Detecting Co-Salient Objects in Large Image Sets

Shuze Du and Shifeng Chen

***Abstract*—Co-salient object detection has attracted much more attention recently as it is useful for many problems in vision computing. However, most of existing methods emphasize detecting the common salient objects in a small group of images and the objects of interest in those images have clear borders with respect to the backgrounds. In this work, we propose a novel co-saliency detection method, which aims at discovering the common objects in a large and diverse image set composed of hundreds of images. First, we search a group of similar images for each image in the set. Our method is based on the overlapped groups. We handle each group with an unsupervised random forest to extract the rough contours of the common objects. Then a contrast-based measure is utilized to produce the saliency map for an individual image. For each image in the set, we collect all the maps from the groups that contain the image and fuse them together as the inter-saliency map for the image. The final co-saliency map is computed by combining the inter-saliency map with the single saliency map of this image. Experimental evaluation on an established large dataset demonstrates that our approach attains superior results and outperforms the state-of-the-art methods.**

***Index Terms*—Co-salient object, random forest.**

I. Introduction

SALIENCY [1] detection aims to identify the informative and interesting regions in an image for further processing, which is beneficial to various vision applications, like image resizing [2], video summary [3], video compression [4], [5], and image retrieval [6]. Numerous detecting approaches have emerged during the last two decades, but most of them focus on detecting salient objects in a single image [7], [8].

Co-saliency is a relatively underexplored area. Its goal is to detect the common salient objects in multiple images. Such capability is helpful to object co-segmentation [9] and similar image search [10]. There are a few works on detecting the co-salient objects in a small group of images. In [11], the authors first detected the salient objects in a single image, then made the objects that frequently appear in most images stand out with SIFT matching. Li and Ngan [12] proposed a method to discover co-salient objects from an image pair via a co-multilayer graph. It is hard to extend this method to process more than two images. Fu *et al.* proposed a cluster-based co-saliency model [10], which generated the co-saliency maps by fusing the single image saliency with correspondence-based inter-image saliency. The authors of [13] adopted a hierarchical segmentation model for co-saliency detection.

While the above-mentioned co-saliency detection approaches yield impressing results on a pair of images or a small group, it is difficult to generalize these approaches to the case of hundreds of images. By contrast, we are interested in large image sets of hundreds of images with great diversity. How to detect the common objects in this case is rarely explored so far. As the objects in those images exhibit various colors, textures and shapes, *etc*., previous methods have limited ability to handle the case even for a small group of such images, let alone a large set. In fact, it is not a good way to process a large set simultaneously. In Fu's model [10], it required high computational load to cluster millions of pixels. Moreover, producing reliable correspondences is almost prohibitive as the common patterns less likely happen in most of the images.

This paper develops a novel framework for co-saliency detection in large image sets. The underlying assumption is that the common patterns are easily to be detected in a small group. For each image in a set, inspired by [9], we extract its nearest neighbors from the set to form a group. Recently the random forest based approach [14] provided promising performance on single image saliency detection. It is expanded here to tackle a group of images. The images in a group are decomposed into uniform patches. We consider not only the rarity (uniqueness) of a patch on the single image, but also its commonality (repetitiveness) on the other images inside the group. An active contour model with group similarity [15] is adopted to extract the contours of the salient objects. In our model, the saliency of a patch is evaluated by its contrast to other patches in the image. For an image, the groups containing it will supply complementary information for co-saliency, thus we fuse the saliency maps from these groups as the inter-saliency of the image in the set. In this manner, the object can be detected more precisely while suppressing the background noise. Finally, the co-saliency for the image is estimated by combing the inter-saliency and the single saliency of this image. Fig. 1 gives an illustration of our method. In summary, our first contribution is to propose a group-based strategy to cope with co-salient object detection in large image sets. The second is using a random forest to compute the inter-saliency in each group.

II. The Proposed Method

A. Building Random Forest

Random forest was proposed by Breiman [16]. It often works in a supervised manner and is usually used for classification and regression. Instead, in our model, we build a random forest without supervision as the work [17].

Manuscript received May 04, 2014; revised July 27, 2014; accepted August 04, 2014. Date of publication August 20, 2014; date of current version August 30, 2014. This work was mainly conducted at Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Glenn Easley.

S. Du is with the Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu 610041, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: dushuze09@mails.ucas.ac.cn).

S. Chen is with the Shenzhen Key Laboratory for Computer Vision and Pattern Recognition, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: shifeng.chen@siat.ac.cn).

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/LSP.2014.2347333

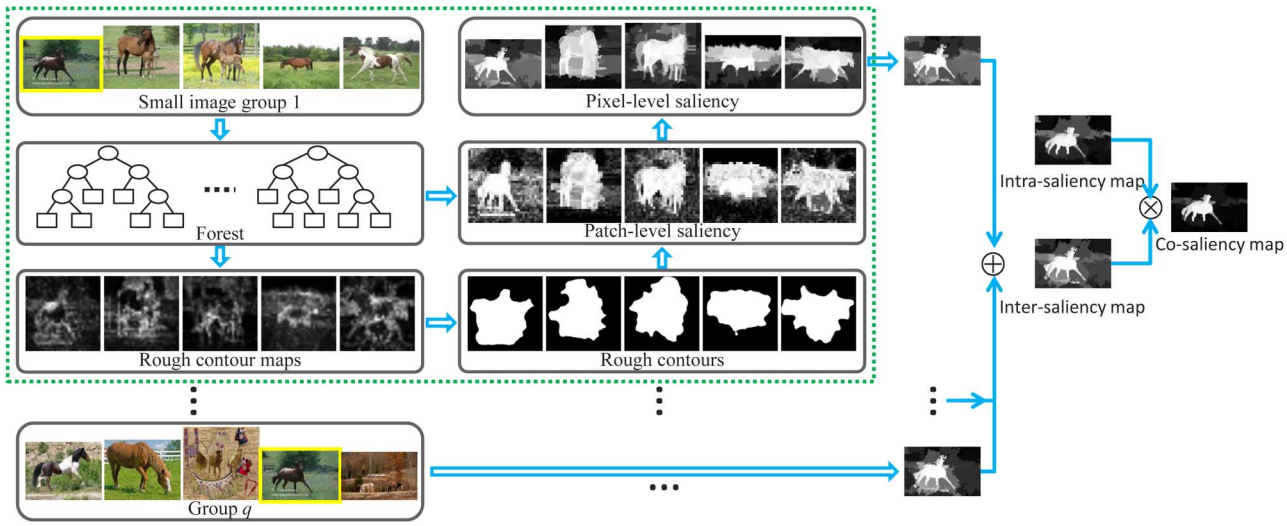

Fig. 1. Illustration of the main phases of our approach to detect co-salient objects in a large set of images. Given a set of images, for each image, a small image group is formed by searching the image's nearest neighbors. Then we perform co-saliency detection in each group. The inter-saliency for image I_j (in a yellow box) is computed by fusing its saliency maps from all groups containing it. We integrate the inter-saliency with the intra-saliency evaluated by a single saliency detection method as I_j's co-saliency in the set. Notice the contours are filled for better show. Best viewed in color.

Given a set of M images $\mathcal{S} = \{I_q\}_{q=1}^{M}$, for image I_q, we search its K nearest neighbors (including I_q itself) from $\mathcal{S}$ using Euclidean distance on Gist [18] descriptors to form a small image group G_q. The images are first scaled into 256×256. We use uniform image patches of size 8×8 as the basic elements. For each image in G_q, we extract all patches on an $m \times m$ grid over the image with $m = 32$. Each patch is represented by the pixel values in Lab and RGB color space, *i.e.*, the values are reorganized into a vector of length 384 (6×8^2). Let $U_q = \{u_{i,j}\}, i = 1, \ldots, N_I$ ($N_I = m^2$ is the number of patches in one image) and $j \in G_q$, be the patch set consisting of N_G patches ($N_G = K \times N_I$) extracted from group G_q, where $u_{i,j}$ is the ith patch extracted from image I_j.

A random forest F is consisted of T independent random trees. Each tree is grown from the complete patch set U_q. A random tree delivers a patch from the root of the tree to one of its leaf nodes. At each node, the associated patch set is divided into two subsets according to a test function. The two subsets form the left and right child nodes, respectively. In our model, the test function is defined as $h(u_{i,j}, k_1, k_2) = [u_{i,j}(k_1) - u_{i,j}(k_2) < \tau]$, where k_1 and k_2 are the two dimensions of $u_{i,j}$ randomly selected for splitting, and $[\cdot]$ is the indicator function. We term $u_{i,j}(k_1) - u_{i,j}(k_2)$ feature difference, then τ is the mean of feature differences of all patches contained in the node. We try 20 times at each node and choose the configuration with the largest variance on feature differences. The tree is recursively built until it reaches the maximum depth d or there is only one patch within a node.

B. Rough Contour Maps

In this section, we introduce how to compute the contour maps with the built random forest. Different from [14], which simply computes a patch's rarity (frequency of occurrences) in a single image, we jointly consider the patch's rarity in an image and commonality (repetitiveness) between different images in group G_q. In tree t, at the leaf node l_t that $u_{i,j}$ arrives, there is a patch subset V_{l_t} of the whole set U_q. Let $V_{l_t,j}$ denote the set of patches generated from image I_j, *i.e.*, $V_{l_t,j} = \{u_{i,j} | u_{i,j} \in V_{l_t}\}$. The commonality criterion for the patches in leaf l_t is based on the distribution of the patches. Following the idea of [19], we utilize the entropy to measure the distribution, which is defined as:

$$H(l_t) = -\sum_{j \in G_q} p(j) \log p(j), \tag{1}$$

where $p(j)$ is the proportion of patches extracted from image I_j, *i.e.*, $p(j) = |V_{l_t,j}|/|V_{l_t}|$. $H(l_t)$ is maximized when $p(j)$ is a uniform distribution and is 0 if all the patches come from a single image. The larger value of $H(l_t)$ indicates that the patches at this node are shared by more images in group G_q, otherwise the lower value indicates the majority of patches appearing in the minority of images.

Compared to the patches inside the objects or the backgrounds, the patches lying on the common contours usually have not only low rarity in its image, but also relatively high repetitiveness between the images in image group G_q. For patch $u_{i,j}$, the probability that it straddles the contour of the common object can be computed from:

$$S_{j,q}^{c}(u_{i,j}) = \left(\frac{1}{T} \sum_{t=1}^{T} \frac{\exp(\alpha \cdot H(l_t))}{|V_{l_t,j}|} \right) \cdot d_s(z_{i,j}, c_j), \tag{2}$$

where $\exp(\alpha \cdot H(l_t))$ is the *commonality* term, α weighs the entropy in order to balance the commonality and the *rarity* $|V_{l_t,j}|$ of patch $u_{i,j}$, and $d_s(z_{i,j}, c_j) = \exp(-(z_{i,j} - c_j)^2/\sigma^2)$ is the distance of patch $u_{i,j}$'s normalized location $z_{i,j}$ to the image center c_j ($\sigma^2 = 0.32$ in our implementation). In most cases, the entropy $H(l_t)$ is much smaller than the rarity item $|V_{l_t,j}|$. The constant α is empirically set to 2 during our tests. The contour map $S_{j,q}^{c}$ is normalized to the range [0, 1].

C. Contour Detection

Given the rough contour maps $S_{j,q}^{c}$, our goal is to find a group of contours corresponding to the object boundaries in the images. To accelerate contour detection for an image group, we adopt an active contour model from Zhou *et al.* [15], which took into account the group similarity of the contours, instead of detecting the contours individually. The similarity is measured by

the rank of the matrix composed of multiple contours and used as a constraint on existing active models.

Denote a parameterized curve (contour) in the (x, y) plane as a vector of n points $C = [x_1, \ldots, x_n, y_1, \ldots, y_n]^\top \in \mathbb{R}^{2n}$. In our case, there are K curves $C_1, C_2, \ldots, C_K$ corresponding to the images in group G_q, respectively. The work shows that the vectors of these contours would form a low-rank matrix, which allows the global change of contours. The model is formally described by

$$\min_{\mathbf{X}} \sum_{i=1}^{K} f_i(C_i), \text{ subject to } \operatorname{rank}(\mathbf{X}) \leq Q, \tag{3}$$

where $\mathbf{X} = [C_1, \ldots, C_K]$ and Q is a given constant. $f_i(C_i)$ is the region-based energy defined as $f_i(C_i) = \int_{R_1} (I_i(x,y) - u_1)^2 \mathrm{d}x\mathrm{d}y + \int_{R_2} (I_i(x,y) - u_2)^2 \mathrm{d}x\mathrm{d}y + \beta \cdot \mathrm{Length}(C_i)$, here, R_1 and R_2 denote the regions inside and outside the contour, and u_1 and u_2 are the average intensity in R_1 and R_2, respectively. Refer to [15] for how to solve this objective function. A few examples of rough contours are shown in Fig. 1.

D. Saliency Estimation in a Group

We now describe our approach to estimate saliency maps for images in group G_q. Some existing methods [10], [11] exploited the correspondences to measure the co-saliency. The images in our case, however, have diverse colors, textures, and so forth, making it hard to establish reliable correspondences. Therefore, we aim to detect the object in the common contour for a single image. The contour estimated above is only a rough estimation of the object contour. As in [14], we compute a patch's saliency by estimating how likely the patch belongs to the object. Since the patch inside the object often has a large contrast with the patches from the background, we consider this likelihood from the view of contrasts between the inside patches and the exterior patches. If the proportion of pixels of a patch contained inside the contour exceeds a threshold γ (here, γ is set to 0.9), we regard it as the *inside* patch, otherwise the *outside* patch. In this way, we collect two patch sets, the inside Ω_{in} and the outside Ω_{out}, respectively.

If two patches $u_{i,j}$ and $u_{k,j}$ reach the same leaf node l_t in tree t, then the similarity $\pi_t(u_{i,j}, u_{k,j})$ between $u_{i,j}$ and $u_{k,j}$ equals $1/|V_{l_t,j}|$, and 0 otherwise. Their similarity in the whole forest is computed by $\pi(u_{i,j}, u_{k,j}) = \sum_{t \in F} \pi_t(u_{i,j}, u_{k,j})$. Then the possibility that the inside patch $u_{i,j} \in \Omega_{in}$ belongs to the object is defined as

$$S^o_{j,q}(u_{i,j}) = \exp\left(-\frac{1}{|\Omega_{out}|} \sum_{u_{k,j} \in \Omega_{out}} \pi(u_{i,j}, u_{k,j}) \right), \tag{4}$$

where the values of $S^o_{j,q}(u_{i,j})$ are normalized to the range [0, 1]. If $u_{i,j}$ is more similar to the outside patches, it will be assigned a lower saliency value. The probability of the outside patch $u_{k,j} \in \Omega_{out}$ being a part of the object is evaluated from

$$S^o_{j,q}(u_{k,j}) = \exp\left(\frac{1}{|\Omega_{in}|} \sum_{u_{i,j} \in \Omega_{in}} \pi(u_{k,j}, u_{i,j}) \right). \tag{5}$$

Patch $u_{k,j}$'s saliency will be high when it is similar to the inside patches. Also, we scale the values of $S^o_{j,q}(u_{k,j})$ to [0, 1]. Likewise, the other images in group G_q can be processed.

E. Co-Saliency Maps

After computing the saliency maps $S^o_{j,q}$ in all the groups, the maps for image I_j are then fused as its inter-saliency in image set $\mathcal{S}$. Specially, suppose image I_j appears in groups $G_j, G_{q_1}, \ldots, G_{q_a}$, then the inter-saliency for patch $u_{i,j}$ is obtained by:

$$S^{inter}_j(u_{i,j}) = S^o_{j,j}(u_{i,j}) + S^o_{j,q_1}(u_{i,j}) + \cdots + S^o_{j,q_a}(u_{i,j}). \tag{6}$$

We scale map S^{inter}_j to its original size. To recover some of the object boundary information, similar to [6], we segment image I_j using a graph-based segmentation method [20]. We average the saliency values in a segment region as the saliency for that region, thus producing the saliency map $\hat{S}^{inter}_j$ for image I_j.

Finally, we also consider the saliency cues (intra-saliency) computed from image I_j alone. We employ the model of [14] to evaluate the intra-saliency map S^{intra}_j of image I_j, which is also based on the random forest. By fusing the intra-saliency and inter-saliency cues for image I_j, the co-saliency value at pixel p_x of image I_j can be evaluated by:

$$S^{co}_j(p_x) = \hat{S}^{inter}_j(p_x) \cdot S^{intra}_j(p_x). \tag{7}$$

This fusion scheme is capable of suppressing the background effectively, hence making the saliency map more precise (see the right half of Fig. 1).

III. Experimental Results

We evaluate our algorithm on the Internet dataset collected by Rubinstein *et al.* [9], in which the images are downloaded by using the Bing search engine. The dataset is composed of 3 large sets of images: car, horse, and airplane. For each set, we remove the images that do not contain the object of interest at all. In the end, there are 1, 208 car, 810 horse, and 470 airplane images in the corresponding sets.

In all experiments, we use $K = 5$ (K is the number of images in each group), and build 20 trees ($T = 20$) for each group with maximum depth $d = 10$. As in [7], [8], we evaluate our method by using the precision recall curve. To be specific, we binarize the saliency maps at a fixed threshold and compare the obtained binary maps with the ground truth masks to compute the mean precision and recall. By sliding the threshold from 0 to 255, a precision recall curve is then plotted.

We compare the proposed random forest based co-saliency (RFCS) with five state-of-the-art salient object detection methods designed for a single image, including RC [7], RARE2012 [21], BMS [8], HS [22] and RFSR [14]. Our method is further compared with Fu's co-saliency model (Fu_CS) proposed in [10]. We extend this model to the large image sets with our strategies as described in Section II, that is, we use the model to process each group of images and then fuse the corresponding maps for an image from all the groups as its co-saliency map.

Results on the car, horse and airplane set are shown in Figs. 2(a), 2(b) and 2(c), respectively. It can be seen that our co-saliency outperforms all the other methods consistently. Furthermore, our inter-saliency (RF_inter) is better than the single image based approaches, showing the benefit of considering multiple groups of images when detecting the salient object in an image. When combining the inter-saliency with the intra-saliency computed from the random forest based model

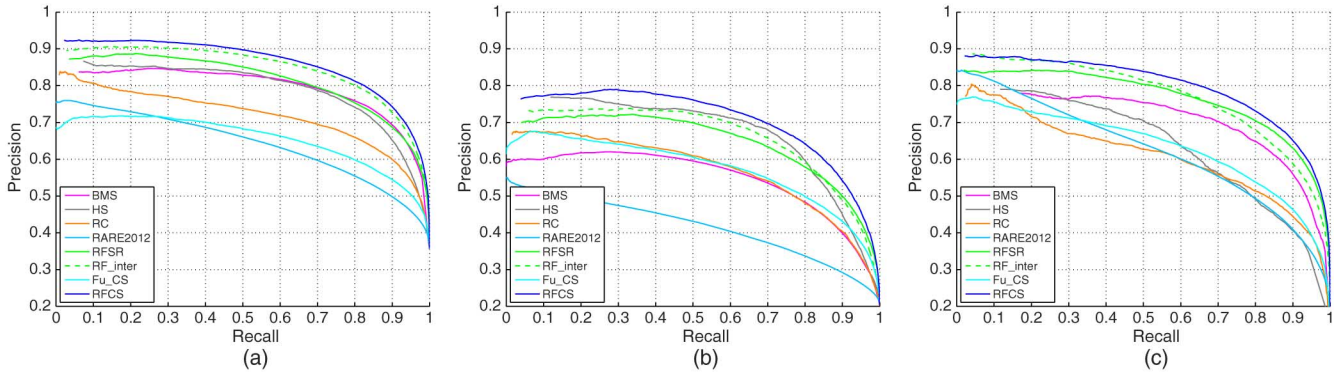

Fig. 2. Precision-recall curves for different methods. (a) Results on the car set. (b) Results on the horse set. (c) Results on the airplane set.

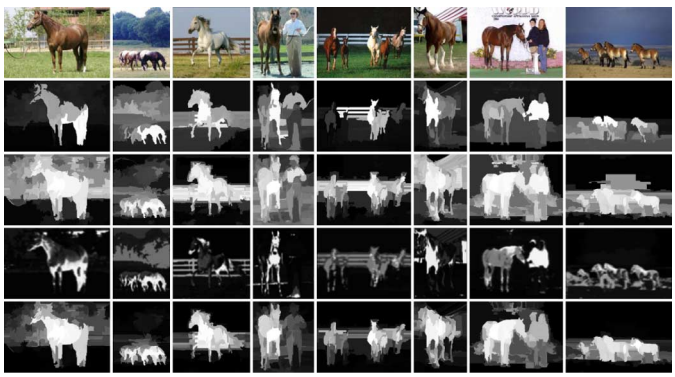

Fig. 3. Qualitative comparisons of different methods on some horse images (Row 1). Rows 2-5: maps generated by HS, RFSR, Fu_CS, and our model (RFCS), respectively. Best viewed in color.

(RFSR [14]), the performance of our model is further improved. This means that our method highlights the co-salient objects while suppressing the background, manifesting the importance of the combination for co-saliency detection. In addition, we observe that Fu's model does not perform well on the three sets, largely because the pixels inside the objects from multiple images are difficult to be clustered together. A visualization of some representative saliency maps out of the horse set is provided in Fig. 3. Note that our model can depress the humans in the images, and they will be considered as salient objects in the single image approaches. Our method is more robust to the complex textures, whereas the single image methods will label them as salient regions. Fu's model suppresses the background effectively at the expense of the salient objects.

IV. Conclusions

This paper has proposed a novel random forest based framework of co-saliency detection, which detects the common salient objects from a large image set with great diversity. We divide the set into some small groups and detect the co-salient objects in each group. The inter-saliency map for an image is generated by fusing all its corresponding saliency maps in these groups, which is integrated with the intra-saliency of the image to produce the co-saliency map. Our method achieves excellent performance on three large benchmark sets. In the future, we intend to pursue the benefits of the proposed method in image retrieval.

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, Nov. 1998.
[2] R. Achanta and S. Susstrunk, "Saliency detection for content-aware image resizing," in *Proc. IEEE ICIP*, 2009, pp. 1005–1008.
[3] M. Decombas, F. Dufaux, and B. Pesquet-Popescu, "Spatio-temporal grouping with constraint for seam carving in video summary application," in *IEEE DSP*, 2013, pp. 1–8.
[4] M. Decombas, F. Dufaux, E. Renan, B. Pequet-Popesu, and F. Capman, "Improved seam carving for semantic video coding," in *Proc. IEEE MMSP*, 2012, pp. 53–58.
[5] M. Mancas, D. D. Beul, N. Riche, and X. Siebert, "Human attention modelization and data reduction," in *Video Compression*. Rijeka, Croatia: Intech, 2012.
[6] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *Proc. CVPR*, 2013, pp. 3238–3245.
[7] M.-M. Cheng, G.-X. Zhang, N. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. CVPR*, 2011, pp. 409–416.
[8] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proc. ICCV*, 2013, pp. 153–156.
[9] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu, "Unsupervised joint object discovery and segmentation in internet images," in *Proc. CVPR*, 2013, pp. 1939–1946.
[10] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, Oct. 2013.
[11] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. CVPR*, 2011, pp. 2129–2136.
[12] 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.
[13] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 88–92, Jan. 2014.
[14] S. Du and S. Chen, "Salient object detection via random forest," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 51–54, Jan. 2014.
[15] X. Zhou, X. Huang, J. S. Duncan, and W. Yu, "Active contours with group similarity," in *Proc. CVPR*, 2013, pp. 2969–2976.
[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[17] G. Yu, J. Yuan, and Z. Liu, "Unsupervised random forest indexing for fast action search," in *Proc. CVPR*, 2011, pp. 865–872.
[18] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.
[19] S. O'Hara and B. A. Draper, "Scalable action recognition with a subspace forest," in *Proc. CVPR*, 2012, pp. 1210–1217.
[20] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, 2004.
[21] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Process.: Image Commun.*, vol. 28, no. 6, pp. 642–658, 2013.
[22] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. CVPR*, 2013, pp. 1155–1162.