means and standard deviations of Q_p for 1,000 segmentations in our evaluation database. Error bars show the standard deviations with respect to λ, where λ is set within the range of [500, 1200] with an interval of 50. The results are sorted in an ascending order of means.

As shown in Figure 12, there are 963 of the 1,000 segmentations with deviation smaller than 0.05. Only 1 segmentation has deviation larger than 0.1 and the deviation is 0.106. Since Q_p is a number between -1 and 1, the results show that our measure is insensitive to the change of λ within a large range.

To evaluate the effect of initial labeling on the final evaluation score, we carry out the proposed algorithm 50 times with random initialization of labeling. Figure 13 shows the plot of the means and standard deviations of

Q_p for the 1,000 segmentations where λ is fixed to be 800. For 973 of the 1,000 segmentations, the deviation of Q_p is less than 0.05, and the largest deviation is 0.078. These results show that Q_p does not change much by varying the initial labelings. Therefore, we carry out experiments using the proposed algorithm with random initialization.

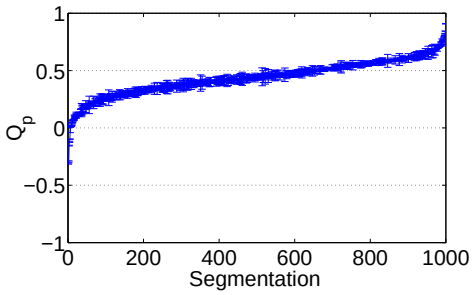


Fig. 13. Means and standard deviations of Q_p for 1,000 segmentations in our evaluation database. Error bars show the standard deviation of Q_p with respect to 50 random initial labelings. The results are sorted in ascending order of means.

6.2 Evaluation with meta-measure

By verifying different hypotheses on the evaluation outputs, meta-measures [28], [34] have been proposed to compare the goodness of the segmentation evaluation measures. For example, one hypothesis is that a good measure should be able to discriminate two pairs of human segmentations, one pair from the same image while another pair from different images. More specifically, the quality score of a human segmentation evaluated by segmentations from the same image should be better than that evaluated by segmentations from a different image. Based on this principle, a meta-measure [28] is defined to count the number of human segmentation pairs coming from the same images which are misjudged as less similar than other pairs from different images. Pont-Tuset and Marques [34] further extended this meta-measure to discriminate machine segmentations from different images.

We evaluate the proposed measure using an approach similar to that in [28], [34]. The meta-measure is defined as the percentage of human labeled segmentations from the same images that are determined as more similar than the machine segmentations from different images. Namely, the meta-measure is used to compare each human labeled segmentation of a certain image with two other groups of segmentations: (i) human labeled segmentations of the same image and (ii) machine segmentations of a different image. The rationale is that the evaluation score generated by case (i) should be better than that by case (ii), and we use the percentage of comparisons that agree with this principle as the meta-measure result. This comparison incorporates the discriminations of segmentations from the same and the different images, and segmentations created by different sources (human labeled and machine generated).

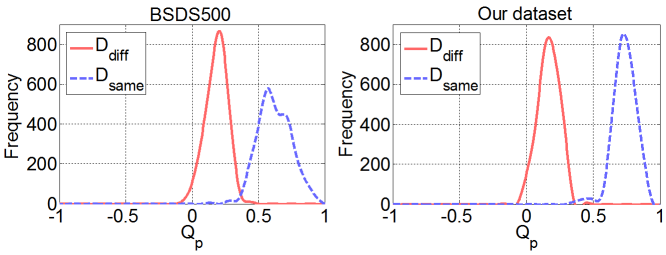


Fig. 14. Distributions of Q_p for segmentations of the same images (in blue) and different images (in red) on the BSDS500 and proposed datasets, respectively.

We evaluate all the human segmentations in the BSDS500 and our proposed segmentation datasets. The same amount of machine segmentations are randomly created by the EG, MS, CTM and TBES methods for each image in the databases. Each human segmentation is evaluated by the rest of human segmentations of the same image, as well as the same number of machine segmentations from a different image. Let D_{same} and D_{diff} be the distributions of scores in case (i) and case (ii), respectively. Figure 14 shows the distributions for the two types of segmentations using the Q_p measure. The meta-measure is computed as the percentage of comparisons outside the overlap between D_{same} and D_{diff} (the percentage of overlap is reported as Bayes risk in [28]). Table 3 shows the meta-measure results for different segmentation quality evaluation measures. On both databases, the proposed measure performs favorably better than other measures, which demonstrates the effectiveness of using composed references for evaluation. Except for the F-measure, all the other measures obtain higher scores on the proposed dataset than the BSDS500 database, which suggests the merits of using more reference segmentations for segmentation evaluation.

6.3 Evaluation with proposed segmentation dataset

We first use an example to illustrate the effectiveness of the proposed measure for assessing segmentation quality. Figure 15 shows 5 segmentations of a given image generated by the EG method [16] with different parameters. In this experiment, 10 subjects are asked to rank the segmentation results. Most participants agree that the best segmentation is Figure 15(d) since it preserves the main structure of the object with minimal number of misclassified boundaries, followed by (c), (e), (a) and (b). The quality scores measured by different methods are shown on the right side in Figure 15. These plots indicate that the proposed measure matches human perception best, while other measures either do not reflect the best segmentation (i.e., F-measure) or do not well differentiate the best segmentation and the others (i.e., PRI and BDE). The values of the SC measure are in

TABLE 3
Evaluation results with the meta-measure.

Measures	PRI	GCE	VOI	BDE	F-measure	$SC(S \to G)$	$SC(G \to S)$	Q_P
BSDS500	0.911	0.929	0.967	0.921	0.882	0.962	0.956	0.984
Proposed dataset	0.959	0.981	0.991	0.947	0.838	0.974	0.979	0.994

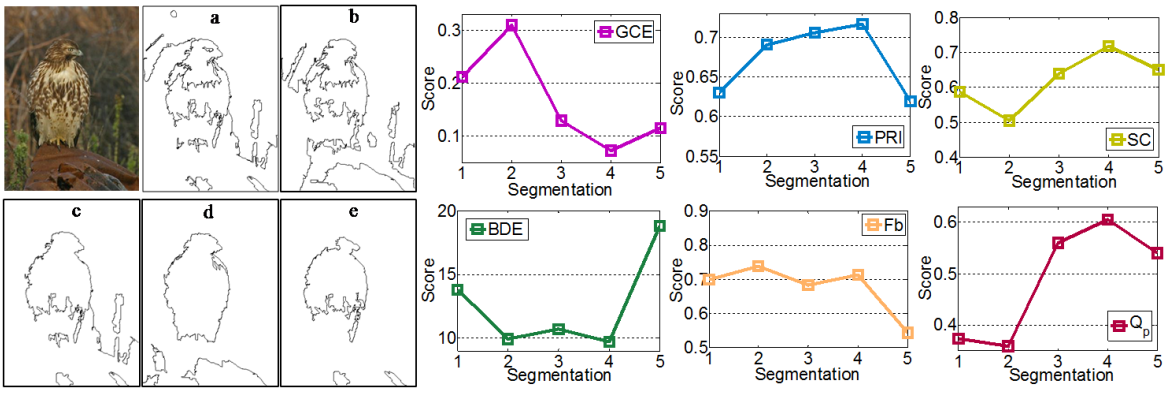


Fig. 15. Quality scores of the segmentations in $a \sim e$ by the EG algorithm [16] using different measures, i.e., GCE, PRI, $SC(S \to G)$, BDE, F_b (F-measure on boundary) and Q_p.

TABLE 4
Evaluation results by different measures.

Measures		PRI	GCE	VOI	BDE	F-measure	$SC(S \to G)$	$SC(G \to S)$	Ave(Q_p)	Min(Q_p)	Max(Q_p)	Q_P
Part A (200 pairs)	Correct No.	156	156	148	146	146	133	147	165	160	164	168
	Rate (%)	0.78	0.78	0.74	0.73	0.73	0.67	0.74	0.83	0.80	0.82	0.84
Part B (300 pairs)	Correct No.	241	143	182	237	232	165	215	120	137	118	251
	Rate (%)	0.80	0.48	0.61	0.79	0.77	0.55	0.72	0.4	0.46	0.39	0.83

a relatively small range, which do not reflect the differences of segmentations well. In contrast, the proposed measure is effective for modeling segmentation quality as it adaptively evaluates image structures on different levels.

We then quantitatively examine the segmentation measures using the proposed evaluation dataset presented in Section 5.2. In addition, we evaluate three other measures: Q_p based on average Ave(Q_p), minimum Min(Q_p), and maximum Max(Q_p) of human annotated scores. Table 4 shows the number of correct evaluations (i.e., the ones which are consistent with human judgments) for all evaluated measures. On both Part A and Part B of our dataset, the proposed measure by using reference (Q_p) outperforms existing measures. For GCE, VOI, Ave(Q_p), Min(Q_p) and Max(Q_p), the performance on Part A and Part B varies significantly. When the number of labeled segmentations per image is small (e.g., Part B), there is a significant decline in the number of correct evaluations. The PRI, BDE and F-measure measures have opposite trends on the two parts but with less variation. The proposed measure has the highest correction rate and comparable results on both sets, which can be attributed to the composition of references with the proposed evaluation measure.

Another important factor is whether a measure is effective when it is difficult to evaluate a segmentation pair. We compare the false evaluation rates with respect to the confidence rate of human subjects (See Figure 11 for distribution of confidence rate). Figure 16 shows the results where we uniformly quantize the confidence rate into 5 bins, and count the falsely evaluated pairs in each bin. When the confidence rate is low (i.e., less than 80%), Q_p has lower false evaluation rates than the others. When the confidence rate is high (e.g., in the range between 0.9 and 1), the advantages of our measures are not significant. This can be explained by the fact that if a segmentation is clearly good or bad (which usually results in high confidence in subject evaluation), the task is easier and many existing measures perform well.

7 Conclusions

We proposed a framework for evaluating segmentation quality with multiple human labeled segmentations to take into account both local structure and global consistency of segmentations. To achieve this goal, a reference segmentation was adaptively constructed for a given segmentation and used in conjunction with the proposed

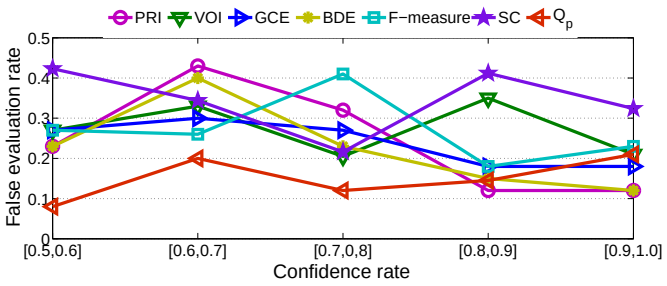


Fig. 16. False evaluation rate of segmentation pairs under different confidence levels.

measures to compute quality score. In addition, we presented a segmentation dataset and segmentation evaluation dataset to facilitate quantitative quality assessment. The segmentation dataset contains images with more labeled segmentations than BSDS, which is important for objective evaluation. The evaluation dataset is diverse in segmentation quality and contains extensive subjective evaluation results. Both datasets are publicly available to the research community. Extensive experiments on the proposed datasets and the BSDS dataset demonstrate the effectiveness of our framework in evaluating segmentation quality.

ACKNOWLEDGMENT

This work was supported by the HK RGC GRF Grant (No. PolyU 5315/12E), the NSFC (Nos.61202190, 61571359), the National Key Basic Research Program (2016YFA0202003) and US NSF CAREER Grant (No. 1149783).

REFERENCES

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, 2009. 8

[2] A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. *ACM Transactions on Graphics*, 23(3):294–302, 2004. 4

[3] S. Alpert, M. Galun, R. Basri, and A. Brandt. Image segmentation by probabilistic bottom-up aggregation and cue integration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 4, 8

[4] P. Arbelaez, M. Maire, C. C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011. 1, 3, 4, 10

[5] O. Boiman and M. Irani. Similarity by composition. In *Advances in Neural Information Processing Systems*, pages 177–184, 2006. 4

[6] M. Borsotti, P. Campadelli, and R. Schettini. Quantitative evaluation of color image segmentation results. *Pattern Recognition Letters*, 19(8):741–747, 1998. 1

[7] K. Bowyer, C. Kranenburg, and S. Dougherty. Edge detector evaluation using empirical roc curves. *Computer Vision and Image Understanding*, 84(1):77–103, 2001. 4

[8] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1222–1239, 2001. 5, 10

[9] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3241–3248, 2010. 4

[10] H. Christensen and P. Phillips. *Empirical evaluation methods in computer vision*. World Scientific Publishing Company, 2002. 1

[11] D. Comanicu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:603–619, 2002. 3, 8, 9

[12] I. Endres and D. Hoiem. Category independent object proposals. In *European Conference on Computer Vision*, pages 575–588, 2010. 4

[13] F. Estrada and A. Jepson. Quantitative evaluation of a novel image segmentation algorithm. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1132–1139, 2005. 4

[14] M. Everingham, H. Muller, and B. Thomas. Evaluating image segmentation algorithms using the pareto front. In *European Conference on Computer Vision*, pages 34–48, 2002. 3

[15] A. Falcao, J. Udupa, S. Samarasekara, and S. Sharma. User-steered image segmentation paradigms: Live wire and live lane. *Graphical Models and Image Processing*, 60:233–260, 1998. 8

[16] P. Felzenszwalb and D. Huttenlocher. Efficient graph based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004. 3, 9, 11, 12

[17] E. Fowlkes and C. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983. 3

[18] J. Freixenet, X. Munoz, D. Raba, J. Mart, and X. Cuf. Yet another survey on image segmentation: region and boundary information integration. In *European Conference on Computer Vision*, pages 408–422, 2002. 4, 10

[19] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *2013 Journal of Robotics Research (IJRR)*, 32(11):1231–1237, 2008. 3

[20] B. Goldluecke and D. Cremers. Introducing total curvature for image processing. In *IEEE International Conference on Computer Vision*, pages 1267–1274, 2011. 5

[21] A. Hoover, G. Jean-Baptiste, X. Jiang, P. Flynn, H. Bunke, and D. Goldgof. An experimental comparison of range image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(7):673–689, 1996. 3

[22] Q. Huang and B. Dom. Quantitative methods of evaluating image segmentation. In *International Conference on Image Processing*, pages 53–56, 1995. 3

[23] A. Ion, J. Carreira, and C. Sminchisescu. Image segmentation by figure-ground composition into maximal cliques. In *IEEE International Conference on Computer Vision*, pages 2110–2117, 2011. 4

[24] X. Jiang, C. Marti, C. Irniger, and H. Bunke. Distance measures for image segmentation evaluation. *EURASIP Journal on Applied Signal Processing*, 2006:209, 2006. 1, 3

[25] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollã, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 3

[26] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 8

[27] T. Malisiewicz and A. Efros. Improving spatial support for objects via multiple segmentations. In *British Machine Vision Conference*, pages 1–10, 2007. 4

[28] D. Martin. *An Empirical Approach to Grouping and Segmentation*. PhD thesis, EECS Department, University of California, Berkeley, 2002. 1, 3, 4, 6, 9, 10, 11

[29] 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 *IEEE International Conference on Computer Vision*, pages 416–424, 2001. 1, 2, 3, 6, 7, 8, 9

[30] M. Meila. Comparing clusterings: an axiomatic view. In *International Conference on Machine Learning*, pages 577–584, 2005. 3, 10

[31] C. Odet, B. Belaroussi, and H. Benoit-Cattin. Scalable discrepancy measures for segmentation evaluation. In *International Conference on Image Processing*, pages 785–788, 2002. 4

[32] B. Peng and O. Veksler. Parameter selection for graph cut based image segmentation. In *British Machine Vision Conference*, pages 153–162, 2008. 9

[33] B. Peng and L. Zhang. Evaluation of image segmentation quality by adaptive ground truth composition. In *European Conference on Computer Vision*, pages 287–300, 2012. 1

[34] J. Pont-Tuset and F. Marques. Measures and meta-measures for the supervised evaluation of image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2131–2138, 2013. 4, 9, 11

[35] R. Potts. Some generalized order-disorder transformation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48:106–

109, 1952. 5
[36] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 60(336):846–850, 1971. 3, 4
[37] S. Rao, H. Mobahi, A. Yang, S. Sastry, and Y. Ma. Natural image segmentation with adaptive texture and boundary encoding. In *Asian Conference of Computer Vision*, pages 135–146, 2009. 9
[38] X. Ren and J. Malik. Learning a classification model for segmentation. In *IEEE International Conference on Computer Vision*, pages 10–17, 2003. 1
[39] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Segmenting scenes by matching image composites. In *Advances in Neural Information Processing Systems*, pages 1580–1588, 2009. 4
[40] M. Sampat, Z. Wang, S. Gupta, A. Bovik, and M. Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE Transactions on Image Process*, 18(11):2385–2401, 2009. 6
[41] T. Schoenemann, F. Kahl, S. Masnou, and D. Cremers. A linear framework for region-based image segmentation and inpainting involving curvature penalization. *International Journal of Computer Vision*, 99(1):53–68, 2012. 5
[42] M. Shin, D. Goldgof, and K. Bowyer. Comparison of edge detector performance through use in an object recognition task. *Computer Vision and Image Understanding*, 84(1):160–178, 2001. 4
[43] J. Stuhmer, P. Schroder, and D. Cremers. Tree shape priors with connectivity constraints using convex relaxation on general graphs. In *IEEE International Conference on Computer Vision*, pages 2336–2343, 2013. 5
[44] R. Unnikrishnan and M. Hebert. Measures of similarity. In *In IEEE Workshop on Applications of Computer Vision*, pages 394–400, 2005. 1, 3
[45] R. Unnikrishnan, C. Pantofaru, and M. Hebert. A measure for objective evaluation of image segmentation algorithms. In *Workshop on Empirical Evaluation Methods in Computer Vision, IEEE Conference on Computer Vision and Pattern Recognition*, page 34, 2005. 1, 3
[46] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):929–944, 2007. 1, 3, 4, 10
[47] S. van Dongen. Performance criteria for graph clustering and markov cluster experiments. Technical Report INS-R0012, Centrum voor Wiskunde en Informatica, 2002. 3
[48] S. Vicente, V. Kolmogorov, and C. Rother. Graph cut based image segmentation with connectivity priors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 5
[49] P. Viola and W. Wells. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–C154, 1997. 4
[50] Z. Wang and A. Bovik. *Modern image quality assessment*. Morgan and Claypool Publishing Company, New York, 2006. 1
[51] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Process*, 13(4):600–612, 2004. 4
[52] A. Yang, J. Wright, Y. Ma, and S. Sastry. Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 11(2):212–225, 2008. 9
[53] H. Zhang, J. Fritts, and S. Goldman. An entropy-based objective segmentation evaluation method for image segmentation. In *SPIE Sotrage and Retrieval Methods and Applicaitons for Multimedia*, pages 38–49, 2004. 1
[54] H. Zhang, J. Fritts, and S. Goldman. A co-evaluation framework for improving segmentation evaluation. In *SPIE Signal Processing and Target Recognition*, pages 420–430, 2005. 9

Bo Peng is an Assistant Professor in the School of Information Science & Technology, Southwest Jiaotong University, Chengdu, China. She received the M.S. degree from the Department of Computer Science, University of Western Ontario (UWO) in 2008 and the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University in 2012. From Aug. 2011 to Jan. 2012, she worked as a Research Assistant in the Department of Computing, The Hong Kong Polytechnic University. Her research interests include image segmentation and segmentation quality evaluation.

Lei Zhang (M04, SM14) received his B.Sc. degree in 1995 from Shenyang Institute of Aeronautical Engineering, Shenyang, P.R. China, and M.Sc. and Ph.D degrees in Control Theory and Engineering from Northwestern Polytechnical University, Xian, P.R. China, respectively in 1998 and 2001, respectively. From 2001 to 2002, he was a research associate in the Department of Computing, The Hong Kong Polytechnic University. From January 2003 to January 2006 he worked as a Postdoctoral Fellow in the Department of Electrical and Computer Engineering, McMaster University, Canada. In 2006, he joined the Department of Computing, The Hong Kong Polytechnic University, as an Assistant Professor. Since July 2015, he has been a Full Professor in the same department. His research interests include Computer Vision, Pattern Recognition, Image and Video Processing, and Biometrics, etc. Prof. Zhang has published more than 200 papers in those areas. As of 2016, his publications have been cited more than 20,000 times in the literature. Prof. Zhang is an Associate Editor of IEEE Trans. on Image Processing, SIAM Journal of Imaging Sciences and Image and Vision Computing, etc. He is a "Highly Cited Researcher" selected by Thomson Reuters. More information can be found in his homepage http://www4.comp.polyu.edu.hk/~cslzhang/.

Xuanqin Mou has been with the Institute of Image Processing and Pattern Recognition (IPPR), Electronic and Information Engineering School, Xian Jiaotong University, since 1987. He has been an Associate Professor since 1997, and a Professor since 2002. He is currently the director of IPPR, and the director of the National Data Broadcasting Engineering and Technology Research Center of China. He served as the member of the 12th Expert Evaluation Committee for the National Natural Science Foundation of China, the Member of the 5th, 6th and 7th Executive Committee of China Society of Image and Graphics, the Vice President of Shaanxi Image and Graphics Association. He has authored or co-authored more than 200 peer-reviewed journal or conference papers. He has been granted as the Yung Wing Award for Excellence in Education, the KC Wong Education Award, the Technology Academy Award for Invention by the Ministry of Education of China, and the Technology Academy Awards from the Government of Shaanxi Province, China.

Ming-Hsuan Yang is an associate professor in Electrical Engineering and Computer Science at University of California, Merced. He received the PhD degree in computer science from the University of Illinois at Urbana-Champaign in 2000. Prior to joining UC Merced in 2008, he was a senior research scientist at the Honda Research Institute working on vision problems related to humanoid robots. He coauthored the book Face Detection and Gesture Recognition for Human-Computer Interaction (Kluwer Academic 2001) and edited special issue on face recognition for Computer Vision and Image Understanding in 2003, and a special issue on real world face recognition for IEEE Transactions on Pattern Analysis and Machine Intelligence. Yang served as an associate editor of the IEEE Transactions on Pattern Analysis and Machine Intelligence from 2007 to 2011, and is an associate editor of the International Journal of Computer Vision, Image and Vision Computing and Journal of Artificial Intelligence Research. He received the NSF CAREER award in 2012, the Senate Award for Distinguished Early Career Research at UC Merced in 2011, and the Google Faculty Award in 2009. He is a senior member of the IEEE and the ACM.