Co-Salient Object Detection From Multiple Images

Hongliang Li, *Senior Member, IEEE*, Fanman Meng, and King Ngi Ngan, *Fellow, IEEE*

***Abstract*—In this paper, we propose a novel method to discover co-salient objects from a group of images, which is modeled as a linear fusion of an intra-image saliency (IaIS) map and an inter-image saliency (IrIS) map. The first term is to measure the salient objects from each image using multiscale segmentation voting. The second term is designed to detect the co-salient objects from a group of images. To compute the IrIS map, we perform the pairwise similarity ranking based on an image pyramid representation. A minimum spanning tree is then constructed to determine the image matching order. For each region in an image, we design three types of visual descriptors, which are extracted from the local appearance, e.g., color, color co-occurrence and shape properties. The final region matching problem between the images is formulated as an assignment problem that can be optimized by linear programming. Experimental evaluation on a number of images demonstrates the good performance of the proposed method on co-salient object detection.**

***Index Terms*—Attention model, co-saliency, minimum spanning tree, similarity.**

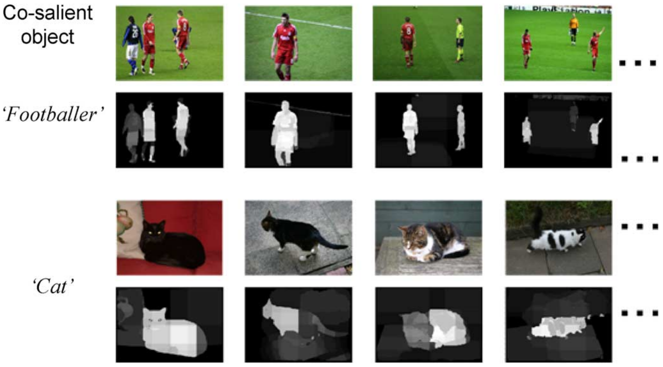


Fig. 1. Examples of visual co-saliency detection. First and Third rows: Two sets of images that contain co-salient objects *footballer* and *cat*, respectively. Second and Fourth rows: Our co-saliency maps.

I. Introduction

HUMANS have an extraordinary ability to rapidly scan a set of images and fixate their attention on the most valuable information (e.g., similar entity). This fixation ability can be viewed as visual co-attention which can be carried out both in a fast, saliency-driven and bottom-up manner, as well as in a top-down processing and memory-dependent manner [1], [2].

Visual co-saliency is a subjective perceptual quality that makes similar objects in a group of images stand out from their neighbors and grab our attention by visually co-salient stimuli [1]. A co-salient region usually exhibits the following properties, i.e., 1) a salient region in an image should be prominent or noticeable with respect to its surroundings. 2) high similarity can be observed for such regions with respect to certain features (e.g., intensity, color, texture or shape). Co-saliency detection is a key attentional mechanism by allocating the perceptual and cognitive resources to the most relative common data while ignoring other dissimilar contents. Two examples can be found in Fig. 1, where co-salient objects *footballer* and *cat* attract more attention than others in the image groups. The corresponding co-saliency maps by our method can be found in the second and the last rows of Fig. 1.

In this paper, we focus on a co-saliency driven attention model which is to simulate the attention search process in a group of images. Two saliency maps namely intra-image saliency (IaIS) and inter-image saliency (IrIS) are defined to model the proposed visual co-saliency via a linear combination. The first term utilizes multiscale segmentation to detect the salient objects from each image, while the second term is to measure the co-salient objects from a set of images. In our work, we build an image pyramid representation to perform pairwise similarity ranking, which is then employed to construct a minimum spanning tree for image matching. Three types of visual descriptors (i.e., color, color co-occurrence and shape descriptors) are defined to represent the region aspects in an image. The final region matching problem between the images can be solved by linear programming as an assignment problem. Experimental evaluation on many public image datasets shows that the proposed method can detect co-saliency effectively, and can be easily extended to the image co-segmentation task.

This paper is organized as follows. The related work is briefly described in Section II. Section III introduces our proposed algorithm for co-saliency detection. Experimental results are provided in Section IV to demonstrate the effectiveness of our approach. Finally, Section V concludes this paper.

II. Related Work

In the last decade, a number of methods have been presented to identify the visual saliency from an image, which can be applied in many fields, such as image search [3], image retargeting [4] and segmentation [5]. Generally, visual saliency can be classified into two categories, i.e., local and global models. The local saliency model focuses on the extraction of local contrast features, which aims to find the local salient region that stands out

Manuscript received October 10, 2012; revised January 27, 2013 and April 15, 2013; accepted April 15, 2013. Date of publication June 27, 2013; date of current version November 13, 2013. This work was supported in part by NSFC (No. 61271289), National High Technology Research and Development Program of China (863 Program, No. 2012AA011503), and The Ph.D. Programs Foundation of Ministry of Education of China (No. 20110185110002). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Cees Snoek.

H. Li and F. Meng are with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 610073, China (e-mail: hlli@uestc.edu.cn; fanmanmeng@yahoo.com).

K. N. Ngan is with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong, and also with the School of Electronic Engineering, University of Electronic Science and Technology of China, Chengdu 610073, China (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/TMM.2013.2271476

from its neighborhood. Based on a biologically-plausible model proposed by Koch and Ullman [6], Itti *et al.* [2] first presented a saliency-based visual attention model for rapid scene analysis, which combined multiscale image features into a single topographical saliency map. This model was successfully extended to segment video objects of interest such as the facial saliency model [7] and the focused saliency model [8]. Inspired by the biologically plausible and the center-surround mechanisms, various local saliency models have been proposed to measure the local saliency from different aspects, such as graph based visual saliency [9], site entropy rate saliency [10], feature learning based saliency [11] and local region contrast methods [12], [13]. Unlike the local model, global saliency model is to measure a pixel's contrast to all pixels within an image. The global contrast feature can be computed from the frequency or spatial domain, such as log-spectrum (SR) [14], frequency tuned (FT) [15], symmetric surrounds [16], context-aware (CA) [17] and histogram/region based contrast [18], [19]. Recently, some works aim to incorporate the high level knowledge to detect the saliency from a single image [20]–[22], which show good performance compared with low-level feature based methods.

Unlike the single image saliency models, visual co-saliency is to discover co-salient objects from a set of images. It is the subjective perceptual quality that implies a selection and/or ranking by importance and makes similar or common objects in a set of images stand out from their neighbors. It is known that common object detection from a set of images has become one of the most important tasks in computer vision, such as common pattern discovery [23], [24], image matching [25] and co-recognition [26]. Generally, the visual co-saliency map can be simply obtained by performing the single image saliency detection for all images. However, the inherent similarity cue between the images will be ignored, which may result in poor performance after the 'blind' co-saliency detection. Recently, a co-saliency model has been proposed to simulate the attention search process for an image pair [1]. This model first combines three existing saliency techniques to generate a local saliency map. Then a co-multilayer graph and the normalized single-pair SimRank algorithm are employed to find the co-salient objects from the image pair. Good performance can be observed for the co-saliency detection of a pair of images. However, high computational load is needed, which is difficult to apply to a number of images. Although the similar goal of detecting the co-saliency can be observed, we can see that the proposed method is significantly different from the previous work [1] in the entire framework. Firstly, the proposed method mainly focuses on co-saliency detection from a group of images instead of an image pair [1]. Secondly, the proposed method proposes a new method to measure the intra-saliency using multiscale segmentation voting, while the work [1] simply combines the existing saliency detection techniques. Thirdly, unlike the work [1], the proposed method designs a different framework to compute the inter-saliency map via a minimum spanning tree. Finally, the proposed method defines three powerful types of visual descriptors, i.e., color, color co-occurrence and shape properties, which are distinctly different from the previous work [1].

In addition, a similar work with our approach is called 'cosegmentation' that aims to segment the common regions from images [27]–[34]. This method can be traced back to the work of Rother [27] which cosegmented the common parts of an image pair by histogram matching. Mukherjee *et al.* [28] extended this work by adding the histogram constraint as the regularized term in a MRF energy function for the simultaneous segmentation. Instead of penalizing the difference of the two foreground histograms, Hochbaum and Singh [29] turned to reward their similarity using a maximum flow procedure on an appropriately constructed graph, which leaded to a polynomial time algorithm for cosegmentation. Joulin *et al.* [30] proposed a discriminative clustering framework for image cosegmentation, which combines the existing tools for bottom up image segmentation such as the spectral clustering technique and positive definite kernels. Recently, a number of cosegmentation methods have been proposed to extract the common regions from multiple images based on different optimization models [33]–[37]. Compared with the co-saliency mode, distinct differences can be observed between co-saliency and cosegmentation tasks, i.e., cosegmentation usually performs in supervised or weakly supervised manner, where several object-like proposals should be obtained before object cosegmentation. In other words, cosegmentation can extract a specific common object by combining the object classifier even the object is not salient in the images. In addition, the co-salient result can be easily extended to the cosegmentation task if we take the co-saliency map as the object-like proposal. Table I summarizes the differences between these related works and the co-salient object detection.

TABLE I
COMPARISON OF A NUMBER OF DIFFERENT SALIENT OBJECT EXTRACTION MODELS

	Intra-Image	Inter-Image	Saliency
Local Contrast Model (e.g., [2], [9])	√	–	√
Global Contrast Model (e.g., [19])	√	–	√
Co-segmentation Model (e.g., [27], [33])	–	√	–
Pairwise Saliency Model [1]	√	Pairwise	√
The Proposed Method	√	√	√

III. PROPOSED METHOD

The co-saliency defined in our paper is obtained by computing the intra-image saliency and inter-image saliency maps. The first is used to identify the salient regions within each image. The second aims to measure the saliency of a set of images. The framework of our proposed method is illustrated in Fig. 2.

A. Intra-Image Saliency (IaIS)

In the current literature, most saliency models focus on the local or global contrast mechanism for the salient region detection. Generally, good results can be achieved when the salient object is surrounded by a uniform background. However, it is still a challenging task to identify the salient object from complex scenes. In order to achieve robust saliency detection, a new saliency detection method is proposed in our work, which aims to improve detection performance based on the multiscale segmentation voting.

1) Multiscale Object Segmentation: Multiple or over segmentation is widely used in saliency detection [38], image categorization [27] and object recognition [39] as an important pre-

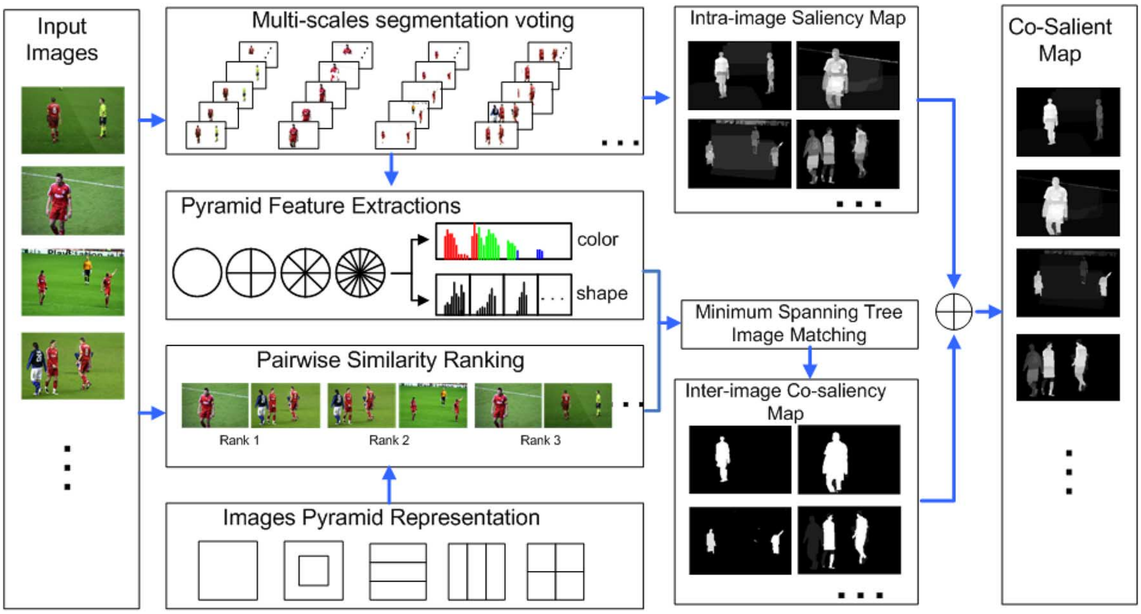

Input Images
Multi-scales segmentation voting
Pyramid Feature Extractions
color
shape
Pairwise Similarity Ranking
Rank 1
Rank 2
Rank 3
Images Pyramid Representation
Intra-image Saliency Map
Minimum Spanning Tree Image Matching
Inter-image Co-saliency Map
Co-Salient Map

Fig. 2. The framework of our proposed method.

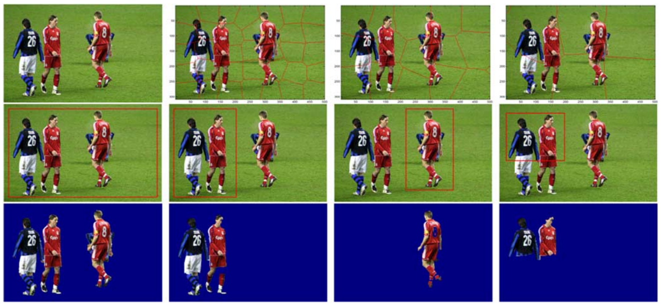


Fig. 3. First row: Superpixel results by the method [40]. Second row: Four candidate segmentation windows with red color. Third row: Corresponding segmentation results by Grab-Cut.

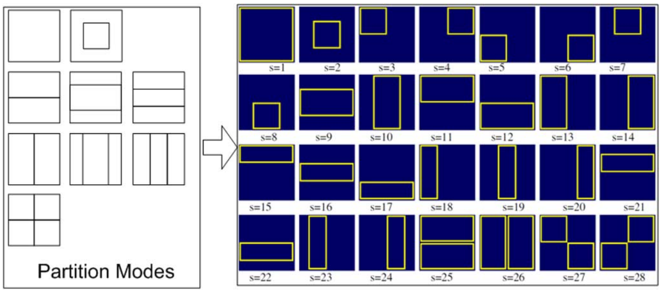


Fig. 4. The predefined segmentation windows. Left: Basic partition modes. Right: Generated segmentation windows (yellow region).

cessing step. After over segmentation, an image can be divided into a lot of homogeneous regions called "superpixels", which can reduce the computational cost and avoid undersegmentation [40]. An example can be found in the first row of Fig. 3, where 30, 10 and 5 superpixels are created from the original image. Since there is no semantic information linked to these superpixels, it is unknown that which superpixel belongs to the salient object. In order to discover the possible salient object, we propose a new method to produce the object superpixels from an image. The idea is motivated by an observation that the salient object or its part can be successfully segmented from its surroundings using the interactive segmentation algorithms such as graph-cut [41], lazy snapping [42] or grab-cut [43]. As illustrated in the second row of Fig. 3, we select four rectangular windows to perform image segmentation based on the Grab-Cut algorithm [43]. The first window includes most image regions except for the image boundary, while the others enclose small parts of the image. The results are presented in the third row of Fig. 3 which show that only the object regions are extracted from the predefined windows. It means that most objects can be segmented when we properly choose the candidate segmentation windows.

In order to extract most objects, we define multiscale segmentation windows which are shown in Fig. 4. Note that the left part presents the basic partition modes which are used to create the candidate windows shown in the right part. Here s denotes the scale number. It is seen that different positions and different sizes of windows are taken into account to perform the object segmentation. For example, the first window (i.e., $\mathrm{s}=1$) selects the whole image as the segmentation window, while the second window (i.e., $\mathrm{s}=2$) only takes the center region as the candidate window. To deal with multiple objects, we define several mixed windows that are listed in the last four cases. For each window, we run the Grab-Cut [43] to perform the binary segmentation, which assigns the foreground pixels with label one and background with zero. Assume L_s denote the segmentation label at the sth scale, we have

$$L_s(p) = \begin{cases} 1, & \text{if} \quad p \in \text{Foreground} \\ 0, & \text{if} \quad p \in \text{Background} \end{cases} \tag{1}$$

2) Intra-Image Saliency Map: After the multiscale image segmentation, we extract the objects from an image at various scales. In other words, an object or its parts may appear in many

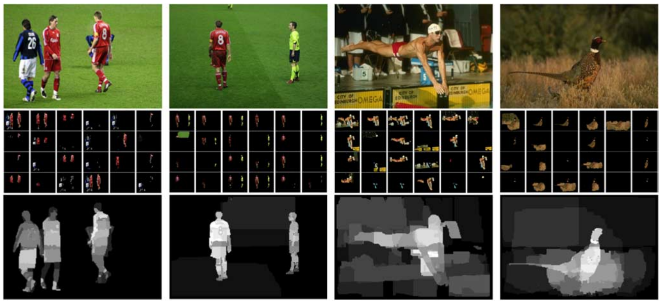


Fig. 5. Examples of the intra image saliency map. First row: Original images. Second row: Multiscale image segmentation. Third row: Saliency maps.

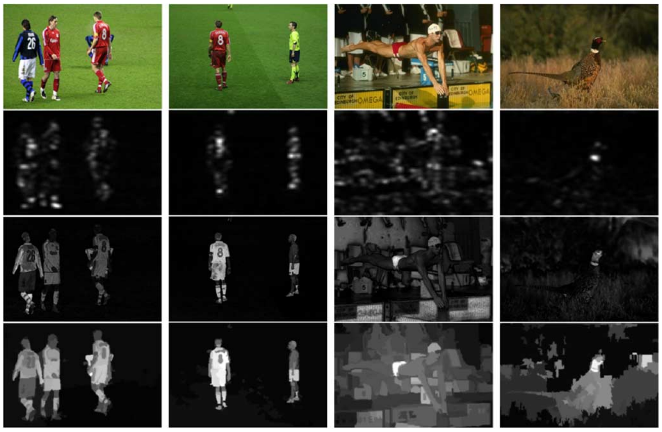


Fig. 6. Saliency map by the existing methods. First row: Original images. Rows 2–4: Saliency maps by SR [14], FT [15] and RC [19], respectively.

scales due to the overlapped candidate windows. Here, we define that a pixel is said to be salient if this pixel is voted as an object pixel with many times. Assume $IaIS(p)$ denotes the intra image saliency value at pixel p. We have

$$IaIS(p) = \frac{1}{\mathcal{Z}_1}\text{Vot}(p) = \frac{1}{\mathcal{Z}_1}\sum_{s=1}^{N}\delta\left(L_s(p)\right) \tag{2}$$

where $\text{Vot}(p)$ is defined as the voting number by the multscale segmentations, N denotes the total number of defined scales, and $\mathcal{Z}_1$ is a normalized constant to ensure the *IaIS* value in the range of [0, 1]. From (2), we can see that if a pixel is voted as the object label for all scales, it will have the maximum salient value.

As shown in Fig. 5, four test images are given in the first row. The second row presents multiscale segmentation results. The intra image saliency maps by (2) are shown in the last row, where most objects are highlighted with large salient values. In addition, we compute the saliency map using the existing methods, i.e., spectral residual (SR) [14], frequency tuned (FT) [15] and global contrast (RC) [19], which are shown in Fig. 6. For the first two images with simple backgrounds, most methods achieve good performance and extract those football players successfully. However, for the last two images with complex backgrounds, our method tends to produce more robust results than the existing methods.

B. Inter-Image Saliency (IrIS)

Visual system relies on several heuristics to direct attention to important locations and objects [44]. The subject is searching for a favorite object, and the attention is geared to react when it appears [45]. Therefore, if a group of images includes a co-salient object (e.g., *cat*), the object region in each image should be highlighted with distinct saliency value. In this work, we define such saliency as inter-image saliency (IrIS) which aims to describe the co-salient object from a group of images. Unlike the intra-image saliency that measures object saliency within an image, inter-image saliency is to discover the co-attention object that simultaneously stands out in the image group. As illustrated in Fig. 2, there are three stages included in our IrIS detection, i.e., pairwise similarity ranking, pyramid feature extraction and pairwise image matching.

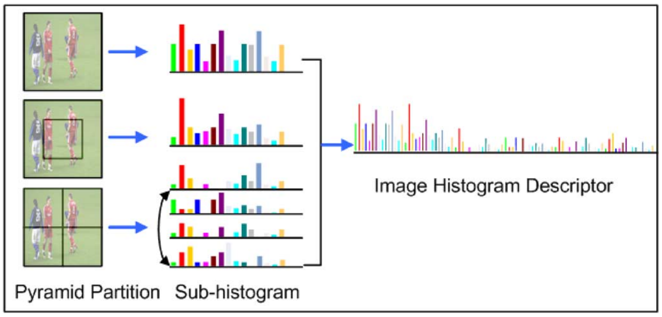


Fig. 7. The image descriptor based on a pyramid representation.

1) Pairwise Similarity Ranking: In order to identify the co-salient object from multiple images, we should solve the image correspondence problem to find out which parts of an image correspond to which parts of another image. If the co-salient objects share the similar color information, some methods, e.g., image pair saliency detection [1], can be employed to solve the correspondence problem between the images. However, co-salient objects usually exhibit the color and shape diversities in images, which can be described as follows.

(i) The co-salient objects, especially for those man-made objects, are allowed to exhibit different *colors*.

(ii) The co-salient objects may appear with different *shapes* when they are captured in different views or positions.

It is known that object matching highly depends on the feature description, while the feature description is usually sensitive to the object diversity. Thus, it is still a challenging task to extract co-salient objects with various color or shape features from multiple images. Some work simplifies this diversity problem by assuming the objects are with similar color [1][34].

However, it is interesting to see that co-salient objects appear to exhibit local feature consistence for some instances although they are allowed to have various color and shape features. As shown in Fig. 2, some cat instances appear with black color, while some others are calico cats. Based on this phenomenon, we first compute a pairwise similarity for the image group. Then the images ranking is performed to generate the local consistent images, i.e., the similar pair of images.

Assume there are N images that need to perform the co-salient object detection, which are denoted by I_k, $k = 1, \ldots, N$, respectively. Firstly, we compute the RGB color histogram for each image based on a pyramid structure, which is shown in the first column of Fig. 7. We use three levels to perform the pyramid decomposition. The whole image is set to the first level, while the center-surround partition is the second level. The third level is created by a quadtree partition of an image. Prior to the histogram computation, we quantize all the pixels in I_k, $k = 1, \ldots, N$ into D (=100 in our work) color words using the k-means clustering algorithm. For each level, we count the occurrence number for each word to produce the histogram. Let Ω_m denote the set of pixels in the mth pyramid level. The histogram descriptor at the mth level can be expressed by

$$h_m(b) = \frac{1}{|q|} \sum_{q \in \Omega_m} \delta(y_q = b) \quad b = 1, \ldots, D \tag{3}$$

where y_q denotes the clustering label of pixel q after the k-means clustering, and $|q|$ is the total number of pixels at the mth pyramid level and is used to normalize the sum of the histogram into one. As shown in Fig. 7, the final image histogram descriptor is built by concatenating the histograms of all levels together, which can be expressed by

$$Hist = \frac{1}{3}[h_1, h_2, h_3] \tag{4}$$

Here, the histogram dimension is set to 600 (i.e., $100 + 100 + 4 \times 100$) in our work.

Secondly, we measure the similarity between images based on the obtained histograms. The chi-square distance χ^2 is used to evaluate the pairwise similarity, which is described as follows.

$$\begin{aligned} S_\chi(I_j, I_k) &= 1 - \chi^2(I_j, I_k) \\ &= 1 - \frac{1}{2} \sum_{b=1}^{|b|} \frac{\left(Hist_{I_j}(b) - Hist_{I_k}(b)\right)^2}{Hist_{I_j}(b) + Hist_{I_k}(b)} \end{aligned} \tag{5}$$

where $|b|$ denotes the total number of histogram bins (=600). Finally, we rank the image pairs according to the similarity values in a descending order. From (5), we can see that the first ranked image pairs exhibit the similar color content.

2) Image Feature Extraction: After the pairwise similarity ranking, we can perform the image matching to discover which parts are co-salient between images. Here, we implement the image matching at the region level. In order to perform image region matching, three types of visual descriptors are built to describe the region properties, i.e., color word, color co-occurrence word and shape word. The first two visual descriptors are designed to capture the image appearance from the color distribution, while the third is used to describe the image appearance in terms of the shape property.

To create a color descriptor, we use RGB color space to represent the color feature. All the pixels in an image pair are quantized into 50 clusters using the k-means clustering algorithm. Each cluster center is called a codeword. As mentioned in the previous section, each image has been partitioned into a number of regions after multiscale image segmentation. For each region, we compute the histogram by counting the number of codewords at each bin (i.e., cluster) according to (3). The color descriptor h_{cl} of a region is represented by the bins of the histogram.

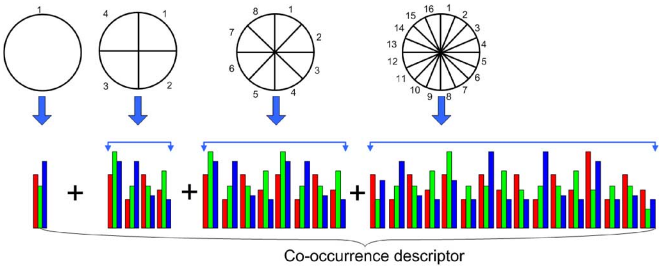


Fig. 8. The construction progress of our co-occurrence descriptor.

To create a co-occurrence descriptor, we first extract the image contour based on the globalized probability of boundary (gPb) [46], which can detect and localize the candidate contour using a combination of local and global cues. Inspired by the fact that color variation usually occurs at image boundaries, we can describe the color co-occurrence around boundary pixels. In this work, we compute the co-occurrence descriptors from the salient boundary points (i.e., $gPb \geq 0.2$). As shown in Fig. 8, we first design a polar location grid with the radius of 8 pixels, which centers at the boundary point. Then we divide this polar grid into 4, 8, 16 subregions in an angular direction, which yield a four-level pyramid structure. For each subregion at a pyramid level, we compute the mean RGB color value of all the pixels in this subregion, which are then set to the bins of a co-occurrence descriptor. This produces a 87 bins histogram descriptor. In order to achieve the rotation-invariant property, we determine the start subregion based on the maximum intensity change. Assume two points pt_1 and pt_2 equally divide the polar grid into two parts which are denoted by $\overrightarrow{pt_1pt_2}$ and $\overrightarrow{pt_2pt_1}$ in a clockwise direction. There can be 8 possible divisions of the polar grid into 2 equal parts, which result in 16 possible endpoints for pt_1 selection. The start subregion is defined as

$$\begin{aligned} pt &= \arg\max_{pt_1} \left(Intensity(\overrightarrow{pt_1pt_2}) - Intensity(\overrightarrow{pt_2pt_1}) \right), \\ pt_1 &= 1, 2, \ldots, 16 \end{aligned} \tag{6}$$

where $Intensity(\overrightarrow{pt_1pt_2})$ denotes the average intensity value for the part $\overrightarrow{pt_1pt_2}$. All the descriptors in an image pair are also quantized into 50 clusters (e.g., codeword) using the k-means clustering algorithm. The color co-occurrence descriptor h_{cc} for a region is represented by the bins of the histogram which is generated by counting the number of codewords at each bin (i.e., cluster) according to (3).

The third descriptor used in our work is the shape descriptor. Unlike the above color related descriptors, the shape descriptor focuses on measuring intensity variations of boundary points. We also employ the salient boundary points to build the shape descriptors. The construction process is illustrated in Fig. 9, which consists of two parts, i.e., gPb coefficients and normalized AC coefficients. For each subregion at a pyramid level, we compute the mean gPb values of all boundary pixels in this subregion, which are then set to the bins of a shape descriptor. In addition, we compute the mean intensity value for each subregion in the last level, and perform a discrete cosine transform (DCT). The obtained 15 AC coefficients are normalized by the DC coefficient, and then combined with the gPb coefficients to produce the final shape descriptor. To describe a region, we quantize all shape descriptors in an image pair into 50 clusters using the k-means clustering algorithm. The shape descriptor h_s of a

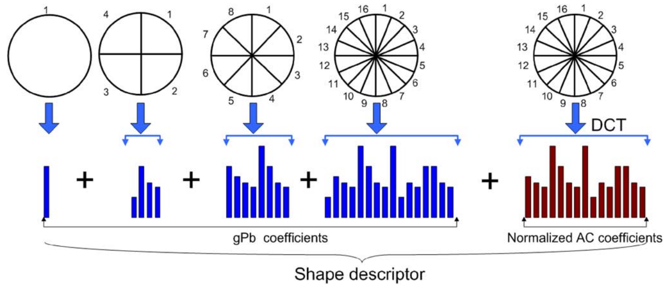


Fig. 9. The construction progress of our defined shape descriptor.

region is generated by measuring the frequency of codewords at each bin according to (3).

3) Minimum Spanning Tree: After multiscale segmentation for a set of images $I_1, I_2, \ldots, I_N$, each image has been divided into K segments. Here $\mathcal{G}_k = \{R_{k,j}, j = 1, 2, \ldots, K\}$ denotes the segment set of the image I_k, where $R_{k,j}$ corresponds to the jth segment. For simplicity, we use R_k to denote a segment in the image I_k, i.e., $R_k \in \mathcal{G}_k$. Let $\mathcal{R}_{1:N}$ denote a group of regions, i.e., $R_1, R_2, \ldots, R_N$, which are taken from the image set. In other words, we can build a region group $\mathcal{R}_{1:N}$ by simply choosing one region from each image. Thus, the co-salient map for a region $R \in \mathcal{R}_{1:N}$ can be defined as

$$S_{co}(R) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} S_\chi\left(R_i, R_j)\right) \qquad (7)$$

with

$$S_\chi(R_i, R_j) = 1 - \chi^2(R_i, R_j). \qquad (8)$$

Here the similarity of a group of regions is obtained by summating the pairwise similarity among these regions. It is seen that each region in a region group $\mathcal{R}_{1:N}$ will have the same co-salient value, i.e., the summation of all pairwise similarities. However, the total number of region groups will reach to $N \times (N-1) \times \cdots \times 2 \times 1 \times K^N$, which is an impossible task for current computation capability especially for large N and K.

It is known that the more similar a pair of image, the more accurate the image matching. Therefore, we identify the co-salient objects by performing the pairwise similarity computation for those images with high similarities. To achieve this goal, we first construct a minimum spanning tree (MST) based on the ranked similarities. Given an undirected graph $G = (V, E)$ with nodes $v \in V$ and edges $e \in E$, where the nodes $V = \{V_1, V_2, \ldots, V_N\}$ denote a set of images. Two nodes v_i and v_j are connected by an undirected edge e_{ij} which has the weight $w(e_{ij}) = 1 - S_\chi(I_i, I_j)$. It is known that a minimum spanning tree is a subgraph that connects all the vertices together, which has the weight less than the weight of every other spanning tree. Here, this tree can be easily obtained by cutting the image pairs with high ranking scores to build a spanning tree [48].

Let E_{mst} denote the edge set for the MST graph. Based on (7), we measure the co-saliency of a region in the region group $\mathcal{R}_{1:N}$ by

$$S_{co}(R) = \sum_{e_{ij} \in E_{mst}} S_\chi(R_i, R_j) \quad R \in \mathcal{R}_{1:N}. \qquad (9)$$

From (9), we can see that the region matching is only performed on the image pairs with edge links in MST graph.

4) Image Pairwise Matching: After the MST construction, we are ready to measure the pairwise similarity so as to infer the co-salient region $S_{co}(R)$ with the high matching score from a group of images. Given a pair of images I_i and I_j that have an edge in the MST graph (i.e., $e_{ij} \in E_{mst}$), we first compute the correspondence matrix C between regions, where each element c_{uv} denotes the distance between regions R_{iu} and R_{jv}. Here, R_{iu} and R_{jv} denote two regions in the image I_i and I_j, respectively. Based on the feature extraction, each region can be represented by three types of descriptors, i.e., color descriptor h_{cl}, color co-occurrence descriptor h_{cc} and shape descriptor h_s. Here we further combine h_{cl} and h_{cc} into a color descriptor h_c based on the scheme used in (4). For a pair of regions R_{iu} and R_{jv}, the distance c_{uv} between regions is defined as:

$$\begin{aligned} c_{uv} = S_\chi(I_i, I_j) \times \chi^2\left(h_c(R_{iu}), h_c(R_{jv})\right) + (1 - S_\chi(I_i, I_j)) \\ \times \chi^2\left(h_s(R_{iu}), h_s(R_{jv})\right) \end{aligned} \qquad (10)$$

where χ^2 is used to evaluate the histogram distance between tow regions. It is observed that a weighted χ^2 distance is used to generate the matrix C. For a pair of images with high similarity, large weight will be imposed to the first term (i.e., color distance). Otherwise, the image pair will be evaluated mainly based on the shape feature due to the distinct color difference.

Let x_{uv} be an indicator variable. If region R_{iu} corresponds to region R_{jv}, we have $x_{uv} = 1$, otherwise $x_{uv} = 0$. The one-to-one constraint matching between this image pair can be expressed as the following optimization problem

$$\begin{aligned} & \min_x \sum_{u=1}^{K} \sum_{v=1}^{K} c_{uv} x_{uv} \\ & s.t. \quad \forall \quad u, \quad \sum_{v=1}^{K} x_{uv} = 1 \\ & \qquad \forall \quad v, \quad \sum_{u=1}^{K} x_{uv} = 1. \end{aligned} \qquad (11)$$

From (11), we can see that this problem is actually an assignment problem, which can be solved via linear programming.

Using the image matching (11), we can obtain K groups of regions. We compute the total matching similarities for all the groups, and then sort them in a descending order. Generally, we can select the region group with the maximum matching score to yield a final inter-image saliency map. However it is difficult to deal with multiple co-salient object detection. To achieve a robust performance, we consider the first M groups to generate our inter-image saliency map. Let $\mathcal{R}^l_{1:N}, l = 1, 2, \ldots, M$ denote the sorted groups according to the matching scores. Assume p denotes a pixel in an image I_i, the inter-image saliency value of pixel $p \in I_i$ can be computed by

$$IrIS_i(p) = \sum_{l=1}^{M} \omega_l \lambda_l(p) S_{co}\left(\mathcal{R}^l_i\right) \qquad (12)$$

with

$$\omega_l = \exp\left(-\frac{(l-1)^2}{\sigma^2}\right), \quad l = 1, 2, \ldots, M \qquad (13)$$

$$\lambda_l(p) = \begin{cases} 1, & \text{if} \quad p \in R^l_i \\ 0, & \text{otherwise} \end{cases} \qquad (14)$$

Algorithm 1: Co-salient object discovering

```
Input: A group of Images $I_i$ $i = 1, 2, ..., N$
Output: A group of co-saliency maps $CoS_i$
        $i = 1, 2, ..., N$
foreach $I_i$ do
    Multiscale Object Segmentation to generate $K$ regions
    Compute the intra-image saliency $IaIS_i$ by (2);
end
foreach image pair $(I_j, I_k)$ $j \neq k$ do
    Compute image descriptors $Hist_j$ and $Hist_k$ after the pyramid decomposition
    Calculate $\chi^2(Hist_j, Hist_k)$
end
Build a minimum spanning tree (MST) with the edge set $E_{mst}$
foreach $e_{ij} \in E_{mst}$ do
    foreach $R_i \in I_i$ do
        Compute $h_c(R_i)$ and $h_s(R_i)$
    end
    foreach $R_j \in I_j$ do
        Compute $h_c(R_j)$ and $h_s(R_j)$
    end
    Compute $c_{uv}$ based on $\chi^2(h_c(R_{iu}), h_c(R_{jv}))$ and $\chi^2(h_s(R_{iu}), h_s(R_{jv}))$
    Build region correspondence matrix $\boldsymbol{C}$
    Solve the matching problem by the linear programming
    Record $S_\chi(R_{iu}, R_{jv})$ with $x_{uv} = 1$
end
Compute inter-image saliency $IrIS_i(p) = \sum_{l=1}^{M} \omega_l \lambda_l(p) S_{co}(\mathcal{R}_i^l)$
foreach $I_i$ do
    Compute the final co-saliency map $CoS_i$ by (15)
end
```

where ω is a weighting coefficient, and the parameter σ adjusts the range (i.e., distance) similarity.

C. Object Co-Saliency Map

Based on intra-image and inter-image saliency maps, we are ready to extract the co-salient object from a group of images $I_i, i = 1, \ldots, N$. Let $CoS_i(p)$ denote the co-salient value of a pixel p in the image I_i. By combining the two saliency maps (2) and (12), we have

$$CoS_i(p) = \alpha \cdot IaIS_i(p) + (1 - \alpha) \cdot IrIS_i(p) \quad i = 1, 2, \ldots, N \tag{15}$$

where α is a constant that is used to control the impact of the IaIS and IrIS on the image co-saliency. From (15), we can see that the co-saliency map is built by a linear combination of the IaIS and IrIS [1], which means that a pixel with high co-saliency value will not only exhibit strong intra-image saliency but also inter-image saliency. The contributions of the IaIS and IrIS are controlled by the weight α. The detailed steps of our proposed method are presented in Algorithm 1.

IV. Experiments

In this section, we evaluate the performance of our proposed method on a number of image groups. Four public image datasets are used for the extensive experiments and comparisons with the state-of-the-art methods. Some subjective and objective assessments of detection results are reported.

A. Parameter Settings and Evaluation Metrics

We first introduce the parameter settings in our experiments. For the multiscale object segmentation, we design 28 candidate windows to run the Grab-Cut segmentation. In order to avoid the effects of possible false segmentations, we sort the segmentation results according to the number of foreground pixels, and remove the first two and last two segmentation results, i.e., those segmentations with seldom or numerous foreground pixels in the segmentation window. For the final object co-saliency map, we chose $\alpha = 0.5$ as the weight in (15). Here, we set $\sigma = 0.5$ to compute the control parameter in (13), which shows good performance from our empirical study.

In order to evaluate the quality of our proposed method, we perform an objective comparison based on the extracted co-saliency map and the hand-annotated ground-truth mask. The comparison between an algorithm's output and the ground truth is performed on three evaluation metrics, i.e., *Precision* (Pre), *Recall* (Rec), and *F-measure* (F) [1]. Given a group of images, the *Precision* is defined as the ratio of correctly segmented object regions to all the segmented regions, while the *Recall* is computed by the ratio of correctly extracted object regions to the ground-truth masks. A weighted mean of precision and recall [15] is employed to calculate *F-measure*, which can be expressed as

$$F_\beta = \frac{(1 + \beta^2) Pre \times Rec}{\beta^2 \times Pre + Rec}. \tag{16}$$

Here, we set $\beta^2 = 0.3$ that was also recommended in the work[15].

B. Experiments on ICoseg Database

We first evaluate our proposed method on the public image dataset ICoseg,[1] which consists of 38 groups (643 images) along with pixel ground-truth hand annotations [32]. Each group contains a common object with the similar color, e.g., bear, pandas, kite.

We compare our result with four state-of-the-art methods for saliency detection, i.e., frequency model SR [14], FT [15], computational model SER [10] and global contrast model RC [19]. To perform a fair comparison, most methods are implemented based on the source codes or executable codes by authors. Note that all the results are computed using the default parameters given by the source codes. We implement our method with the Matlab code.

Some experimental results are illustrated in Fig. 10, which contains two image groups, i.e., *Red Sox* and *Cheetah*. For the first image group *Red Sox*, we can see that most existing methods can provide good performance when the common object (i.e., players) appears in the simple backgrounds, such as the first three images in the upper part of Fig. 10. But for those images that not only include the players but also the complex backgrounds, many false detections can be observed for the existing saliency models. In addition, for the second image group *Cheetah*, the common object usually exhibits more complex texture with respect to the first group, which

[1] http://chenlab.ece.cornell.edu/projects/touch-coseg/

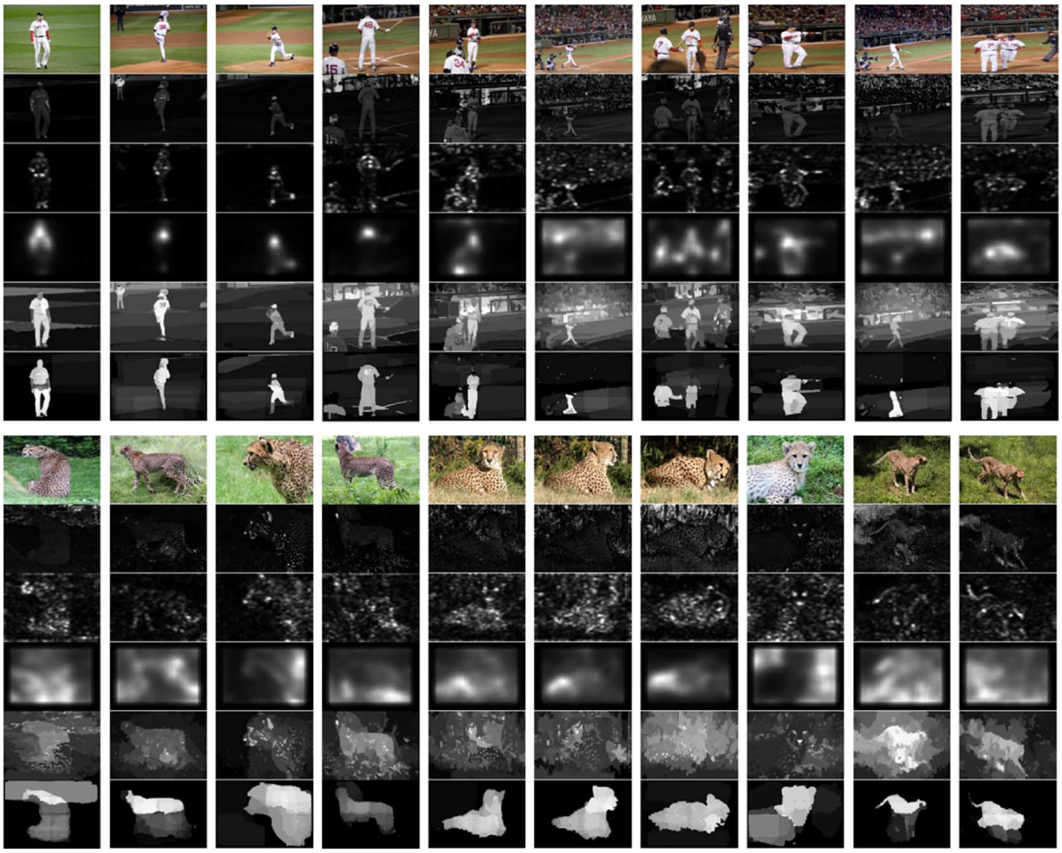


Fig. 10. Evaluation results for two ICoseg image groups. Top and Bottom: Some results for image groups *Red Sox* and *Cheetah*, respectively. Row 1: Some original images. Rows 2–6: Results for FT, SR, SER, RC, and Our method, respectively.

leads to a poor performance for the existing methods. However, as shown in the last row of Fig. 10, it is seen that our method is able to discover co-salient objects successfully from each image group.

To provide a fair comparison, we follow the evaluation strategies used in [1], [15]. Firstly, for each image, we employ an adaptive threshold in [14], [15] to obtain the binary saliency map, which is computed by two times the mean saliency of a given image. We compute the average value of the evaluation metrics for each image group. Table II gives the comparison results on all ICoseg image groups which show that our method measures the co-saliency more accurately with the high Precision, Recall and F-measure. Compared with the method SER [10], our method achieves about 41.03%, 50.77% and 53.76% improvements of Recall, Precision and F-measure, respectively. Compared with the method RC [19], our method yields about 32.85%, 5.74% and 25.83% gains of Recall, Precision and F-measure, respectively. Secondly, we vary this threshold from 0 to 255, and calculate the precision and recall at each value of the threshold. It provides a reliable comparison of how well various saliency maps highlight salient regions in images. We compute the area under the precision versus recall curve (PRC-Area) using the method [9]. The last term in Table II presents the mean area result for each image group, which shows that our method ourperforms the state-of-the-art methods. About 1.44%, 25.21%, and 33.38% gains can be achieved by our proposed method compared with RC [19], FT [15] and SER [10], respectively.

C. Experiments on MSRC Database

We next evaluate our proposed method on the public MSRC object class recognition dataset[2] which consists of 20 groups (591 images) along with pixel-wise labelled annotations. Different objects, such as cow, aeroplane, car and flowers, are included in the dataset. Compared with the ICoseg image dataset, different colors are allowed for the common objects within the image groups. Here, we further compare our proposed method with the similar work FCO [36] which computed the co-saliency map by combining the single-view saliency and repeatedness together. Since the authors have not released their code for open evaluation, we implement the co-saliency algorithm according to the original paper [36]. Note that in our experiment the single-view saliency is replaced by the more effective saliency model RC [19] instead of SER [17] used in [36].

Fig. 11 shows the comparison results of two MSRC image groups, i.e., *sign* and *flower*, where the common objects exhibit distinct diversities in color or shape property. Note that since most existing methods identify the salient object from a single image, such diversity will produce more effect on the co-saliency model. Compared with existing methods, our proposed method still achieves good performance on salient object detection, which highlights the common objects with large saliency values.

[2] http://research.microsoft.com/en-us/projects/objectclassrecognition/

TABLE II
EVALUATION RESULTS FOR ICOSEG IMAGE DATASET

Images	Recall					Precision					F-measure					PRC-Area				
	Our	*RC*	*FT*	*SR*	*SER*	*Our*	*RC*	*FT*	*SR*	*SER*	*Our*	*RC*	*FT*	*SR*	*SER*	*Our*	*RC*	*FT*	*SR*	*SER*
Alaskan-Bear	**0.577**	0.294	0.258	0.188	0.273	0.689	**0.750**	0.502	0.309	0.466	**0.660**	0.552	0.412	0.269	0.401	**0.704**	0.650	0.400	0.313	0.473
Red-Sox	**0.754**	0.607	0.566	0.613	0.679	**0.800**	0.667	0.426	0.508	0.523	**0.789**	0.652	0.452	0.529	0.552	**0.827**	0.669	0.495	0.584	0.658
Stongehenge	**0.639**	0.149	0.061	0.345	0.407	**0.932**	0.829	0.238	0.649	0.649	**0.843**	0.404	0.143	0.539	0.571	**0.903**	0.714	0.181	0.598	0.763
Salisbury	**0.417**	0.079	0.059	0.198	0.229	**0.879**	0.675	0.255	0.662	0.692	**0.700**	0.247	0.144	0.430	0.471	**0.795**	0.557	0.310	0.620	0.679
Liverpool	**0.650**	0.611	0.603	0.379	0.608	0.601	**0.604**	0.568	0.356	0.466	**0.612**	0.605	0.575	0.361	0.492	0.622	**0.739**	0.618	0.349	0.604
Ferrari	**0.748**	0.356	0.437	0.166	0.415	**0.906**	0.861	0.723	0.304	0.683	**0.864**	0.649	0.628	0.255	0.595	**0.897**	0.865	0.651	0.305	0.730
LeszekZadlo	**0.718**	0.097	0.155	0.312	0.305	**0.713**	0.602	0.363	0.409	0.330	**0.714**	0.274	0.277	0.381	0.324	**0.745**	0.503	0.524	0.390	0.367
Inde-du	**0.282**	0.005	0.004	0.177	0.134	**0.464**	0.016	0.007	0.240	0.184	**0.404**	0.011	0.006	0.222	0.170	**0.479**	0.276	0.138	0.250	0.263
Egypt	**0.565**	0.140	0.079	0.118	0.163	**0.742**	0.418	0.226	0.158	0.213	**0.692**	0.286	0.158	0.147	0.199	**0.709**	0.519	0.259	0.198	0.295
Elephants	**0.716**	0.226	0.236	0.138	0.091	**0.859**	0.846	0.303	0.193	0.124	**0.821**	0.518	0.284	0.177	0.115	**0.798**	0.620	0.349	0.204	0.206
Goose	**0.598**	0.340	0.421	0.245	0.213	**0.935**	0.927	0.902	0.548	0.565	**0.827**	0.663	0.714	0.426	0.409	**0.877**	0.865	0.844	0.481	0.498
Pandas-Tai	**0.410**	0.134	0.249	0.135	0.048	0.910	**0.929**	0.792	0.423	0.130	**0.710**	0.392	0.527	0.284	0.094	**0.829**	0.807	0.696	0.398	0.360
Helicopter	0.931	**0.932**	0.854	0.668	0.925	0.847	0.829	**0.934**	0.537	0.542	0.865	0.851	**0.914**	0.562	0.600	0.865	**0.920**	0.894	0.560	0.847
Planes	**0.964**	0.932	0.911	0.848	0.938	0.391	0.282	**0.432**	0.298	0.256	0.454	0.336	**0.492**	0.351	0.307	0.747	0.672	**0.768**	0.450	0.482
Huntsville	**0.990**	0.935	0.963	0.914	1.000	0.211	0.223	**0.258**	0.186	0.221	0.258	0.271	**0.310**	0.228	0.270	0.617	0.772	**0.925**	0.420	0.679
Cheetah	**0.432**	0.120	0.128	0.262	0.219	**0.803**	0.715	0.419	0.771	0.588	**0.670**	0.333	0.275	0.532	0.423	**0.822**	0.649	0.449	0.689	0.634
Pandas	**0.347**	0.125	0.204	0.121	0.080	0.898	**0.974**	0.894	0.493	0.367	**0.657**	0.380	0.502	0.288	0.201	0.862	**0.908**	0.835	0.513	0.536
Brighton-kite	0.591	**0.724**	0.670	0.323	0.571	0.545	0.605	**0.657**	0.265	0.420	0.555	0.629	**0.660**	0.276	0.447	0.827	**0.924**	0.799	0.270	0.517
Kitekid	**0.384**	0.250	0.309	0.238	0.289	0.868	**0.998**	0.952	0.704	0.795	**0.672**	0.591	0.643	0.485	0.566	0.790	**0.907**	0.777	0.567	0.746
Margate-Kite	**0.461**	0.311	0.278	0.264	0.212	**0.891**	0.881	0.767	0.634	0.565	**0.733**	0.619	0.545	0.479	0.408	0.778	**0.828**	0.592	0.604	0.613
Colt-Park	0.809	**0.916**	0.801	0.389	0.912	**0.609**	0.588	0.577	0.247	0.526	**0.646**	0.641	0.617	0.270	0.583	0.686	**0.950**	0.762	0.298	0.731
Gymnastics-1	0.861	**0.978**	0.886	0.584	0.661	**0.960**	0.715	0.621	0.420	0.626	**0.935**	0.762	0.667	0.449	0.634	0.935	**0.977**	0.770	0.436	0.739
Gymnastics-2	0.833	**0.847**	0.819	0.444	0.760	0.678	**0.749**	0.615	0.398	0.607	0.708	**0.770**	0.652	0.408	0.637	0.846	**0.914**	0.649	0.415	0.691
Gymnastics-3	0.826	**0.878**	0.798	0.474	0.671	**0.961**	0.918	0.788	0.517	0.706	**0.926**	0.908	0.790	0.507	0.697	0.912	**0.944**	0.805	0.500	0.770
Skating-Rich	0.329	0.311	**0.360**	0.212	0.345	0.814	**0.913**	0.893	0.490	0.762	0.608	0.631	**0.665**	0.376	0.595	0.752	**0.878**	0.798	0.461	0.705
Skating-ISU	0.976	**0.978**	0.853	0.706	0.944	**0.764**	0.654	0.680	0.476	0.511	**0.804**	0.708	0.714	0.514	0.572	**0.937**	0.916	0.793	0.479	0.776
Woman-Soccer1	**0.666**	0.553	0.546	0.322	0.596	0.712	**0.882**	0.754	0.408	0.628	0.701	**0.776**	0.693	0.385	0.620	0.757	**0.840**	0.692	0.369	0.666
Woman-Soccer2	**0.761**	0.212	0.124	0.357	0.283	**0.559**	0.388	0.205	0.416	0.313	**0.596**	0.326	0.178	0.401	0.305	**0.574**	0.447	0.285	0.382	0.337
Monks	**0.398**	0.248	0.303	0.130	0.355	0.837	**0.982**	0.857	0.321	0.702	**0.667**	0.583	0.603	0.240	0.573	0.773	**0.843**	0.708	0.329	0.680
Hot-Balloons	**0.602**	0.516	0.455	0.336	0.484	**0.745**	0.468	0.537	0.393	0.592	**0.706**	0.478	0.516	0.379	0.563	**0.898**	0.758	0.625	0.418	0.737
EricaJoy	0.638	**0.652**	0.469	0.361	0.464	**0.987**	0.962	0.982	0.671	0.708	**0.876**	0.867	0.784	0.560	0.631	0.898	**0.951**	0.912	0.653	0.683
Christ	**0.509**	0.346	0.278	0.314	0.363	**0.898**	0.866	0.817	0.763	0.691	**0.763**	0.643	0.565	0.574	0.572	**0.886**	0.869	0.781	0.725	0.684
Speedskating	0.575	**0.706**	0.472	0.436	0.694	0.260	**0.332**	0.212	0.204	0.287	0.298	**0.378**	0.243	0.233	0.332	0.324	**0.622**	0.254	0.214	0.470
Track	**0.609**	0.272	0.224	0.302	0.373	0.592	**0.781**	0.565	0.490	0.621	**0.596**	0.545	0.418	0.428	0.538	0.607	**0.637**	0.446	0.486	0.632
Windmill	0.671	**0.955**	0.530	0.375	0.378	**0.460**	0.382	0.312	0.252	0.192	**0.496**	0.444	0.344	0.272	0.217	**0.561**	0.530	0.412	0.286	0.235
Kendo-Kendo	0.531	0.634	**0.747**	0.245	0.338	0.917	**0.992**	0.982	0.419	0.523	0.785	0.878	**0.916**	0.360	0.465	0.775	**0.972**	0.950	0.371	0.484
Kendo-EKC	**0.747**	0.453	0.430	0.226	0.302	0.945	**0.963**	0.943	0.441	0.507	**0.890**	0.765	0.740	0.362	0.439	0.902	**0.948**	0.876	0.382	0.521
brown_bear	**0.434**	0.203	0.197	0.168	0.255	**0.839**	0.711	0.518	0.441	0.569	**0.691**	0.451	0.376	0.320	0.443	**0.963**	0.701	0.518	0.426	0.613
Average	**0.630**	0.474	0.440	0.343	0.447	**0.748**	0.707	0.591	0.432	0.496	**0.689**	0.548	0.504	0.375	0.448	**0.776**	0.765	0.620	0.431	0.582

The objective performance on all MSRC image groups are illustrated in Table III, where the results by adaptive thresholds are shown at the first three terms in Table III, respectively. Compared with the method SER [10], It is seen that our method provides about 50.52%, 22.10%, and 32.35% improvements of Recall, Precision and F-measure, respectively. It is noted that similar precision on average can be observed between our method and the global contrast model RC [19]. But our method achieves about 74.22% and 31.69% gains of Recall and F-measure, respectively. Compared with FCO model [36], our method also provides distinct improvements in the Recall, Precision and F-measure with 71.30%, 4.12% and 31.23% gains on average, respectively. It means that our method is able to discover most salient object regions from image groups under the similar precision levels. The last term in Table III shows the mean area result for each image group, which shows that our method still outperforms the state-of-the-art methods, which yields 3.60%, 40.10%, 35.83%, 15.16% and 5.70% gains compared with RC [19], FT [15], SR [14] SER [10], and FCO [36], respectively.

To further investigate the performance of our proposed method, we combine the previous saliency methods, i.e., SR [14], FT [15], SER [10] and RC [19], with our inter-saliency to generate a final co-saliency map. The results are given in Table IV which shows that all the previous saliency methods can achieve improvements with respect to Recall, Precision and F-measure. For example, about 0.1249, 0.1633, 0.2732 and 0.1926 gains of F-measure on ICoseg dataset can be achieved for RC [19], FT [15], SR [14], and SER [10], respectively. For MSRC dataset, the corresponding gains of F-measure are also about 0.1046, 0.1642, 0.1509 and 0.1042, respectively. Compared with the previous saliency methods with the IrIS saliency, our proposed method still achieves more accurate saliency detection with the high Precision, Recall and F-measure. In addition, we only consider our IaIS map to fairly compare with the other methods, which can be found in Table IV. Compared with the RC saliency model, our IaIS achieves about 10.9%, 0.71% and 11.83% improvements of Recall, Precision and F-measure on the ICoseg dataset, respectively. The corresponding improvements will reach to 59.9%, 1.83% and 28.86% for the MSRC dataset. The experimental results further demonstrate that our IaIS can compete with the state-of-the-art methods for the single image saliency detection.

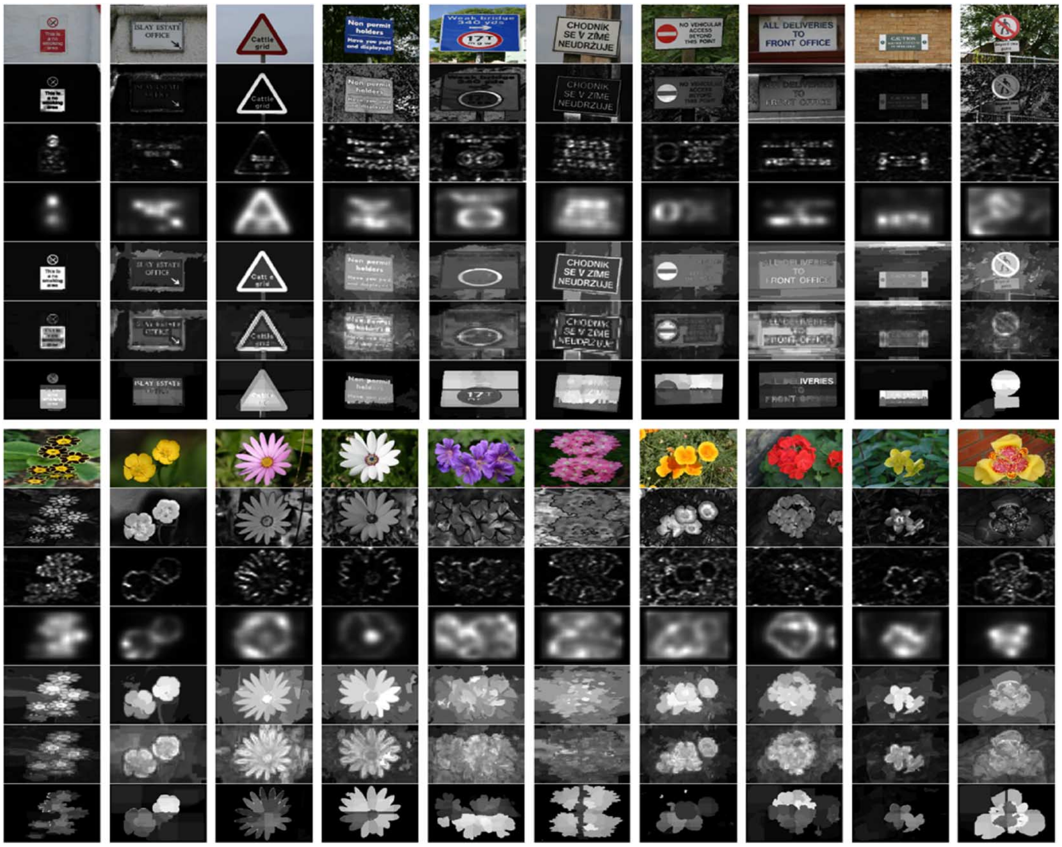


Fig. 11. Evaluation results for two MSRC image groups. Top and Bottom: Some results for the image group *sign* and *flower*, respectively. Row 1: Some original images. Rows 2–7: Results for FT, SR, SER, RC, FCO, and Our method, respectively.

TABLE III
EVALUATION RESULTS FOR MSRC IMAGE DATASET

Images	Recall						Precision						F-measure						PRC-Area					
	Our	*RC*	*FT*	*SR*	*SER*	*FCO*	*Our*	*RC*	*FT*	*SR*	*SER*	*FCO*	*Our*	*RC*	*FT*	*SR*	*SER*	*FCO*	*Our*	*RC*	*FT*	*SR*	*SER*	*FCO*
Multi-cow	**0.602**	0.492	0.387	0.327	0.447	0.478	**0.658**	0.640	0.541	0.510	0.575	0.634	**0.644**	0.599	0.496	0.452	0.539	0.589	0.697	**0.729**	0.632	0.595	0.684	0.719
Tree	0.292	0.082	0.093	0.241	**0.316**	0.109	0.888	0.874	0.523	0.832	**0.956**	0.863	0.603	0.271	0.253	0.531	**0.652**	0.332	0.860	0.805	0.478	0.797	**0.915**	0.819
Building	**0.353**	0.150	0.116	0.217	0.200	0.147	0.860	**0.871**	0.408	0.645	0.596	0.842	**0.646**	0.413	0.258	0.443	0.409	0.403	**0.831**	0.726	0.426	0.592	0.627	0.690
Aeroplane	**0.594**	0.296	0.200	0.430	0.580	0.309	0.627	**0.708**	0.460	0.650	0.679	0.641	0.619	0.535	0.354	0.581	**0.653**	0.513	0.663	0.629	0.308	0.608	**0.711**	0.587
Single-cow	**0.673**	0.466	0.375	0.307	0.372	0.443	0.942	**0.985**	0.759	0.677	0.660	0.909	**0.863**	0.784	0.614	0.530	0.560	0.731	0.898	**0.906**	0.685	0.571	0.639	0.865
Face	**0.418**	0.123	0.122	0.156	0.194	0.101	**0.604**	0.452	0.236	0.278	0.326	0.388	**0.548**	0.279	0.194	0.236	0.282	0.234	**0.605**	0.464	0.276	0.300	0.397	0.435
Car	**0.329**	0.144	0.195	0.182	0.270	0.198	0.823	**0.894**	0.707	0.590	0.774	0.862	**0.611**	0.406	0.441	0.389	0.541	0.486	0.774	**0.825**	0.586	0.528	0.742	0.806
Bicycle	0.273	0.125	0.176	0.239	**0.305**	0.152	0.710	0.742	0.495	0.723	**0.764**	0.739	0.519	0.347	0.349	0.493	**0.567**	0.390	0.621	0.664	0.445	0.653	**0.710**	0.664
Sheep	**0.671**	0.476	0.405	0.293	0.289	0.489	**0.942**	0.941	0.817	0.737	0.597	0.918	**0.862**	0.768	0.662	0.546	0.479	0.763	**0.916**	0.899	0.729	0.620	0.603	0.890
Flower	**0.458**	0.175	0.182	0.228	0.297	0.168	**0.914**	0.913	0.732	0.668	0.825	0.864	**0.743**	0.463	0.432	0.462	0.585	0.442	**0.893**	0.879	0.676	0.618	0.826	0.843
Sign	**0.474**	0.230	0.192	0.283	0.408	0.235	**0.917**	0.826	0.547	0.732	0.898	0.794	**0.754**	0.517	0.384	0.535	0.703	0.512	**0.888**	0.773	0.595	0.667	0.863	0.765
Bird	**0.655**	0.578	0.512	0.356	0.387	0.576	0.705	**0.801**	0.704	0.511	0.458	0.786	0.692	**0.736**	0.648	0.465	0.440	0.725	0.751	0.759	0.647	0.502	0.525	**0.772**
Book	**0.245**	0.084	0.142	0.165	0.200	0.102	0.968	**0.993**	0.819	0.812	0.970	0.976	**0.576**	0.285	0.389	0.426	0.514	0.329	0.911	0.886	0.734	0.777	**0.925**	0.871
Chair	**0.560**	0.255	0.209	0.219	0.325	0.298	**0.809**	0.710	0.443	0.493	0.695	0.695	**0.733**	0.503	0.352	0.383	0.550	0.532	**0.762**	0.681	0.435	0.439	0.671	0.662
Cat	**0.545**	0.355	0.316	0.293	0.282	0.362	0.819	**0.875**	0.688	0.625	0.575	0.842	**0.734**	0.654	0.541	0.496	0.464	0.645	**0.784**	0.771	0.592	0.567	0.601	0.773
Dog	**0.506**	0.273	0.284	0.216	0.183	0.333	0.711	**0.717**	0.530	0.425	0.347	0.716	**0.650**	0.522	0.441	0.347	0.287	0.566	**0.674**	0.657	0.502	0.379	0.392	0.668
Road	**0.151**	0.002	0.021	0.030	0.016	0.007	**0.208**	0.007	0.038	0.048	0.025	0.019	**0.191**	0.004	0.032	0.042	0.022	0.013	**0.255**	0.147	0.176	0.150	0.157	0.154
Boat	**0.667**	0.469	0.378	0.423	0.575	0.406	**0.350**	0.338	0.245	0.279	0.332	0.311	**0.393**	0.361	0.267	0.303	0.367	0.329	0.578	0.571	0.497	0.520	**0.617**	0.559
Body	**0.546**	0.413	0.377	0.265	0.329	0.377	0.466	**0.696**	0.444	0.287	0.338	0.672	0.482	**0.601**	0.426	0.281	0.336	0.569	0.508	**0.647**	0.418	0.263	0.357	0.598
Water	**0.072**	0.024	0.032	0.044	0.061	0.013	**0.144**	0.074	0.082	0.111	0.128	0.039	**0.117**	0.050	0.060	0.082	0.102	0.026	0.248	0.212	0.238	0.246	**0.251**	0.217
Average	**0.454**	0.261	0.236	0.246	0.302	0.265	**0.703**	**0.703**	0.511	0.532	0.576	0.675	**0.599**	0.455	0.380	0.401	0.453	0.457	**0.706**	0.681	0.504	0.520	0.611	0.668

D. Experiments on Images With Larger Variations

We collect six image groups with large variations from the public image datasets (e.g., PASCAL VOC 2008, UIUC Butterfly, Stanford-40 and Flickr). Some image examples can be found in Fig. 12(a), which includes *Running*, *Baby*, *Butterfly*, *Dog2*, *Riding-horse*, and *Toy-baby* image groups. It is seen that the first two image groups are taken from the road, room or countryside. The butterfly images contain a similar object (i.e., admiral butterfly) in a cluttered background. The objective comparison results are illustrated in Fig. 12(b), which shows that our proposed method also achieves good performance compared with the existing methods. Compared with the RC method, our method yields about 19.14% and 8.27% gains of F-measure and mean area on average for all images. Furthermore, two image groups with multiple objects are evaluated, which are shown in

TABLE IV
COMPARISON RESULTS BY COMBINING THE EXISTING METHODS WITH OUR IRIS

	Icoseg Dataset						MSRC Dataset					
	FT+IrIS	SR+IrIS	SER+IrIS	RC+IrIS	IaIS	Our	FT+IrIS	SR+IrIS	SER+IrIS	RC+IrIS	IaIS	Our
Recall	0.6143	0.6077	0.6203	0.6159	0.5261	0.6302	0.3937	0.4017	0.4124	0.3930	0.4170	0.4542
Precision	0.7298	0.6987	0.6873	0.7469	0.7123	0.7479	0.6637	0.6630	0.6645	0.7029	0.7157	0.7032
F-measure	0.6671	0.6485	0.6408	0.6726	0.6126	0.6892	0.5438	0.5521	0.5568	0.5596	0.5862	0.5991

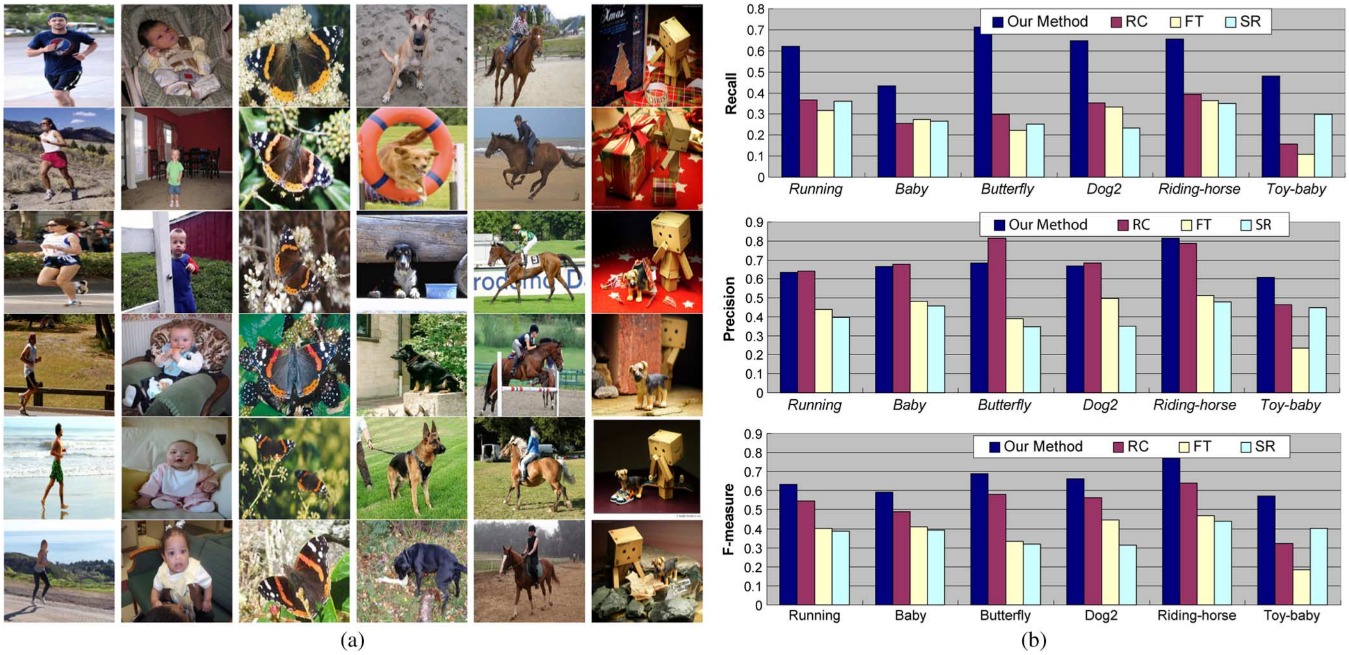


Fig. 12. Some examples taken from six image groups. (a) From left to right: *Running*, *Baby*, *Butterfly*, *Dog2*, *Riding-horse*, and *Toy-baby*. (b) Evaluation results for five image groups for Recall, Precision and F-measure, respectively.

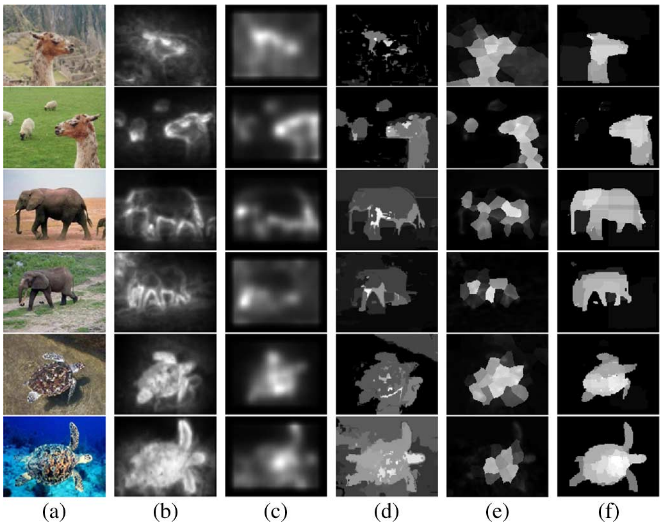


Fig. 13. Experimental results for image pairs. (a): Original image pairs, i.e., *llama*, *elephant*, *hawksbill*. (b)-(f): Results by CA [17], SER [10], RC [19], IPCO [1], and our method.

the last two columns of Fig. 12(a). The first contains two different objects (i.e., toy baby and toy dog), where some occlusions and lighting variations can be observed. From objective evaluation metrics in Fig. 12(b), we can see that good performance can be achieved by our proposed method. Compared with RC method [19], about 78.36% and 20.74% gains of F-measure on average can be obtained for our method, respectively.

E. Experiments on Image Pair Database

In this experiment, we compare our result with the recent image pair co-saliency (IPCS) detection method [1] which measures visual saliency for a pair of images. Three saliency models used in [1], i.e., SER [10], CA [17] and RC [19], are also considered for the evaluation. The comparisons are performed on the public image pair dataset[3] given in [1], which consists of 105 pairs of images, such as human objects, flowers, buses, cars, boats and various animals.

Some examples are shown in Fig. 13, where the original image pairs are presented in the first column. The corresponding results are presented in Figs. 13(b)–(f), respectively. It is shown that good performance for co-salient object detection can be achieved by our proposed method. For the image pairs *elephant* and *hawksbill*, the object appears with different sizes or views making it a challenging image to detect. However, our method provides more accurate results to identify the co-salient regions.

[3] http://ivipc.uestc.edu.cn/hlli/projects/cosaliency.html

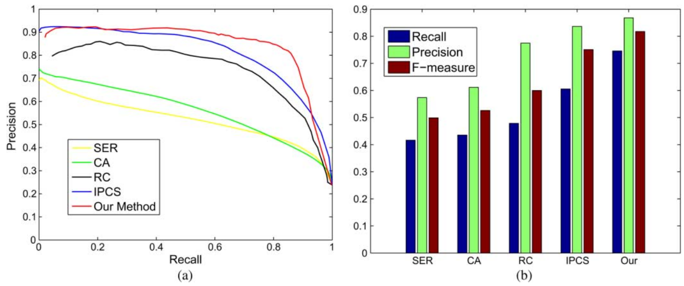


Fig. 14. Evaluation results for 105 image pairs given in [1]. (a) Precision-recall curves for varying thresholds. (b) Precision-recall bars for adaptive thresholds.

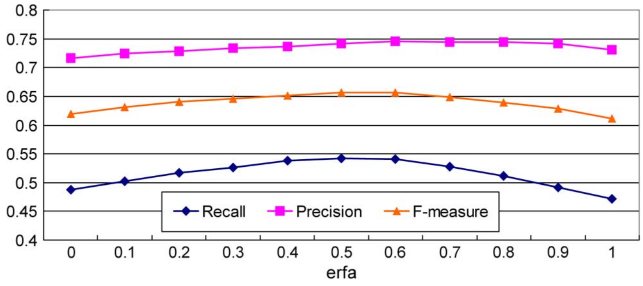


Fig. 15. The curve of the evaluation metric verse the parameter α.

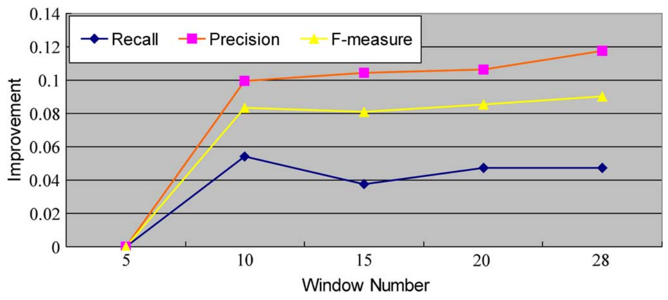


Fig. 16. The relative performance of the evaluation metric verse the selected windows.

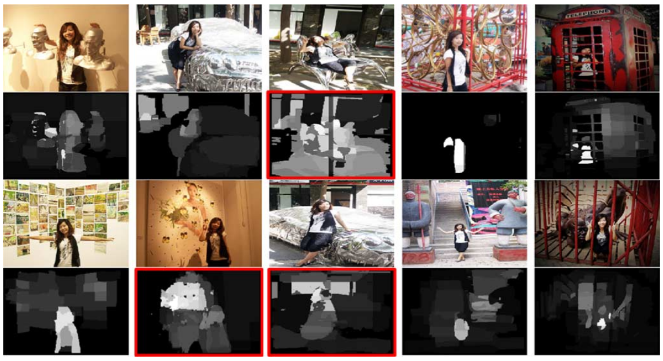


Fig. 17. Evaluation results for *girl* image group with significantly cluttered images. Rows 1 and 3: Some original images. Rows 2 and 4: Co-saliency results.

We follow the evaluation strategies used in [1]. The precision versus recall curve is plotted in Fig. 14(a), which shows that our method achieves high precision for most recall values especially for the recall in the range of [0.2 0.9]. Then, we employ an adaptive threshold in [15] to obtain the binary saliency map. As shown in Fig. 14(b), the comparison results show that our method measures the co-saliency more accurately with the highest Precision, Recall and F-measure. Compared with the method [1], our method achieves about 23.14%, 3.73% and 8.84% improvements of Recall, Precision and F-measure, respectively.

F. Discussion

In this work, we use the linear fusion to generate the final co-saliency map, which is controlled by the weight parameter α. From (15), we can see that the large weight will be imposed to the inter-image saliency with the decreasing α. To investigate the effect of α on the co-saliency detection, we also compute *Precision*, *Recall* and *F-measure* metrics based on the adaptive threshold by changing α from 0 to 1. Fig. 15 plots the curve of the evaluation metrics for all the images (namely ICoseg + MSRC datasets), which shows that good performance can be achieved when the weight α is in the range of [0.4, 0.6]. The result also demonstrates the effectiveness of our proposed co-saliency model.

In order to measure the co-saliency map, the multiscale segmentation windows (see Fig. 4) are defined based on different positions and different sizes of the windows. There are total 28 segmentation windows used in our work. To investigate influence of the number of windows on the final accuracy, we compute *Precision*, *Recall* and *F-measure* metrics based on the adaptive threshold for different windows (i.e., 5, 15, 20, 28). Each window scale was randomly selected from the total window set for many times. Fig. 16 plots the curve of the relative improvement on evaluation metrics for all the images (namely ICoseg + MSRC datasets). Here, the 5-window scale was used as a reference and the relative differences for the remaining window scales were calculated. We can see that the worst performance can be found when there are only 5 windows used. The main reason is that a small number of windows cannot provide enough object candidates for co-salient object detection. Furthermore, a slight improvement can be observed when we increase candidate windows.

In addition, we investigate the effect of three types of visual descriptors (i.e., color-based, color co-occurrence-based, and shape-based visual descriptors) on the final performance. Table V shows the relative performance values for different visual descriptors. For each evaluation metric, the worst descriptor was used as a reference and the relative differences for the remaining descriptors were calculated. It is seen that the color word and shape word respectively provide large gains for the ICoseg and MSRC datasets if only one visual descriptor is considered. For the ICoseg dataset, the combined descriptors achieve the highest gains compared with other descriptors, i.e., 4.79%, 5.97% and 4.79% for Recall, Precision and F-measure, respectively. For the MSRC dataset, the combined descriptors also provide higher gains of Recall and F-measure except for the Precision.

Furthermore, it should be noted that the false detection will occur when the objects have no clear shape or color within the image group, especially for the significantly cluttered images. An example can be found in Fig. 17, where some false co-saliency detections can be observed (i.e., marked with red windows) due to the cluttered background. The main reason is

TABLE V
RELATIVE PERFORMANCE ON THREE VISUAL WORDS FOR ALL ICOSEG AND MSRC DATASETS

ICoseg	%			MSRC	%		
Descriptors	Recall	Precision	F-measure	**Descriptors**	Recall	Precision	F-measure
Shape only	+1.92	0	0	Shape only	+0.44	**+1.27**	**+0.83**
Color only	+2.63	+2.31	+1.17	Color only	**+1.38**	+0.01	+0.41
Co-occurrence	0	+2.34	+0.4	Co-occurrence	0	+0.16	0
All-combined	**+4.79**	**+5.97**	**+4.79**	All-combined	+1.06	0	+0.48

that our method is based on the assumption that a co-salient region should exhibit high similarity with respect to certain features (e.g., intensity, color, texture or shape).

V. CONCLUSION

In this paper, we have presented a new method to discover co-salient objects from a set of images. This method aims to simulate the attention search process and predict the human fixed objects within an image group. The proposed co-saliency model is built based on two saliency maps, i.e., the intra-image saliency (IaIS) and the inter-image saliency (IrIS) maps. The first term is designed to identify the salient objects from a single image according to the multiscale image segmentation voting, while the second term is to discover the co-salient objects from a set of images. To compute the IrIS map, we first construct an image pyramid representation to perform a pairwise similarity ranking. The ranked results are then used to build a minimum spanning tree for image pairwise matching. To describe the region aspects of local appearance in an image, we design three types of visual descriptors in terms of color, color co-occurrence and shape properties. Finally, we solve the final region matching problem between images by linear programming. Experimental evaluation on a number of image groups demonstrates the good performance of the proposed method on the co-salient object detection.

REFERENCES

[1] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, Dec. 2011.

[2] 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.

[3] J. Huang, X. Yang, X. Fang, W. Lin, and R. Zhang, "Integrating visual saliency and consistency for re-ranking image search results," *IEEE Trans. Multimedia*, vol. 13, no. 4, pp. 653–661, 2011.

[4] Y. Fang, W. Lin, B.-S. Lee, C.-T. Lau, Z. Chen, and C.-W. Lin, "Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 187–198, 2012.

[5] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, "Unsupervised salient object segmentation based on kernel density estimation and two-phase Graph Cut," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1275–1289, 2012.

[6] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, pp. 219–227, 1985.

[7] H. Li and K. N. Ngan, "Saliency model based face segmentation in head-and-shoulder video sequences," *J. Visual Commun. Image Represent. (Elsevier Science)*, vol. 19, no. 5, pp. 320–333, 2008.

[8] 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, Nov. 2007.

[9] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Adv. Neural Inf. Process. Syst.*, vol. 19, pp. 545–552, 2007.

[10] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 13–18, 2010, pp. 2368–2375.

[11] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MN, 2007, pp. 1–8.

[12] R. Achanta, F. Estrada, P. Wils, and Süsstrunk, "Salient region detection and segmentation," in *Proc. Int. Conf. Computer Vision Systems*, 2008, pp. 66–75.

[13] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Int. Conf. Multimedia*, New York, NY, USA, 2003, pp. 374–381.

[14] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[15] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Ssstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1597–1604.

[16] R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," in *Proc. 17th IEEE Int. Conf. Image Processing (ICIP)*, Hong Kong, Sep. 2010, pp. 2653–2656.

[17] S. Goferman and L. Zelnik-Manor, "Context-aware saliency detection," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 2376–2383.

[18] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM Multimedia*, 2006, pp. 815–824.

[19] M. M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, "Global contrast based salient region detection," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Colorado Springs, CO, USA, 2011, pp. 409–416.

[20] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 16–21, 2012, pp. 478–485.

[21] M. Wang, J. Konrad, P. Ishwar, K. Jing, and H. Rowley, "Image saliency: From intrinsic to extrinsic context," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 417–424.

[22] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 16–21, 2012, pp. 853–860.

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

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

[25] A. Toshev, J. Shi, and K. Daniilidis, "Image matching via saliency region correspondences," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[26] 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. Computer Vision (ECCV)*, 2008, pp. 144–157.

[27] 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 Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2006, vol. 1, pp. 993–1000.

[28] L. Mukherjee, V. Singh, and C. R. Dyer, "Half-integrality based algorithms for cosegmentation of images," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 2028–2035.

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

[30] A. Joulin, F. Bach, and J. Ponce, "Discriminative clustering for image co-segmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2010, pp. 1943–1950.

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

[32] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoseg: Interactive Co-segmentation with Intelligent Scribble Guidance," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3169–3176.

[33] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. Int. Conf. Computer Vision*, Nov. 2011, pp. 169–176.

[34] F. Meng, H. Li, G. Liu, and K. N. Ngan, "Object co-segmentation based on shortest path algorithm and saliency model," *IEEE Trans. Multimedia*, vol. 14, no. 5, pp. 1429–1441, 2012.

[35] S. Vicente, C. Rother, and V. Kolmogorov, "Object cosegmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 2217–2224.

[36] K. Chang, T. Liu, and S. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 2129–2136.

[37] L. Mukherjee, V. Singh, and J. Peng, "Scale invariant cosegmentation for image groups," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Jun. 2011, pp. 1881–1888.

[38] Y. Lu, W. Zhang, H. Lu, and X. Xue, "Salient object detection using concavity context," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, Nov. 6–13, 2011, pp. 233–240.

[39] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, San Francisco, CA, USA, Jun. 13–18, 2010, pp. 73–80.

[40] A. Levinshtein, A. Stere, K. N. Kutulakos, D. J. Fleet, and S. J. Dickinson, "Turbopixels: Fast superpixels using geometric flows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2290–2297, Dec. 2009.

[41] 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.Computer Vision (ICCV)*, Vancouver, BC, Canada, Jul. 7–14, 2001, vol. 1, pp. 105–112.

[42] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, "Lazy snapping," in *Proc. SIGGRAPH*, 2004, pp. 303–308.

[43] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut-interactive foreground extraction using iterated graph cuts," in *Proc. SIGGRAPH 2004*, Los Angeles, CA, USA, Aug. 10–12, 2004.

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

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

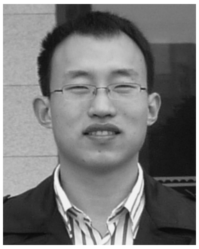
[46] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, "Using contours to detect and localize junctions in natural images," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[47] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2012, pp. 542–549.

[48] J. B. Kruskal, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem," *Proc. Amer. Math. Soc.*, vol. 7, no. 1, pp. 48–50, 1956.

Hongliang Li (SM'12) received his Ph.D. degree in Electronics and Information Engineering from Xi'an Jiaotong University, China, in 2005. From 2006 to 2008, he joined the visual signal processing and communication laboratory (VSPC) of the Chinese University of Hong Kong (CUHK) as a Postdoctoral Fellow. 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, image and video coding, and visual attention.

Dr. Li has authored or co-authored numerous technical articles in well-known international journals and conferences. He is a co-editor of a Springer book titled "Video segmentation and its applications". Dr. Li was involved in many professional activities. He is a member of the Editorial Board of the Journal on Visual Communications and Image Representation. He served as TPC members in a number of international conferences, e.g., ICME 2012–2013, ISCAS 2013, PCM2009, and VCIP2010, and General co-chair of the 2010 ISPACS. He serves as a local chair of the 2014 IEEE International Conference on Multimedia and Expo (ICME).



Fanman Meng received the B.Sc. degree in computer science and technology and the M.Sc. degree in computer software and theory in 2006 and 2009 respectively. Since September 2009, he has been working toward the Ph.D. degree in the Intelligent Visual Information Processing and Communication Laboratory (IVIPC) at University of Electronic Science and Technology of China (UESTC). His research interests include image segmentation, object detection and visual attention.

King Ngi Ngan (F'00) received the Ph.D. degree in Electrical Engineering from the Loughborough University in 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 served as associate editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, Journal on Visual Communications and Image Representation, EURASIP Journal of Signal Processing: Image Communication, and Journal of Applied Signal Processing. He chaired and co-chaired a number of prestigious international conferences on image and video processing including the 2010 IEEE International Conference on Image Processing, and served on the advisory and technical committees of numerous professional organizations. He has published extensively including 3 authored books, 6 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.

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