Spatiotemporal Saliency Detection and Its Applications in Static and Dynamic Scenes

Wonjun Kim, *Student Member, IEEE,* Chanho Jung, *Student Member, IEEE,* and Changick Kim, *Senior Member, IEEE*

***Abstract*—This paper presents a novel method for detecting salient regions in both images and videos based on a discriminant center-surround hypothesis that the salient region stands out from its surroundings. To this end, our spatiotemporal approach combines the spatial saliency by computing distances between ordinal signatures of edge and color orientations obtained from the center and the surrounding regions and the temporal saliency by simply computing the sum of absolute difference between temporal gradients of the center and the surrounding regions. Our proposed method is computationally efficient, reliable, and simple to implement and thus it can be easily extended to various applications such as image retargeting and moving object extraction. The proposed method has been extensively tested and the results show that the proposed scheme is effective in detecting saliency compared to various state-of-the-art methods.**

***Index Terms*—Discriminant center-surround hypothesis, ordinal signature, salient regions, visual attention.**

I. Introduction

THE HUMAN visual system has a remarkable ability to quickly grasp salient regions in static and dynamic scenes without training. Therefore, human can easily understand scenes based on this selective visual attention. In the field of computer vision, many models have been proposed to accomplish this task automatically over the last few decades. First of all, to find regions coherent with visual attention, high level information such as the sky, faces, and humans has been employed as a useful indicator in the literature [1]–[3]. The drawback of the use of high level information is hard generalization since it is not available in every image. To solve this problem, various bottom-up methods driven by low level features have been proposed, referred to as *saliency detection* [4]–[6]. Most of them use a set of basic visual features such as color, texture, and frequency components to build a saliency map, which suppresses unnoticeable regions and emphasizes salient regions, given as a density map as shown in Fig. 1. Note that pixels with higher intensity values denote that corresponding pixels are visually important. This saliency map can be used for a wide range of applications such as object detection [7], image and video summarization [8], image retargeting [9], [10], video surveillance [11], and so on. For example, various multimedia contents (i.e., images and videos) consumed on hand-held devices should be appropriately adapted to small displays for raising viewing experience. This can be achieved by using image retargeting based on extracting salient regions.

Although previous approaches for detecting such salient regions are very diverse, most of them fail to minimize false positives which occur in highly textured background areas. For example, cluttered leaves and grass regions in the backgrounds are highly likely to be detected as salient areas (see Fig. 1). Moreover, only a few methods have been proposed for saliency detection in videos [15], [16]. They combine motion features into a saliency model, but still suffer from high computational cost.

In this paper, to overcome the problems mentioned above, we propose a novel unified framework for detecting salient regions in both images and videos. Our approach, which is extended from our previous work [17], provides an enhanced saliency map based on the center-surround framework in the spatial and temporal domain, which can be described as follows. Given an image or a video, we are interested in finding salient regions in real time. To this end, we first propose a simple and robust importance measure, so-called self-ordinal resemblance measure (SORM) defined by computing distances between ordinal signatures of edge and color orientation histograms (E&COH) obtained from the center and its surrounding regions in the spatial domain. It should be emphasized that ordinal signatures of local feature distributions are used for measuring the spatial saliency rather than directly using sample values of local feature distributions employed in the most previous methods [18], [19]. It is because ordinal measures can tolerate variations of local feature distributions due to quantization and thus effectively suppress false positives occurring in complex and cluttered backgrounds. Then the temporal saliency, by simply computing the sum of absolute difference (SAD) between temporal gradients of the center and the surrounding regions, is combined with the spatial saliency to generate our spatiotemporal saliency map.

The main contributions of the proposed method can be summarized as follows.

1) We define the edge and color orientation histogram as our features, along with temporal gradients to generate the spatiotemporal saliency map.

Manuscript received July 6, 2010; revised September 2, 2010; accepted October 14, 2010. Date of publication March 10, 2011; date of current version April 1, 2011. This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency), under Grant NIPA-2011-(C1090-1111-0003). This paper was recommended by Associate Editor S. Pankanti.

The authors are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea (e-mail: jazznova@kaist.ac.kr; peterjung@kaist.ac.kr; cikim@ee.kaist.ac.kr).

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/TCSVT.2011.2125450

2) We propose to use ordinal signatures of our features for measuring the spatial saliency. It has been shown that ordinal measures are robust to various modifications of original signals [20], [21] and thus they can be employed to cope with variations of local feature distributions due to quantization.
3) We provide a simple, but powerful unified framework for detecting salient regions in both images and videos. Thus, our method can be easily extended to various applications dealing with static and dynamic scenes.
4) The proposed scheme is fast enough to meet real-time requirements.

The rest of this paper is organized as follows. Previous methods are briefly summarized in Section II. The technical details about the steps outlined above are explained in Section III. Various images and videos are tested to justify the efficiency and robustness of our proposed method in Section IV, and its applications in static and dynamic scenes are introduced in Section V. Our conclusion follows in Section VI.

II. Backgrounds

In the human visual system (HVS), the brain and the vision systems work together to identify the relevant regions, which are indeed salient regions in the image. In this sense, Itti *et al*. [4] proposed a biologically inspired saliency model. Their saliency map is generated based on the linear combination of normalized feature maps obtained from three basic components: intensity, color, and orientation. Although this model has been shown to be successful in predicting human fixations and useful in object detection, it is criticized since the objective function designed to be optimized is not specified and many parameters need to be tuned by users. Liu and Gleicher [5] proposed a region enhanced saliency detection. Although they provided the meaningful region information based on the scale-invariant saliency, image segmentation needs to be incorporated, which is time-consuming. Oliva *et al*. [6] proposed a model of attention guidance based on global scene configuration. They employed a bottom-up approach using a Bayesian framework to determine where the specific objects should be located. Similarly, Zhang *et al*. [13] also proposed a bottom-up scheme using natural statistics incorporating top-down information to estimate the probability of a target at each pixel position. They used local features derived from independent component analysis. However, such methods using filter responses require many parameters to be estimated. More notably, the authors of [14] generated a saliency map by using a spectral residual (SR). In this approach, the spectral residual is simply defined based on the difference between the log spectrum of the given image and the averaged log spectrum of sample natural images. It performs quite well across diverse images. However, the role of spectral residual for saliency detection is unclear. Moreover, the authors of [22] pointed out that it is the phase spectrum, not the amplitude spectrum, of the Fourier transform of the image that finds the location of salient areas. Recently, spatial relations between neighbor pixels have drawn much attention to detect saliency. This is because only a local area in the given image can be perceived quickly by human observers with the high resolution due to the foveation characteristics of the HVS [23]. Therefore, many studies have tried to model the visual saliency based on the relation between local features obtained from neighboring pixels. Gao *et al*. [18] defined a discriminant center-surround saliency based on the idea that local image features are stimuli of interest when they are distinguishable from the background. They used the difference of Gaussian and Gabor filters as features. In [24], the authors proposed nonparametric saliency detection by using the local steering kernel (LSK) as a feature. Although the shape information is well defined in the LSK features, it is not suitable for real-time applications due to the high dimensionality.

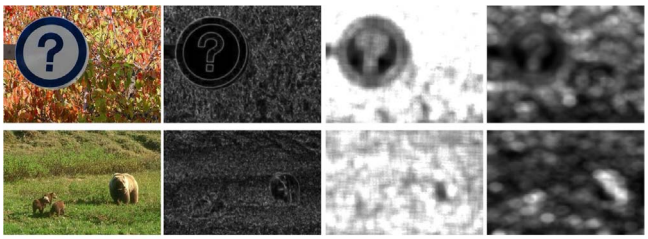

Fig. 1. Examples of saliency map. (a) Original image. (b) Contrast-based method [12]. (c) Saliency using natural statistics (SUN) [13]. (d) Spectral residual method [14]. Note that highly textured backgrounds are hard to suppress in traditional methods.

For detecting saliency in dynamic scenes, Guo and Zhang [22] proposed a spatiotemporal saliency model called phase spectrum of quaternion Fourier transform (PQFT). They applied the Fourier transform to a quaternion image composed of intensity, color, and motion features for taking the phase spectrum, which is the key to compute the location of salient areas. Although this method is simple and fast, it still suffers from highly textured backgrounds. Wolf *et al*. [25] generated the visual saliency map by using the motion detection based on the fact that moving objects in video draw most of the viewers' attention and are content-wise important. Mahadevan and Vasconcelos [26] adopted the dynamic textures to conduct center-surround computations. They provided useful results in highly variable backgrounds, but the algorithm is not deployable in real-time without further investigation due to its high computational complexity.

Unlike existing methods based on the center-surround framework, we propose a novel solution for detecting salient regions by using SORM instead of directly using sample values of local feature distributions, to be robust to variations of local feature distributions. Furthermore, we combine temporal gradients with our SORM for generating the spatiotemporal saliency map. We will provide further details about the proposed algorithm in the following subsections.

III. Proposed Method

A. Feature Representation

In this subsection, we introduce our features defined in a local region for the center-surround computations. For the spatial saliency, we define the edge and color orientation histogram (E&COH), which is nonparametric and represents exceedingly the underlying local structure of the image data.

The key idea behind E&COH is to incorporate the magnitude and orientation of edge and color in a 1-D histogram. The procedure to form each histogram for edges and colors is briefly summarized as follows. First of all, edge orientations are computed over all pixels belonging to the local region centered at the ith pixel and then quantized into K levels in the range of [0°, 180°]. Note that the sign of edge orientations is neglected due to the symmetric characteristic. Then, the histogram is generated by accumulating the edge magnitude for its orientation over the pixels within each local region. The outcome histogram is called the edge orientation histogram (EOH) and this EOH is formulated as follows:

$$\mathbf{F}_i^E(n) = (f_{i,0}^E(n),\ f_{i,1}^E(n),\ f_{i,2}^E(n),\ \cdots,\ f_{i,K}^E(n))$$
$$\text{where}\quad f_{i,q}^E(n) = \sum_{\substack{(x,\,y)\,\in\,B_i \\ \theta(x,\,y,\,n)\,\in\,q}} m(x,\,y,\,n). \tag{1}$$

Here $m(x, y, n)$ and $\theta(x, y, n)$ denote the edge magnitude and the quantized orientation at the pixel position (x, y) of the nth frame, respectively. B_i is a set of neighboring pixels centered at the ith pixel.

Next, we define the color orientation histogram (COH) using the HSV color model in a similar way with EOH generation. Since the hue value represents the different color tone according to the angle ranged in the HSV color space, it can be quantized and thus used as the indices of the COH bin. Also, the color magnitude can be represented by the saturation value. That is, purer colors have larger magnitudes for the corresponding color tone in the COH. In detail, the COH is generated by accumulating the saturation value to the corresponding hue value. Unlike the edge orientation, since the color tones are different through the range of 0°–360°, the entire range needs to be quantized into H levels. The COH can be defined as follows:

$$\mathbf{F}_i^C(n) = (f_{i,0}^C(n),\ f_{i,1}^C(n),\ f_{i,2}^C(n),\ \cdots,\ f_{i,H}^C(n))$$
$$\text{where}\quad f_{i,q}^C(n) = \sum_{\substack{(x,\,y)\,\in\,B_i \\ v(x,\,y,\,n)\,\in\,q}} s(x,\,y,\,n). \tag{2}$$

Similar to (1), $s(x, y, n)$ and $v(x, y, n)$ denote the saturation value and the quantized hue value at the pixel position (x, y) of the nth frame, respectively. It is important to note that we use one additional bin with $q = 0$ for both orientation histograms, respectively, to handle pixels having zero edge magnitudes and no color attributes (i.e., acromatic colors) separately.

For the temporal saliency, the temporal gradients are computed by using the intensity differencing between frames as follows:

$$\mathbf{F}_i^T(n) = (f_{i,1}^T(n),\ f_{i,2}^T(n),\ \cdots,\ f_{i,P}^T(n))$$
$$\text{where}\quad f_{i,j}^T(n) = |I_{i,j}(n) - I_{i,j}(n-\tau)|\ . \tag{3}$$

Here P denotes the total number of pixels in the local region generated by the ith pixel and thus j indicates the index of those pixels. τ is the user-defined latency coefficient and it is set to 3 in our implementation.

In summary, we have built three feature vectors for the input image: edge orientation histogram, color orientation histogram, and temporal gradients. Finally, our features defined at the ith pixel of the nth frame, $\mathbf{F}_i(n) = [\mathbf{F}_i^E(n), \mathbf{F}_i^C(n), \mathbf{F}_i^T(n)]$, are fed into the center-surround framework for generating the spatiotemporal map (see Fig. 2). Note that N denotes the total number of local regions in a center-surround window $\mathbf{A}$, which is represented by the dotted rectangle in Fig. 2. Further details will be explained in the following subsections.

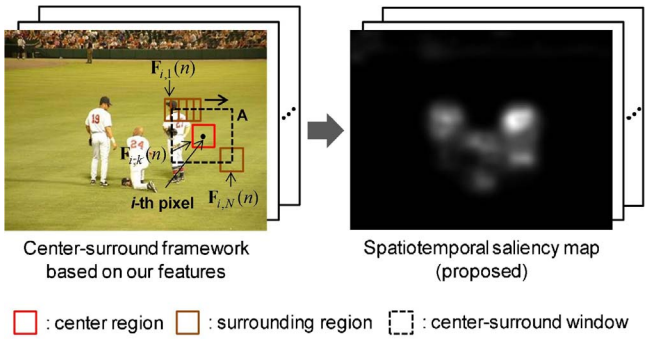

Fig. 2. Overall procedure of the proposed method.

B. Spatial Saliency by Self-Ordinal Resemblance Measure

Most traditional methods using the center-surround framework employ various similarity measures such as Kullback–Leibler divergence [26], Bhattacharyya distance [27], and Euclidean distance [28] for finding visual saliency in static images. Basically, they collect various image features in local regions and then build an underlying feature distribution relevant to visual importance. However, most of them are apt to fail in tolerating variations of local feature distributions especially near high contrast edges due to quantization, yielding drastic changes in local image statistics even though color variations are visually neglectable [29] (see Fig. 3). To solve this problem, we propose to use the ordinal measure of local image features obtained from the center and the surrounding regions for the spatial saliency. Note that it is known to be robust to various signal modifications as mentioned. For example, two EOHs obtained from the center and its close surrounding regions, which are positioned near high contrast edges, are shown in Fig. 3. Although two local regions have very similar structural information, sample values of EOHs are quite different due to quantization whereas the ordinal signatures obtained from corresponding EOHs are equivalent to each other, which is desirable.

Let us define the $1 \times (K + 1)$ rank matrix by ordering the elements of the EOH defined in (1) at the ith pixel of the nth frame as $R(\mathbf{F}_i^E(n))$, where K denotes the quantization levels as mentioned. For example, the rank matrices of two EOHs in Fig. 3 can be equally represented as $\mathbf{R}(\mathbf{F}_i^E(n)) = [5, 2, 1, 3, 4, 6]$ and also each element of the rank matrix can be expressed as $\mathbf{R}^0(\mathbf{F}_i^E(n)) = 5, \cdots, \mathbf{R}^5(\mathbf{F}_i^E(n)) = 6$. Then the distance from the rank matrix of the center region to the rank matrices of surrounding regions is computed as follows:

$$sM_i^E(n) = \frac{1}{Z^E}\sum_{l=1}^{N} d_l^E$$
$$\text{where}\quad d_l^E = \sum_{q=0}^{K} |\mathbf{R}^q(\mathbf{F}_l^E(n)) - \mathbf{R}^q(\mathbf{F}_i^E(n))| \tag{4}$$

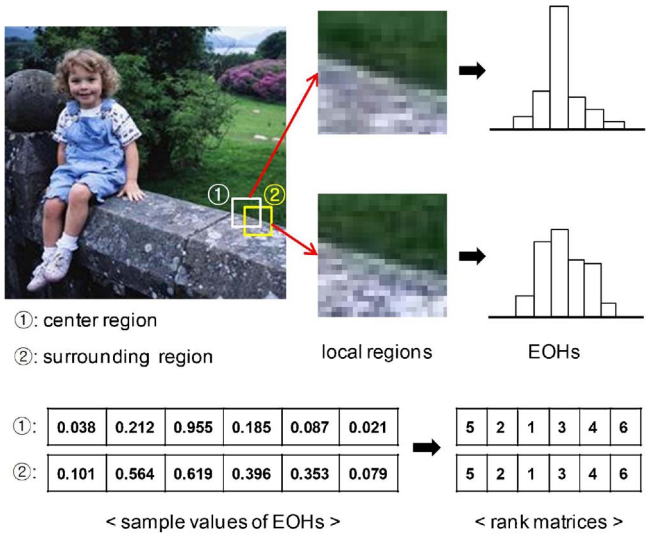

Fig. 3. Two different EOHs obtained from the center region (①) and its close surrounding region (②) (near high contrast edges). Note that sample values of EOHs are normalized for better numerical representation. Even though two local regions contain very similar structural information, sample values are quite different due to quantization whereas their ordinal signatures are equivalent to each other.

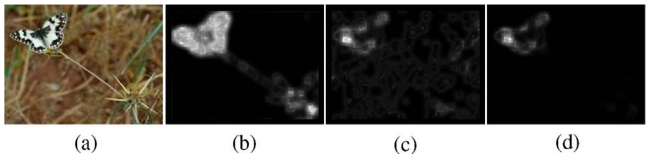

Fig. 4. (a) Original image. (b) Edge saliency. (c) Color saliency. (d) Spatial saliency by the proposed method.

Here, N and Z^E denote the number of local regions in a center-surround window as mentioned and the maximum distance between two rank matrices. Similarly, saliency based on COH can also be defined as

$$sM_i^C(n) = \frac{1}{Z^C}\sum_{l=1}^{N} d_l^C \quad \text{where} \quad d_l^C = \sum_{q=0}^{H} |\mathbf{R}^q(\mathbf{F}_l^C(n)) - \mathbf{R}^q(\mathbf{F}_i^C(n))| \tag{5}$$

Finally, the spatial saliency at the ith pixel of the nth frame is computed by the combination of edge and color saliency as follows:

$$sM_i^S(n) = sM_i^E(n) \times sM_i^C(n). \tag{6}$$

It will have a maximum value on locations where both edge and color saliency agree and also vary proportionately in regions that favor only one of edge and color saliency. The example of our spatial saliency is shown in Fig. 4. We can see that non-salient regions are effectively suppressed.

C. Combining with Temporal Saliency of Videos

In spatiotemporal saliency detection, we believe that the temporal trail of the intensity distribution should be combined with the spatial saliency for dealing with video sequences efficiently since the video is a process evolving with time. In this sense, distinct from most of previous methods considering only static images for detecting saliency, we adopt temporal gradients defined in (3) to detect the temporal saliency based on the center-surround framework as follows:

$$sM_i^T(n) = \frac{1}{Z^T}\sum_{l=1}^{N} |\mathbf{F}_l^T(n) - \mathbf{F}_i^T(n)| \tag{7}$$

where Z^T denotes the normalization factor. Thus, the temporal saliency can be simply computed by using SAD between temporal gradients of the center and the surrounding regions. This added motion channel allows our method to represent the spatiotemporal saliency in order to select regions relevant to visual attention in videos as well as images. It is worth noting that we quickly obtain the motion information in video sequences by modeling the distributions of temporal gradients from the center and the surrounding regions in an implicit manner, as compared to the explicit motion estimation mostly involving the computation of optical flow whose estimation is difficult and time consuming.

Finally, the proposed methodology for measuring the spatiotemporal saliency relies on the combination of the spatial and temporal saliencies, $sM_i^S(n)$ and $sM_i^T(n)$ as follows:

$$sM_i^{ST}(n) = \alpha \cdot sM_i^S(n) + (1-\alpha)\cdot sM_i^T(n) \tag{8}$$

where $\alpha \in [0, 1]$ denotes the weighting factor for balancing between the spatial and temporal saliency. For example, a low value of α will focus on more on the temporal saliency while $\alpha = 0.5$ provides the balanced saliencies. In general view, since moving objects are more attractable than static objects and backgrounds in video sequences [25], we set α to 0.3. Note that the spatiotemporal saliency values defined in (8) are scaled from 0 to 255 for the grey-scale representation. For better visual effects, we smoothed the saliency map by using a Gaussian filter as in [14]. Our spatiotemporal map can also deal with static images easily by setting α to one.

D. Scale-Invariant Saliency Map

Even though there exist diverse visual resolutions related to salient regions, HVS automatically selects the appropriate resolutions by computing the view distance between the observer and the scene. Similar to HVS, our proposed scheme can be operated under various visual resolutions. To model HVS process in selecting visual resolutions, the proposed method is applied to input images with various sizes ranging from 32×32 to 128×128 pixels. Note that a smaller size of saliency map represents a coarser resolution. Examples of the saliency maps with various resolutions are shown in Fig. 5. When the size of the saliency map is 32×32 pixels, the global features are emphasized whereas detailed local features are suppressed. In other words, the proposed method performs as an observer looks at the scene from a long distance. In contrast to that, the local features are well retained in the scene at a finer resolution (i.e., the size of the spatial saliency map is 128×128 pixels). More specifically, a group of people is regarded as one object at a coarser resolution whereas each person is successfully separated at a finer resolution in Fig. 5. While, based on this

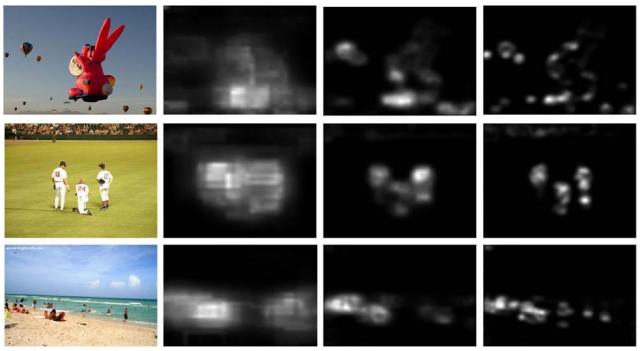

Fig. 5. Examples of the saliency maps computed by the proposed method with different resolutions. (a) Input image. (b) 32×32 pixels. (c) 64×64 pixels. (d) 128×128 pixels. The scale of salient objects decreases from top to bottom. Note that the saliency maps are resized to the size of input images (400 × 300 pixels).

observation, the hierarchical selectivity framework for saliency detection has been proposed [22], [30], it is very difficult to determine the optimal hierarchical level due to the diversity of salient regions in an image. Moreover, our concern is not what the detailed fixation process is as in [22] and [30] but how to effectively suppress non-salient regions and build a reliable saliency map. Therefore, we construct a scale-invariant saliency map through a multi-scale analysis. To do this, we combine saliency maps obtained from three different scales mentioned above (i.e., from 32 × 32 to 128 × 128 pixels) as follows:

$$s\tilde{M}_i^{ST}(n) = \sum_{l=1}^{3} \omega_l \cdot sM_{i,l}^{ST}(n), \qquad \sum_{l=1}^{3} \omega_l = 1 \tag{9}$$

where l denotes the scale index (i.e., 32 × 32 → 128 × 128 maps to 1 → 3) and thus $sM_{i,l}^{ST}(n)$ is a saliency map generated at the scale l, rescaled to the size of input image. Note that we assign the larger weight for the finer resolution. This is based on the assumption that humans look at the objects at the finer resolution in a very careful manner compared to the coarser resolution. In our implementation, we set the weight values as $\omega_1 = 0.1$, $\omega_2 = 0.3$, and $\omega_3 = 0.6$, respectively. Fig. 6 illustrates our scale-invariant saliency maps obtained from images in Fig. 5(a). Note that the proposed method is fully automatic and involves none of selecting the optimal scale (i.e., the single and fixed scale) and manual interactions.

For the sake of completeness, the overall procedure of the proposed method is summarized in Algorithm 1.

Algorithm 1 A novel unified framework for detecting salient regions in both images and videos based on center-surround framework

Step 1: Compute $\mathbf{F}_i^E(n)$, $\mathbf{F}_i^C(n)$, and $\mathbf{F}_i^T(n)$.

Step 2: Compute the spatiotemporal saliency at each scale as follows.

$$sM_i^{ST}(n) = \alpha \cdot sM_i^S(n) + (1-\alpha) \cdot sM_i^T(n)$$

Step 3: Compute a scale-invariant saliency map as follows.

$$s\widetilde{M}_i^{ST}(n) = \sum_{l=1}^{3} \omega_l \cdot sM_{i,l}^{ST}(n)$$

Step 4: Resizing saliency map to the same size of input image.

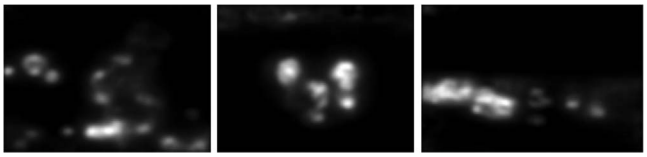

Fig. 6. Examples of the scale-invariant saliency maps for images shown in Fig. 5(a).

IV. Experimental Results

A. Performance Evaluation in Static Images

In this subsection, we show several experimental results on detecting saliency in various static images. The experiments are conducted on a database composed of about 5000 images provided by Microsoft Research Asia, Beijing, China (MSRA database) [31]. Images in MSRA database are taken in indoor and outdoor environments and the image resolution is 400 × 300 pixels for horizontally oriented images and 300 × 400 pixels for vertically oriented images, respectively. These images contain various salient objects, which include face, human, car, animal, sign, and so on. In our implementation, the EOH and COH are computed in a local region whose size is 5 × 5 pixels with nine and seven bins including an additional bin (i.e., $K = 8$ and $H = 6$), respectively, and thus the maximum distance between two rank matrices are $Z^E = 40$ and $Z^C = 24$. For the center-surround computation, the size of the center-surround window defined in Fig. 2 is set to 7 × 7 pixels.

To confirm the superiority of our proposed method, we compare our approach with the state-of-the-art methods, which are saliency tool box (STB) [32], saliency using natural statistics (SUN) [13], contrast-based method (CB) [12], and SR approach [14]. Results of saliency detection are shown in Figs. 7 and 8. More specifically, region information for further applications is not enough in the saliency map generated by STB as shown in Figs. 7(b) and 8(b). Moreover, relative large objects are hardly captured in this method as shown in the first and third images of Fig. 7 (also see the results for the second and fourth images in Fig. 8). In the results of SUN, CB, and SR methods, highly textured backgrounds belonging to non-salient regions are not suppressed even though these are visually uncompetitive compared to salient regions. In particular, highly textured backgrounds are extremely emphasized rather than salient objects by using these methods as shown in the third image of Fig. 7. Moreover, false positives near high contrast edges in the background are hard to be eliminated (see results for the second image in Fig. 7 and the fifth image in Fig. 8). As compared to these results, it is easy to see that our method provides visually acceptable saliency, which is consistent with visual attention [see Figs. 7(f), 8(f)].

Moreover, to confirm the robustness and efficiency of our ordinal measure in detecting the spatial saliency, it is compared with the case of directly using sample values of E&COH. For

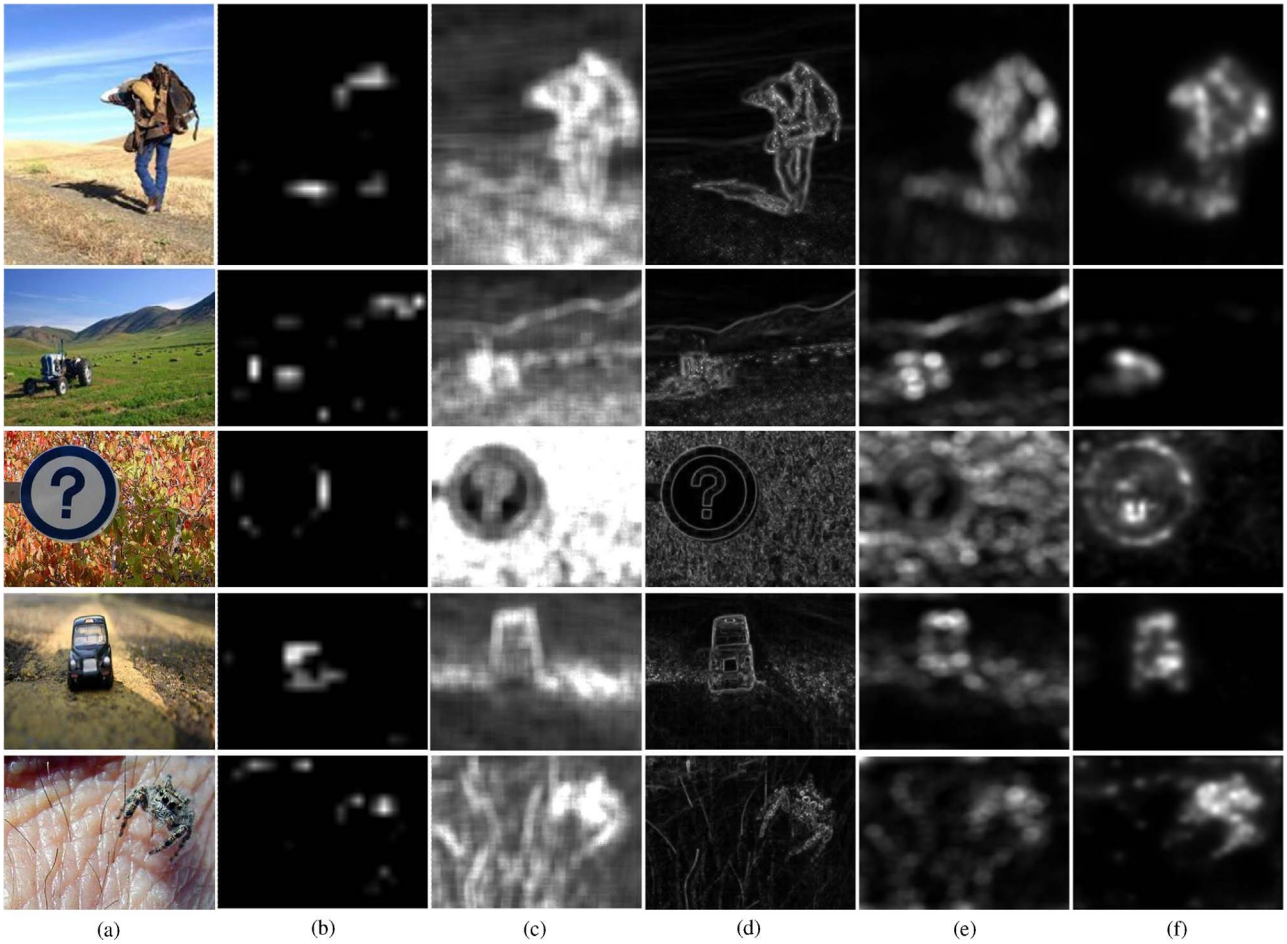

Fig. 7. Results of saliency detection. (a) Original image. (b) Saliency tool box (STB) [32]. (c) Saliency using natural statistics (SUN) [13]. (d) Contrast-based method (CB) [12]. (e) Spectral residual approach (SR) [14]. (f) Proposed method. Our method provides visually acceptable saliency as compared to other methods.

TABLE I

PERFORMANCE COMPARISON OF METHODS ON FINDING PROTO-OBJECTS

Measure \ Method	CB method [12]	SR approach [14]	Proposed method
Recall	0.41	0.64	0.82
Precision	0.35	0.36	0.51

this comparison, sample values of E&COH are normalized by using the $L2$ norm [3] to cope with uneven illumination conditions. Euclidean distance between sample values of local feature distributions obtained from the center and its surrounding regions is used to measure the similarity. As shown in Fig. 9, many pixels are falsely detected as salient pixels near high contrast edges belonging to non-salient regions when Euclidean distance measure is directly applied to sample values of E&COH due to quantization. In contrast to that, ordinal signatures for sample values of E&COH can tolerate variations of local feature distributions and thus efficiently suppress false positives occurring near high contrast edges as explained in Section III-B. Therefore, it is thought that our SORM can provide a more efficient way of building a reliable saliency map.

Furthermore, the proposed spatial saliency can be employed to explicitly represent proto-objects, which are simply segmented by thresholding the saliency map. To determine a threshold value efficiently, we employ nonparametric significance testing [24]. As explained in [24], instead of assuming a type of underlying distribution, we compute the probability density function empirically by using sample $s\tilde{M}_i^{ST}(n)$ and we set the threshold value to achieve 95% confidence level in deciding whether given values are in the extremely right tails of the estimated distribution. This idea is based on the assumption that most local regions do not contain salient objects. To quantitatively evaluate the performance for detecting proto-objects based on our saliency, the proposed method is tested on our dataset composed of 30 images containing various salient objects (see Fig. 10). Note that the ground truth images are manually generated. After the nonparametric significance testing, we obtain the binarized object map. Some examples of proto-objects are shown in Fig. 10. The detection accuracy is evaluated by using recall and precision defined as follows:

$$\text{Recall} = \frac{Card(G \bigcap P)}{Card(G)} \qquad \text{Precison} = \frac{Card(G \bigcap P)}{Card(P)} \tag{10}$$

where G denotes the set of salient pixels in the ground truth images and P denotes that in the binarized object map, respectively. Also, $Card(A)$ indicates the size of the set A. Since non-salient backgrounds need to be suppressed strongly

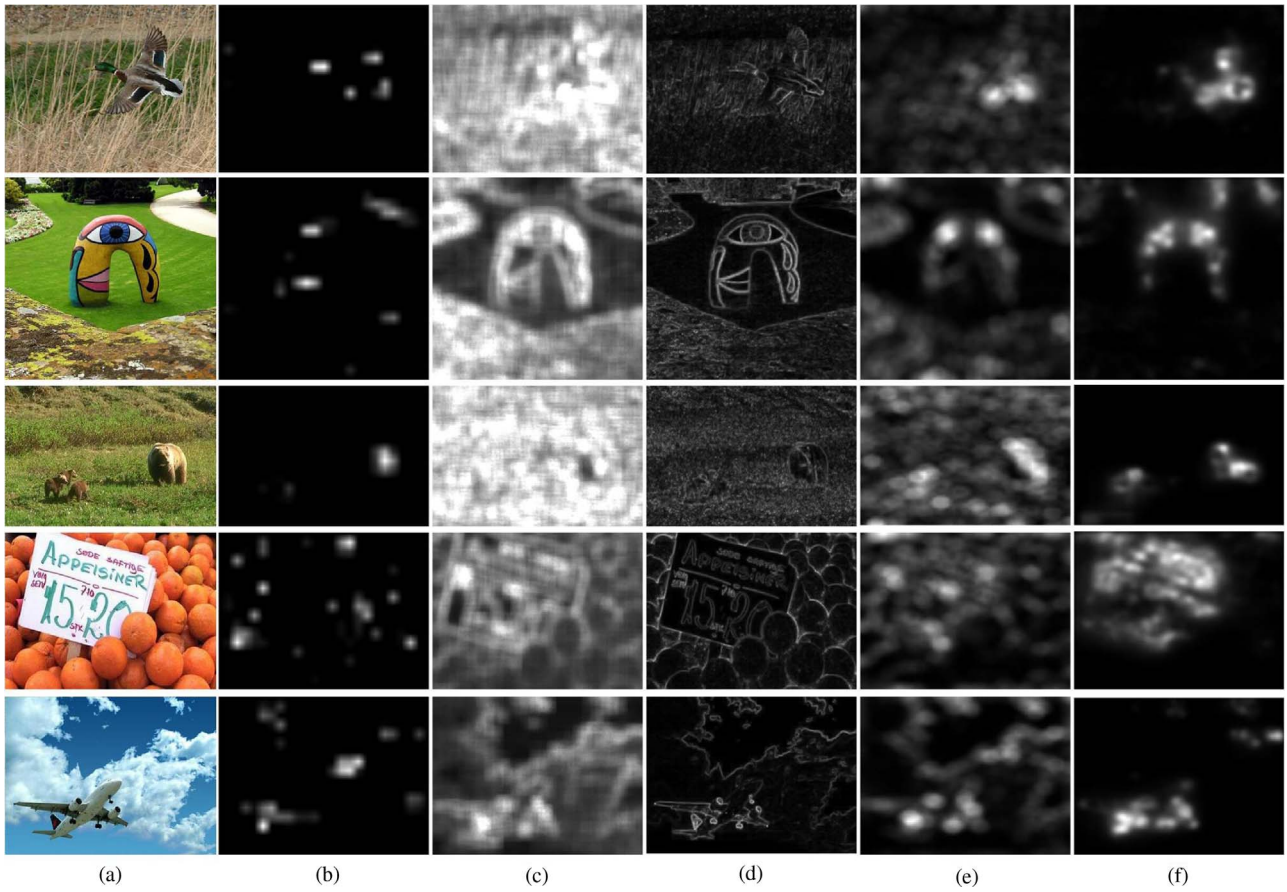

Fig. 8. Results of saliency detection. (a) Original image. (b) Saliency tool box (STB) [32]. (c) Saliency using natural statistics (SUN) [13]. (d) Contrast-based method (CB) [12]. (e) Spectral residual approach (SR) [14]. (f) Proposed method. Our method provides visually acceptable saliency as compared to other methods.

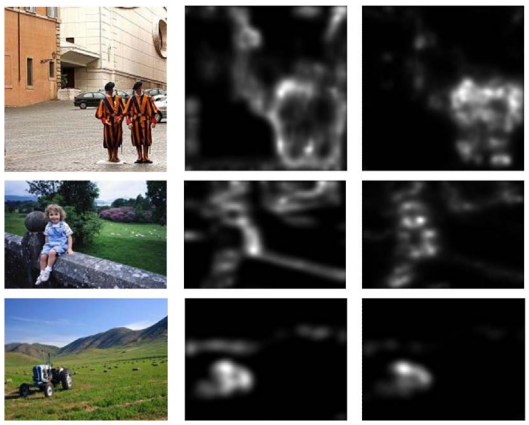

Fig. 9. (a) Original images. (b) Saliency by Euclidean distance of sample values of E&COH. (c) Saliency by ordinal measure of sample values of E&COH (proposed).

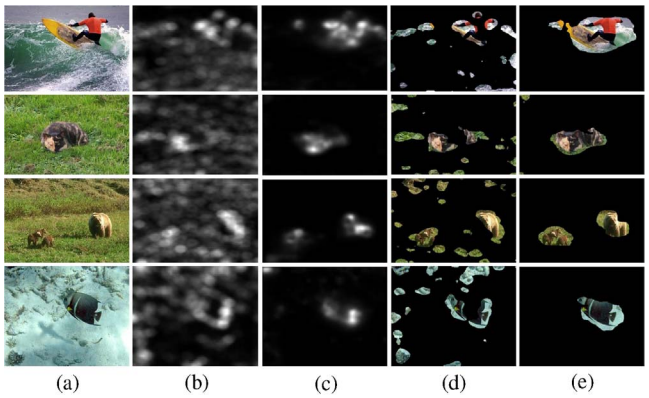

Fig. 10. (a) Original images. (b) Saliency by SR approach. (c) Saliency by the proposed method. (d) Proto-objects by SR approach. (e) Proto-objects by the proposed method.

to detect proto-objects, the high recall and precision are more desirable. As shown in Table I, we observe that our method outperforms previous techniques such as CB method [12] and SR approach [14].

B. Performance Evaluation in Video Sequences

To evaluate the performance of the proposed method in dynamic scenes, we use two videos from our dataset and PETS2001 dataset [33] captured in indoor and outdoor environments, respectively, with the image size of 320×240 pixels. Note that we keep the visual resolution of the saliency map to 64×64 pixels here since the scales of salient objects are almost fixed in our dataset and PETS2001 dataset of QVGA resolution. For computing the temporal saliency values defined in (7), the normalization factor Z^T is set to 49 since the temporal gradients are normalized to [0, 1]. To justify the efficiency of our spatiotemporal saliency, we compare the proposed method with PQFT model [22] in Fig. 11. More

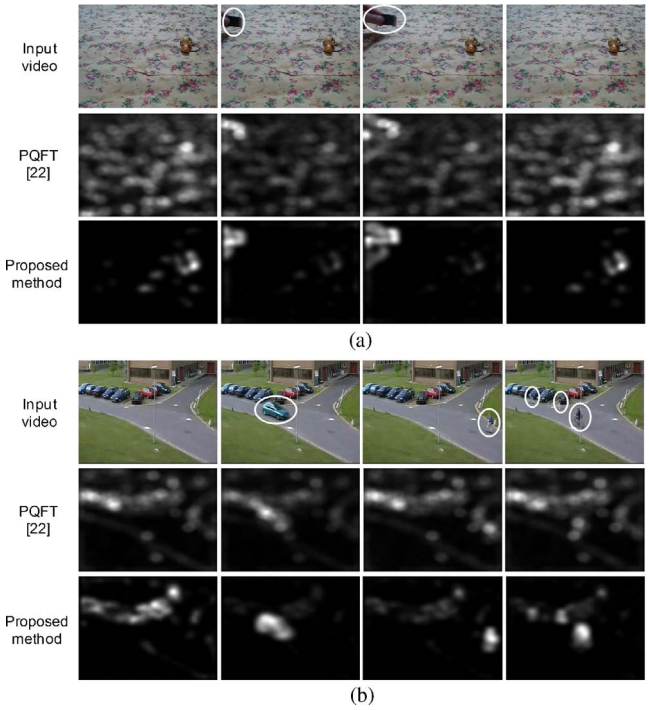

Fig. 11. Spatiotemporal saliency in video sequences. (a) Our video captured in indoor environments. (b) PETS2001 dataset captured in outdoor environments. Moving objects are represented by using white rectangles. Note that our method outperforms the PQFT model [22] in chasing visual attention.

specifically, since there are no moving objects in the beginning part of the first video, our model selects a small doll as salient areas whereas PQFT model pays attention to highly textured backgrounds, which are less salient as shown in Fig. 11(a). Also, both models capture the moving object as salient areas successfully, but highly textured backgrounds are rarely suppressed in PQFT model. In Fig. 11(b), ours and PQFT model select the parked cars as salient areas at the first part of the video, on the other hand, moving people is not effectively emphasized as salient objects in PQFT model even though they attract visual attention above all things in the given scene. In contrast to that, our proposed scheme selects moving car and people as the most salient areas while retaining static salient areas with relative small importance. Thus, our spatiotemporal saliency map can provide reduced search regions and also save search time for object detection and recognition tasks in video sequences.

The framework of the proposed method for evaluating performance has been implemented by using Visual Studio 2005 (C++) under FFMpeg library, which has been utilized for MPEG and Xvid decoding. The experiments are performed on the low-end PC (Core2Duo 3.0 GHz). The processing speed of the proposed method achieves averagely 23 ms per frame (i.e., about 43 frames/s), which is comparable to that of PQFT model, and thus it can be sufficiently applied for real-time applications.

V. Applications in Static and Dynamic Scenes

In the previous section, extensive experiments are conducted on various image and video datasets to prove the efficiency of our spatiotemporal saliency compared to other models proposed in the literature. In this section, based on useful results driven by the proposed method, we extend our model to applications of image retargeting and moving object extraction to show its plentiful possibilities in the field of image and video processing.

A. Static Scenes: Image Retargeting

Image retargeting is the process of adapting a given image to fit the size of arbitrary displays based on the energy map, which is generated by computing the pixel importance. In the field of image retargeting, $L1$-norm and $L2$-norm of gradient magnitude are most widely used to build the energy map. However, these lead to severe distortion when the scene is cluttered or the background is complex. To cope with this problem, our saliency model can be employed as a reliable energy map for image retargeting since it effectively suppresses non-salient regions such as highly textured backgrounds, which are hard to handle in traditional methods. More specifically, pixels represented by low intensities in our saliency map are removed iteratively by dynamic programming, which is widely used for image retargeting, to achieve the target sizes. To confirm the robustness of image retargeting based on our saliency, we compare a novel method proposed in the literature [34], named as seam carving, with our method. Note that we employ the improved seam carving, which incorporates forward energy criterion for better performance [9]. The seam carving method provides desirable viewing effects up to a certain target height or width. However, if the reducing ratio further increases when pixels with high energy are scattered over the whole image, it starts to remove a connected path of pixels across salient objects, resulting in drastic distortions in the resized images. In contrast to that, the proposed saliency map suppresses effectively pixels having high energy in non-salient regions and thus the shape of important objects is preserved more efficiently. Some results of image retargeting for small-display devices based on our saliency map are shown in Fig. 12. Note that we use the same size of input images (e.g., 400×300 pixels for horizontally oriented images and 300×400 pixels for vertically oriented images) for building the saliency map to measure pixel importance at every pixel position. First of all, images are reduced in horizontal direction by 100, 150, and 200 pixels in Fig. 12(a) and (b). The improved seam carving leads to the loss of many pixels belonging to salient objects such as car and church due to trees, yielding high energy over the whole image. Similar to this, images are also reduced in vertical direction by 100, 150, and 200 pixels in Fig. 12(c) and (d). Salient objects are carved out when the reducing ratio is over a certain level since highly textured grass and small scattered stones yield high energies at the background, respectively. Compared to retargeting results by the improved seam carving, it is easy to see that our method provides visually acceptable retargeting results while preserving viewers' experience. Therefore, it is thought that our saliency map is a useful indicator for the image adaptation on small-display devices.

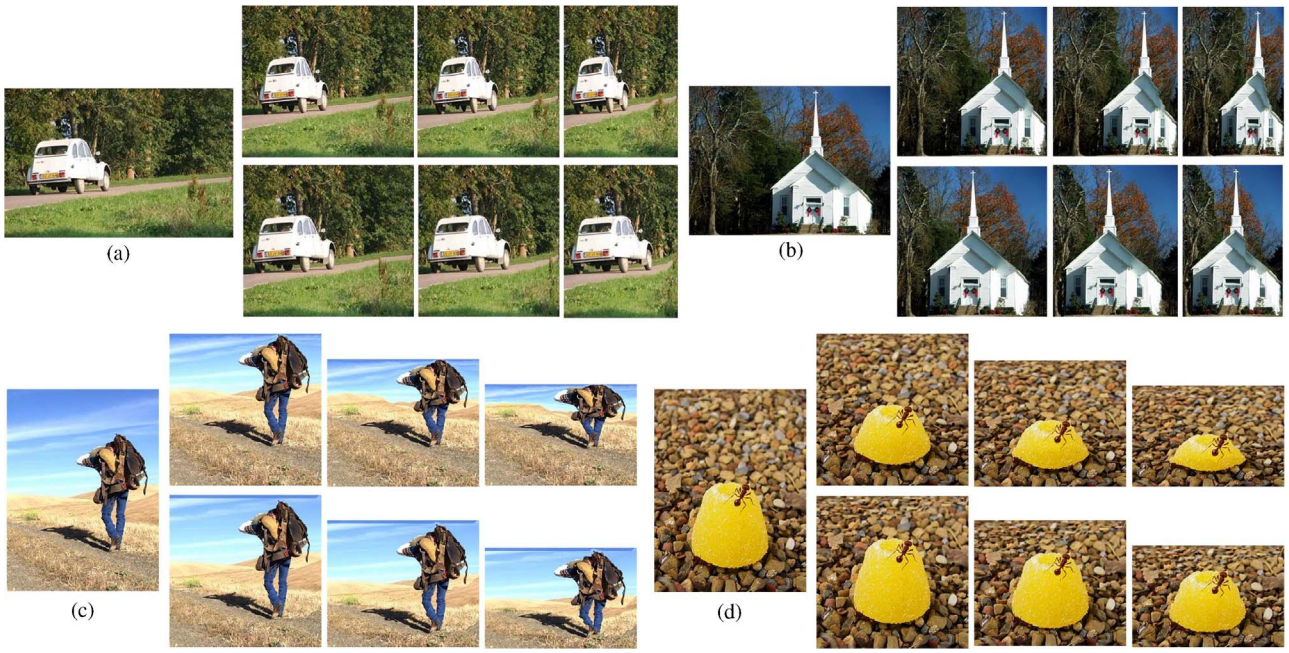

Fig. 12. Retargeting results with various reducing ratios, where the original width [(a) and (b)] or height [(c) and (d)] is reduced by 100, 150, and 200 pixels. In each sub-figure, the top row shows the retargeting results by the improved seam carving [9] and the bottom row is the retargeting results by using the proposed saliency map.

B. Dynamic Scenes: Moving Object Extraction

In the extended outdoor scenes, varying illuminations often cause to high-level false positives, which are hard to handle in the traditional background subtraction frameworks. More specifically, since most surveillance systems are based on the statistical model of color values at each pixel, their performance is drastically dropped in such outdoor conditions with varying illuminations. In this subsection, we address this limitation through our spatiotemporal saliency. Detecting moving objects can be formulated to the detection of salient motion, which is discriminant from global motion. From this saliency point of view, background subtraction thus boils down to ignoring the locations determined as non-salient motions. Since our temporal saliency is computed based on the relative contrast of temporal gradients collected from the center and the surrounding regions, it is useful to capture the salient motion without the explicit motion estimation. Our spatiotemporal saliency map has various advantages over the traditional background subtraction models. First of all, the proposed saliency is completely unsupervised and thus does not require training and initializing the background model. Moreover, there is no need to conduct the additional task for estimating illumination in the scene in our saliency-based background subtraction.

For the performance evaluation of moving objects extraction, we employ two challenging datasets, which are PETS2001 dataset [33] and OTCBVS dataset [35]. Both of them are taken in outdoor environments and contain the dynamic illumination changes. The illumination changes gradually in PETS2001 dataset whereas it changes very fast by passing clouds across buildings in OTCBVS dataset. Thus, it is regarded as the most difficult dataset. The image resolution of each dataset is 320×240 pixels. We compare our saliency-based approach with other methods, the classical GMM [36] and the ensemble algorithm [37], which are widely used for background subtraction. The background subtraction results are shown in Fig. 13. Note that our saliency map is binarized by nonparametric significance testing as mentioned in Section IV-A. Also, the morphological filer (i.e., opening operation with rectangle structure of 3×3 pixels) is applied to the results of the classical GMM and the ensemble algorithm. Since the classical GMM method relies on the statistical information of each pixel, it fails to correctly detect moving objects in both datasets. In the results of the ensemble algorithm, we can see that false positives are reduced. However, it cannot tolerate the significant changes of edge magnitude on the same regions due to sudden illumination changes as shown in Fig. 13(c). Compared to these results, moving objects are correctly detected by our saliency-based scheme even in dynamic outdoor environments with varying illuminations. For the quantitative comparison, recall and precision defined in (10) are employed. To do this, the bounding boxes of moving objects are used as the ground truth. If at least 35% of pixels within each bounding box are classified as foreground pixels,

TABLE II

PERFORMANCE EVALUATION FOR MOVING OBJECT EXTRACTION

Method \ Dataset	PETS2001		OTCBVS	
	Recall	Precision	Recall	Precision
Classical GMM	0.34	0.38	0.34	0.26
Ensemble GMM	0.35	0.56	0.34	0.38
Proposed method	0.36	0.82	0.35	0.68

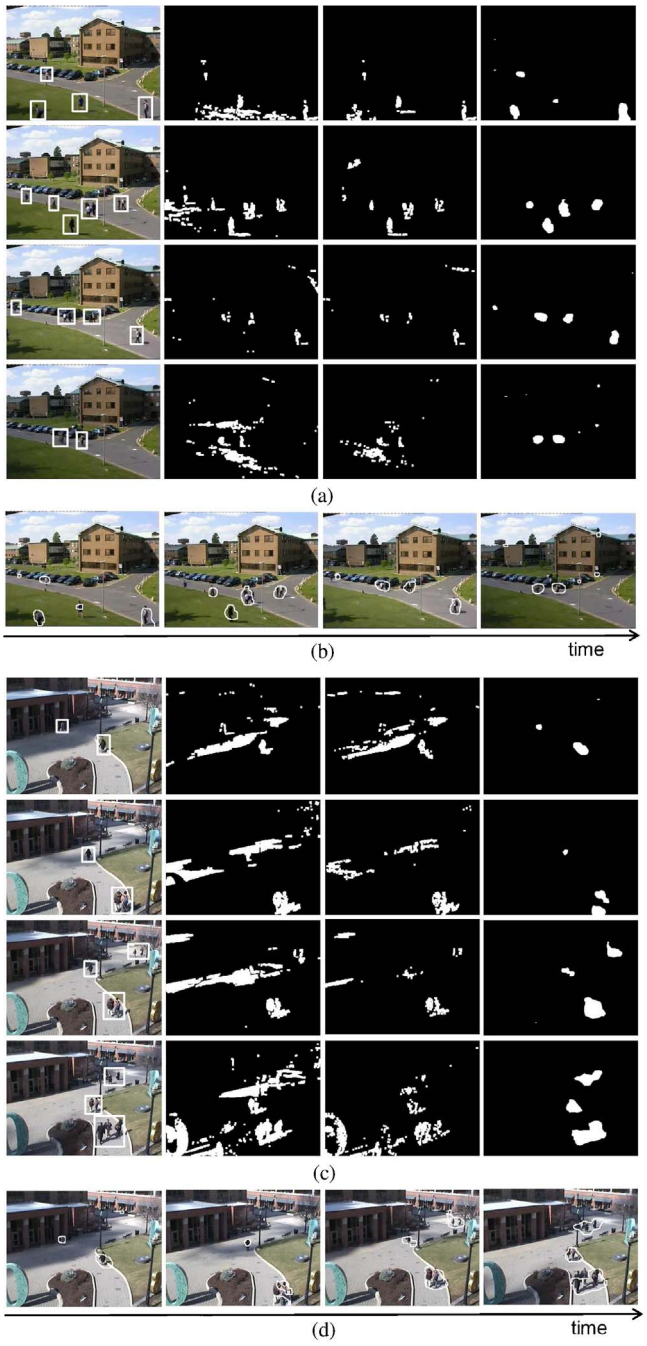

Fig. 13. (a), (c) Background subtraction results in PETS2001 and OTCBVS datasets, respectively. First column: input video. Second column: classical GMM [36]. Third column: ensemble algorithm [37]. Fourth column: our method. Note that moving objects are represented by using white rectangles. (b), (d) Sequences of moving object map by our spatiotemporal saliency in PETS2001 and OTCBVS datasets, respectively.

it can be easily detected as the moving object by using the simple post processing (e.g., connected component analysis). Thus, we compare our method with other approaches based on this recall rate. The recall and precision values computed from 15 frames randomly taken throughout the entire dataset (i.e., they are not from consecutive frames and they span the entire set) are shown in Table II. Note that the low precision value indicates that false positives occur more frequently. Based on these experimental results, we confirm that the proposed saliency can be successfully employed for extracting moving objects.

VI. Conclusion

A simple and novel algorithm for detecting saliency in both images and videos has been proposed in this paper. The basic idea is to detect salient regions in natural scenes by measuring how each local region stands out from its surroundings based on a discriminant center-surround hypothesis. To this end, we first defined a set of visual features composed of edge and color orientations and temporal gradients. In the spatial domain, the saliency of each pixel is measured based on the self-ordinal resemblance of E&COH. By using ordinal signatures of sample values of E&COH, our model can cope with variations of local feature distributions efficiently due to quantization, which are hard to be handled in traditional methods. For the spatiotemporal saliency, we combined the spatial saliency with the temporal saliency by simply computing SAD between temporal gradients of the center and the surrounding regions. To justify the efficiency and robustness of our approach, the performance of the proposed algorithm is compared with that of other methods proposed in the literature by using various images and videos. Based on the experimental results, we confirmed that our saliency provides a reasonable starting point for semantic scene understanding. Moreover, the proposed scheme performs in real-time and thus can be extended to various applications in static and dynamic scenes. In particular, to show potential of our spatiotemporal saliency in the field of image and video processing, we applied our model to image retargeting and moving object extraction. Since highly textured backgrounds are efficiently suppressed in the proposed method, image retargeting is successfully conducted without visual distortion on the salient regions. In the video surveillance, moving objects are also correctly extracted by using our saliency-based background subtraction even in varying illumination conditions. Therefore, we believed that plentiful possibilities of our work lie in many promising applications in the field of multimedia processing such as image and video quality measure [38], object segmentation [39], and so on. Our future work is to investigate these issues based on the proposed saliency detection framework for more advanced and intelligent applications.

References

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

[2] J. Luo and S. P. Etz, "A physical model-based approach to detecting sky in photographic images," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 201–212, Mar. 2002.

[3] Y.-T. Chen and C.-S. Chen, "Fast human detection using a novel boosted cascading structure with meta stages," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1452–1464, Aug. 2008.

[4] 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.

[5] F. Liu and M. Gleicher, "Region enhanced scale-invariant saliency detection," in *Proc. IEEE ICME*, Jul. 2006, pp. 1477–1480.

[6] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," in *Proc. IEEE ICIP*, vol. 1. Sep. 2003, pp. 253–256.
[7] D. Gao and N. Vasconcelos, "Integrated learning of saliency, complex features, and object detectors from cluttered scenes," in *Proc. IEEE Conf. CVPR*, vol. 2. Jun. 2005, pp. 282–287.
[8] W.-H. Chen, C.-W. Wang, and J.-L. Wu, "Video adaptation for small display based on content recomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp. 43–58, Jan. 2007.
[9] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–8, Aug. 2008.
[10] J.-S. Kim, J.-H. Kim, and C.-S. Kim, "Adaptive image and video retargeting technique based on Fourier analysis," in *Proc. IEEE Comput. Vision Pattern Recognition*, Jun. 2009, pp. 1730–1737.
[11] V. Mahadevan and N. Vasconcelos, "Background subtraction in highly dynamic scenes," in *Proc. IEEE Comput. Vision Pattern Recognition*, Jun. 2009, pp. 1–6.
[12] Y. F. Ma and H. J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 374–381.
[13] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vision*, vol. 8, no. 7, pp. 1–20, 2008.
[14] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Comput. Vision Pattern Recognition*, Jun. 2007, pp. 1–8.
[15] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.
[16] D.-Y. Chen, H.-R. Tyan, D.-Y. Hsiao, S.-W. Shih, and H.-Y. M. Liao, "Dynamic visual saliency modeling based on spatiotemporal analysis," in *Proc. IEEE ICME*, Jun. 2008, pp. 1085–1088.
[17] W. Kim, C. Jung, and C. Kim, "Saliency detection: A self-ordinal resemblance approach," in *Proc. IEEE ICME*, Jul. 2010, pp. 1260–1265.
[18] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *J. Vision*, vol. 8, no. 7, pp. 1–18, 2008.
[19] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Salient region detection by modeling distributions of color and orientation," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 892–905, Aug. 2009.
[20] C. Kim, "Content-based image copy detection," *Signal Process. Image Commun.*, vol. 18, no. 3, pp. 169–184, Mar. 2003.
[21] C. Kim and B. Vasudev, "Spatiotemporal sequence matching for efficient video copy detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 127–132, Jan. 2005.
[22] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.
[23] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1397–1410, Oct. 2001.
[24] H. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vision*, vol. 8, no. 12, pp. 1–27, 2009.
[25] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video retargeting," in *Proc. IEEE ICCV*, Oct. 2007, pp. 1–6.
[26] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.
[27] O. Michailovich, Y. Rathi, and A. Tannenbaum, "Image segmentation using active contours driven by the Bhattacharyya gradient flow," *IEEE Trans. Image Process.*, vol. 16, no. 11, pp. 2787–2801, Nov. 2007.
[28] H. Liu, S. Jiang, Q. Huang, C. Xu, and W. Gao, "Region-based visual attention analysis with its application in image browsing on small displays," in *Proc. ACM Int. Conf. Multimedia*, 2007, pp. 305–308.
[29] P. Noriega and O. Bernier, "Real time illumination invariant background subtraction using local kernel histograms," in *Proc. Br. Mach. Vision Assoc.*, 2006, p. III:979.
[30] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artif. Intell.*, vol. 146, no. 1, pp. 77–123, May 2003.
[31] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Y. Shum, "Learning to detect a salient object," in *Proc. IEEE CVPR*, Jun. 2007, pp. 1–8.
[32] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Netw.*, vol. 19, no. 9, pp. 1395–1407, 2006.
[33] J. Ferryman, ed., in *Proc. 2nd IEEE Int. Workshop PETS*, Dec. 2001 [Online]. Available: ftp://ftp.pets.rdg.ac.uk/pub/PETS2001
[34] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Trans. Graph.*, vol. 26, no. 3, pp. 267–276, 2007.
[35] J. Davis and V. Sharma, "Background-subtraction using contourbased fusion of thermal and visible imagery," *Comput. Vision Image Understand.*, vol. 106, nos. 2–3, pp. 162–182, 2007 [Online]. Available: http://www.cse.ohio-state.edu/otcbvs-bench
[36] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE CVPR*, vol. 2. Jun. 1999, pp. 246–252.
[37] B. Klare and S. Sarkar, "Background subtraction in varying illuminations using an ensemble based on an enlarged feature set," in *Proc. IEEE CVPRW*, Jun. 2009, pp. 66–73.
[38] A. Ninassi, O. Lemeur, P. Lecallet, and D. Barba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *Proc. IEEE ICIP*, vol. 2. Oct. 2007, pp. 169–172.
[39] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Proc. IEEE ICPR*, Dec. 2008, pp. 1–4.

Wonjun Kim (S'10) received the B.S. degree in electronic engineering from Sogang University, Seoul, Korea, and the M.S. degree in electronic engineering from Information and Communications University, Daejeon, Korea, in 2006 and 2008, respectively. He is currently pursuing the Ph.D. degree from the Computational Imaging Laboratory, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon.

His current research interests include object detection, image/video processing, computer vision, and pattern recognition.

Chanho Jung (S'10) received the B.S. and M.S. degrees in electronic engineering from Sogang University, Seoul, Korea, in 2004 and 2006, respectively. He is currently pursuing the Ph.D. degree from the Computational Imaging Laboratory, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea.

From 2006 to 2008, he was a Research Engineer with the Digital TV Research Laboratory, LG Electronics, Seoul. His current research interests include image/video understanding, computer vision, pattern recognition, and image processing.

Changick Kim (SM'11) was born in Seoul, Korea. He received the B.S. degree in electrical engineering from Yonsei University, Seoul, the M.S. degree in electronics and electrical engineering from the Pohang University of Science and Technology, Pohang, Korea, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 1989, 1991, and 2000, respectively.

From 2000 to 2005, he was a Senior Technical Staff Member with Epson Research and Development, Inc., Palo Alto, CA. From 2005 to 2009, he was an Associate Professor with the School of Engineering, Information and Communications University, Daejeon, Korea. Since March 2009, he has been with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, where he is currently an Associate Professor. His current research interests include multimedia signal processing, 3-D video processing, image/video understanding, intelligent media processing, and video coding for Internet protocol television.