Cluster-based Co-saliency Detection

Huazhu Fu, Xiaochun Cao, Zhuowen Tu

***Abstract*—Co-saliency is used to discover the common saliency on the multiple images, which is a relatively under-explored area. In this paper, we introduce a new cluster-based algorithm for co-saliency detection. Global correspondence between the multiple images is implicitly learned during the clustering process. Three visual attention cues: contrast, spatial, and corresponding, are devised to effectively measure the cluster saliency. The final co-saliency maps are generated by fusing the single image saliency and multi-image saliency. The advantage of our method is mostly bottom-up without heavy learning, and has the property of being simple, general, efficient, and effective. Quantitative and qualitative experimental results on a variety of benchmark datasets demonstrate the advantages of the proposed method over the competing co-saliency methods. Our method on single image also outperforms most the state-of-the-art saliency detection methods. Furthermore, we apply the co-saliency method on four vision applications: co-segmentation, robust image distance, weakly supervised learning, and video foreground detection, which demonstrate the potential usages of the co-saliency map.**

***Index Terms*—saliency detection, co-saliency, co-segmentation, weakly supervised learning.**

I. Introduction

Saliency detection could be considered as a preferential allocation of computational resources [1]–[5]. Most of existing saliency algorithms formulate on detecting the salient object from an individual image [6]–[8]. Recently, the multiple image correspondence based on a small image set has become one of the popular and challenging problems, and meanwhile the co-saliency is proposed. Co-saliency detection in [9] is firstly defined as discovering the unique object in a group of similar images. However, the requirement of the similar images, captured within the same burst of shots, narrows its applications. An alternative concept is more favourite, which targets to extract the common saliency from the multiple images [10]–[12]. The extracted co-saliency map under later concept is more useful in various applications, such as co-segmentation [13]–[15], common pattern discovery [16], [17], object co-recognition [18], [19], image retrieval [20], and image summaries [21], [22]. The objective of this paper focuses on the later definition and proposes an efficient cluster-based method for detecting the common saliency on the multiple images.

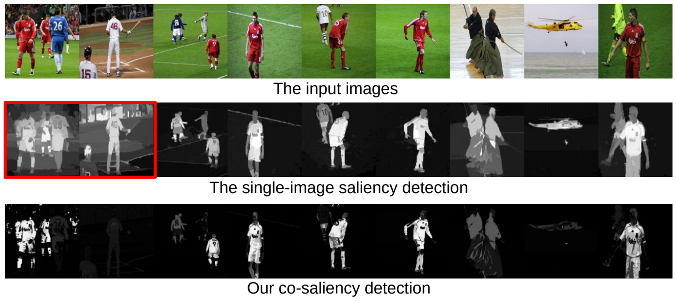

The input images
The single-image saliency detection
Our co-saliency detection

Fig. 1. Given a group of images (first row), the state-of-the-art saliency detection method in [3] (second row) might be confused with complex background, and lacks the relevance information on the multiple images. In contrast with single image saliency, our co-saliency (third row) provides the recurring co-salient objects (the player in red).

Fig. 1 illustrates the co-saliency example, where the single image algorithm [3] (second row) extracts salient objects in each image, but at the cost of having confusions with the complex backgrounds. For example, the audiences in the first two images, highlighted by a red rectangle, are wrongly labeled as saliency. Single image saliency method also ignores the relevance cues on the multiple images. In contrast, our co-saliency utilizes repetitiveness property as additional constraint, and discovers the common salient object on the multiple images, e.g. the red player as shown in the bottom row of Fig. 1.

The goal of our work is to develop a novel cluster-based algorithm for co-saliency detection. Our method employs the clustering to persevere the global correspondence among multiple images, and generates final co-saliency maps by fusing three effective bottom-up cues. A nice thing about our method is mostly bottom-up without heavy learning, and has the property of being simple, general, efficient, and effective. This paper includes the following properties: 1) we propose a cluster-based co-saliency detection method. 2) Based on cluster-based method, a number of bottom-up cues are adopted to measure the saliency, which are originated from a well-known property of the human visual perception. 3) Our method not only has encouraging performances on co-saliency detection, but also significantly outperforms the state-of-the-art methods on single image saliency detection. 4) We present four applications, including co-segmentation, robust image distance, weakly supervised learning, and video foreground detection, to demonstrate the potential usages of the co-saliency.

This paper is organized as follows: after a brief introduction of related works in Section I-A, Section II gives the detailed

This work was supported by National Basic Research Program of China (2013CB329305), National High-tech R&D Program of China (2012AA011503), 100 Talents Programme of The Chinese Academy of Sciences, and Tianjin Key Technologies R&D program (No.11ZCKFGX00800). Z. Tu is supported by NSF CAREER award IIS-0844566, NSF award IIS-1216528, and NIH R01 MH094343.

H. Fu is with School of Computer Science and Technology, Tianjin University, Tianjin 300072, China (E-mail: hzfu@tju.edu.cn).

X. Cao is with State Key Laboratory Of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China (E-mail: caoxiaochun@iie.ac.cn).

Z. Tu is with Laboratory of Neuro Imaging, Department of Neurology, University of California, Los Angeles, CA 90095, USA (E-mail: ztu@loni.ucla.edu).

implementation of our method, including the two-layer clustering, the cluster-based cues, and the cue integration. Then the quantitative and qualitative experiments on a variety of benchmark datasets are shown in Section III. Moreover, four applications of co-saliency are proposed in Section IV. Finally, some concluding remarks are presented in Section V.

A. Related Works

1) Co-saliency detection: Solutions utilizing additional companion images as cues are proved to be effective [10]–[12]. Chen [10] proposes a method to find the co-saliency between a pair of images by enhancing the similar and preattentive patches. Li and Ngan [11] model the co-saliency as a linear combination of the single image saliency map and the multi-image saliency map, by employing a complex co-multilayer graph. However, it is hard to generalize these two methods [10], [11] to the case of multiple images. [12] considers the single-view saliency map and concentrates on those salient parts that frequently repeat in most images. Nevertheless, this method only defines the co-saliency as a prior for the co-segmentation task and the advantage of the co-saliency is not distinctly illuminated. Moreover, the computational requirement of [12] is very expensive. In contrast, we propose a simple yet effective cluster-based method to detect the co-saliency from multiple images. Compared with the existing techniques, our approach has a distinct advantage of being efficient and effective.

2) Co-segmentation: A closely related research area to co-saliency detection is 'co-segmentation', which aims to segment out the similar objects from two/multiple images [13], [14], [23]. Compared with the co-segmentation, our co-saliency detection implies a priority based on the lower level concepts, more specifically human visual attention. Furthermore, co-segmentation has three differences to co-saliency detection: first, similar but non-salient background in images could interfere the correspondence procedure for the unsupervised co-segmentation approaches [24], [25]. Second, some co-segmentation methods [14], [26] need user inputs to guide the segmentation process under ambiguous situations. Third, co-segmentation systems are often computationally demanding, especially, on a large number of images. In practice, applications such as image retargeting [27], [28], object location and recognition [29], only need to roughly but quickly localize the common objects from the multiple images. Unlike co-segmentation, our co-saliency detection method automatically discriminates the common salient objects. Thanks to its simplicity and efficiency, our approach is able to be used as a pre-processing step for subsequent high-level image understanding tasks. Nevertheless, we evaluate the proposed co-saliency method on a number of co-segmentation datasets, finding out that, despite the lack of complex learning, the performance of our co-saliency is rather competitive with many of the recent co-segmentation methods.

II. Our Proposed Approach

As stated above, the co-saliency map can be used in various vision applications. However, co-saliency detection has not received many attentions and a limited number of the existing methods [10], [12] are not able to produce the satisfactory results. In this paper, we regard the common object to be co-saliency if it satisfies the following aspects:

- (Intra-saliency) Co-saliency should follow the laws of the visually salient stimuli in the individual image, which is efficient for distinguishing the salient object against the background.
- (Inter-saliency) Co-saliency exhibits high similarity on the multiple images, and hence the global repetitiveness feature/distribution should be employed to highlight the common patterns.
- Furthermore, as the pre-processing step for subsequent applications, the co-saliency detection should be easy to implement and fast to compute.

In this paper, we propose a two-layer cluster-based method to detect co-saliency on the multiple images. Fig. 2 shows the flowchart of our cluster-based method. Given a set of images, our method starts by two-layer clustering. One layer groups the pixels on each image (single image), and the other layer associates the pixels on all images (multi-image). We then compute the saliency cues for each cluster, and measure the cluster-level saliency. The measured features include the uniqueness (on single/multi-image), the distance from the image center (on single/multi-image) and the repetitiveness (on multi-image). We call them contrast, spatial, and corresponding cues, respectively. At last, based on these cluster-level cues, our method computes the saliency value for each pixel, which is used to generate the final saliency map.

A. Cluster-based Method

The cluster-based idea is inspired by the global-contrast methods [3], [30]–[32] on the single image. These methods quantize the feature channels of pixels into the histogram format to measure the spatial contrast dissimilarity, and evaluate the saliency of the pixel with respect to the other pixels in the entire image. But the estimated feature distributions using histogram are discontinuities at the bin edges. Instead, we employ clustering to avoid the discontinuities at the bin edges of histograms, and obtain a highly cohesive global constraint. Simultaneously, clustering on the multiple images provides the global corresponding relationship for the all images. In our method, we are not constrained to specific choice of the clustering methods, and herein K-means is used.

There are two challenges in the clustering process. How to predefine a suitable cluster number, and how much does the misclassified pixel (e.g. a background pixel is grouped with saliency pixels) harm the saliency detection? Fewer clusters cause the pixels within the same cluster to have the same saliency values without sufficient discrimination. To avoid this 'discrete' clustering, we adopt a probability framework to soft assign the co-saliency value to each pixel, which is discussed in Section II-C. Furthermore, we find that the cluster number is not the major factor for our method. The effect of the cluster number on our method with this soft assigning is tested in Section III-D.

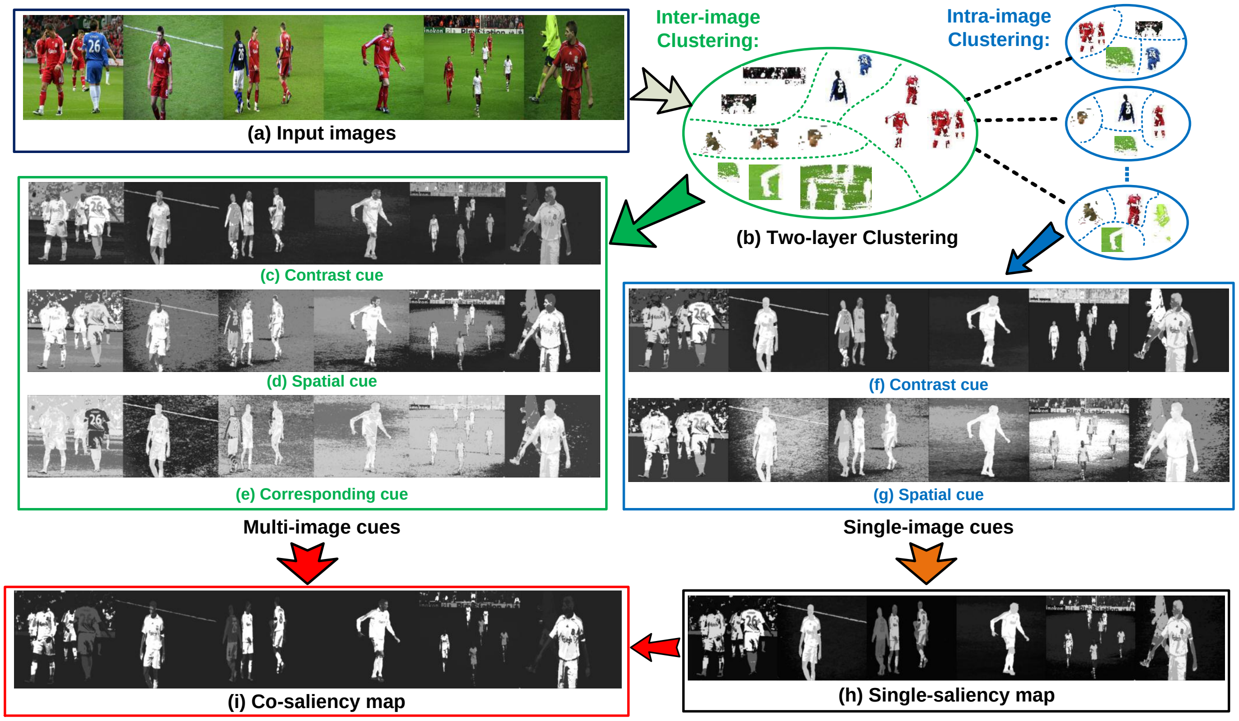


Fig. 2. The framework of our cluster-based co-saliency detection method. (a) The input images. (b) Two-layer clustering of the pixels on intra-image and inter-image layers. Contrast cues (c, f) and spatial cues (d, g) are extracted for both intra-image and inter-image clusters, while corresponding cues (e) is computed only for inter-image clusters. At last, the single image saliency map (h) and co-saliency map (i) are generated from these cues.

B. Cluster-based Saliency Cues

In this section, three cluster-based cues are introduced to measure the cluster-level saliency. The first two are contrast and spatial cues, which are previously used in the single image saliency detection. We extend these two cues into our cluster-based pipeline, and utilize them on both single image and multi-image saliency weighting. We also present a corresponding cue for discovering the common objects appearing on the multiple images. The main property of our cluster-based method is that the visual attention cues appear on cluster-level rather than the individual pixel-level. After clustering single or multiple images, the cluster-level analysis is the same between the single image and multi-image.

Notations: The pixel is denoted by $\{p_i^j\}_{i=1}^{N_j}$ with index i in the image I^j, where the N_j denotes the j th image lattice. $\{z_i^j\}_{i=1}^{N_j}$ denotes the normalized location of the pixel p_i^j in the image I^j. Given M images $\{I^j\}_{j=1}^{M}$, we obtain K clusters[1] $\{C^k\}_{k=1}^{K}$. The clusters are denoted by a set of D-dimensional vectors $\{\mu^k\}_{k=1}^{K}$, in which μ^k denotes the prototype (cluster center) associated with the cluster C^k. And the function b: $\mathbb{R}^2 \rightarrow \{1...K\}$ associates the pixel p_i^j and the cluster index $b(p_i^j)$.

1) Contrast cue: Contrast cue represents the visual feature uniqueness on the single or multiple images. Contrast is one of the most widely used cues for measuring saliency in single image saliency detection algorithms [1], [3], [7], since the contrast operator simulates the human visual receptive fields. This rule is also valid in the case of cluster-based method for the multiple images, while the difference is that contrast cue on the cluster-level better represents the global correspondence relationship than the pixel/patch level.

The contrast cue $w^c(k)$ of cluster C^k is defined using its feature contrast to all other clusters:

$$w^c(k) = \sum_{i=1, i\neq k}^{K} \left(\frac{n^i}{N} \|\mu^k - \mu^i\|_2 \right), \tag{1}$$

where a L_2 norm is used to compute the distance on the feature space, n^i represents the pixel number of cluster C^i, and N denotes the pixel number of all images. This definition favours the large cluster to play more influence. The formulation (1) is similar to Histogram-based Contrast in [3]. However, there are two differences: first, [3] evaluates the saliency value using a simplified histogram, while we employ the cluster. Our cluster-based method perseveres a high coherence. Second, the contrast cue is employed only as one of three basic cues in our co-saliency method. The visual example between [3] and our contrast cue on the single image is shown in Fig. 7.

The contrast cue is valid on both single and multiple images, as shown as Fig. 2(c) and (f). The advantage of the contrast cue is that the rare clusters, e.g. the players, are intuitively more salient. However, the power of contrast cue degrades in the situation of the complex background (e.g. the audience). In addition, it does not address the locating of the common patterns on the multiple images.

2) Spatial cue: In human visual system, the regions near the image center draw more attention than the other regions [32]–[34]. When the distance between the object and the

[1] In practise, we employ two independent class numbers, K_1 and K_2, for single and multiple images, respectively. Here, for cluster-level, we do not discriminate them, and use K to denote the cluster number.

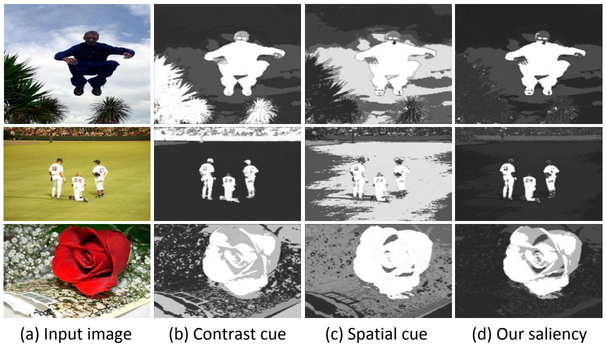


Fig. 3. Some examples of the single image saliency detection using our contrast cue and spatial cue. (a) Input image. (b) Contrast cue is expert in discriminating the most salient object. (c) Spatial cue is good at handling the textured background around the image boundaries. (d) Our final single image saliency map joints two cues and obtains a satisfactory saliency map.

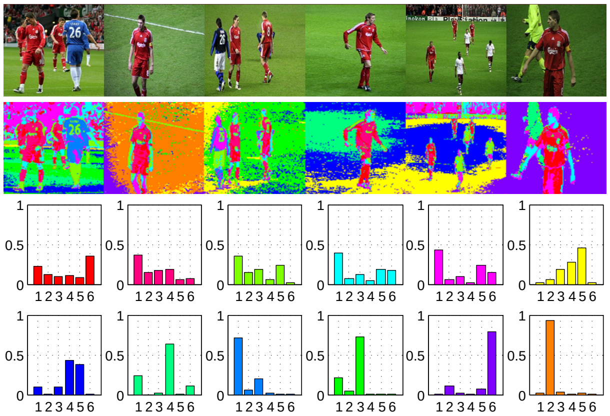


Fig. 4. Illustration of the corresponding cue. Top: the input image. Middle: the inter-image clustering result with cluster number $K = 12$. Bottom: the M-bin histogram for each cluster. These 12 clusters are ranked by their variances, where the bin color is the same with the cluster color in the second row. M equals the image number, i.e. 6 in this example.

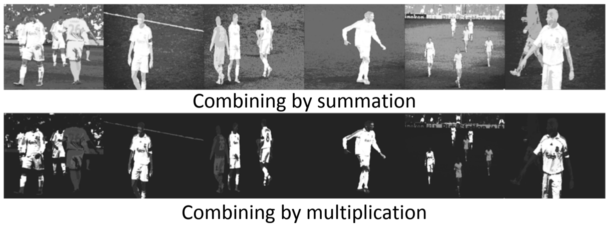


Fig. 5. The saliency map using summation and multiplication, where the multiplication effectively reduces the noisy saliency pixels caused by negative effects of each cue.

image center increases, the attention gain is depreciating. This scenario is known as 'central bias rule' in single image saliency detection. We extend this concept to the cluster-based method, which measures a global spatial distribution of the cluster. The spatial cue $w^s(k)$ of cluster C^k is defined as:

$$w^s(k) = \frac{1}{n^k}\sum_{j=1}^{M}\sum_{i=1}^{N_j}\left[\mathcal{N}\left(\|z_i^j - o^j\|^2 \mid 0, \sigma^2\right)\cdot \delta[b(p_i^j) - C^k]\right], \tag{2}$$

where $\delta(\cdot)$ is the Kronecker delta function, o^j denotes the center of image I^j, and Gaussian kernel $\mathcal{N}(\cdot)$ computes the Euclidean distance between pixel z_i^j and the image center o^j, the variance σ^2 is the normalized radius of images. And the normalization coefficient n^k is the pixel number of cluster C^k. Different from the single image model, our spatial cue w^s represents the location prior on the cluster-level, which is a global central bias on the multiple images.

The same as the contrast cue, the spatial cue is also valid on both single and multiple images, as shown in Fig. 2(d) and (g), where the red players located in center have higher spatial weighting than the blue players. Fig. 3 illustrates the differences between the contrast and spatial cues, where the contrast cue selects the most salient object, while the spatial cue eliminates the textured and 'salient' background, especially those away from the image center. On one hand, the spatial cue addresses the negative effects of the contrast cue and suppresses the confusion of complex background (e.g. the tree in first row, and the audience in second row). On the other hand, the centrally placed background (e.g. the playground in the second row) might have inaccurate spatial bias. Benefiting from both cues, our single image saliency method provides pleasing saliency maps as shown in Fig. 3(d).

3) Corresponding cue: Being different from contrast and spatial cues, the third cue of our method, corresponding cue, is presented to measure how the cluster distribute on the multiple images. The repetitiveness, describing how frequent the object recurs, is an important global property of the common saliency. In fact, the clustering on inter-image approximately perseveres the global correspondence on the multiple images. Fig. 4 gives an example of clustering distribution, where the common object (e.g. the red player) distributes almost equally in each image. Based on this observation, we employ the variances of clusters to roughly measure how widely is the cluster distributed among the multiple input images.

Firstly, a M-bin histogram $\hat{\mathbf{q}}^k = \{\hat{q}_j^k\}_{j=1}^M$ is adopted to describe the distribution of cluster C^k in M images:

$$\hat{q}_j^k = \frac{1}{n^k}\sum_{i=1}^{N_j}\delta[b(p_i^j) - C^k],\ j = 1...M. \tag{3}$$

where n^k is the pixel number of cluster C^k, which enforces the condition $\sum_{j=1}^M \hat{q}_j^k = 1$. Then, our corresponding cue $w^d(k)$ is defined as:

$$w^d(k) = \frac{1}{\text{var}(\hat{\mathbf{q}}^k) + 1}, \tag{4}$$

where var$(\hat{\mathbf{q}}^k)$ denotes the variance of histogram $\hat{\mathbf{q}}^k$ of the cluster C^k. The cluster with the high corresponding cue represents that the pixels of this cluster evenly distribute in each image.

Fig. 2(e) shows the corresponding cue, where the soccer players in red, frequently appearing in all images, have the higher distribution score than those in blue. However, the similar background also has a higher corresponding score. Thanks to the contrast and spatial cues, these background regions are discouraged in the final co-saliency maps.

C. The Co-saliency Maps

So far, three bottom-up cues in our cluster-based method are introduced[2]. Each cue, if used independently, has its advantages and, of course, disadvantages. A common fusion is formulated as a linear summation [1], [7] or point-wise multiplication [35] of static salient features. Fig. 5 illustrates the difference between summation and multiplication fusions. The multiplication is better in depressing the noises than summation. And summation is better in getting higher recall. For saliency detection, however, the precision is more important than recall [31]. In our work, our also prefer a precise, rather than a large, saliency map. Therefore, we employ the multiplication operation to integrate the saliency cues.

Before combining saliency cues, we normalize each cue map to standard Gaussian using the distribution of scores across all clusters. Then the cluster-level co-saliency probability $p(k)$ of cluster k is defined as:

$$p(C^k) = \prod_i w_i(k), \tag{5}$$

where $w_i(k)$ denotes saliency cues.

Now that the cluster-level co-saliency value is computed, which provides the discrete assignment. Then we smooth the co-saliency value for each pixel. The saliency likelihood of the pixel x belonging to the cluster C^k satisfies a Gaussian distribution $\mathcal{N}$ as:

$$p(x \mid C^k) = \mathcal{N}\left(\|v_x, \mu^k\|_2 \mid 0, \sigma_k^2\right), \tag{6}$$

where v_x denotes the feature vector of pixel x, and the variance σ_k of Gaussian uses the variance of cluster C^k. Hence, the marginal saliency probability $p(x)$ is obtained by summing the joint saliency $p(C^k)p(x|C^k)$ over all clusters:

$$p(x) = \sum_{k=1}^{K} p(x, C^k) = \sum_{k=1}^{K} p(x \mid C^k) p(C^k). \tag{7}$$

Finally, the pixel-level co-saliency is obtained, as shown in Fig. 2(i). Our method is summarized in Algorithm 1.

III. Experiments

We evaluate our co-saliency detection method on two aspects: the single image saliency detection, and the co-saliency detection on the multiple images. We compare our method with the state-of-the-art methods on a variety of benchmark datasets. And then a discussion is given to analyze the effectiveness of each saliency cue, the running time, and the cluster number. In our implementation, CIE Lab color and Gabor filter are employed to represent the feature vector. The Gabor filter responses with 8 orientations. The bandwidth is chosen to be 1 and one scale is extracted. We compute the magnitude map of Gabor filter by combining 8 orientations as the texture feature. K-means is used in two-layer clustering. The cluster numbers in Algorithm 1 are set to $K_1 = 6$ for intra image (single image), and $K_2 = \min\{3M, 20\}$ for inter image (multiple images), where M denotes the image number.

Algorithm 1: Cluster-based Co-saliency Detection.

```
Input: The input image set $\{I^j\}_{j=1}^{M}$, intra cluster number
       $K_1$, and inter cluster number $K_2$.
1. Single image saliency detection:
for each image do
    Clustering image into $K_1$ clusters;
    for each cluster do
        Computing the contrast cue using Eq. (1) and
        spatial cue using Eq. (2);
    end
    Combining two saliency cues using Eq. (5);
    for each pixel do
        Obtaining the final single saliency map using Eq.
        (7);
    end
end
2. Multiple image co-saliency detection:
Clustering all images into $K_2$ clusters;
for each cluster do
    Computing the contrast cue using Eq. (1), spatial cue
    using Eq. (2), and corresponding cue using Eq. (4);
end
Combining the single saliency map and three
cluster-based cues using Eq. (5);
for each pixel do
    Assigning the final co-saliency map using Eq. (7);
end
Output: The single saliency map and co-saliency map.
```

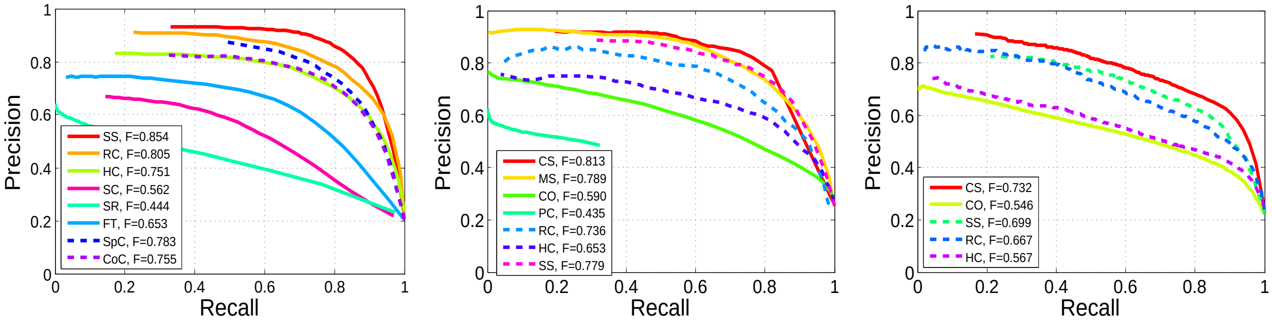
A. Single-image saliency detection

First, we evaluate our method on the single image saliency detection. We employ the publicly available MSRA1000 saliency database provided by [31], which is one of the largest saliency image databases (1000 images) and has pixel-level ground truth in the form of accurate manual labels. We compare our single image saliency method (SS) with five state-of-the-art detection methods: Spatiotemporal Cues (SC) in [30], Frequency-tuned saliency (FT) in [31], Spectral residual (SR) in [6], Region-based Contrast (RC) and Histogram-based Contrast (HC) in [3]. Fig. 6(a) shows the results using naive thresholding on the dataset, where the F is calculated by:

$$\text{F-measure} = \frac{(1+\beta^2)\text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \tag{8}$$

where we use $\beta^2 = 0.3$ as in [3], [31] to weight precision more than recall. Moreover, we also provide two individual saliency cues of our method: contrast cue (CoC) and spatial cue (SpC), as shown the dotted curves in Fig. 6(a). Our contrast cue formulates similarly to the HC [3], while ours employs the cluster instead of histogram. Hence, the performance of contrast cue ($F = 0.755$) has the similar results to the HC [3] ($F = 0.751$). Interestingly, the spatial cue outperforms the contrast cue, because most natural images are satisfying the central bias rule in the photography. This observation agrees with the success of the recent proposed methods [32], [33], [36]. Although the spatial cue itself can not compete with the

[2] In fact, there totally have five cues: three cues on the multiple images and two cues on the single image.



(a) Saliency on single image (b) Co-saliency on image pair (c) Co-saliency on multi-image

Fig. 6. Comparison between our cluster based saliency detection methods, including the single image saliency (SS) and co-saliency (CS) methods, and other state-of-the-art works. (a) The Precision/Recall curves for naive thresholding of saliency maps on MSRA1000 dataset. We compared our work with the state-of-the-art single image saliency detection methods including RC [3], HC [3], SC [30], SR [6], FT [31]. For better understanding of the contribution of each individual cues, we also provide two curves of individual saliency cues: contrast cue (CoC) and spatial cue (SpC). (b) The Precision/Recall curves of co-saliency map on co-saliency pairs dataset. Our co-saliency detection method (CS) are compared with CO [12], MS [11] and PC [10]. (c) The Precision/Recall curves of saliency detection on iCoseg dataset.

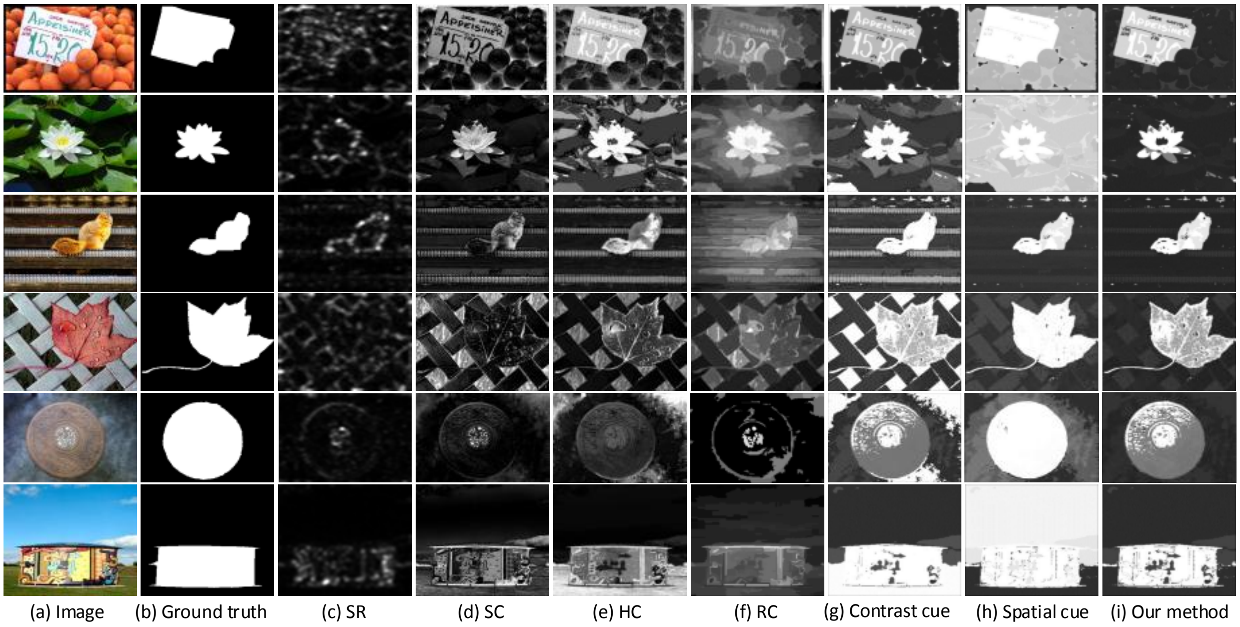


Fig. 7. Visual comparison of single image saliency detection on MSRA1000 dataset. (a) Input image. (b) Ground truth. Saliency maps: (c) SR [6]. (d) SC [30]. (e) HC [3]. (f) RC [3]. (g) Our contrast cue. (h) Our spatial cue. (i) Our final single image saliency.

RC [3], our cluster-based method on the single image based on the spatial and contrast cues outperforms RC [3]. And our F-measure is 0.854, which is around 5% better than RC [3] ($\mathbf{F} = 0.805$).

Visual comparison of different saliency results obtained by various methods is shown in Fig. 7. For the first two images, the contrast cue contributes more than the spatial cue. However, the spatial cue still assigns the salient objects with a higher saliency score than the background. For the case of the complex and textured background or the low contrast foreground, such as the third to fifth images in Fig. 7, most existing saliency detection methods lose their power. Since these methods only employ the low-level feature to detect saliency, they are easily affected by the high-contrast noise on background (the third and fourth images) and low-contrast foreground (the fifth image). However, as argued before, the center bias rule, based on spatial rule instead of feature contrast, is effective for this case. And our method, which combines the contrast and spatial cues, is robust and obtains the better saliency map. In the last image of Fig. 7, contrast cue itself can not locate the salient object accurately, since the grassland has high contrast. Spatial cue itself provides high score for the sky as this image is disobeying the central bias rule. In other words, the salient object is lower than the image center. Nevertheless, our final single image map integrates the benefits of both cues, and only the salient object, the tent, satisfies these two saliency cues.

B. Co-saliency detection

Most existing co-saliency detection methods focus on a pair of images, which are designed to detect salient objects in common. We compare our co-saliency method (CS) with three previous methods: the co-saliency (CO) in [12], the preattentive co-saliency (PC) in [10], and the multi-image saliency (MS) in [11]. The dataset uses the Co-saliency Pairs dataset [11], which includes 210 images (105 image pairs). Each image pair contains one or more similar objects with different backgrounds. Fig. 6(b) shows the Precision/Recall curves for naive thresholding on this dataset. We also plot the single image saliency methods: RC, HC in [3], and our single image method (SS), shown as the dotted curves. Similar to the observation in Section III-A, our single image method (SS) wins among all single image methods. Moreover, our SS outperforms co-saliency methods PC [10] and CO [12], and is comparable to MS [11]. The main reason is that each image inside co-saliency pairs dataset has the obvious foreground, which reduces the contribution of the second image. However, our co-saliency method still improves from the SS ($F = 0.779$) and MS [11] ($F = 0.789$) to $F = 0.813$.

Fig. 8 shows some visual results of saliency detection on image pairs. Overall, our method provides visually acceptable saliency, which is consistent with visual attention. In the results of RC [3], highly textured backgrounds belonging to non-salient regions are not suppressed, e.g. the first two rows in Fig. 8. Relative large objects are hardly captured by MS [11] as shown in the second and fifth rows in Fig. 8. One potential reason is that MS [11] employs the local contrast of each patch and the patch size is not adaptable to the global constraint. As a result, the inside patches of large object lack salient property against their surrounding patches. In contrast, our method relieves this limitation by clustering them as one entire group, and obtains the better results on large object. The complex background, such as the third and sixth, also hurts the saliency detection in RC [3] and MS [11]. Our spatial cue provides the robustness to the complex background. Therefore, our method offers the best results for the complex background. The last two pairs demonstrate the difference between the single and co-saliency detection. The single image saliency detects all the salient objects for each image. The power of co-saliency extracts the common saliency from the multiple images, such as the yellow boat and red peony.

At last, we employ the CMU Cornell iCoseg dataset [22] to test our co-saliency method on the multiple images (image number $\gg 2$), which is the largest publicly available co-segmentation benchmark with 643 images in 38 groups. Since the co-saliency methods, MS [11] and PC [10], are not valid on more than two images, we only compare the CO [12] and our co-saliency method (CS). The same as above, we also plot the Precision/Recall curves of the single image saliency methods: RC, HC in [3], and our single image method (SS). Fig. 6(c) shows the curves of these methods. Our single image method wins among all single image methods with $F = 0.699$. Without surprise, our co-saliency method obtains the best performance on the multiple images with $F = 0.732$. The iCoseg dataset is provided for the co-segmentation, where the common objects may not have the bottom-up saliency properties. Therefore, the Precision/Recall scores of all methods are lower than those on co-saliency pairs dataset. Some co-saliency detection results are shown in Fig. 9, where the image set includes the common salient object with non-salient background (first two samples) and complex background (last two samples). Our method obtains the nice co-saliency maps, utilizing the overall constraints on the multiples images.

C. The running time

Our approach adopts the bottom-up cues to measure the co-saliency without heavy learning. Simultaneously, the cluster-based method, comparing with the individual pixel operator, achieves an efficient storage and computation. In our method, we are not constrained to specific choices of the clustering methods, and K-means is used. The experiment is run on a laptop with Dual Core 2.8 GHz processor and 4GB RAM. The code is implemented in matlab without optimization. We randomly select the images in the same category of the iCoseg dataset with various image numbers, and resize the images to the resolution 128×128 to evaluate the running time. For each image number, we run 20 times independently, and report the mean value and the variance. Fig. 10(a) shows the running time in seconds of our entire method (red) and the clustering process (blue) running on the multiple images with cluster number $K = \min\{3M, 20\}$, where M denotes the image number. The running time increases consistently with respect to the image number. For the large image numbers (i.e. $M > 5$), the clustering (K-means) is the main consumer of our method. Typically, our method takes about 3 seconds for a pair of images with image size 128×128, and 4 seconds for 4 images. In contrast, the co-saliency method [12] takes about 50 seconds for generating co-saliency maps for 4 images[3]. And the method [11] spends about 450 seconds for an image pair. Our proposed method obtains a substantial improvement in running time with the better performance.

A similar research to our method is co-segmentation, which extracts the pixel-level segmentation from two/multiple images. However, the computational requirements of co-segmentation are very expensive. For example, the reported running time of [24] is 30 seconds for two images. For another example, the reported running time of [23] is 4 to 9 hours for 30 images. In contrast, our proposed method processes 300 images within only 22 mins, and offers a competitive segmentation result. The more details between co-segmentation and our method are provided in Section IV-A.

D. Discussion

In this section, we discuss three factors related to our approach: the effectiveness of each saliency cue, the cluster number, and the degenerated cases.

1) The effectiveness of each saliency cue: Our method employs three bottom-up cues to measure the co-saliency. To evaluate the effectiveness of each cue, we test seven Precision/Recall curves on the iCoseg dataset: co-saliency (CS), single-saliency (SS), contrast cue (CS-CoC), spatial cue (CS-SpC), and corresponding cue (CS-Corr) on multi-image, contrast cue (SS-CoC) and spatial cue (SS-SpC) on

[3] This running time is evaluated on our platform, since the paper [12] only reports the co-segmentation time without the co-saliency detection time.

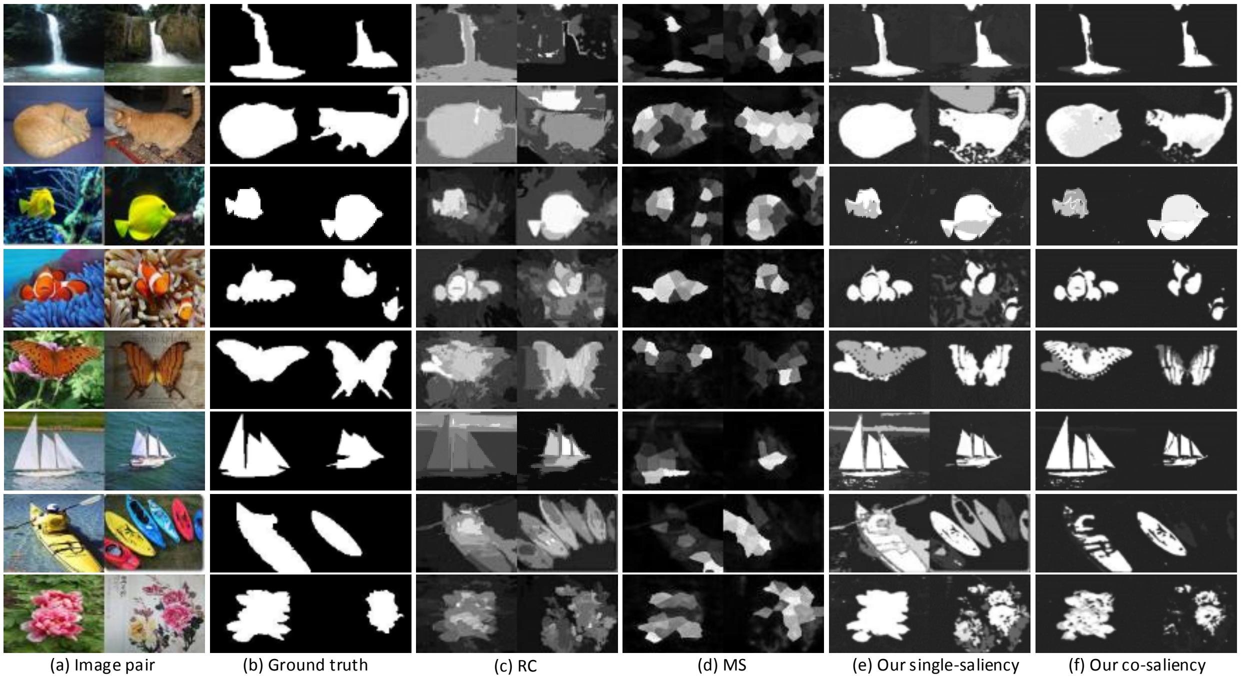


Fig. 8. Visual results of saliency detection on the Co-saliency Pairs dataset. (a) Input image pair. (b) Ground truth. (c) Saliency map by RC [3]. (d) Saliency map by MS [11]. (e) Our single image saliency. (f) Our co-saliency.

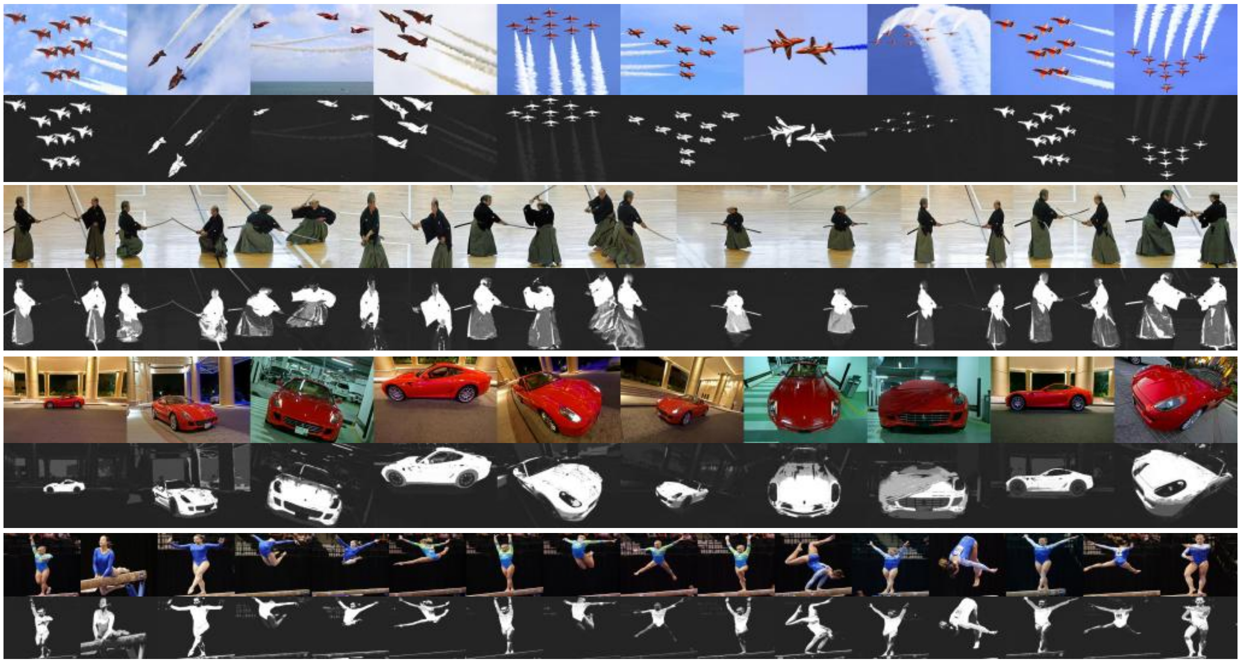
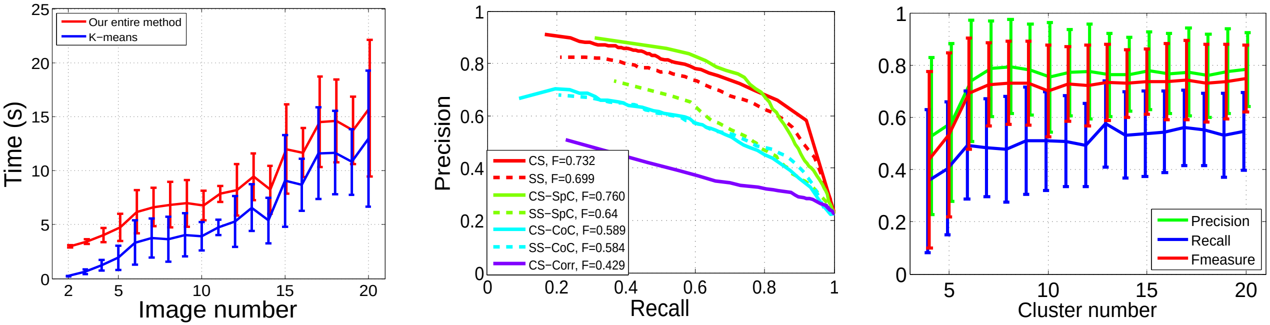


Fig. 9. Some visual results of our co-saliency detection on the iCoseg dataset. Our co-saliency map provides the accurate common object mask on the multiple images.

single image. From the results in Fig. 10(b), we have the following four observations. First, the saliency detection in the multi-image case (solid curves) mostly performs better than the result in the single image (dotted curves). Since the global correspondence utilizes additional images to constrain the saliency detection problem, and makes it easier to decide which object is the most salient one among many possible candidates. Second, the contrast cue performs similarly on the single and multi-image cases for the iCoseg dataset. This is due to the fact that the images of the same category in this dataset are captured from the similar scenes. As a result, the global contrast cue on the multi-image is close to the contrast cue on the single image. Third, the spatial cue (CS-SpC) is the most useful one for this dataset, and performs even better than co-saliency (CS). This is because foreground objects in the iCoseg dataset are mostly located in the image center. However, this



(a) (b) (c)

Fig. 10. The analysis of our saliency map on the iCoseg dataset with various conditions. (a) The running time for various image numbers with image sizes 128×128 and cluster number $K = \min\{3M, 20\}$, where M denotes the image number. (b) Precision/Recall curves with different saliency cues. (c) Results with various cluster numbers.

location prior is not always valid in practice. One example is the last row in Fig. 7. Fourth, the corresponding cue (CS-Corr) itself has lower performance. This is expected, since the merit of the corresponding cue is to enforce the ‘common’ property on the multiple images, rather than distinguishing the saliency. In other words, the corresponding cue mainly helps on deciding which object is the common salient one among many possible candidates. Therefore, the similar backgrounds confuse the corresponding cue to wrongly select the salient pixels inside those areas.

2) The cluster number: Thanks to the soft assigning co-saliency value in Eq. (7), the cluster number is not the major factor in our saliency detection process. Here we only vary the cluster number of inter image clustering. We evaluate all the categories of the iCoseg dataset with various cluster numbers, and report the mean value and the variance. Fig. 10(c) shows the performance of saliency detection with respect to various cluster numbers. The co-saliency results in terms of precision, recall, and F-measure, are stable when the cluster number goes beyond 15. On the other hand, a large cluster number leads to the increasing computational requirements. Generally, we chose a loose upper bound of cluster number with $K = \min\{3M, 20\}$ in the experiments, where M denotes the image number.

3) The degenerated cases: There are some degenerated cases for our saliency detection. The first failure case is that the object is composed of multiple components as shown in Fig. 11 (a) and (c). Our saliency method only highlights the salient component rather than the entire object. The main reason is that our saliency detection is based on bottom-up cues without heavy learning, which could not provide the object-level constraint. Our method also degenerates when the non-salient background involves the similar appearance (e.g. color) as the salient areas. Fig. 11 (b) shows this case, where the cloth of child is similar to the walls of the house in the background. This hardness can also be caused by the protective color of animal, as shown in Fig 11 (d-e).

IV. Applications of the Co-saliency

In the past several years, the single image saliency map has been widely used in many image processing applications. However, the co-saliency is still a relatively under-explored technology. In this section, we discuss four potential applications, which benefit from the co-saliency detection results.

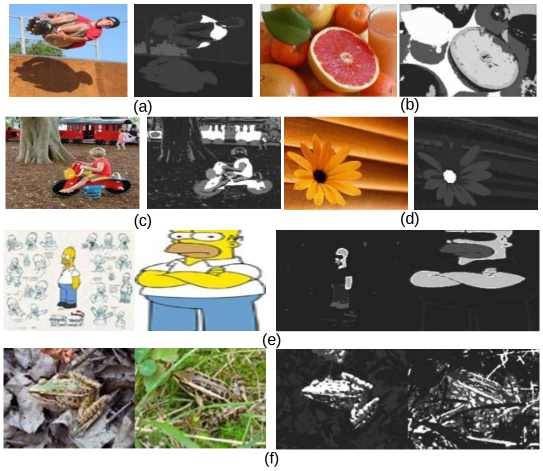


Fig. 11. Some challenging examples for our saliency detection. We show both examples on single image saliency detection (a-b), and multiple image saliency detection (c-e).

A. Co-segmentation

A directly related application is motivated by the recent trend of co-segmentation. Most co-segmentation tasks are formulated as an energy optimization problem, including a within-image energy term and a global constraint term [13], [14], [24], [37]. These complex energy functions often cost significantly. More importantly, the co-segmentation focuses on the “similarly looking object” in a small number of images, and tends to wrongly label the “similarly background”, especially in the fully automated system. A common solution to this high level knowledge is to use manually input strokes [22], [38] or bounding boxes [39], [40]. In contrast, the co-saliency map provides an initial highlight of similarly looking object, which replaces the user interaction.

In this experiment, we utilize a bilayer segmentation method that estimates the foreground and background pixels of the input image by a Markov random field function. The energy

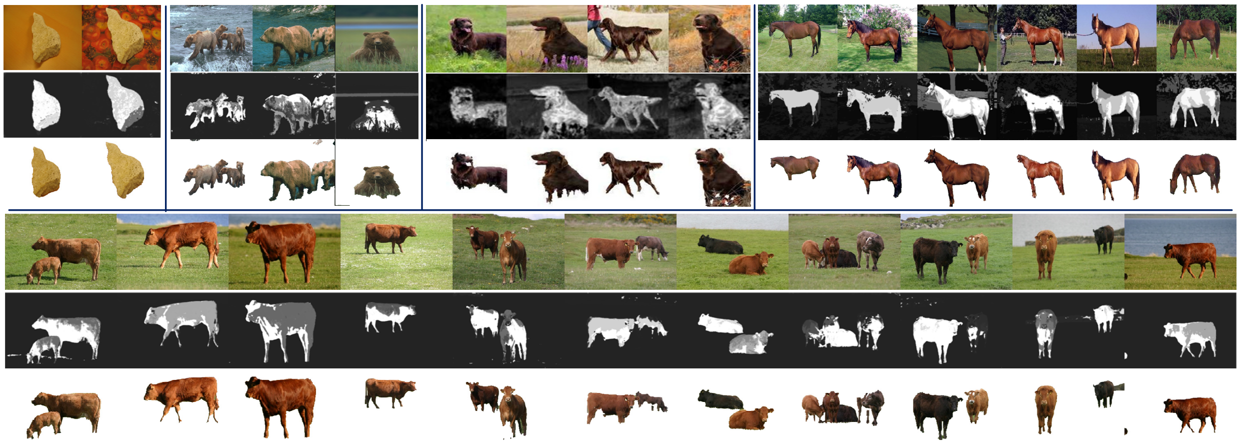


Fig. 12. Segmentation results using our co-saliency map. First and fourth rows are the input image set, the second and fifth rows are our co-saliency map, and the third and sixth rows show the segmentation results using our method.

TABLE I
SEGMENTATION ACCURACY (%) ON THE MSRC DATASET.

Class	Plane	Cow	Cat	Face	Bike	Avg.
[23]	73.8	81.6	74.4	84.3	63.3	75.48
[12]	87.66	91.36	86.68	87.27	76.76	85.95
Our method	86.75	90.39	83.53	84.83	72.28	83.55

function E is defined as:

$$E = \sum_{i \in I} p(x_i) + \sum_{\{p,q\} \in \mathcal{N}} V_{p,q}(x_p, x_q), \tag{9}$$

where $\mathbf{x} = \{x_i | i \in I\}$ denotes the binary-valued label of image I, $p(x_i)$ is the co-saliency value of pixel i in Eq. (7), $V_{p,q}$ is the smoothness penalty, which measures the cost of assigning different labels to two neighboring pixels, $\mathcal{N}$ is a set of pairs of adjacent pixels. The weight $V_{p,q}$ of the smoothness term is given by [41]:

$$V_{p,q} = \lambda \exp\left(-\beta ||z_p - z_q||^2\right), \tag{10}$$

where z_p is the RGB color appearance of the pixel p, $\beta = \left(2\left\langle (z_p - z_q)^2 \right\rangle\right)^{-1}$, $\langle\cdot\rangle$ denotes expectation over the image, and λ is the weight for the contrast sensitive term. Fig. 12 shows some segmentation results using our co-saliency map. The number of images in a group of image varies from 2, 3, 4, 6 to 11. In general, we see that our co-saliency map provides an accurate saliency mask (e.g., stone and bear). In other images, which have a shared background in some images (e.g. horse and cow), our method automatically extracts the salient foreground.

Table I shows the comparison between our method and [12], [23] on the MSRC dataset [42]. The performance is measured by its accuracy, i.e. the proportion of correctly classified pixels (foreground and background) to the total number of pixels. The column named 'Avg.' in the table denotes the average score on all categories. Our method outperforms the co-segmentation [23] with about 8% improvement. The co-segmentation method [12] employs the co-saliency map as the prior and segments the foreground with a global energy minimization model, which reaches the best segmentation accuracy (Avg. $=$ 85.95%). Our method obtains a slightly lower accuracy (Avg. $=$ 83.55%) without any extra global energy term in Eq. (9). For some categories, such as cat, face and bike, the common objects appear in the wide range of illumination and appearance, which makes our algorithm hard to group them into one cluster. However, the global energy minimization model of [12] also brings an expensive computational requirement, which needs more than 40 seconds for 4 images of the resolution 128 × 128. In contrast, our method has the advantage of cheap computation, which only needs about 5 seconds to obtain the segmentation results without significantly reducing the quality.

B. Robust Image Distance

The other interesting application of the co-saliency map is the robust image distance. Image visual distance measuring is a fundamentally problem and is widely employed in the reranking of image retrieval [43], [44] and image content clustering [45], [46]. Recently, the object sensitive image pair-distance has been demonstrated helpful to the image retrieval based on global feature comparison [15], [20], [47], [48]. For example, Rother et. al. [13] employ the co-segmentation energy as a distance measure between an image pair. However, this method is limited to its segmentation cue, which is short of a general formula for other common visual features. Inspired by [13], we provide a more efficient and general robust image distance based on the co-saliency map. The traditional visual distance between two images I_1 and I_2 is denoted by $D(I_1, I_2)$. Given an image pair, the co-saliency segments each image into the co-saliency foreground I^f and the background I^b using Eq. (9). Then we introduce a saliency weighting rate r^f as:

$$r^f = \text{Size}(I_1^f + I_2^f) \times \text{Mean}(I_1^f + I_2^f), \tag{11}$$

where Size$(\cdot)$ and Mean$(\cdot)$ denote the pixel number and co-saliency mean value of foreground I^f. The background weighting rate r^b is defined similarly. Finally, our robust image distance $D'(I_1, I_2)$ based on co-saliency map is defined as:

$$D'(I_1, I_2) = \frac{r^f}{r^f + r^b} D(I_1^f, I_2^f) + \frac{r^b}{r^f + r^b} D(I_1^b, I_2^b), \tag{12}$$

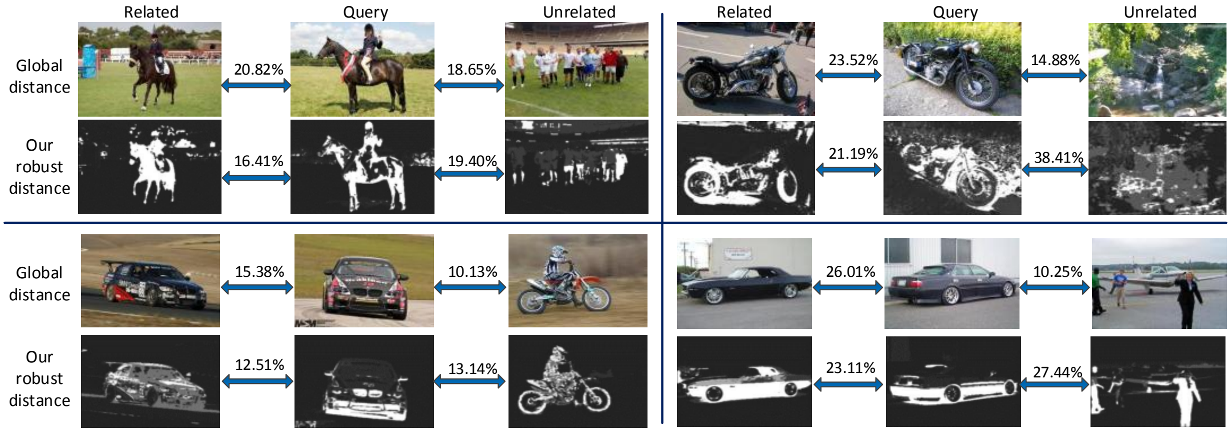


Fig. 13. Robust image distance based on the co-saliency map. The middle image is the query image, the left image is the matched image with similar object, and the right image is the unrelated image but with similar global color statistics. With the help of the co-saliency map, our method reduces the distance from the matched image, and increases the distance from the unrelated image.

where

$$I_i = I_i^f + I_i^b,\ i = 1, 2.$$

Simply put, our method makes the co-saliency region (object) play the more important role in the image distance computing.

Fig. 13 illustrates some samples of comparison between the traditional distance and our robust image distance based on the co-saliency map. Our method is not limited to specific choice of the distance measure. In this experiment, color histogram Chi-squared distance is used. In the global image distance, two unrelated images have the smaller distance caused by similar backgrounds. In contrast, our robust distance focuses more on the salient objects, and relieves the affection of backgrounds. Comparing with the distance based on co-segmentation [13], [15], our proposed framework also has a favourite running time, more general form, and is easy to implement. Any distance measure can be integrated into our framework.

C. Weakly Supervised Learning

Weakly supervised learning discriminates a new object class from training images [29] with weak annotations. Different from the fully supervised scenario, the location of objects is not given. Existing approaches discover possible regions of object instances and output a set of windows containing the common object [50]–[52]. However, some classifier models need the full labeling map on pixel-level, e.g. auto-context [49]. Benefiting from our co-saliency detection, these full label classifiers could be learned without any user intervention. Fig. 14 gives an illustration of auto-context learning with our co-saliency map. Firstly, the training images are selected by weakly supervised selecting. Next, our co-saliency detection method provides the co-saliency map as the full labeling map. With the co-saliency map, the auto-context model is learned and can be used to recognize the same type of objects in the challenging images as shown in Fig.14(d).

D. Video Foreground Detection

Video is treated as a sequence of images, and sometimes the foreground could be defined as the saliency object [53]–[55]. The saliency of video also conforms to the feature contrast property [7], [30], [56]. Simultaneously, a reasonable assumption is that the foreground object in video may recur in most frames, which fits the corresponding cue of the multiple images. For the spatial cue of our method, we do statistics about the position of foregrounds on video saliency dataset

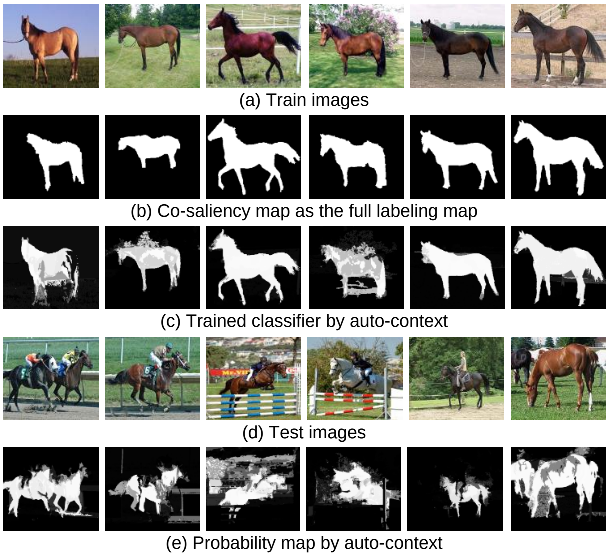

(a) Train images
(b) Co-saliency map as the full labeling map
(c) Trained classifier by auto-context
(d) Test images
(e) Probability map by auto-context

Fig. 14. Classifying results using our co-saliency map and auto-context [49]. (a) Training images collected by weak supervision. (b) Our co-saliency detection map as the full labeling map. (c) Trained classifier by auto-context. (d) Test images. (e) Probability maps by auto-context recognition

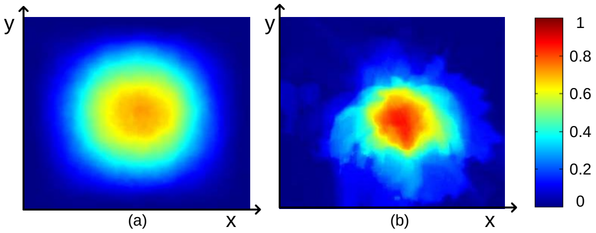


Fig. 15. Center bias map estimation on (a) single image and (b) video. The xy-axes denote the normalized image spatial coordinates, and the color denotes the saliency distribution of center bias map.

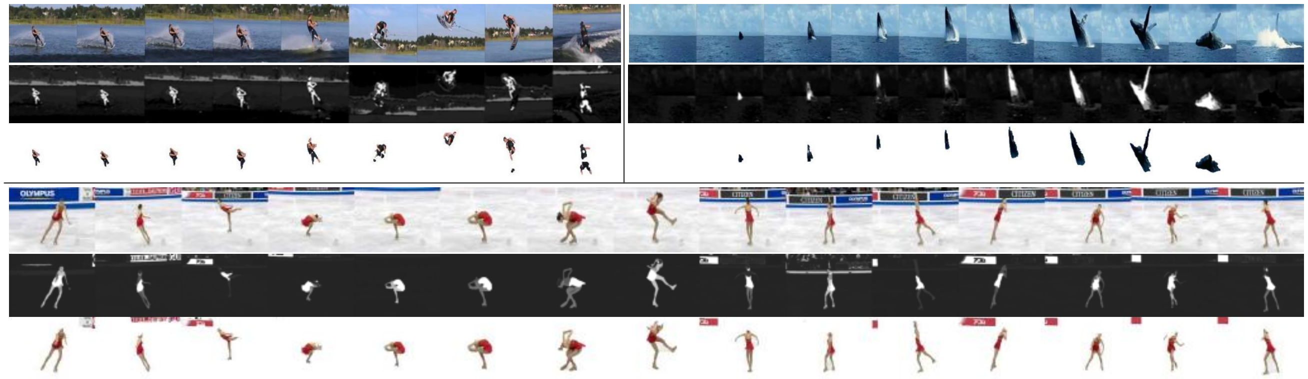


Fig. 16. The results of foreground video segmentation using our co-saliency detection method. First and fourth rows are the input video sequence, the second and fifth rows are the co-saliency map, and the third and sixth rows show the segmentation results.

[54], which includes 10 videos with pixel-level ground truth. The video frames are normalized. Fig. 15 shows the location map of center bias rule, where the left is the estimated result on single saliency dataset [31] and the right shows the result on video saliency dataset [54]. The xy-axes denote the normalized image spatial coordinates, and the color denotes the saliency distribution of center bias map, which is similar to [33], [34]. It can be seen that the center bias rule (spatial cue) is still valid for the saliency of video, same as the saliency of the single image. Therefore, our method could be directly used to discover the foreground on video. Fig. 16 shows the results of our co-saliency detection method on videos [57], where the foregrounds are extracted well by our co-saliency maps. Moreover, as a global association on the multiple images, our method has sufficient robustness to the outlier frames, where the foreground disappears, such as the first and last frames of the second sample in Fig. 16. Note that we only use the color and texture features to cluster the pixel and extract co-saliency map. However, other spatio-temporal features, e.g. optical flow, could be easily introduced into our framework to improve the detection result on video.

V. Conclusion

In this paper, we have presented an efficient and effective cluster-based co-saliency detection method. A global association constraint is enforced by clustering, avoiding the heavy learning. Contrast and spatial cues worked well for a lot datasets, since the objects in these datasets are well centered in the images and occupy a large portion of them. Corresponding cue effectively discovered the common objects on the multiple images using clustering distribution. The combined cue by multiplication obtained the encouraging results on a wide variety of datasets on both single image saliency and co-saliency detection. Our co-saliency detection, as an automatic and rapid pre-processing step, is useful for the many vision applications.

In the future, we plan to use more visual features to improve the co-saliency detection results, and investigate motion features to detect co-saliency on video. Also, it is desirable to apply saliency detection algorithms to handle the large scale dataset and develop the object-driven system.

References

[1] 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, 1998. 1, 3, 5

[2] W. Lee, T. Huang, S. Yeh, and H. Chen, “Learning-based prediction of visual attention for video signals,” *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3028–3038, 2011. 1

[3] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, “Global contrast based salient region detection,” in *CVPR*, 2011, pp. 409–416. 1, 2, 3, 5, 6, 7, 8

[4] A. Toet, “Computational versus psychophysical bottom-up image saliency: A comparative evaluation study,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 11, pp. 2131–2146, 2011. 1

[5] R. Valenti, N. Sebe, and T. Gevers, “What are you looking at? improving visual gaze estimation by saliency,” *Int. J. Comput. Vision*, vol. 98, no. 3, pp. 324–334, 2012. 1

[6] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in *CVPR*, 2007, pp. 1–8. 1, 5, 6

[7] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum, “Learning to detect a salient object,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, 2011. 1, 3, 5, 11

[8] C. Lang, G. Liu, J. Yu, and S. Yan, “Saliency detection by multitask sparsity pursuit,” *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1327–1338, 2012. 1

[9] D. Jacobs, D. Goldman, and E. Shechtman, “Cosaliency: where people look when comparing images,” in *ACM symposium on User interface software and technology*, 2010, pp. 219–228. 1

[10] H. Chen, “Preattentive co-saliency detection,” in *ICIP*, 2010, pp. 1117–1120. 1, 2, 6, 7

[11] H. Li and K. Ngan, “A co-saliency model of image pairs,” *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, 2011. 1, 2, 6, 7, 8

[12] K. Chang, T. Liu, and S. Lai, “From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model,” in *CVPR*, 2011, pp. 2129–2136. 1, 2, 6, 7, 10

[13] C. Rother, T. Minka, A. Blake, and V. Kolmogorov, “Cosegmentation of image pairs by histogram matching - incorporating a global constraint into mrfs,” in *CVPR*, 2006, pp. 993–1000. 1, 2, 9, 10, 11

[14] D. Hochbaum and V. Singh, “An efficient algorithm for co-segmentation,” in *ICCV*, 2009, pp. 269–276. 1, 2, 9

[15] S. Vicente, C. Rother, and V. Kolmogorov, “Object cosegmentation,” in *CVPR*, 2011, pp. 2217–2224. 1, 10, 11

[16] H. Tan and C. Ngo, “Common pattern discovery using earth mover’s distance and local flow maximization,” in *ICCV*, 2005, pp. 1222–1229. 1

[17] J. Yuan and Y. Wu, “Spatial random partition for common visual pattern discovery,” in *ICCV*, 2007, pp. 1–8. 1

[18] A. Toshev, J. Shi, and K. Daniilidis, “Image matching via saliency region correspondences,” in *CVPR*, 2007, pp. 1–8. 1

[19] M. Cho, Y. Shin, and K. Lee, “Co-recognition of image pairs by data-driven monte carlo image exploration,” in *ECCV*, 2008, pp. 144–157. 1

[20] L. Yang, B. Geng, Y. Cai, A. Hanjalic, and X.-S. Hua, “Object retrieval using visual query context,” *IEEE Trans. Multimedia*, vol. 13, no. 6, pp. 1295–1307, 2011. 1, 10

[21] S. Goferman, A. Tal, and L. ZelnikManor, “Puzzle-like collage.” *Comput. Graph. Forum*, vol. 29, no. 2, pp. 459–468, 2010. 1

[22] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, “Interactively co-segmentating topically related images with intelligent scribble guidance,” *Int. J. Comput. Vision*, vol. 93, no. 3, pp. 273–292, 2011. 1, 7, 9

[23] A. Joulin, F. Bach, and J. Ponce, “Discriminative clustering for image co-segmentation,” in *CVPR*, 2010, pp. 1943–1950. 2, 7, 10

[24] L. Mukherjee, V. Singh, and J. Peng, “Scale invariant cosegmentation for image groups,” in *CVPR*, 2011, pp. 1881–1888. 2, 7, 9

[25] G. Kim, E. Xing, L. Fei-Fei, and T. Kanade, “Distributed cosegmentation via submodular optimization on anisotropic diffusion,” in *ICCV*, 2011, pp. 169–176. 2

[26] L. Mukherjee, V. Singh, and C. Dyer, “Half-integrality based algorithms for cosegmentation of images,” in *CVPR*, 2009, pp. 2028–2035. 2

[27] H. Wu, Y. Wang, K. Feng, T. Wong, T. Lee, and P. Heng, “Resizing by symmetry-summarization,” *ACM Trans. Graph.*, vol. 29, no. 159, pp. 1–10, 2010. 2

[28] Y. Fang, Z. Chen, W. Lin, and C. Lin, “Saliency detection in the compressed domain for adaptive image retargeting,” *IEEE Trans. Image Process.*, vol. 21, no. 9, pp. 3888–3901, 2012. 2

[29] M. H. Nguyen, L. Torresani, F. de la Torre, and C. Rother, “Weakly supervised discriminative localization and classification: a joint learning process,” in *CVPR*, 2009, pp. 1925–1932. 2, 11

[30] Y. Zhai and M. Shah, “Visual attention detection in video sequences using spatiotemporal cues,” in *ACM Multimedia*, 2006, pp. 815–824. 2, 5, 6, 11

[31] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, “Frequency-tuned salient region detection,” in *CVPR*, 2009, pp. 1597–1604. 2, 5, 6, 12

[32] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, “Visual saliency detection by spatially weighted dissimilarity,” in *CVPR*, 2011, pp. 473–480. 2, 3, 5

[33] B. W. Tatler, “The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions,” *Journal of Vision*, vol. 7, no. 14, pp. 1–17, 2007. 3, 5, 12

[34] T. Judd, K. Ehinger, F. Durand, and A. Torralba, “Learning to predict where humans look,” in *ICCV*, 2009, pp. 2106–2113. 3, 12

[35] C. Lang, T. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, “Depth matters: Influence of depth cues on visual saliency,” in *ECCV*, 2012, pp. 101–115. 5

[36] L. Liu, R. Chen, L. Wolf, and D. CohenOr, “Optimizing photo composition,” *Computer Graphic Forum*, vol. 29, no. 2, pp. 469–478, 2010. 5

[37] S. Vicente, V. Kolmogorov, and C. Rother, “Cosegmentation revisited: Models and optimization,” in *ECCV*, 2010, pp. 465–479. 9

[38] Y. Li, J. Sun, C.-K. Tang, and H.-Y. Shum, “Lazy snapping,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 303–308, 2004. 9

[39] C. Rother, V. Kolmogorov, and A. Blake, ““GrabCut”: interactive foreground extraction using iterated graph cuts,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, 2004. 9

[40] V. Lempitsky, P. Kohli, C. Rother, and T. Sharp, “Image segmentation with a bounding box prior,” in *ICCV*, 2009, pp. 277–284. 9

[41] A. Blake, C. Rother, M. Brown, P. Pérez, and P. Torr, “Interactive image segmentation using an adaptive GMMRF model,” in *ECCV*, 2004, pp. 428–441. 10

[42] J. Shotton, J. Winn, C. Rother, and A. Criminisi, “Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context,” *Int. J. Comput. Vision*, vol. 81, no. 1, pp. 2–23, 2009. 10

[43] R. Zhang and Z. Zhang, “Effective image retrieval based on hidden concept discovery in image database,” *IEEE Trans. Image Process.*, vol. 16, no. 2, pp. 562–572, 2007. 10

[44] T. Deselaers and V. Ferrari, “Global and efficient self-similarity for object classification and detection,” in *CVPR*, 2010, pp. 1633–1640. 10

[45] J. Goldberger, S. Gordon, and H. Greenspan, “Unsupervised image-set clustering using an information theoretic framework,” *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 449–458, 2006. 10

[46] M. Blaschko and C. Lampert, “Correlational spectral clustering,” in *CVPR*, 2008, pp. 1–8. 10

[47] W. Jiang, G. Er, Q. Dai, and J. Gu, “Similarity-based online feature selection in content-based image retrieval,” *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 702–712, 2006. 10

[48] P. Wang, J. Wang, G. Zeng, J. Feng, H. Zha, and S. Li, “Salient object detection for searched web images via global saliency,” in *CVPR*, 2012, pp. 3194–3201. 10

[49] Z. Tu and X. Bai, “Auto-context and its application to high-level vision tasks and 3d brain image segmentation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1744–1757, 2010. 11

[50] G. Kim and A. Torralba, “Unsupervised detection of regions of interest using iterative link analysis,” in *NIPS*, 2009, pp. 961–969. 11

[51] T. Deselaers, B. Alexe, and V. Ferrari, “Localizing objects while learning their appearance,” in *ECCV*, 2010, pp. 452–466. 11

[52] J. Zhu, J. Wu, Y. Wei, E. Chang, and Z. Tu, “Unsupervised object class discovery via saliency-guided multiple class learning,” in *CVPR*, 2012, pp. 3218–3225. 11

[53] M. Yang, J. Yuan, and Y. Wu, “Spatial selection for attentional visual tracking,” in *CVPR*, 2007, pp. 1–8. 11

[54] F. Ken, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, “Saliency-based video segmentation with graph cuts and sequentially updated priors,” in *IEEE International Conference on Multimedia and Expo*, 2009, pp. 638–641. 11, 12

[55] Y. Lee, J. Kim, and K. Grauman, “Key-segments for video object segmentation,” in *ICCV*, 2011, pp. 1995–2002. 11

[56] E. Rahtu, J. Kannala, M. Salo, and J. Heikkil, “Segmenting salient objects from images and videos,” in *ECCV*, 2010, pp. 366–379. 11

[57] M. Grundmann, V. Kwatra, H. Mei, and I. Essa, “Efficient hierarchical graph-based video segmentation,” in *CVPR*, 2010, pp. 2141–2148. 12

Huazhu Fu received the B.S. degree from Nankai University in 2006, and the M.E. degree in Electromechanical Engineering from Tianjin University of Technology in 2010 with the Best Thesis Award. He is currently a Ph.D. candidate with School of Computer Science and Technology at Tianjin University. His research interests include: weakly supervised leaning, multiple image correspondence, saliency detection and segmentation.

Xiaochun Cao is a professor at the Institute of Information Engineering, Chinese Academy of Sciences since 2012. He received the B.E. and M.E. degrees both in computer science from Beihang University (BUAA), China, and the Ph.D. degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the university-level Outstanding Dissertation Award. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a professor at Tianjin University. He has authored and coauthored over 50 journal and conference papers. In 2004 and 2010, Dr. Cao was the recipients of the Piero Zamperoni best student paper award at the International Conference on Pattern Recognition.

Zhuowen Tu is an assistant professor in the lab of neuro imaging (LONI), Department of Neurology, with a joint appointment in the Department of Computer Science, University of California Los Angeles. He is also affiliated with the Bioengineering IDP program and Bioinformatics IDP program at UCLA. He took leave of absence to work at Microsoft Research Asia from March of 2011 to December 2012. He received his PhD from the Ohio State University and his M.E. from Tsinghua University. Zhuowen Tu received NSF CAREER award in 2009 and David Marr prize in 2003.