A Co-Saliency Model of Image Pairs

Hongliang Li, *Member, IEEE*, and King Ngi Ngan, *Fellow, IEEE*

***Abstract*—In this paper, we introduce a method to detect co-saliency from an image pair that may have some objects in common. The co-saliency is modeled as a linear combination of the single-image saliency map (SISM) and the multi-image saliency map (MISM). The first term is designed to describe the local attention, which is computed by using three saliency detection techniques available in literature. To compute the MISM, a co-multilayer graph is constructed by dividing the image pair into a spatial pyramid representation. Each node in the graph is described by two types of visual descriptors, which are extracted from a representation of some aspects of local appearance, e.g., color and texture properties. In order to evaluate the similarity between two nodes, we employ a normalized single-pair SimRank algorithm to compute the similarity score. Experimental evaluation on a number of image pairs demonstrates the good performance of the proposed method on the co-saliency detection task.**

***Index Terms*—Attention model, co-saliency, similarity, SimRank.**

I. Introduction

VISUAL attention is an effective simulation of perceptual behavior, which aims to find a salient object from its surroundings by computing a spatial saliency map. In the past several years, attention models have been successfully applied to many fields, such as object recognition [1]–[3], image segmentation and understanding [4], adaptive coding [5], object tracking [6], image database querying, video retrieval and summary [7], [8].

Effective attention can be carried out as a bottom up process without any information about the sought for objects [9]. A saliency-based visual attention model for rapid scene analysis was first presented in [10], which combined multiscale image features into a single topographical saliency map. This model was successfully applied to object extraction from color images [11], which formulated the attention towards objects as a Markov random field (MRF) by integrating computational visual attention mechanisms with attention object growing techniques. This model was also extended to segment video objects of interest such as the facial saliency model [12] and the focused saliency model [13].

In order to extract visual attention effectively, a lot of methods have been presented recently to deal with salient points detection. Based on the center-surround mechanism [10], a visual saliency measure called Site Entropy Rate is proposed to compute the average information transmitted from a node (neuron) to all the others during the random walk on the graphs/network [14]. This method uses the information maximization principle to construct a new computational model for visual saliency. In [15], a new type of saliency, namely context-aware saliency, is proposed to detect the image regions that represent the scene. Instead of identifying fixation points or detecting the dominant object, this work is based on four principles (i.e., local and global considerations, visual organization rules and high-level factors) observed in the psychological literature [16]–[19], which helps to produce compact, appealing, and informative summaries. Recently, a global contrast based method is proposed to measure the region saliency [20].

Most of existing saliency models focus on detecting salient objects from an image rather than an image pair. It is known that same (similar) objects detection from multiple images has become one of the most important and challenging problems in computer vision and multimedia applications, such as common pattern discovery [21], [22] , image matching and co-recognition [23]. It can be seen as a combination of similar objects identification and co-classification tasks. The first task is to measure the degree of similarity (e.g., structural similarity [24]) between images using a feature matching method. The second task aims to extract the object by grouping together similar pixels, which provides closed boundary or mask of the similar object. Inspired by the work on simultaneous object segmentation, our goal in this paper is to extract the similar objects from image pairs.

A similar work with our approach is called 'cosegmentation' that aims to segment the similar object from images. This method can be traced back to the work of Rother [25], which relies on a generative model for cosegmenting the common parts of an image pair. To match the appearance histograms of the common parts, this method presented trust region graph cuts to minimize an energy with a MRF term encoding spatial coherency and a global constraint. Inspired by [25], the histogram constraint was added in [26] as the regularized term in a segmentation objective function. MRF energy term is used for the simultaneous segmentation together with a penalty on the sum of squared differences of the foreground region histograms. Instead of penalizing the difference (distance) of the two foreground histograms, a cosegmentation algorithm was proposed to reward their similarity by using a maximum flow procedure on an appropriately constructed graph [27]. This method chooses a suitable measure of his-

Manuscript received October 13, 2010; revised January 31, 2011; accepted May 02, 2011. Date of publication May 19, 2011; date of current version November 18, 2011. This work was supported in part by the NSFC (No.60972109), in part by the Program for New Century Excellent Talents in University (NCET-08-0090), and in part by Sichuan Province Science Foundation for Youths (No. 2010JQ0003). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zhou Wang.

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

K. N. Ngan is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: knngan@ee.cuhk.edu.hk).

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

Digital Object Identifier 10.1109/TIP.2011.2156803

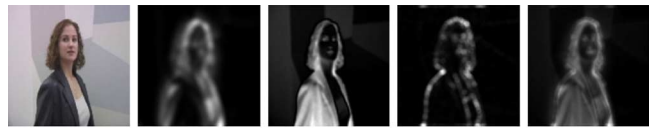
togram consistency, which leads to a polynomial time algorithm for cosegmentation.

Recently, a discriminative clustering framework for image cosegmentation is proposed in [28], which combines existing tools for bottom up image segmentation such as the spectral clustering technique and positive definite kernels. The main idea in this work is to train a supervised classifier for maximal separation of the foreground and background classes. The optimal solution of the supervised learning problem is obtained by using an efficient convex relaxation. The comparisons of existing models and corresponding optimization techniques are discussed in [29], which show that these methods [25]–[27] differ only in the distance measure between the two color histograms. Based on the analysis, this work is also extended to the binary image segmentation by incorporating a Dual Decomposition optimization technique. In addition, some works focus on a simplified cosegmentation method based on the user interaction, such as [30], which discusses how to decide the seed image among a group of related images.

In this paper, we introduce a perceptual model to describe the similar entity (e.g., a region or object) within an image pair. We will refer to such entity in a pair of images as co-saliency, which is defined as follows: 1) Each region in the pair should have strong local saliency with respect to its surroundings. 2) The region pair should exhibit high similarity of certain features (e.g., intensity, color, texture, or shape). Given an image pair, co-saliency is closely related to how we perceive visual stimuli and fixate on the most valuable information from the image pair. Compared with image cosegmentation that partitions an image pair into different segments (i.e., similar regions and backgrounds), saliency as a concept from human vision implies a selection and/or ranking by importance. More precisely, co-saliency is the subjective perceptual quality that makes similar objects in an image pair stand out from their neighbors and capture our attention by visually salient stimuli.

We build the co-saliency model to simulate the attention search process for an image pair, which can be obtained by a linear combination of the SISM and MISM. The first term SISM focuses on a local saliency description. Three saliency detection techniques, namely Itti's model [10], frequency-tuned saliency [32] and spectral residual saliency [33], are used to generate a robust single-image saliency map. Apart from the SISM, an important task of our method is to find the co-salient objects from the image pair. A co-multilayer graph is designed by performing the image pyramid decomposition, where the similarity between two nodes can be obtained by computing the distance of the node-pair. Notice that two types of region descriptors, i.e., color and texture, are used to represent the region aspects of local appearance. Finally, we use a normalized single-pair SimRank algorithm to compute similarity scores. Experimental evaluation on a number of image pairs shows that our proposed method can detect co-saliency effectively. In addition, an extension to the cosegmentation is also addressed, which demonstrates the advantage of our proposed method.

This paper is organized as follows. Section II introduces our proposed co-saliency algorithm. Experimental results are provided in Section III to demonstrate the effectiveness of our approach. Finally, Section IV concludes the paper.



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

Fig. 1. Example of the single-image saliency map. (a) Original image *amira*. (b) Saliency map by [10]. (c) Saliency map by [32]. (d) Saliency map by [33]. (e) Our single-image saliency map.

II. Proposed Method

The co-saliency defined in our paper is obtained by computing the single-image saliency and multi-image saliency maps. The first is used to identify the salient regions within each image. The second aims to measure the saliency for a pair of images.

A. Single-Image Saliency Map (SISM)

It is widely recognized that salient object detection is very helpful in computer vision and image processing [31]. However, it is still a challenging task to solve this problem. In the current literature, there is no method that can detect the saliency accurately for all images. In order to achieve robust saliency detection, a weighted saliency detection method is proposed in our work, which aims to improve detection performance by combining several saliency maps linearly. The motivation of this work is based on the idea of the voting algorithm, which is to compute the number of times that each model is selected. Assume I denotes an input image, while S_l represents the corresponding single-image saliency map. We have

$$S_l = \sum_{j=1}^{J} w_j \cdot \mathcal{N}(S_{l_j}) \tag{1}$$

where $\mathcal{N}(S_{l_j})$ denotes the jth normalized saliency map where each pixel has the salient value in the range $[0, 1]$. Here, w_j denotes the weight with $\sum_{j=1}^{J} w_j = 1$. From (1), we can see that if a pixel is identified as a salient pixel by most of algorithms, it will have a high single-image saliency value. Otherwise, it can be regarded as a background pixel.

In our work, we calculate three types of saliency maps, namely Itti's model saliency [10], frequency-tuned saliency (FTA) [32], and spectral residual saliency (SRA) [33]. The first is the well-known saliency model which mimics the visual search process of human. The saliency map is computed using multiscale image features in a bottom-up manner. The second estimates the center-surround contrast using color and luminance features based on a frequency-tuned approach. The third saliency model employs the log-spectrum of an input image, and extracts the spectral residual of an image in the spectral domain. The three saliency models perform the saliency detection in different ways. Advantages of each method are expected based on such combination.

An example of the single-image saliency map is illustrated in Fig. 1, where the original image *amira* is shown in Fig. 1(a). Fig. 1(b) to Fig. 1(d) show the saliency maps extracted by the

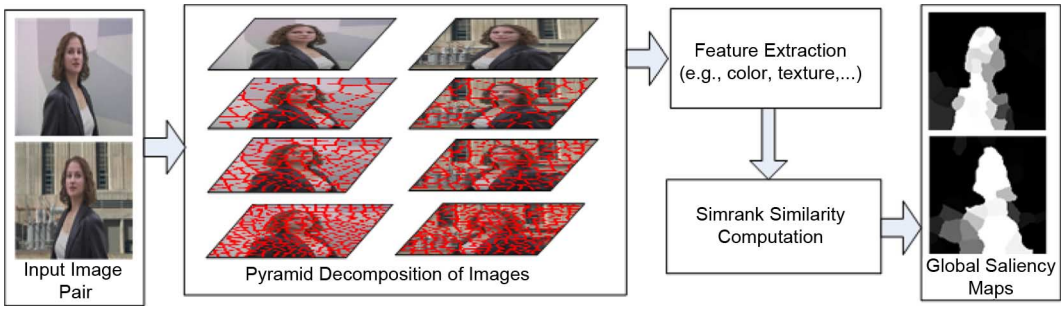


Fig. 2. Block diagram of the multi-image saliency extraction.

methods [10], [32], and [33], respectively. Difference in selected regions can be observed from the results produced by these models. For example, the gray coat is identified as a salient object by [32], but it fails to be detected by the method [33]. The similar case can also be found for the skin regions. However, most of boundaries of the girl are identified as salient regions for all methods. The single-image saliency map generated by combining these maps, as shown in Fig. 1(e), contains all possible salient regions.

B. Multi-Image Saliency Map (MISM)

Unlike the single-image saliency map that is used to describe the region saliency within an image, the goal of MISM is to extract the multi-image saliency information from multiple images. Given a pair of images, the multi-image saliency is defined as the inter-image correspondence, which can be obtained by feature matches. It is known that the visual system relies on several heuristics to direct attention to important locations and objects [34]. The subject is searching for a particular or interesting object, and the attention is geared to react when it appears [35]. Therefore, if the two images contain a similar object, the object region in each image should be assigned high visual saliency values. It means that more visual attention will be attracted by this object. Otherwise, low multi-image saliency values should be considered for the dissimilar regions. According to this principle, the multi-image saliency map of the image I_i is defined as

$$S_g(I_i(p)) = \max_{q \in I_j} Sim(I_i(p), I_j(q)) \tag{2}$$

where p and q denote entities (e.g., pixels or regions) in images I_i and I_j, respectively. $Sim()$ represents a function that measures the similarity between two entities.

The block diagram of our proposed multi-image saliency detection is given in Fig. 2, which mainly consists of four stages, namely pyramid decomposition, feature extraction, SimRank optimization, and multi-image saliency computation.

1) Pyramid Decomposition of an Image Pair: This stage is used to obtain a pyramid of images with decreasing resolutions. As a first step of the MISM computation described in Fig. 2, an initial over-segmentation is performed by partitioning an image into multiple regions. Each image is divided into a sequence of increasingly finer spatial regions by repeatedly partitioning the regions at each level of resolution. This is a pyramid decomposition because each region at one level may be divided into several subregions at the next level.

Given an image pair, we decompose each into multiple segmentations using the methods [36] and [37]. Pixels are grouped into "superpixels", which are roughly homogeneous in size and shape. A region at finer pyramid resolution level should be computed with respect to the boundary of the region at the coarse level. Here, we call the region at the coarse level as a parent-region, while the divided sub-regions are called as children-regions. It is worth mentioning that there is no specific requirement for the over-segmentation algorithm. Any segmentation algorithm such as the efficient graph segmentation [38] and Normalized Cuts [39] can also be used.

2) Region Feature Extraction: Two properties are used as descriptors of regions, i.e., color and texture descriptors. The color descriptor is used to describe the region appearance from the aspect of color variations, while the second descriptor is designed to describe the region appearance in terms of texture property. The block diagram of region feature extraction is illustrated in Fig. 3, which is described in the following paragraphs.

In the proposed method RGB, L*a*b* and YCbCr color spaces are used together to represent the color feature. Each color space is adjusted to range from 0 to 1. To create the color visual descriptor of a region, we first represent a pixel as a 9-dimensional (9-D) color vector by combining components of RGB, L*a*b* and YCbCr color spaces. Then all pixels in the image pair are quantized into N clusters by using the k-means clustering algorithm. Each cluster center is called a codeword. For each region, we simply compute the histogram by counting the number of codewords at each bin (i.e., cluster). The color descriptor for a region is represented by the N bins of the histogram. It is noticed that three color spaces are used to build the color histogram. By concatenating three color spaces, we try to consider more valuable color information from a higher dimensional color space. In other words, we will seek a sparser representation with an overcomplete set of basis functions. It is shown that overcomplete representations have greater robustness in the presence of noise, can be sparser, and can have greater flexibility in matching structure in the data [40]. The similar idea has been applied to solve the color tracking problem [41].

The other region descriptor used in our work is the texture descriptor. Unlike the above color descriptor, the texture descriptor is created only from RGB color space. Given an image

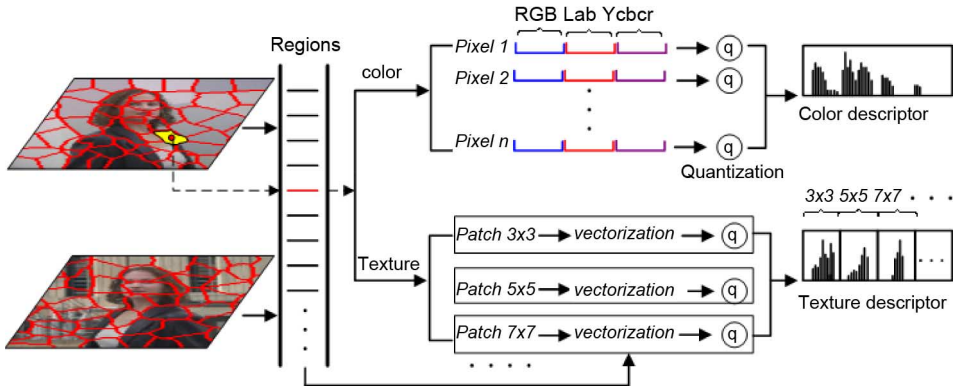


Fig. 3. Block diagram of region feature extraction (e.g., the region with yellow color).

pair, we first extract $p \times p$ patches from color images. The basic idea is that patch feature not only considers the local variation, but also the texture structure in terms of the neighborhood. Each patch is concatenated to form a single vector of size p^2. We then perform the k-means clustering over all vectors to generate M clusters. Patchwords are then defined as the centers of the clusters. We measure the frequency of patchwords and create the texture descriptor by combining a series of histograms of patchwords. Each component histogram represents the probability of occurrence of each patch type (one bin per patchword). We concatenate the component histograms to generate the final texture descriptor

$$f^t(k) = [H_{3\times3}(k), H_{5\times5}(k), H_{7\times7}(k), \ldots] \tag{3}$$

where $H_{i\times i}(k)$ denotes the histogram computed for the kth region of size $i \times i$. The descriptor dimension is the sum of all patchwords. The texture descriptor is normalized to sum to unity.

3) The Co-Multilayer Graph Representation: After feature extraction, we are ready to measure the similarity so as to infer the co-salient object from a pair of images. We begin by designing a co-multilayer graph $G = (V, E)$ with nodes $v \in V$ and edges $e \in E$, where the nodes $V = \{V^I \cup V^{II}\}$ denote a set of regions. Two nodes v_i and v_j are connected by the directed links e_{ij} and e_{ji}, which have weights $w(e_{ij})$ and $w(e_{ji})$, respectively. An example of our co-multilayer graph model is shown in Fig. 4, which contains three-level pyramid decomposition for a pair of images. Each region is represented as a node, which connects with other nodes by the directed edges. Note, each node not only has links with the neighboring layers within an image, it but also connects with other image nodes. For example, the node in V_1^I labelled as yellow color connects with the nodes in V_0^I and V_2^I. It also has links with other nodes in V_0^{II} and V_2^{II}.

In order to represent the edges, one must define a function that assigns a weight to each edge of the graph. Given N nodes, we can get $N(N-1)/2$ links between nodes. In this work, we only consider edges between the nodes within adjacent layers.

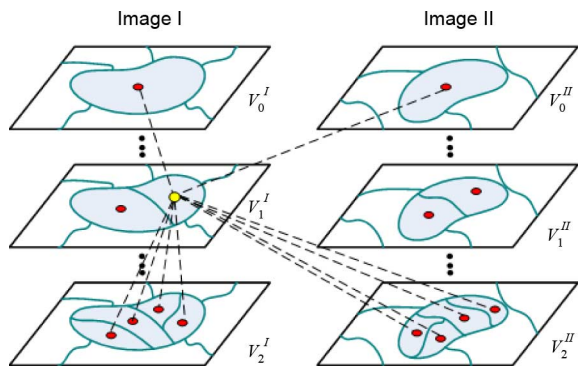


Fig. 4. Our co-multilayer graph model.

Let i and j denote two nodes and l_i and l_j represent their level numbers then the weight w_{ij} for the edge e_{ij} is defined as

$$w_{ij} = \begin{cases} \exp\left(-\theta_f d(f_i, f_j)\right), & \text{if } l_i - l_j = -1 \\ & \text{or } l_i - l_j = 0 \\ 0, & \text{if } \|l_i - l_j\| > 1 \quad \text{or} \quad l_i > l_j \end{cases} \tag{4}$$

with

$$d(f_i, f_j) = \chi^2(f_i, f_j) = \sum_{z=1}^{Z_f} \frac{(f_i(z) - f_j(z))^2}{f_i(z) + f_j(z)} \tag{5}$$

where f_i and f_j denote the color or texture descriptor for regions i and j, respectively. Z_f denotes the dimensional number of the descriptor. θ_f is a constant that controls the strength of the weight. $\chi^2()$ denotes the chi-square distance.

4) Normalized Simrank Similarity Computation: Based on the defined co-multilayer graph, we next measure the similarity between two region nodes. In our work, SimRank, a link-based similarity measure, is used to compute the similarity score of two region nodes. As defined in area of data mining [42], the basic intuition of SimRank is "two objects are similar if they are referenced by similar objects", which was first appeared in the

data mining work [42]. Let $s(a,b)$ denote the similarity score between objects a and b, which is defined as

$$s(a,b) = \frac{C}{|In(a)||In(b)|} \sum_{i=1}^{|In(a)|} \sum_{j=1}^{|In(b)|} s(In_i(a), In_j(b)) \quad (6)$$

where C is a decay factor between 0 and 1. $|In(a)|$ and $|In(b)|$ denote the numbers of in-neighbors $In(a)$ and $In(b)$ for nodes a and b, respectively. The similarity can be regarded as "propagating" from pair to pair, which is obtained in terms of iteration computation.

From (6), we can see that the SimRank score depends on the number of in-neighbors, which means that different in-neighbors for nodes a and b will result in different SimRank scores even if the region pair (a,b) shares the same region appearance. However, they are expected to exhibit the similar SimRank scores due to the similar region property. Therefore, in this work, we employ the normalization of the SimRank score to measure the similarity, i.e.,

$$s^*(a,b) = \frac{s(a,b)}{\max(s(a,a), s(b,b))}. \quad (7)$$

From (7), we have $s^*(a,b) = 1$ when the nodes a and b share the same sub-region nodes.

Substituting (7) into (2), the multi-image saliency map can be rewritten as

$$S_g(I_i(p)) = \max_{q \in I_j} s^*(I_i(p), I_j(q)) \quad (8)$$

where p and q denote the region nodes in an image pair (I_i, I_j). Given a pyramid decomposition, we can choose region nodes from different levels. In our work, we select the node-pair (p,q) from the same level.

C. Co-Saliency Map

Our goal is to extract the co-saliency map from an image pair. We have presented the methods for computing single-image and multi-image saliency maps. Now we are ready to extract the co-saliency from an image pair (I_i, I_j). Let SS_i and SS_j denote the co-saliency maps for the image pair (I_i, I_j). $R\{I\}$ represents a set of regions in the image I. By combining two saliency maps (1) and (8), we have

$$\begin{aligned} SS(I_i(p)) &= \alpha_1 \cdot S_l(I_i(p)) + \alpha_2 \cdot S_g(I_i(p)) \\ &= \alpha_1 \cdot S_l(I_i(p)) + \alpha_2 \cdot \left(\alpha_3 \cdot S_g^c(I_i(p)) + \alpha_4 \cdot S_g^t(I_i(p))\right) \\ &= \beta_1 \cdot S_l(I_i(p)) + \beta_2 \cdot S_g^c(I_i(p)) + \beta_3 \cdot S_g^t(I_i(p)), \\ &\qquad \text{for all } p \in R\{I_i\} \end{aligned} \quad (9)$$

where β_j is a constant with $\beta_1 + \beta_2 + \beta_3 = 1$ that is used to control the impact of the SISM and MISM on the image co-saliency. S_g^c and S_g^t denote the MISMs obtained by color and texture descriptors, respectively. The detailed parameter descriptions can be found in Table I. From (9), we can see that the co-saliency map is built by a linear combination of the SISM and MISM. It means that a region with high co-saliency value will not only exhibit strong single-image saliency but also multi-image saliency. The contributions of the SISM and MISM are controlled by the weighted coefficients β_j. From our empirical study, good performance can be achieved when β_1 takes value between 0.5 and 0.8.

III. Experiments

In this section, we verify the performance of our proposed co-saliency algorithm on several image pairs, which were used in [27] together with additional image pairs that we collected from various databases such as Microsoft Research Cambridge image database, the Caltech-256 Object Categories database, and PASCAL VOC dataset. Some subjective and objective assessments of detection results are reported.

A. Parameters Setting

Before the co-saliency detection, we first introduce the parameters setting in our experiments. To compute the SISM given in (1), we adopt the same weight for each saliency map, i.e., $w_j = 1/3$. For the pyramid representation, we adopt four-level decompositions. The number of regions from 0 to 3 levels is set to 1, 40, 100, and 200, respectively. The original image is treated as the first resolution level. The multi-image saliency map is computed from the level 1 that contains about 40 regions. For each region, we calculate the texture descriptor by (3) based on 3×3 and 5×5 patches. Here, the values of N (the number of code words) and M (the number of patch words) are set to 100 for the color and texture histograms by using the k-means algorithm. To compute the adjacency matrix, we set the constant $\theta = 1.0$ in (4), which shows good performance for most of test images. A fast single-pair SimRank algorithm [43] is used to compute the similarity efficiently. As stated in the previous section, the final co-saliency map is a linear combination of SISM and MISMs. In our work, we chose $\beta_1 = 0.2$, $\beta_2 = 0.4$, and $\beta_3 = 0.4$ as the weights in (9).

B. Detection Results of Image Pairs

We collected 210 images (i.e., 105 image pairs), which consist of human objects, flowers, buses, cars, boats, and various animals, etc. Each image pair contains one or more similar objects with different backgrounds. Most of image data are first scaled to the maximal width or height with no more than 200 pixels. We then manually segmented all image pairs into co-salient objects and backgrounds, which are labeled as one and zero in the ground-truth mask, respectively. Fig. 5 shows some test images and the corresponding ground-truth masks. More images and masks can be found from our future website.[1]

In Fig. 6, we show our co-saliency detection results at different stages for a set of image pairs with similar objects in the foreground but different backgrounds. The original images are shown in the first column. The second and the third columns correspond to the SISM and MISM, respectively. Our co-saliency results are given in the fourth column. The final column shows the co-saliency image by simply multiplying the co-saliency map with the original color image. It can be seen that our method is able to extract co-saliency successfully from test image pairs. For example, for the first image pair (i.e., *banana*), the second banana image contains the complex background, which results

[1][Online]. Available: http://ivipc.uestc.edu.cn/project/cosaliency/

TABLE I
PARAMETER DESCRIPTION

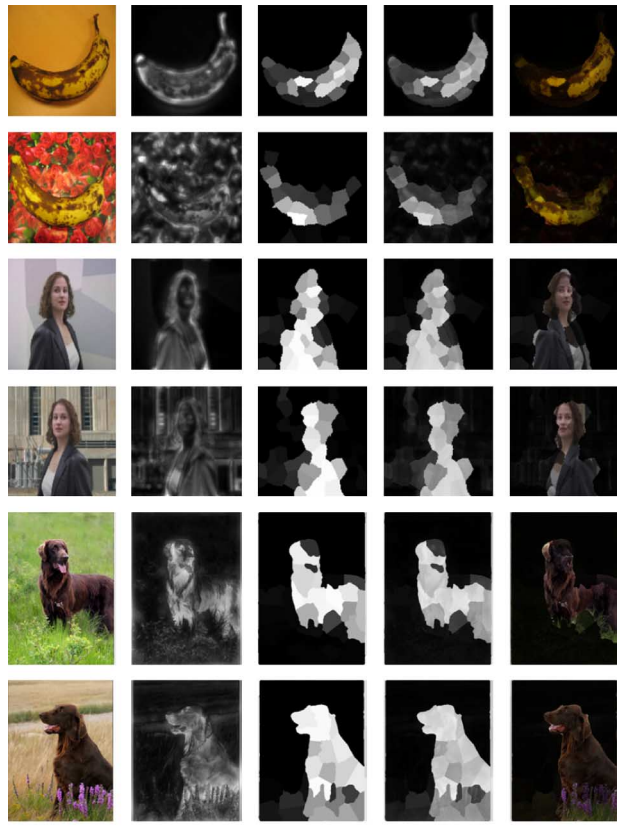
Symbols	Parameters
I_i	The ith image
SS	Co-saliency map
S_l	Single-image saliency map
S_g^c	MISM by color descriptors
S_g^t	MISM by texture descriptors
β_1	Weight for the SISM
β_2, β_3	Weights for the MISMs

(a) (b)

Fig. 5. (a) Original image pairs. (b) Ground truth masks.

in an unsatisfied result for SISM. But our MISM is able to recognize the similar object in both images, which improves the performance effectively. For the second image pair (*amira*), most of object parts are identified as high co-saliency regions using our algorithm such as the face and coat. For the third image pair (*dog*), better results are achieved for both SISM and MISM, which yield identical co-saliency results.

We next evaluate our method on a set of complex image pairs, which contain foreground objects with higher appearance variations or backgrounds with complex scenes. Some original image pairs are shown in Figs. 7(a), 7(b) and 7(e), 7(f). The corresponding results are represented in Figs. 7(c), 7(d) and 7(g), 7(h), respectively. Experimental results show that good performance for detecting co-salient objects can be achieved by our method. For the first image pair (*llama1*) in Fig. 7, the object is difficult to distinguish from the background making it a challenging image to detect. However, our method correctly identifies co-salient regions for both *llama* images. The second image pair contains a human face with different backgrounds. Although backgrounds are very complex, most of regions with high co-saliency values are extracted with respect to the co-attention object (human face). Unlike the above images, many



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

Fig. 6. (a) The test images (i.e., *banana*, *amira*, and *dog*). (b) SISMs. (c) MISMs. (d) Co-saliency maps by our method. (e) Co-saliency images w.r.t. (d).

image pairs consist of similar objects with high variability in shape, size or view. For example, the *flower*, *elephant*, and *crocodile* image pairs appear with different solutions or views. It is shown that our method still yields good results to identify the co-salient objects.

We also evaluate our method on a set of image pairs containing multiple objects. Some example images are shown in the first two columns in Fig. 8. The results by our methods are illustrated in the last two columns of Fig. 8. Experimental results show that our method can detect co-salient multiple objects from an image pair. For example, two cows with different colors are shown in the first row of Fig. 8, which exhibit different poses. Both of them are identified as salient objects by our method. The similar results can be found for other image pairs, such as a group of ducks and two sheep that are shown in the second and the third rows, respectively. It should be noticed that since the proposed method adopts the color and texture features to describe the region, it is still a challenging work for the co-saliency detection for the objects with higher intra-class variations (e.g., the bears in the fourth row of Fig. 8). We believe that other features such as the shape feature may be useful for improving the performance.

C. Objective Evaluation

In order to evaluate the quality of our proposed method, we perform an objective comparison by computing the salient de-

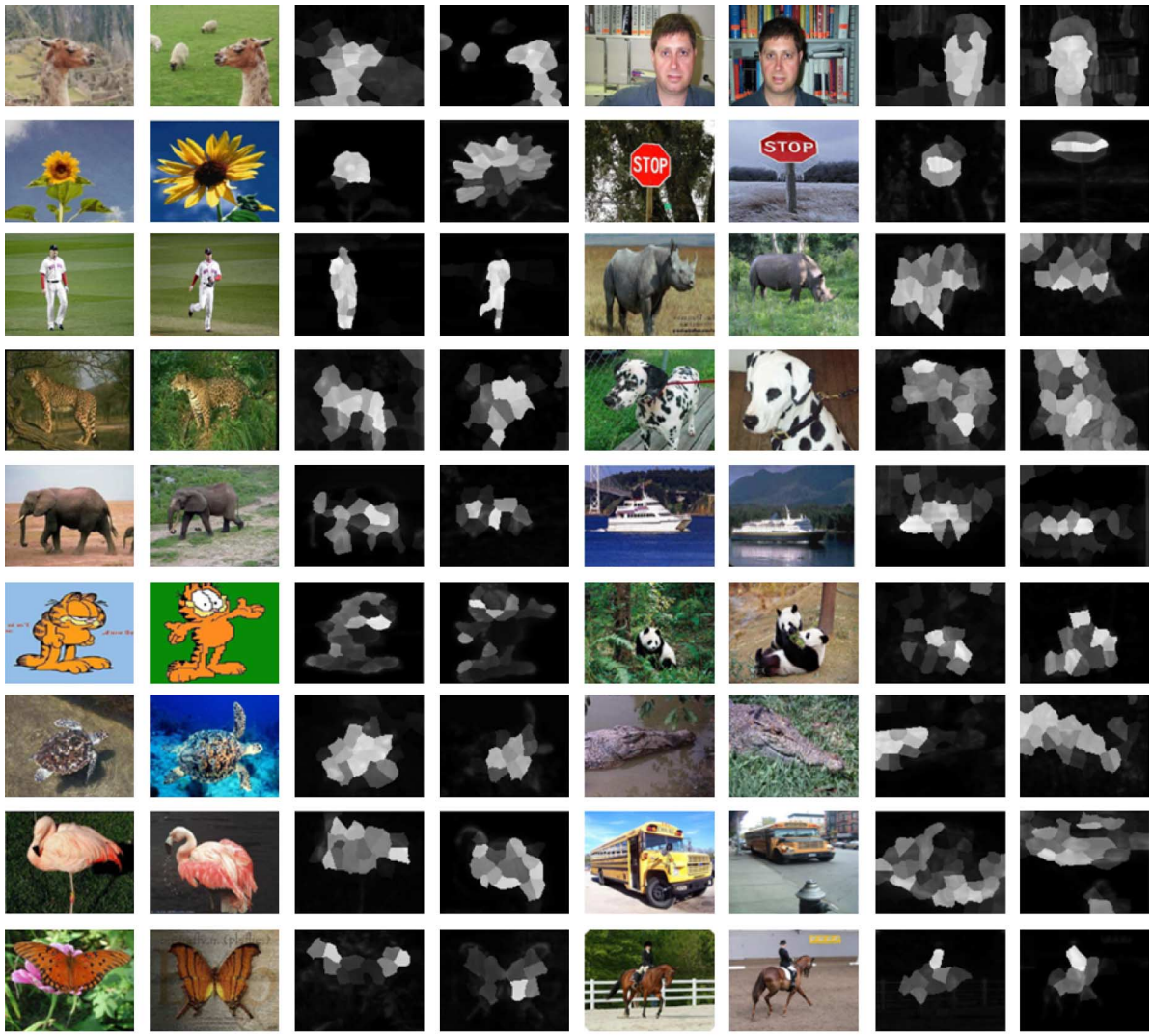


Fig. 7. Experimental results for single objects. (a)-(b) and (e)-(f): Original image pairs. (c)-(d) and (g)-(h): Results by our method.

gree between the extracted co-saliency map and our hand-annotated ground-truth mask. The comparison between an algorithm's output and the ground truth is based on three evaluation metrics, i.e., *Precision* (Pre), *Recall* (Rec), and *F-measure* (F). For a pair of images, the *Precision* is defined as the ratio of correctly extracted regions to all the extracted regions, while the *Recall* is the ratio of correctly extracted regions to the ground-truth regions. *F-measure* is computed by the weighted mean of precision and recall [32], which can be expressed as

$$F_\gamma = \frac{(1+\gamma^2)Pre \times Rec}{\gamma^2 \times Pre + Rec}. \tag{10}$$

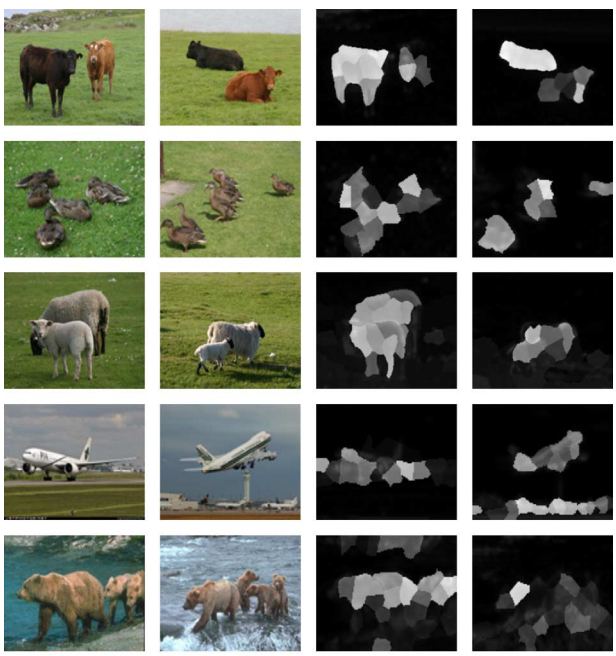
Here, we set $\gamma^2 = 0.3$ that was also used in the work [32].

We compare our result with the existing works for saliency detection [14], [15] and [20]. The first method presents a computational model for visual saliency derived from the information maximization principle. To compute the saliency spots of an image, the model first extracts a number of sub-band feature maps using learned sparse codes [14]. We implement this method using the source code give by the authors, which can be downloaded from http://idm.pku.edu.cn/staff/wangwei/softwares/SiteEntropyRate.rar. The second method aims at detecting the image regions that represent the scene [15]. The unique parts of the background as well as the dominant objects would be marked salient by this method. We also use the source code given by the authors, which can be downloaded from http://webee.technion.ac.il/labs/cgm/Computer-Graphics-Multimedia/Software/Saliency/Saliency.html. The third method introduces a histogram-based approach for saliency detection, while employing a smoothing procedure to control quantization artifacts [20]. We adopt the executable file released by the authors, which can be downloaded from http://cg.cs.tsinghua.edu.cn/people/~cmm/saliency/Saliency.msi. Note that all the results are computed by using the default parameters given by the source codes.

In order to compute the evaluation metrics, an adaptive threshold in [32], [33] is first employed to obtain the binary saliency map. The adaptive threshold value is determined as two times the mean saliency of a given image. For all images, the comparison is shown in Fig. 9(a), which indicates that the

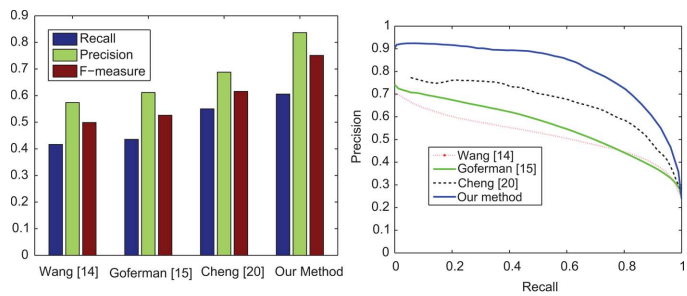
TABLE II
PERFORMANCE EVALUATION BY OBJECTIVE CRITERION.

Image Pair	Results of [14]			Results of [15]			Our Results		
	Recall	*Precision*	*F*	*Recall*	*Precision*	*F*	*Recall*	*Precision*	*F*
banana	0.2760	0.6404	0.4908	0.4252	0.7959	0.6626	0.7025	0.9508	0.8791
amira	0.3989	0.7143	0.6040	0.4360	0.7925	0.6667	0.7675	0.9849	0.9245
kim	0.3005	0.9055	0.6182	0.3503	0.9014	0.6613	0.5297	0.9641	0.8107
stone	0.4941	0.6717	0.6203	0.2624	0.5104	0.4190	0.5550	0.8465	0.7550
dog	0.1919	0.3617	0.3003	0.2937	0.5846	0.4759	0.7106	0.6541	0.6664
llama	0.5390	0.7396	0.6811	0.4669	0.7291	0.6455	0.8036	0.8209	0.8169
face	0.0051	0.0132	0.0096	0.0411	0.1167	0.0819	0.4572	0.9784	0.7746
flower	0.3868	0.4215	0.4129	0.3771	0.4172	0.4072	0.6516	0.9381	0.8517
sign	1.0000	0.7236	0.7729	0.8680	0.6391	0.6805	0.9514	0.9449	0.9464
horse	0.2515	0.6783	0.4874	0.1729	0.6979	0.4104	0.4912	0.9288	0.7704
coke	0.4737	0.9720	0.7821	0.4519	0.9295	0.7472	0.5916	0.9486	0.8326
man	0.7763	0.6430	0.6695	0.9931	0.6999	0.7511	0.8809	0.8785	0.8791



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

Fig. 8. Experimental results for multiple objects. (a)-(b): Original image pairs. (c)-(d): Results by our method.



(a) (b)

Fig. 9. Evaluation results for 210 images. (a) Precision-recall bars for adaptive thresholds. (b) Precision-recall curves for varying thresholds.

similar performances are achieved for the existing methods [14], [15] and [20]. The figure clearly shows that our method detects the co-saliency more accurately with the highest precision, recall and F-measure. Compared with the method [14], our method achieves about 45.41% and 45.81% improvements of recall and precision,respectively. Table II shows the detailed results of some images for our method and the methods [14] and [15]. From Table II. We can see that some detection results are very close to the ground truths by using our method, such as *banana*, *amira* and *coke* etc. In order to evaluate the proposed method sufficiently, we also vary this threshold from 0 to 255, and calculate the precision and recall at each value of the threshold. Note that each saliency map will be normalized in the range of $[0\ 255]$. The curve of the precision versus recall is shown in Fig. 9(b). This curve provides a reliable comparison of how well various saliency maps highlight salient regions in images. It can be seen that our method achieves the highest precision for most of recall values, which demonstrates the best performance on the dataset.

D. Extension to Cosegmentation

As stated in Section II, many approaches were proposed to address the cosegmentation problem in terms of different optimization techniques [29], such as the L1 norm model [25], L2 norm model [26], and the "reward" model [27]. For a given co-saliency map, with saliency values in the range $[01]$, the simplest way to obtain a binary mask for the salient object is to threshold the saliency map at a certain threshold. In this work, we extend our method to the cosegmentation by using an optimized clustering algorithm rather than the simple thresholding method.

Given co-saliency maps, this work uses the iterative graph cut optimization method [44] to perform the cosegmentation within each image. In order to achieve the min-cut optimization, an energy function based on two cost functions (i.e., the data cost and the smoothness cost) should be defined. We use the co-saliency as the data cost for the first graph cut optimization that assigns each node to the object or background label in RGB color space. The color distributions for background/foreground separation are modeled as Gaussian mixture models (GMMs) that can be referred to our previous work [45]. The Gaussian mixture models with 10 components are employed to describe the object and background colors, respectively. The mean and covariance

Fig. 10. Comparison of results of co-segmentation with other methods. First row: Original image pairs including *stone*, *amira*, *llama*, and *horse*. Second row: Results by the method [28]. Third row: Results by the method [27]. Fourth row: Results by our method.

of a component can be estimated based on the k-means algorithm. After the first min-cut optimization, we use the extracted mask to perform the second round graph cut optimization.

Fig. 10 shows the comparison results with other methods. The original image pairs are given in the first row of Fig. 10 including *stone*, *amira*, *llama*, and *horse*. The result by the method [28] is shown in the second row of Fig. 10, which is obtained by using the source code given by authors (see www.di.ens.fr/~joulin). Note that the superpixel code was tested using the work [36]. The parameters μ are set to 0.0015 for *stone*, *amira*, and *horse*, 0.001 for *llama*, respectively. We can see that better results can be found for the *stone* and *horse* image pairs. Due to the similar color between the objects and the backgrounds in the *llama* and *amira* images, many background regions are also detected as the objects. The third row of Fig. 10 shows the result by the method [27], where the results of *amira* and *llama* are given in the original paper. Other results including *stone* and *horse* are obtained by using the source code provided by authors. It can be found that most objects of interest except for *amira* and *horse* images are segmented successfully. Our extension results are shown in the fourth row of Fig. 10, which shows that our cosegmentation extension is comparable to the state of the art methods in [27] and [28].

IV. Discussion and Conclusion

We propose a co-saliency model, which is determined by the weighted combination of the SISM and MISM. The goal of the MISM is to measure the similarity between an image pair, while the SISM is to find the local salient regions within each image. They are expected to contribute to the co-saliency in different aspects. An example of detecting co-saliency from a pair of *moto* images is illustrated in Fig. 11. We compute the evaluation metrics based on the above adaptive thresholding method. If only the MISM is used to generate co-saliency, about 0.7244, 0.2474, and 0.2917 can be obtained for the *Recall*, *Precision*, and *F-measure*, respectively. If only the SISM is used to generate co-saliency, about 0.7671, 0.4375, 0.4857 can be achieved for the *Recall*, *Precision*, and *F-measure*, respectively. By combining them together with $\beta_1 = 0.7$ and $\beta_2 + \beta_3 = 0.3$, we get 0.7598, 0.7440 and 0.7476 for the *Recall*, *Precision*, and *F-measure*, respectively, which show the better performance than the previous cases.

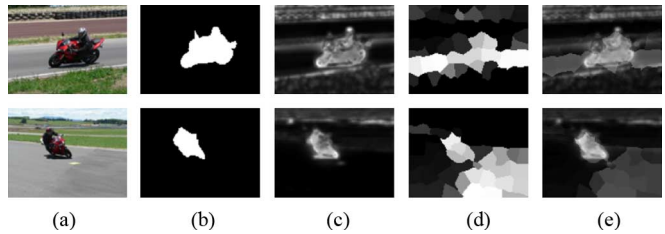


Fig. 11. (a) A pair of *moto* images. (b) Ground truth masks. (c) Result of the SISM. (d) Result of the MISM. (e) Co-saliency results by setting $\beta_1 = 0.7$, $\beta_2 = 0.15$, and $\beta_3 = 0.15$.

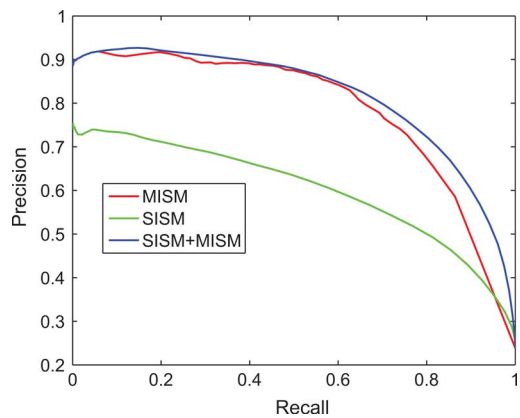


Fig. 12. Precision-recall curves for three cases, i.e., SISM, MISM, and SISM+MISM.

To further investigate the contributions of SISM and MISM on the co-saliency detection, we also compute *Precision*, *Recall* and *F-measure* metrics in different cases, i.e., SISM, MISM, and SISM+MISM by setting $\beta_1 = 1.0, 0.0, 0.5$, respectively. The curve of the precision versus recall for 210 test images is shown in Fig. 12, which shows that the combination of SISM and MISM achieves the highest precision for most of recall values. The result also demonstrates the effectiveness of our proposed co-saliency model.

In conclusion, we have presented a method to identify co-attention objects from an image pair. This method provides an effective way to predict human fixations within multi-images, and robustly highlight co-salient regions. We generate the SISM by computing three visual saliency maps within each image. For the MISM computation, we introduce a co-multilayer graph using a spatial pyramid representation for the image pair. Two types of descriptors (i.e., color and texture visual descriptors) are extracted for each region node, which are then used to compute the similarity between a node-pair. Finally, we employ a fast single-pair SimRank algorithm to measure the similarity based on the normalized SimRank score. Experimental results were obtained by applying the proposed method to several image pairs. It has been shown that our method achieves good performance for the co-salient objects detection. In the future, we hope to incorporate more visual features (e.g., shape and contour features) to further improve the performance. Also extensions to many potential applications such as the image retrieval, semantic object discovery and co-recognition will be investigated.

ACKNOWLEDGMENT

The authors would like to thank all the anonymous reviewers for the constructive comments and useful suggestions that led to improvements in the quality, presentation, and organization of this paper. We thank V. Singh and A. Joulin for providing their source codes and advice for the experiment implementation.

REFERENCES

[1] J. Van de Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 150–156, Jan. 2006.

[2] A. Toshev, J. Shi, and K. Daniilidis, "Image matching via saliency region correspondences," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2007, pp. 1–8.

[3] Y. Sugano, Y. Matsushita, and Y. Sato, "Calibration-free gaze sensing using saliency maps," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 2667–2674.

[4] E. Rahtu, J. Kannala, M. Salo, and J. Heikkil, "Segmenting salient object from images and videos," in *Proc. Eur. Conf. Comput. Vis. (ECCV 2010)*, 2010, pp. 366–379.

[5] Z. Chen, J. Han, and K. N. Ngan, "Dynamic bit allocation for multiple video object coding," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1117–1124, Jun. 2006.

[6] V. Mahadevan, "Saliency-based discriminant tracking," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 1007–1013.

[7] J. You, G. Liu, L. Sun, and H. Li, "A multiple visual models based perceptive framework for multilevel video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 273–285, Mar. 2007.

[8] H. Li and K. N. Ngan, "Automatic video segmentation and tracking for content-based applications," *IEEE Commun. Mag.*, vol. 45, no. 1, pp. 27–33, 2007.

[9] A. Berengolts and M. Lindenbaum, "On the distribution of saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1973–1990, Dec. 2006.

[10] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[11] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.

[12] H. Li and K. N. Ngan, "Saliency model based face segmentation in headand- shoulder video sequences," *J. Vis. Commun. Image Represent.*, vol. 19, no. 5, pp. 320–333, 2008.

[13] H. Li and K. N. Ngan, "Unsupervised video segmentation with low depth of field," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 12, pp. 1742–1751, Dec. 2007.

[14] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 2368–2375.

[15] S. Goferman and L. Zelnik-Manor, "Context-aware saliency detection," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 2376–2383.

[16] C. Koch and T. Poggio, "Predicting the visual world: Silence is golden," *Nature Neurosci.*, vol. 2, pp. 9–10, 1999.

[17] K. Koffka, *Principles of Gestalt Psychology*. London, U.K.: Routledge & Kegan Paul, 1955.

[18] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.

[19] J. Wolfe, "Guided search 2. 0. a revised model of visual search," *Psychonomic Bull. Rev.*, vol. 1, no. 2, pp. 202–238, 1994.

[20] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2011, pp. 409–416.

[21] H.-K. Tan and C.-W. Ngo, "Common pattern discovery using earth mover's distance and local flow maximization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2005, pp. 1222–1229.

[22] J. Yuan and Y. Wu, "Spatial random partition for common visual pattern discovery," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 1–8, 2007.

[23] M. Cho, Y. M. Shin, and K. M. Lee, "Co-recognition of image pairs by data- driven monte carlo image exploration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 144–157.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[25] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, "Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2006, vol. 1, pp. 993–1000.

[26] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 2028–2035.

[27] D. S. Hochbaum and V. Singh, "An efficient algorithm for co-segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2009, pp. 269–276.

[28] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2010, pp. 1943–1950.

[29] S. Vicente, V. Kolmogorov, and C. Rother, "Cosegmentation revisited: Models and optimization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2010, pp. 465–479.

[30] D. Batra, D. Parikh, A. Kowdle, T. Chen, and J. Luo, "Seed image selection in interactive cosegmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2009.

[31] L. Tang and H. Li, "Extract salient objects from natural images," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, 2010, pp. 1–4.

[32] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 1597–1604.

[33] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009.

[34] S. L. Franconeri, A. Hollingworth, and D. J. Simons, "Do new objects capture attention?," *Psycholog. Sci.*, vol. 16, no. 4, pp. 275–281, 2005.

[35] D. Ballard, M. Hayhoe, P. Pook, and R. Rao, "Deictic codes for the embodiment of cognition," *Behavioral Brain Sci.*, vol. 20, no. 4, pp. 723–767, 1997.

[36] X. Ren and J. Malik, "Learning a classification model for segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2003, pp. 10–17.

[37] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2004, pp. 326–333.

[38] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004.

[39] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.

[40] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 2000.

[41] Y. Wu and T. S. Huang, "Color tracking by transductive learning," in *IEEE Comp. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Hilton Head Island, SC, 2000, vol. 1, pp. 133–138.

[42] G. Jeh and J. Widom, "Simrank: A measure of structural- context similarity," in *KDD*, 2002, pp. 538–543.

[43] P. Li, H. Liu, J. X. Yu, J. He, and X. Du, "Fast single-pair SimRank computation," in *Proc. SIAM Int. Conf. Data Mining (SDM 2010)*, 2010, pp. 571–582.

[44] Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in n-d images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2001, pp. 105–112.

[45] H. Li, K. N. Ngan, and Q. Liu, "Faceseg: Automatic face segmentation for real-time video," *IEEE Trans. Multimedia*, vol. 11, no. 1, pp. 77–88, Jan. 2009.

Hongliang Li (M'06) received the Ph.D. degree in electronics and information engineering from Xi'an Jiaotong University, China, in 2005.

From 2005 to 2006, he joined the visual signal processing and communication laboratory (VSPC) of the Chinese University of Hong Kong (CUHK) as a Research Associate. From 2006 to 2008, he was a Postdoctoral Fellow at the same laboratory in CUHK. He is currently a Professor in the School of Electronic Engineering, University of Electronic Science and Technology of China. His research interests include image segmentation, object detection and tracking, image and video coding, and multimedia communication system.

King Ngi Ngan (M'79–SM'91–F'00) received the Ph.D. degree in electrical engineering from the Loughborough Universit, U.K.

He is currently a Chair Professor at the Department of Electronic Engineering, Chinese University of Hong Kong. He was previously a Full Professor at the Nanyang Technological University, Singapore, and the University of Western Australia, Australia. He holds honorary and visiting professorships of numerous universities in China, Australia and South East Asia.

Prof. Ngan is an Associate Editor of the *Journal on Visual Communications and Image Representation*, as well as an Area Editor of EURASIP *Journal of Signal Processing: Image Communication*, and served as an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEM VIDEO TECHNOLOGY and *Journal of Applied Signal Processing*. He chaired a number of prestigious international conferences on video signal processing and communications, and served on the advisory and technical committees of numerous professional organizations. He is a general co-chair of the IEEE International Conference on Image Processing (ICIP) held in Hong Kong in September 2010. He has published extensively including 3 authored books, 5 edited volumes, over 300 refereed technical papers, and edited 9 special issues in journals. In addition, he holds 10 patents in the areas of image/video coding and communications. He is a Fellow of IET, U.K., and IEAust, Australia, and an IEEE Distinguished Lecturer in 2006-2007.