Spatiotemporal Saliency Detection Using Textural Contrast and Its Applications

Journal:	*Transactions on Image Processing*
Manuscript ID:	TIP-08977-2012
Manuscript Type:	Regular Paper
Date Submitted by the Author:	02-May-2012
Complete List of Authors:	Kim, Wonjun; Korea Advanced Institute of Science and Technology (KAIST), Electrical Engineering; Kim, Changick; Korea Advanced Institute of Science and Technology (KAIST), Electrical Engineering;
EDICS:	ARS-IIU Image & Video Interpretation and Understanding < Image & Video Analysis, Synthesis, and Retrieval

Spatiotemporal Saliency Detection Using Textural Contrast and Its Applications

Wonjun Kim, *Student Member, IEEE*, and Changick Kim, *Senior Member, IEEE*.

Abstract

Saliency detection has been extensively studied due to its great possibilities for various computer vision applications. However, most existing methods are easily biased toward edges or corners, which are statistically significant, but not necessarily salient. Moreover, they often fail to find salient regions in complex scenes due to ambiguities between salient regions and highly textured backgrounds. In this paper, we present a novel unified framework for spatiotemporal saliency detection based on *textural contrast*. Our method is simple, robust, yet biologically plausible and it can thus be easily extended to various applications such as image retargeting, object segmentation, and video surveillance. Based on various data sets, we conduct comparative evaluations of 12 representative saliency detection models presented in literature, and the results show that the proposed scheme outperforms other previously developed methods in detecting salient regions of the static and dynamic scenes.

Index Terms

Saliency detection, computer vision applications, human visual attention, textural contrast, comparative evaluations.

I. Introduction

The human visual system (HVS) has an outstanding ability to quickly grasp the most relevant regions at a glance without any prior knowledge. Therefore, we can easily understand contextual information of a given scene based on this selective visual attention in an efficient manner. There are numerous factors contributing such visual saliency. Among them, as reported by many biological experiments, the most important factor is contrast [1]. That is, the relevant element is not the absolute amplitude of visual signals (e.g., intensity, color, etc.) but rather contrast between these amplitudes at a given point and its

Wonjun Kim and Changick Kim are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. E-mail: {jazznova, changick}@kaist.ac.kr

surroundings. The importance of contrast has been strongly supported by a meaningful result in cognitive neuroscience, showing that the receptive field of the retina is performed based on the center-surround cell network (i.e., center-surround contrast) in which cone synaptic input is fed into the center of the receptive field via the dendritic tree and the second input is provided by excitatory surrounds using gap junction between cells [2]. Therefore, computational modeling of this biological system enables various applications (e.g., content-aware image resizing [3], [4], [27], object detection and segmentation [6], [7], [8], adaptive image and video compression [9], video summarization [10], image quality assessment [11], [12], [13], video surveillance [27], and so on), requiring only limited processing resources. For this reason, saliency detection has been extensively studied by researchers in psychology, cognitive neuroscience, and computer vision.

In the field of computer vision, many computational models have been proposed to accomplish this task automatically, and the comprehensive survey on recent developments is also found in [14]. According to the literature review, most of previous methods can be divided into two major groups, i.e., top-down and bottom-up approaches. First of all, the top-down approaches are task-driven and can thus be regarded as solving the problem of visual recognition [15], [16]. In this category, salient visual attributes are defined as descriptors delineating specific objects, such as face, text, etc., for the given task. These approaches mostly require prior knowledge, which is not available in every image, and it thus leads to hard generalization. The majority of saliency detection methods are driven by the biological plausibility of the bottom-up mechanisms. More specifically, most of bottom-up approaches have been proposed based on a set of simple low level features such as luminance, color, and orientation, followed by some center-surround operations. Again, this is because local image features become stimuli of interest when they are best distinguishable from its surroundings that may be of possible interest (it is referred to as the discriminant center-surround hypothesis) [17].

In this paper, we introduce a novel unified framework for detecting salient regions in both images and videos. The key idea behind our approach is to mimic the biological system by formulating two main contrast mechanisms occurring in the retina and the visual cortex. Specifically, we propose to use *textural contrast* defined as the combination of luminance contrast (for retina level) and directional coherence contrast (for visual cortex level), and extend its concept to the spatiotemporal domain with temporal gradients. By incorporating the responses of *textural contrast* into a multiscale framework, we can generate reliable saliency maps for images and videos. Compared to traditional bottom-up models, one important advantage of the proposed method is that it greatly eliminates unwanted fine details whereas highlighting salient regions quite uniformly owing to the ability of providing the contextual information regarding underlying image structures. Note that this work is extended from our previous one [18] and differs in the following respects: 1) we provide more technical details about the appropriateness of *textural contrast* for saliency detection; 2) we extend the concept of the directional coherence presented in [18] to detect saliency from videos as well; 3) we provide comparative evaluations of 12 representative saliency detection models both qualitatively and quantitatively. Moreover, various saliency-inspired applications are also demonstrated.

The rest of this paper is organized as follows. A systematic review of previous bottom-up methods is presented 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 images and videos are introduced in Section V. Conclusion follows in Section VI.

II. A REVIEW OF BOTTOM-UP SALIENCY DETECTION MODELS

In this section, we briefly review several bottom-up models, which are representative in literature, and discuss about their limitations. The main advantage of these models lies in data-driven nature, i.e., do not require any prior knowledge. Most of bottom-up approaches can be further divided into two categories: statistical and spectral methods. Statistical methods are fundamentally based on the center-surround hypothesis. Specifically, they mostly adopt the difference between feature statistics obtained from center and surrounding regions as a measure of saliency. The first computational and statistical model is developed in a center-surround framework by Itti *et al.* [19]. Their saliency map is generated based on the linear combination of normalized feature maps obtained from three basic components: intensity, color, and orientation. Inspired by their success in predicting human fixations, several models, more or less based on different mathematical tools, have been investigated in literature. Ma and Zhang [20] compute the distance between Lab color features obtained from center and surrounding regions on the quantized block image. Harel *et al.* [21] define the graph using pixel positions and weight values proportional to their dissimilarity obtained from orientation, intensity, and its variation. The resulting graphs are treated as Markov chains and their equilibrium distribution is adopted as saliency maps. Achanta *et al.* [22] formulate the problem of detecting salient regions as conducting the band pass filtering, which is simply implemented by computing the difference between mean of colors over the whole image and Gaussian blurred version of the original image. Bruce and Tsotsos [23] propose to employ Shannon's self-information measure and adopt the independent component analysis (ICA) to efficiently estimate the one-dimensional probability density function. In [24], authors compute statistical likelihood of the feature response from each pixel to those of surrounding regions as a measure of saliency. To consider the local structure more precisely, they propose to use the local steering kernel estimated from a collection of spatial gradient vectors. Xu *et al.* [25] propose to utilize the spatial-frequency information to be robust to the complex background. They compute the residual of the Renyi entropy via the pseudo-Wigner-Ville distribution for finding salient regions. Goferman *et al.* [4] incorporate positional information into color contrast between image patches for detecting salient regions. Liu *et al.* [26] propose a set of features including multiscale contrast, center-surround histogram, and color spatial distribution to describe a salient object locally, regionally, and globally. A conditional random field is employed to efficiently combine these features. Authors of [27] exploit ordinal signatures of the feature distribution. The rationale behind this method is that ordinal signatures are robust to small variations occurring in the feature distribution and thus the difference of them between center and surrounding regions indicates salient locations even under highly textured backgrounds. They also propose a framework for spatiotemporal saliency detection by involving the sum of difference obtained from temporal gradients.

On the other hand, spectral methods also have been constantly proposed. These methods attempt to efficiently eliminate background by analyzing filter responses in the frequency domain based on the assumption that less periodicity makes the rare event (i.e., saliency) on the corresponding location in the reconstruction of the original image. It should be emphasized that spectral methods are highly correlated with the human vision mechanism, which is able to grasp salient regions at a glance, since they promptly work in a global view. Hou and Zhang [28] firstly introduce global contrast in the frequency domain to detect salient regions by using spectral residual, which is simply defined by subtracting a smoothed version of the log magnitude spectrum from the original one. However, it is well known that what actually locates saliency is the phase information, rather than the magnitude information. In this sense, Bian and Zhang [29] normalize Fourier coefficients with respect to their magnitudes and only use the phase information to find salient regions. Similarly, Guo and Zhang [9] build the spatiotemporal saliency map using the phase spectrum of the quaternion Fourier transform (PQFT), which is composed of color, intensity, and temporal gradients.

Even though such bottom-up models have been extensively studied, they still suffer from two main limitations: 1) a bias toward edges or corners and 2) vulnerability to cluttered and highly textured background. These limitations are illustrated in Fig. 1. Specifically, previous methods tend to emphasize only high contrast edges and thus easily fail to capture the whole regions of saliency. They also tend to highlight cluttered background rather than salient regions and thus the saliency maps by previous models are expected to be highly unreliable. To tackle these limitations, we propose to use *textural contrast*

Bottom-up saliency models identify the salient regions of an image based on features such as color, intensity and orientation.

In a top-down model, the saliency of an image is computed based on how relevant it is to the current task or goal.

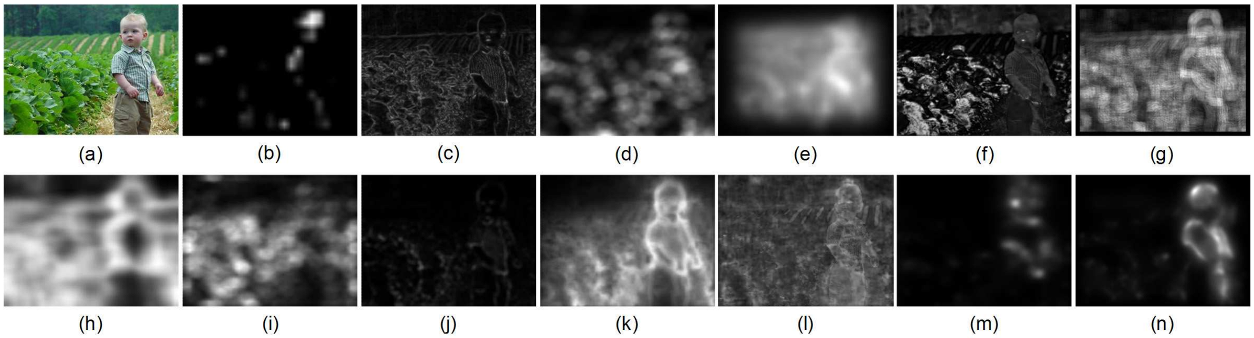


Fig. 1. (a) Input image. Saliency maps generated by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method. Note that pixels in high intensities are highly likely to be salient. For simplicity, we refer to the first author of each method.

for finding salient regions. Moreover, we provide a novel unified framework for spatiotemporal saliency detection by involving directional coherence contrast of temporal gradients. In the following, we will explain the proposed spatiotemporal saliency detection scheme and its excellence in detail.

III. Proposed Method

In general, the brain and the vision systems work together to identify relevant regions in a given scene. We aim to model this biological system focusing on the main visual stream from the retina to the visual cortex. The first type of information captured by our visual system in the retina is luminance contrast. At higher levels of processing in the visual cortex, orientation contrast is involved to understand the context. It is important to note that the conjunction of luminance contrast and orientation contrast makes the corresponding region to be more salient than using either of them separately [1]. Motivated by this fact, we attempt to model such biological mechanisms using *textural contrast* defined by allowing for both luminance contrast and directional coherence contrast. In particular, we propose to exploit the directional coherence to estimate orientation contrast since the use of gradient information in a pixel-wise manner often leads to failure in describing the underlying image structure, especially in cluttered and highly textured regions. We also extend the concept of the directional coherence to the temporal domain for spatiotemporal saliency detection.

A. Spatial saliency by textural contrast

First of all, we define luminance contrast by considering how distinctive the intensity attribute of each pixel is compared to the global one. For the improved dynamic ranges useful for effectively suppressing

Fig. 2. (a) Input image, (b) first-order model, (c) second-order model, and (d) fourth-order model.

Fig. 3. More examples for luminance contrast maps. (a) Input image, (b) first-order model, (c) second-order model, and (d) fourth-order model.

high contrast in the background, the n-th order statistics are applied as follows:

$$S_L^k(i) = \left| \bar{I}^k - \frac{1}{N} \sum_{j \in B_i} I^k(j) \right|^n, \tag{1}$$

where k denotes the frame index. $\bar{I}^k$ denotes the mean of luminance values over the whole image (i.e., the largest surrounding region). B_i and N represent the neighbor region (5×5 pixels in our implementation) centered at the i^{th} pixel position and its size, respectively. The luminance contrast maps generated by using various n values are shown in Fig. 2. From exhaustive experiments, it is carefully observed that the second-order moment (i.e., $n = 2$) yields the best results in suppressing irrelevant regions while sufficiently emphasizing the salient region. More examples are shown in Fig. 3.

Along with luminance contrast, we also attempt to depict the local image structure based on directional

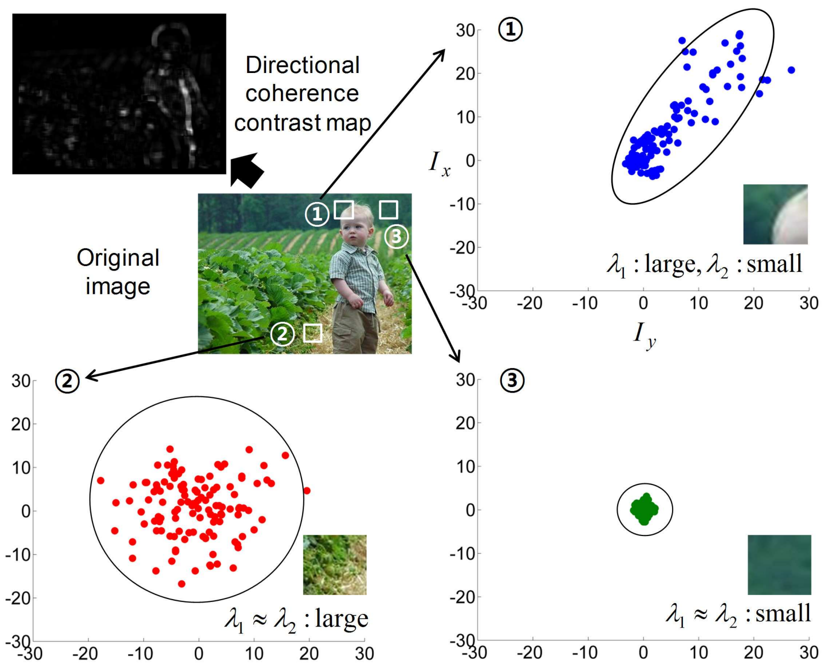


Fig. 4. Gradients obtained from selected image patches are illustrated. Note that λ_1 and λ_2 represent the energy along the dominant orientation of the gradient field and its perpendicular direction, respectively.

coherence contrast obtained from center and surrounding regions. It is important to note that we focus on directional coherence rather than directly using gradient information, which is unreliable in cluttered and highly textured regions. In detail, directional patterns in center and surrounding regions provide a good approximation to the underlying image structure, which is indeed coherent with the visual attention. To do this, we allow for the structure tensor, which efficiently summarizes the dominant orientation and the energy along this direction based on the local gradient field, defined as follows:

$$\mathbf{T}_s^k(i) = \begin{bmatrix} \sum_{j \in B_i} I_x^k(j)^2 & \sum_{j \in B_i} I_x^k(j) I_y^k(j) \\ \sum_{j \in B_i} I_x^k(j) I_y^k(j) & \sum_{j \in B_i} I_y^k(j)^2 \end{bmatrix}, \tag{2}$$

where I_x^k and I_y^k denote the gradient in horizontal and vertical directions at the k^{th} frame, respectively. The usefulness of the structure tensor defined in (2) for our task stems from the fact that the relative discrepancy between two eigenvalues (i.e., $\lambda_1 \geq \lambda_2 \geq 0$) of $\mathbf{T}_s^k(i)$ indicates how intensively gradients in the local region are distributed along the dominant direction (i.e., the degree to which those directions are consistent). For better understanding, we illustrate the distributions of gradients obtained from selected image patches as shown in Fig. 4. As can be seen, the gradients belonging to the textural boundary attracting the visual attention (①) are intensively distributed along the dominant direction compared to

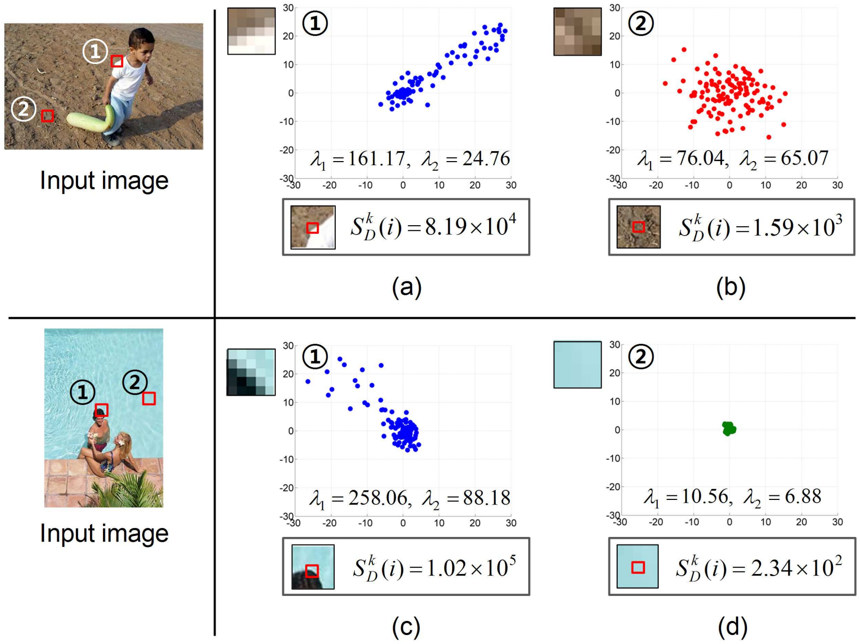


Fig. 5. Directional coherence contrast obtained from (a)(c) salient region, (b) highly textured region, and (d) flat region. Note that $S_D^k(i)$ values from (a) and (c) are much larger than those of (b) and (d) (i.e., irrelevant regions).

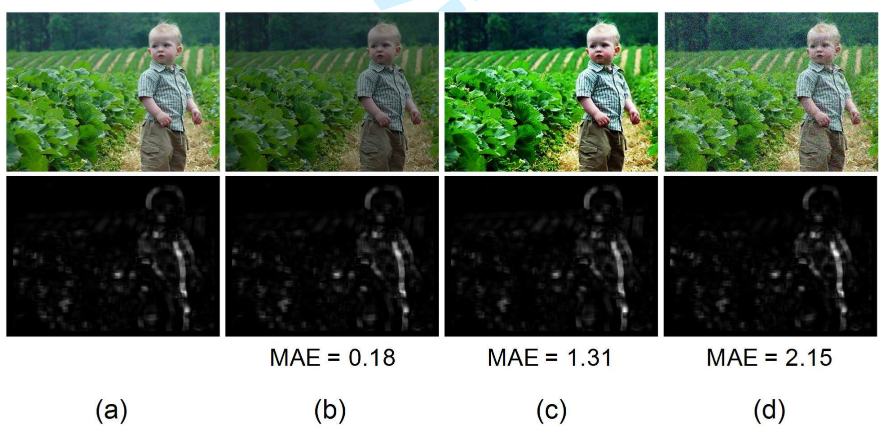


Fig. 6. Directional coherence contrast maps in challenging conditions. (a) Original image, (b) brightness change, (c) contrast change, and (d) white Gaussian noise (0, 0.01). Note that the number marked below each sub-figure denotes the average MAE value, which is obtained from the comparison with the directional coherence contrast map of (a).

those of the highly textured region (②) or the uniformly textured (i.e., flat) region (③). Thus, we define our directional coherence at each pixel position as follows:

$$\xi = (\lambda_1 - \lambda_2)^2. \tag{3}$$

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

Fig. 7. (a) Original image, (b) luminance contrast map, (c) directional coherence contrast map, and (d) our spatial saliency map (single scale).

Here the larger the value ξ is, the higher the directional coherence is. Note that the average of gradients does not guarantee the reliable measure since aligned but oppositely oriented gradients would cancel out in this average. In what follows, directional coherence contrast between center and surrounding regions can be formulated as follows:

$$S_D^k(i) = \sum_{j \in W_i} |\xi^k(j) - \xi^k(i)|, \tag{4}$$

where W_i is a set of neighboring pixels centered at the i^{th} pixel position. Note that the size of W_i is set to 7×7 pixels in our implementation. An example of the directional coherence contrast map (i.e., the gray-scale representation of $S_D^k(i)$) is shown in Fig. 4. More examples for the directional coherence contrast are shown in Fig. 5. We confirm that salient regions yield quite large values compared to irrelevant regions. We also demonstrate some examples of the directional coherence maps in various challenging conditions in Fig. 6. More specifically, those maps provide the reliable image structure even with the drastic change of brightness and contrast (see Fig. 6(b) and (c)). In addition to this, directional coherence contrast is quite robust to the presence of noise (see Fig. 6(d)). For generality, we compute the average MAE (mean absolute error) values obtained from over 500 natural images (see the number marked below each sub-figure in Fig. 6). Note that small MAE values show that directional coherence contrast between center and surrounding regions is invariant to a wide range of variations. Therefore, it is thought that directional coherence contrast is highly desirable to measure visual saliency.

Since salient regions are assumed to contain both luminance contrast and directional coherence contrast as mentioned, our spatial saliency map at the k^{th} frame is thus computed by combining such two responses as follows:

$$S^k(i) = S_L^k(i) \times S_D^k(i). \tag{5}$$

Here each response is smoothed by Gaussian filtering as in [28] and $S^k(i)$ is normalized to [0,255] for gray-scale representation. It is worth noting that the combination strategy defined in (5) produces more

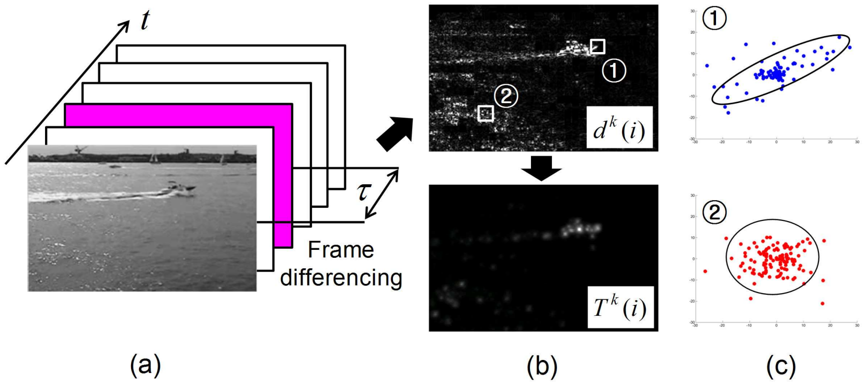


Fig. 8. (a) Input image sequence, (b) results of frame differencing (i.e., temporal gradient, $d^k(i)$) (top) and the proposed temporal saliency map (bottom), and (c) gradient distributions for selected image patches from $d^k(i)$.

desirable saliency maps while effectively suppressing false positives in the background. This is because either of two responses may be high in the irrelevant region. The example of our spatial saliency map at the single scale is shown in Fig. 7.

B. *Combining with temporal saliency*

For the spatiotemporal saliency detection, we need to involve motion stimuli, which can be defined by spatiotemporal orientation (equivalent to the velocity [17]). To compute motion contrast (i.e., motion energy associated with different velocities) strongly attracting the visual attention in videos, we propose to apply the concept of the directional coherence to temporal gradients. First of all, the structure tensor of temporal gradients can be represented similarly with (2) as follows:

$$\mathbf{T}_t^k(i) = \begin{bmatrix} \sum_{j\in B_i} d_x^k(j)^2 & \sum_{j\in B_i} d_x^k(j)d_y^k(j) \\ \sum_{j\in B_i} d_x^k(j)d_y^k(j) & \sum_{j\in B_i} d_y^k(j)^2 \end{bmatrix}, \quad (6)$$

where $d^k(i) = I^k(i) - I^{k-\tau}(i)$ ($\tau = 3$ in our work). Based on this, the temporally directional coherence can be defined by using the difference between two eigenvalues, λ_1 and λ_2, of $\mathbf{T}_t^k(i)$, i.e., $\phi = (\lambda_1 - \lambda_2)^2$. In what follows, we adopt contrast of the temporally directional coherence as the measure of temporal saliency as follows:

$$T^k(i) = \sum_{j\in W_i} |\phi^k(j) - \phi^k(i)|, \quad (7)$$

where W_i is a set of neighboring pixels centered at the i^{th} pixel position as mentioned before. The overall procedure for generating the temporal saliency map is shown in Fig. 8. It is worth noting that our

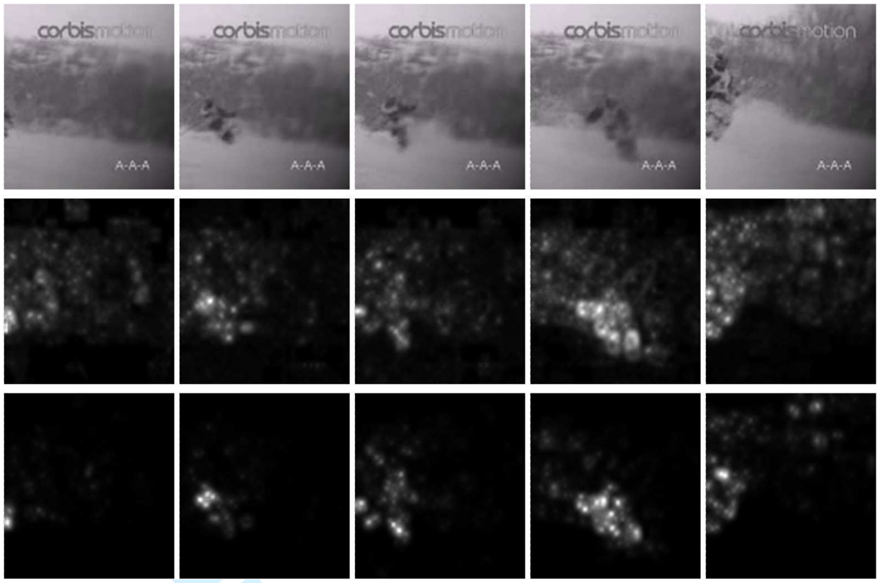


Fig. 9. Performance comparison between the center-surround temporal gradient patterns and the proposed temporal saliency. Ski sequences obtained from [41] (top), results of the center-surround temporal gradient patterns (middle), and our temporal saliency (bottom).

approach for temporal saliency detection has a great ability to suppress irrelevant motions (e.g., rippling water) occurring in the background while still highlighting a region of interest (e.g., a moving boat). This is because temporal gradients by irrelevant motions are generally unstructured in a local regions (see Fig. 8(b)) and they thus yield low contrast of the temporally directional coherence. Moreover, we compare ours with the center-surround temporal gradient patterns, i.e., $\sum_{j \in W_i} |d^k(j) - d^k(i)|$, and results are shown in Fig. 9. We can see that the center-surround temporal gradient patterns often fail to suppress a snowfall in the background whereas the tensor-based analysis allows the proposed temporal saliency to be more closely correlated with visual attention.

Finally, the proposed spatiotemporal saliency map at the single scale can be defined by the combination of the spatial and temporal saliencies, i.e., $S^k(i)$ and $T^k(i)$, as follows:

$$V^k(i) = \alpha \cdot S^k(i) + (1 - \alpha) \cdot T^k(i), \tag{8}$$

where $\alpha \in [0, 1]$ denotes the weighting factor for balancing between the spatial and temporal saliency. In general view, since moving objects are more attractable than static objects and backgrounds in videos [31], we set α to 0.3 (i.e., weigh temporal saliency more than spatial one) in our implementation. For the grey-scale representation, the spatiotemporal saliency values defined in (8) are normalized from 0 to 255.

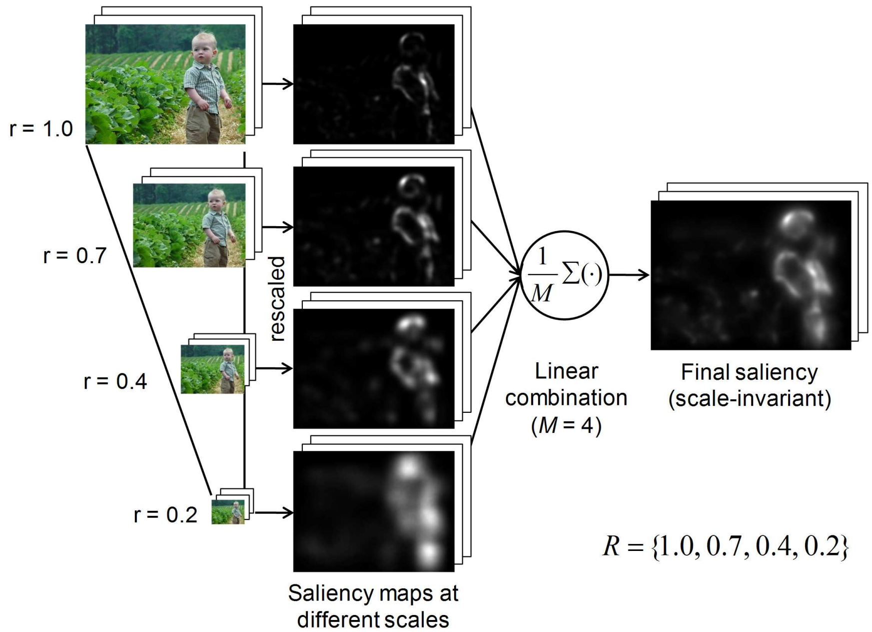


Fig. 10. Scale-invariant spatiotemporal saliency map. Note that the saliency map computed at each scale is resized to the size of the original image.

C. *Scale-invariant spatiotemporal saliency map*

Since the size of salient regions is unknown, saliency maps are usually built on the combination of outputs from different scales [19], [21], [9], [4], [27]. Specifically, let $R = \{r_1, r_2, \cdots, r_M\}$ denote the set of scales to be considered to conduct the multiscale analysis. Note that we treat all image levels equally by taking them into account in a unified solution since no level is more important than others in HVS. Therefore, the scale-invariant spatiotemporal saliency map is finally computed by the linear combination of outputs obtained from each scale with the same weight as follows:

$$\tilde{V}^k(i) = \frac{1}{M}\sum_{r\in R} V_r^k(i), \tag{9}$$

where V_r^k denotes the spatiotemporal saliency map computed by using the scale factor r, which is subsequently rescaled to the size of the original image frame (i.e., finest scale). The scale-invariant spatiotemporal saliency map with $R = \{1.0, 0.7, 0.4, 0.2\}$ is shown in Fig. 10. Specifically, the whole body of the child (i.e., large scale feature) is mostly detected at the coarse scale ($r = 0.2$) while all the details (i.e., small scale features) are captured at the fine scale ($r = 1.0$). By combining outputs from each scale, we can highlight the whole region of salient objects accurately regardless of their sizes through this multi-scale analysis. Thus, we confirm that the proposed method provides well-discriminative

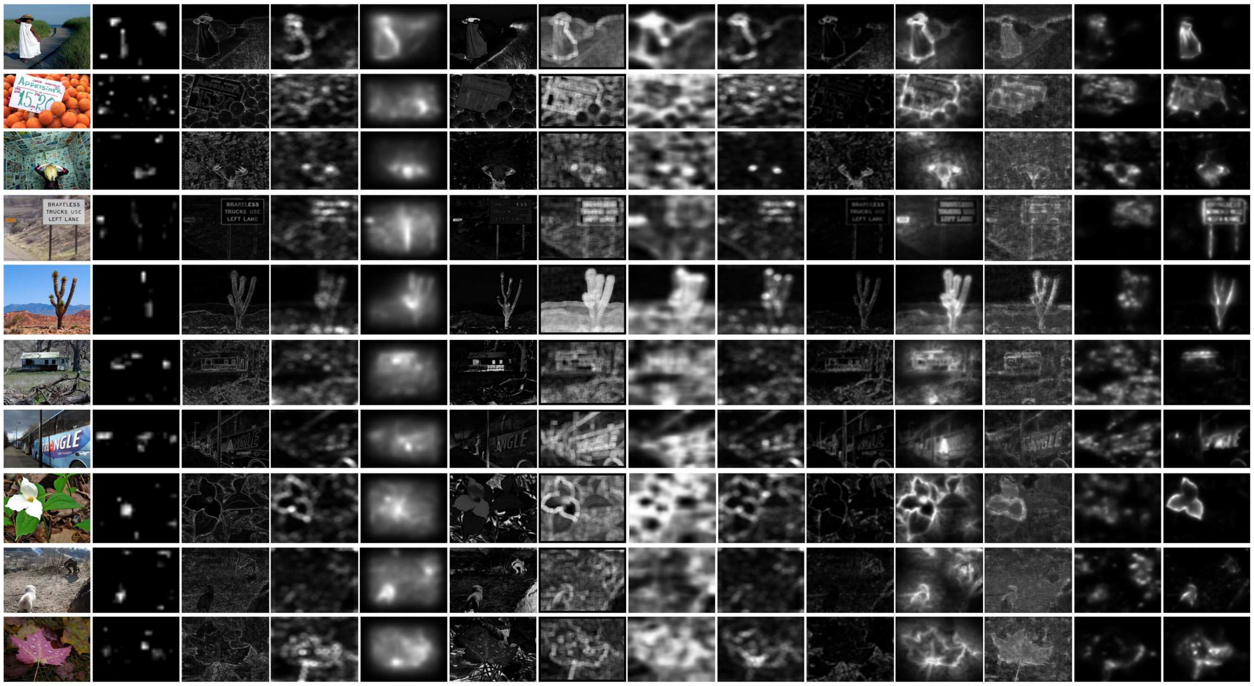


Fig. 11. (a) Input image. Saliency maps generated by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC).

TABLE I

PERFORMANCE VARIATION WITH DIFFERENT BLOCK SIZES

	3×3 pixels	5×5 pixels	7×7 pixels	9×9 pixels
F_β	0.693	0.699	0.703	0.709
sec	0.45	0.58	0.81	1.07

representation for visual saliency while suppressing the non-salient regions (e.g., cluttered and highly textured background).

IV. EXPERIMENTAL RESULTS

A. Performance evaluation in images

In this subsection, we demonstrate the performance of the proposed algorithm for static images. Our experiments were conducted on total 800 images collected from the most popularly used data sets in saliency detection tests, namely MSRA data set [26] and PASCAL VOC data set [32]. Images from both

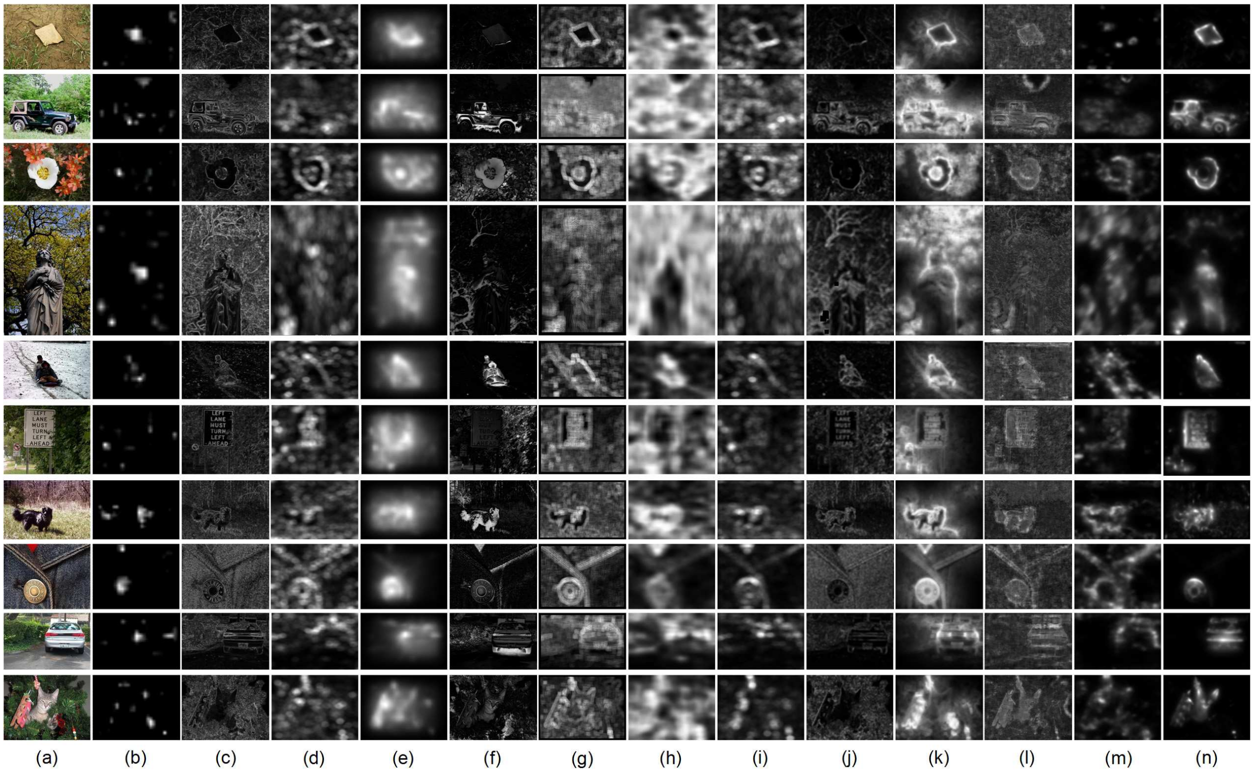


Fig. 12. (a) Input image. Saliency maps generated by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC).

data sets are taken in indoor and outdoor environments and contain a wide range of salient objects such as human, car, train, building, sign, animal, and so on. We used 5×5 pixels of the block (i.e., B_i) for computing luminance contrast and structure tensor, and four scale factors, i.e., $R = \{1.0, 0.7, 0.4, 0.2\}$ as mentioned. The performance variation according to the size of B_i is shown in Table I. Note that the computation of F-measure values will be explained in detail in the latter part of this subsection. By considering both accuracy and processing time (estimated using the image whose size is 400×300 pixels), it is thought that our basic setting (i.e., 5×5 pixels for B_i) is reasonable.

To show the superiority of our proposed method, we compared our approach (we refer to it as TC) with 12 representative models presented in literature, which are proposed by Itti [19], Ma [20], Hou [28], Harel [21], Achanta [22], Bruce [23], Seo [24], Guo [9], Xu [25], Goferman [4], Liu [26], and Kim [27]. Note that we refer to the first author of each method for simplicity. Some experimental results for saliency detection are shown in Fig. 11 and 12. From the results of previous methods, it is easy to

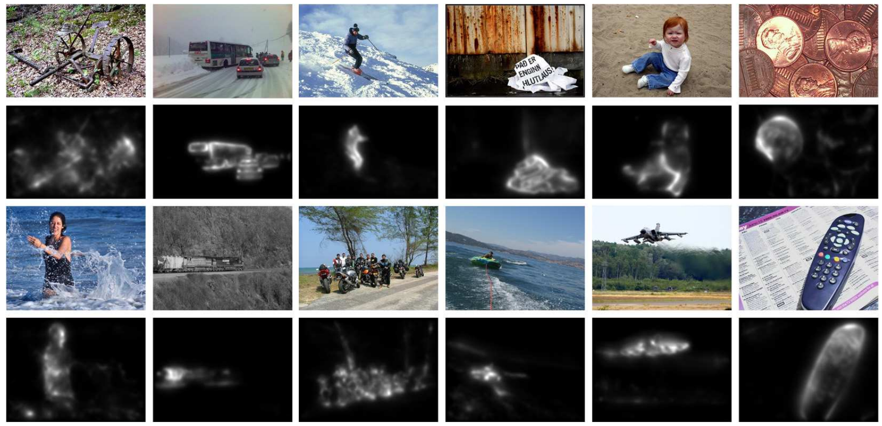


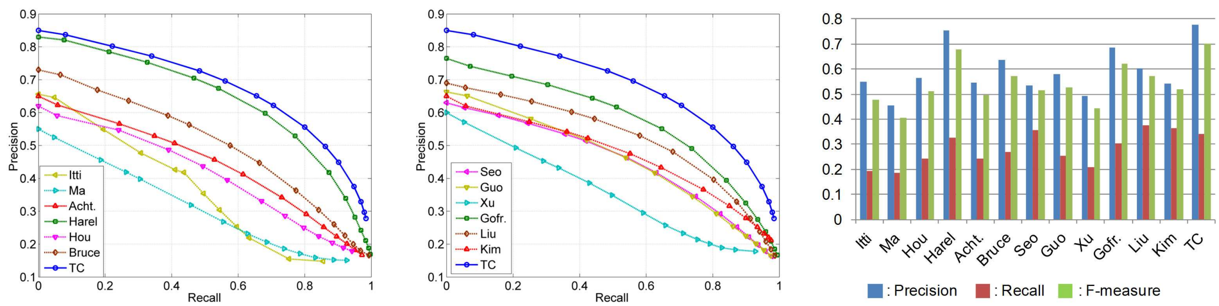
Fig. 13. More examples of saliency maps (odd rows: input images, even rows: saliency maps by the proposed method).

see that a lot false positives are generated in highly textured and cluttered backgrounds. In particular, false positives near high contrast edges in the background are hard to be eliminated by previous models. Even worse, the complicated cluttered backgrounds are more emphasized rather than the salient objects in several results. Also, those methods often fail to capture the whole region of salient objects due to complex colors and textures, which makes the further applications (e.g., image retargeting and object segmentation) unreliable. In contrast to these results, our proposed approach efficiently deals with such challenging conditions, providing visually acceptable saliency strongly coherent with the human visual attention (see Fig. 11(n) and 12(n)). More examples of saliency maps generated by the proposed method are shown in Fig. 13.

For the quantitative evaluation, we compared the binary mask for salient regions, which are obtained by thresholding our saliency map, with that of other methods. Note that the ground truth images for our image data set are manually generated. The detection accuracy is evaluated by using two quantities, i.e., recall and precision, defined as follows:

$$Recall = \frac{TP}{GT}, \quad Precision = \frac{TP}{TP + FP}, \tag{10}$$

where GT denotes the total number of the ground truth pixels in the data set. TP and FP denote the number of true positives and false positives, respectively. Based on these quantities, we plot the ROC curve varying the thresholding value with respect to the whole image data set as shown in Fig. 14(a) and (b). This curve is useful to investigate how reliably each method highlights salient regions while suppressing non-salient ones in various images. More specifically, most of previous methods are vulnerable to highly textured background, thus yielding relatively low precision values at the same recall rate as shown in



(a) (b) (c)

Fig. 14. . (a)(b) ROC curves. (c) Precision-recall bars with F-measures. Note that the proposed method shows the highest F_β values, meaning that our model accurately indicates salient regions without severe false positives.

TABLE II

PERFORMANCE COMPARISON OF PROCESSING SPEED

Method	Itti	Ma	Hou	Harel	Achanta	Bruce	Seo
speed (sec/frame)	1.22	0.62	0.01	0.93	0.03	4.07	0.75
Method	Guo	Xu	Goferman	Liu	Kim	Proposed	-
speed (sec/frame)	0.01	16.32	26.88	2.15	0.06	0.58	-

Fig. 14(a) and (b). Among them, models proposed by Harel [21], Bruce [23], Guo [9], Goferman [4], Liu [26], and Kim [27] perform quite well even though they still provide less reliable visual saliency compared to the proposed method. In contrast to that, it is easy to see that our saliency map clearly outperforms state-of-the-art algorithms. Moreover, we also computed the F-measure defined as follows:

$$F_\beta = \frac{(1+\beta^2)\cdot Precision \cdot Recall}{\beta^2 \cdot Preicsion + Recall}. \quad (11)$$

Here we use $\beta^2 = 0.3$ in our work to emphasize the precision more than recall as in [22]. It is important to note that this F-measure effectively represents the ability to suppress false positives while preserving the salient region. Based on this quantity, we efficiently compared our approach with other methods proposed in literature as shown in Fig. 14(c). As can be seen, our proposed method shows the best detection performance with the highest average values of F_β. Note that the proposed method has the slightly lower recall but has the highest precision, indicating that it is better suitable for further computer vision applications such as image retargeting, object segmentation, etc. In summary, the proposed saliency detection method has the best overall performance (on F-measure) among all the methods.

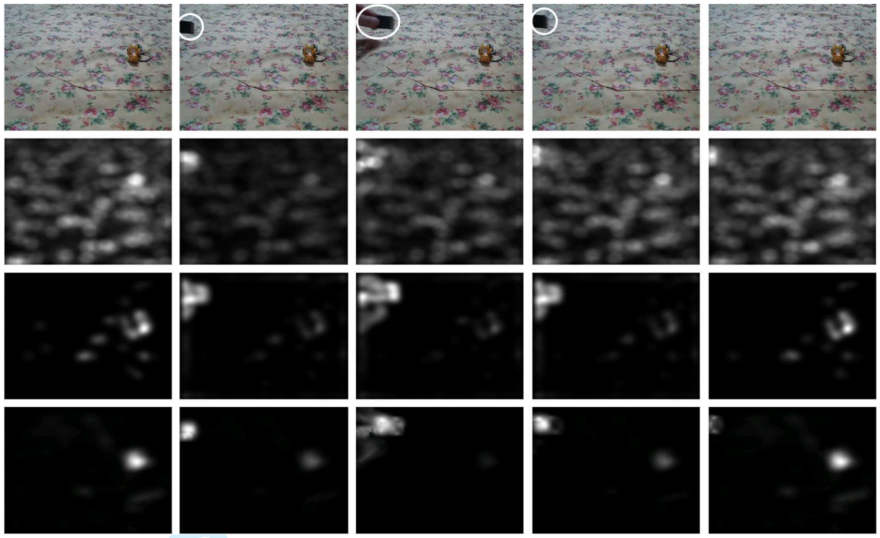
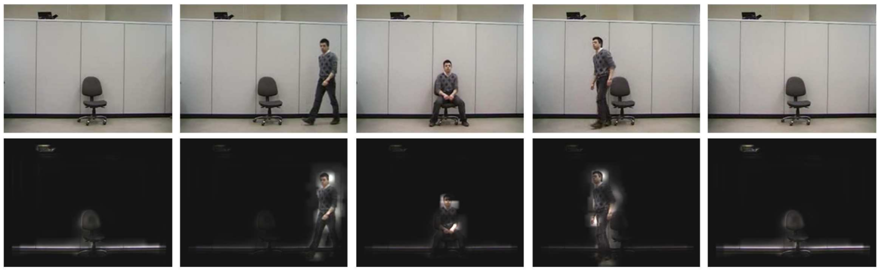


Fig. 15. Spatiotemporal saliency maps of the indoor video (the first row) generated by selected models of Guo [9] (the second row), Kim [27] (the third row), and the proposed method (the fourth row).

The framework of the proposed method has been implemented by using Visual Studio 2005 (C++) on the low-end PC (Core2Duo 3.0GHz). We compared the processing speed of our model with that of above-mentioned 12 competitive methods as shown in Table II. Note that the processing speed is averaged over a number of images of size 400×300 pixels in our database. Specifically, the processing speed of the proposed method is slightly slower than several methods such as models by Hou [28], Achanta [22], Guo [9], and Kim [27], but it provides much better detection performance. With particular regard to methods of Xu [25] and Goferman [4], the processing speed of the proposed method is a lot faster than that of their method with still higher detection accuracy. From experimental results in images, it is clearly thought that our proposed approach can provide an efficient way of building a reliable saliency map.

B. Performance evaluation in videos

To evaluate the performance of the proposed method in videos, we used various image sequences captured in both indoor and outdoor scenes respectively, which are resized to 256×196 pixels. For computing our spatiotemporal saliency map, we employ three scales, which are 128×128, 64×64, and 32×32 pixels. First, to justify the efficiency of our spatiotemporal saliency, we compared the proposed method with two spatiotemporal schemes proposed by Guo [9] and Kim [27] using a simple video obtained from [27] as shown in Fig. 15. Specifically, since there are no moving objects in the beginning part of the given video, Kim [27]'s model and our model successfully select a small doll as salient areas



(a)

(b)

Fig. 16. Spatiotemporal saliency maps generated by the proposed method. Note that the top and the bottom row of each sub-figure show the input image sequences and the corresponding saliency map, respectively.

whereas Guo [9]'s model pays attention to highly textured backgrounds, which are less salient as shown in the first column of Fig. 15. After that, all the methods capture the moving object as salient areas successfully, but highly textured backgrounds are rarely suppressed in Guo [9]'s model. It should be emphasized that the proposed method better captures the salient region (i.e., a small doll) compared to Kim [27]'s saliency model.

In addition, we also demonstrate the spatiotemporal saliency maps for more complicated videos obtained from [33] as shown in Fig. 16. Note that we highlight relevant regions with regard to values of our spatiotemporal saliency map in this example. In Fig. 16(a), a man is walking toward a chair and sits for a while. Then, he leaves his seat. In the beginning part of this video, we can say that the chair and a black bag on the cabinet attract the visual attention. When the man comes on the scene, the proposed method efficiently emphasizes the moving object until he disappears. On the other hand, our model firstly selects several parked cars, a road sign, and a building as salient regions in Fig. 16(b). When new cars enter in the parking lot, the proposed method successfully selects moving cars as the most salient areas while retaining static salient areas with relative small importance. The processing speed of the proposed method achieves averagely about 15 fps on our test videos, and it can thus be sufficiently applied for real-time applications. Based on this, it is thought that our spatiotemporal saliency map can efficiently

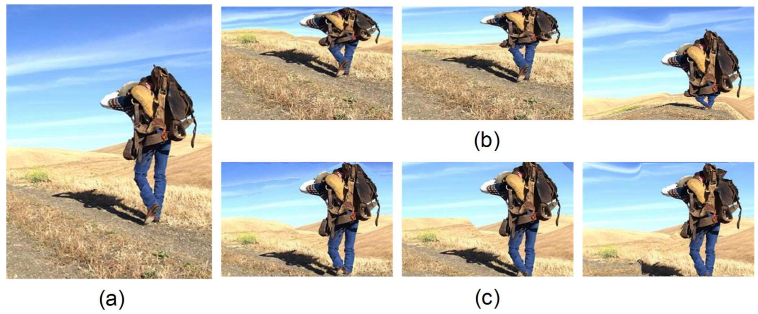


Fig. 17. (a) Input image. (b) Retargeting results by dynamic programming [34], importance diffusion [36], and fisheye-view warping [37] (from left to right) with their own importance measures. (c) Retargeting results by the proposed saliency map with resizing operators used in (b).

provide reduced search regions for object segmentation, recognition, and tracking tasks in various videos, leading to reduction of computational complexity.

V. Applications in images and videos

Owing to the outstanding ability of the proposed method in detecting salient regions as proven in the previous section, we apply our saliency map to three representative applications, i.e., image retargeting (or content-aware image resizing), object segmentation, and background subtraction in dynamic texture scenes, to show its plentiful possibilities in the field of computer vision.

A. Image retargeting

In this subsection, we introduce the most popularly adopted application, i.e., image retargeting, in which the saliency map can be employed. Image retargeting is the process of adaptively resizing a given image to fit the size of arbitrary displays based on the image importance model. For the success of image retargeting, the image importance model needs to be carefully defined since it guides further resizing procedures. To this end, L_1-norm and L_2-norm of gradient magnitude have been popularly used to measure the image importance at each pixel position [34], [35]. However, these often lead to severe distortion when the scene is cluttered or the background is complex. In order to overcome this limitation, the proposed gray-scale saliency map can be employed as a measure of the image importance since it successfully preserves the underlying image structure while suppressing the background. In Fig. 17, we apply our saliency map to various resizing operators, i.e., dynamic programming [34], importance diffusion [36], and fisheye-view warping [37]. It is easy to see that the proposed saliency map performs

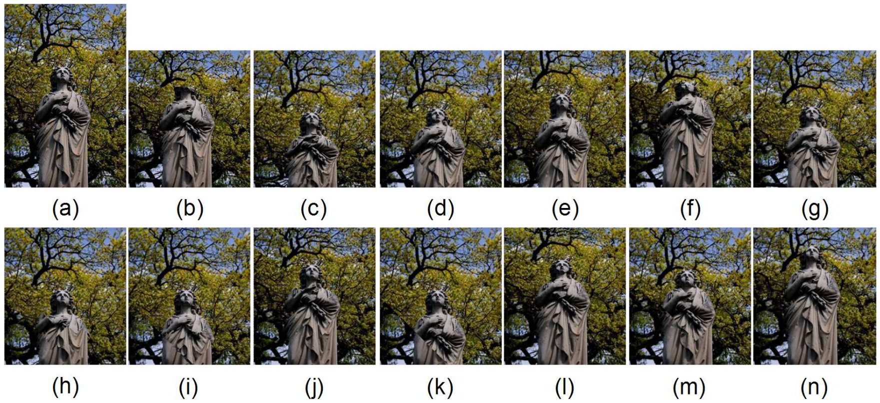


Fig. 18. (a) Input image. Results of image retargeting based on saliency models by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC). Note that the original image is reduced in the vertical direction by 100 pixels.

TABLE III

PERFORMANCE COMPARISON BY USING THE AVERAGE TARGETING RATE (TR) VALUES

Method	Itti	Ma	Hou	Harel	Achanta	Bruce	Seo
TR	0.729	0.581	0.731	0.903	0.786	0.762	0.684
Method	Guo	Xu	Goferman	Liu	Kim	Proposed	-
TR	0.687	0.662	0.834	0.888	0.845	**0.953**	-

well regardless of types of resizing operators. Note that we employ the importance diffusion [36] as a resizing operator to achieve the target sizes in our experiments.

To confirm the appropriateness of the proposed saliency map for image retargeting, we compared ours with other saliency maps generated by above-mentioned 12 methods as shown in Fig. 18 and 19. Retargeting results in the vertical direction are shown in Fig. 18. In this example, original image whose height is 400 pixels is reduced in the vertical direction by 100 pixels. Figure 19 shows the horizontal retargeting results of various images. More specifically, the image importance obtained by previous algorithms provide quite reasonable viewing effects up to a certain target height or width. However, as the reduction ratio further increases, the resizing operator starts to remove a connected path of pixels across salient objects, leading to drastic distortions in the resized image. From Fig. 18 and 19, it is easy to see that our saliency model provides visually acceptable retargeting results while preserving viewers' experience compared to other models. For the quantitative analysis, we compute the targeting rate, which

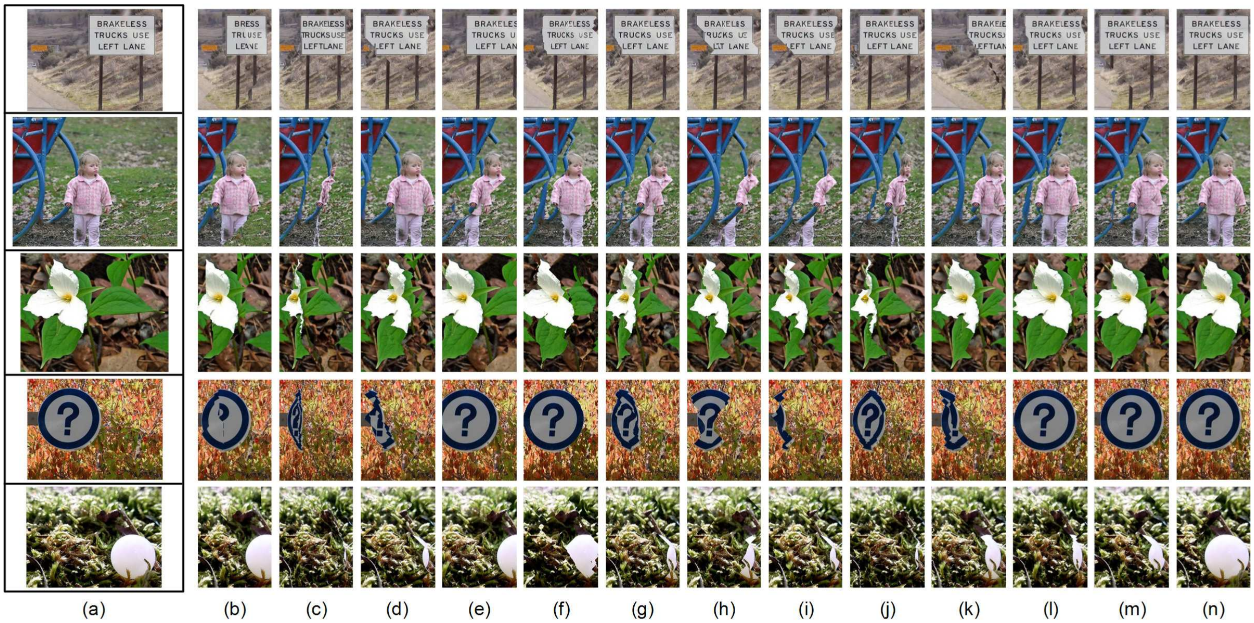


Fig. 19. (a) Input image. Results of image retargeting based on saliency models by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC). Note that each sub-figure is reduced in horizontal direction by 220, 180, 200, 180, and 180 pixels (from top to bottom).

is defined as $TR = TP/GT$ [38] where GT denotes the total number of ground truth pixels belonging to salient objects that should be preserved during the retargeting procedure and TP denotes the amount of preserved ground truth pixels. Table III shows the comparison results of the average TR values on total 100 images randomly collected from our image data set and we can see that our proposed method outperforms other previous ones in image retargeting.

B. Object segmentation

Even though there have been notable research advances for object segmentation, previous methods (e.g., graph-cut [39] and grab-cut [40] schemes) still require the user interaction for seed selection. To be a fully automatic object segmentation, the saliency map can be straightforwardly employed as a seed map since it successfully highlights pixels relevant to the foreground objects in a given scene as shown in Section IV. In this subsection, we show the utility of the proposed saliency map for object segmentation. To do this, we employ the graph-cut scheme [39], which is most widely used for object segmentation. Note that we set the threshold value to satisfy the condition that pixels having higher intensity than three times of the global mean in the saliency map are determined as seeds for foreground objects. Several

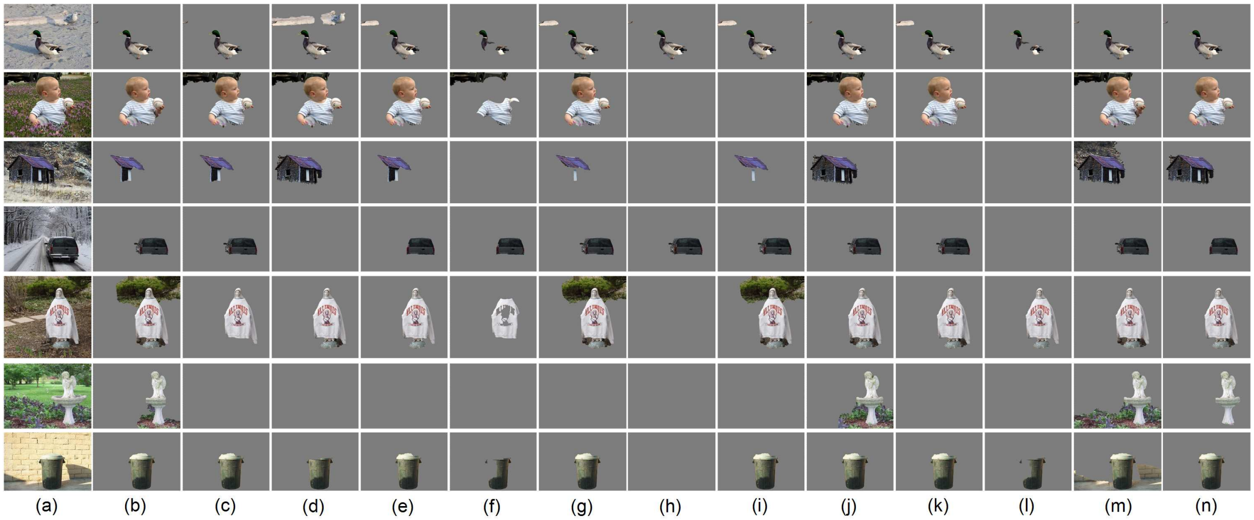


Fig. 20. (a) Input image. Results of object segmentation based on saliency models by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC).

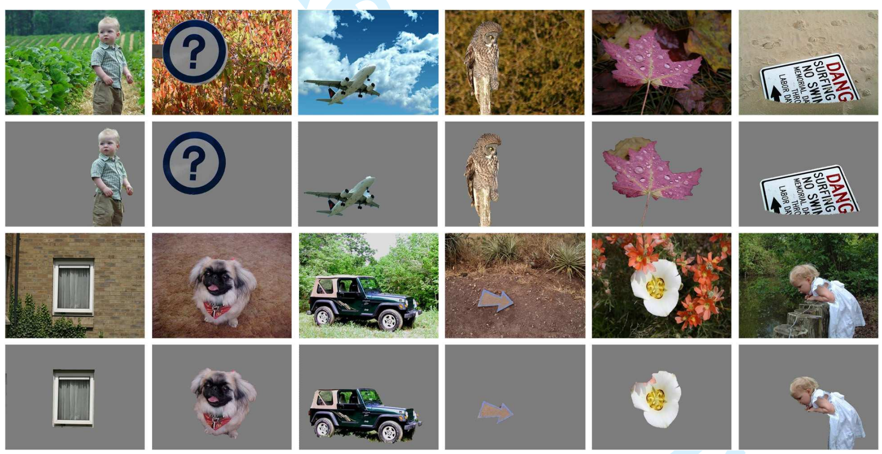


Fig. 21. More segmentation results based on our saliency maps (odd rows: input images, even rows: segmentation results by the proposed method).

segmentation results driven by various saliency maps are shown in Fig. 20. As can be seen, saliency maps generated by previous methods often lead to high-level false negative (i.e., objects are misclassified as background regions) as well as false positive (i.e., background regions are misclassified as objects) rates. In contrast to that, the proposed saliency map provides desirable segmentation results even with the complex background due to its ability of suppressing cluttered and highly textured components. In

TABLE IV

PERFORMANCE COMPARISON OF OBJECT SEGMENTATION

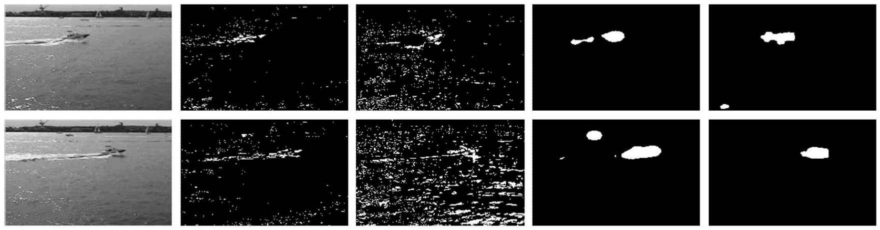
Method	Itti	Ma	Hou	Harel	Achanta	Bruce	Seo
$Recall$	0.807	0.773	0.829	0.818	0.576	0.633	0.314
Method	Guo	Xu	Goferman	Liu	Kim	Proposed	-
$Recall$	0.801	0.792	0.746	0.392	0.848	**0.882**	-
Method	Itti	Ma	Hou	Harel	Achanta	Bruce	Seo
$Precision$	0.708	0.672	0.664	0.721	0.638	0.583	0.264
Method	Guo	Xu	Goferman	Liu	Kim	Proposed	-
$Precision$	0.662	0.629	0.649	0.455	0.664	**0.738**	-
Method	Itti	Ma	Hou	Harel	Achanta	Bruce	Seo
F_β	0.754	0.721	0.737	0.766	0.606	0.608	0.287
Method	Guo	Xu	Goferman	Liu	Kim	Proposed	-
F_β	0.725	0.702	0.694	0.421	0.745	**0.803**	-

particular, our saliency map leads to the successful segmentation whereas most of previous models fail to segment the foreground (i.e., white statue) due to the cluttered background in the sixth row of Fig. 20. More segmentation results based on our saliency map are shown in Fig. 21.

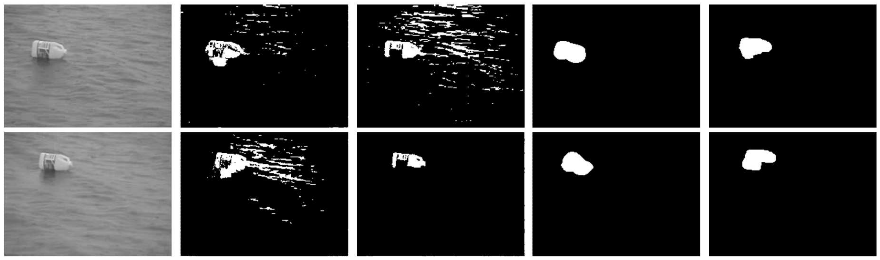
We also provide the quantitative performance comparison in Table IV. In this analysis, we adopt the previously introduced quantities, i.e., recall, precision, and F-measure (see (10) and (11)). Note that we set $\beta = 1.0$ to assign the balanced weights between recall and precision for the F-measure. From Table IV, we can see that the proposed saliency map based segmentation achieves the highest recall, precision, and F-measure. It should be emphasized again that such favorable segmentation results can be obtained since our proposed scheme performs robustly even in cluttered scenes, which fits with the human visual perception.

C. Background subtraction in dynamic texture scenes

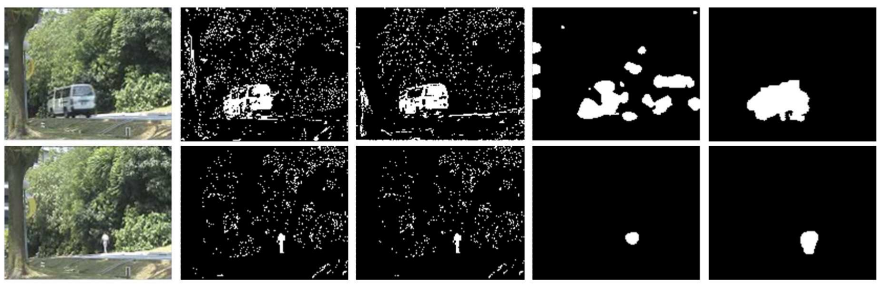
With increasing interest in high-level safety and security, video surveillance systems have been in critical demand. For the success of such systems, background subtraction, one of essential tasks in video surveillance, has been studied in various environments. However, motion patterns of the background (e.g., waving leaves, spouting fountain, rippling water, etc.), which are not static over short time periods, often cause to high-level false positives and it thus is still hard to handle them in the traditional background subtraction frameworks. In this subsection, we address this limitation through our temporal saliency



(a)



(b)

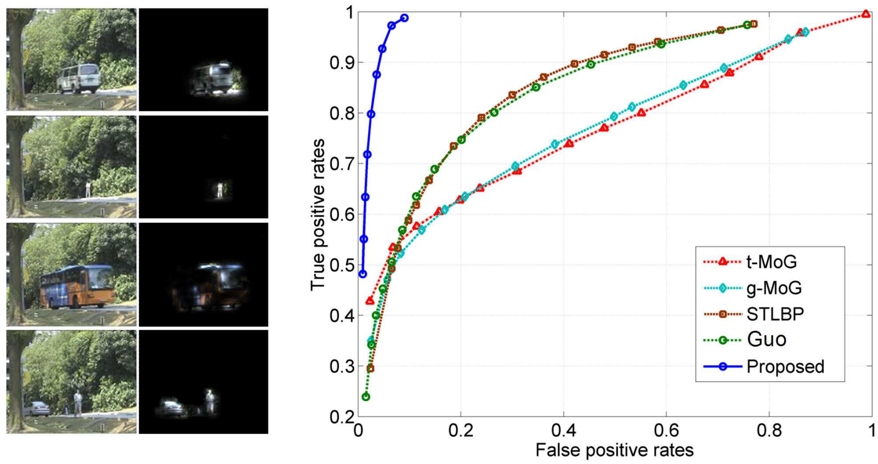


(c)

Fig. 22. 1st column: input frames of each video, i.e., (a) boat, (b) bottle, and (c) campus. Background subtraction results by t-MoG [44], g-MoG [45], Guo [9], and the proposed method (from the 2nd column to the 5th column).

detection scheme. From the visual saliency point of view, the problem of robust background subtraction narrows down to suppressing the locations having non-salient motions. The proposed temporal saliency map has several advantages over the traditional background subtraction models as follows: 1) since our scheme is based on the contrast of the temporally directional coherence, it has a great ability to suppress island-like false positives occurring in irrelevant regions and 2) the proposed method is completely unsupervised and thus does not require training and initializing the background model.

For the performance evaluation of background subtraction in dynamic texture scenes, we first tested our method and several background subtraction algorithms on three short videos, i.e., boat (180×100 pixels) [41], bottle (244×180 pixels) [42], and campus (160×128 pixels) [43] sequences, and the corresponding detection results are shown in Fig. 22. As can be seen, traditional background subtraction models (i.e., traditional mixture of Gaussian (t-MoG) model [44] and generalized mixture of Gaussian



(a) (b)

Fig. 23. (a) Input sequences (left) and some examples of the corresponding temporal saliency maps by the proposed method (right). (b) ROC curves determined by t-MoG [44], g-MoG [45], STLBP [46], Guo [9], and the proposed method.

TABLE V

PERFORMANCE COMPARISON BY FP RATES MEASURED AT TP = 0.8

Method	t-MoG [44]	g-MoG [45]	Guo [9]	Proposed
boat [41]	0.057	0.092	0.015	0.007
bottle [42]	0.027	0.058	0.004	0.006
campus [43]	0.552	0.511	0.264	0.023

(g-MoG) model [45]) yield many false positives due to rippling water and strongly waving leaves whereas the proposed temporal saliency map is quite robust to such background motions. Note that, to make fair comparisons, we report experimental results at a true positive (TP) rate of 0.8, which is good enough to correctly extract moving objects for further applications. We also plot the ROC curve using the campus sequence in Fig. 23. Note that we include the method based on spatiotemporal local binary patterns (STLBP) [46] in the ROC curve. For the quantitative evaluation, the false positive (FP) rates, which are obtained from 20 frames randomly selected from each video, are shown in Table V. From Fig. 22, 23, and Table V, we confirm that the proposed saliency-based scheme is capable of correctly extracting moving objects with low false positive rates in dynamic texture scenes.

VI. Conclusion

A novel method for detecting salient regions in the spatiotemporal domain has been proposed in this paper. The key idea of the proposed method is that the biological mechanism of the bottom-up visual attention can be approximated by exploiting two main contrasts captured in the retina and the visual cortex. To this end, we propose to use textural contrast defined based on combination of luminance contrast and directional coherence contrast, and extend its concept to the spatiotemporal domain with temporal gradients. By incorporating the responses of the proposed contrast mechanism into a multiscale framework, we can generate reliable saliency maps. Based on extensive experimental results, we confirm that the proposed method efficiently highlights relevant regions in images and videos even with the cluttered background. Furthermore, to show the plentiful possibilities of the visual saliency, we apply the proposed saliency map to various real-world applications such as image retargeting, automatic object segmentation, background subtraction in dynamic texture scenes. From the comparison results, it is thought that the proposed method is effective enough to be applicable to various vision-based intelligent applications.

References

[1] R. VanRullen, "Visual saliency and spike timing in the ventral visual pathway," *Journal of Physiology*, vol. 97, pp. 365–377, 2003.

[2] O. S. Packer and D. M. Dacey, "Synergistic center-surround receptive field model of monkey H1 horizontal cells," *Journal of Vision*, vol. 5, pp. 1038–1054, 2005.

[3] R. Achanta and S. Susstrunk, "Saliency detection for content-aware image resizing," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 1005–1008, Nov. 2009.

[4] S. Goferman, L. Z. Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2376–2383, June 2010.

[5] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 446–456, Apr. 2011.

[6] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, pp. 1–4, Dec. 2008.

[7] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): meaningful attention using stochastic image modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 693–708, Apr. 2010.

[8] O. L. Meur and J. -C. Chevet, "Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2801–2813, Nov. 2010.

[9] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[10] S. Marat, M. Guironnet, and D. Pellerin, "Video summarization using a visual attention model," in *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 1784–1788, 2007.

[11] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality ? Applying to image quality metric," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 2, pp. 169–172, Oct. 2007.

[12] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency inspired full-reference quality metric for packet-loss-impaired video," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 81–88, Mar. 2011.

[13] H. Liu and I. Heynderickx, "Visual attention in object image quality assessment: based on eyetracking data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, July 2011.

[14] A, Toet, "Computational versus psychophysical bottom-up image saliency: a comparative evaluation study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.

[15] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 1, pp. 253–256, Sept. 2003.

[16] H. Li and K. N. Ngan, "Saliency model-based face segmentation and 646 tracking in head-and-shoulder video sequences," *Journal of Visual Communication and Image Representation*, vol. 19, no. 5, pp. 320–333, July 2008.

[17] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 2008.

[18] W. Kim and C. Kim, "Saliency detection via textural contrast," *Optics Letters*, vol. 37, no. 9, pp. 1550–1552, May 2012.

[19] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no, 11, pp. 1254–1259, Nov. 1998.

[20] Y. F. Ma and H. J . Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM International Conference on Multimedia*, pp. 374–381, 2003.

[21] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 19, pp. 545–552, 2007.

[22] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1597–1604, June 2009.

[23] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: an information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 2009.

[24] H. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 8, no. 12, pp. 1–27, 2009.

[25] Y. Xu, Y. Zhao, C. Jin, J. Qu, L. Liu, and X. Sun, "Salient target detection based on pseudo-Wigner-Ville distribution and Renyi entropy," *Optics Letters*, vol. 35, pp. 475–477, 2010.

[26] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. -Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[27] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 446–456, Apr. 2011.

[28] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, June 2007.

[29] P. Bian and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency advances in neuro-information processing," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 251–258. 2009.

[30] A. B. Chan, V. Mahadevan, and N. Vasconcelos, "Generalized Stauffer- Grimson background subtraction for dynamic scenes," *Machine Vision and Applications*, vol. 22, no. 5, pp. 751–766, 2011.

[31] L. Wolf, M. Guttmann, and D. Cohen-Or, “Non-homogeneous contentdriven video retargeting,” in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 1–6. Oct. 2007.

[32] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A, Zisserman, “The PASCAL visual object classes (VOC) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[33] http://imagelab.ing.unimore.it/visor/index.asp.

[34] S. Avidan and A. Shamir, “Seam carving for content-aware image resizing,” *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 267–276, 2007.

[35] M. Rubinstein, A. Shamir, and S. Avidan, “Improved seam carving for video retargeting,” *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 1–8, 2008.

[36] S. Cho, H. Choi, Y. Matsushita, S. Lee, “Image retargeting using importance diffusion,” in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 997–980, Nov. 2009.

[37] F. Liu and M. Gleicher, “Automatic image retargeting with fisheye-view warping,” in *ACM Annual Symp. on User Interface Software and Technology*, pp. 153–162, 2005.

[38] W. Kim and C. Kim, “A texture-aware salient edge model for image retargeting,” *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 631–634, Nov. 2011.

[39] Y. Boykov and G. Funka-Lea, “Graph cuts and efficient N-D image segmentation,” *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006.

[40] C. Rother, V. Kolmogorov, and A. Blake, “GrabCut-Interactive foreground extraction using iterated graph cuts,” in *Proc. ACM SIGGRAPH*, pp. 309–314, 2004.

[41] A. B. Chan, V. Mahadevan, and N. Vasconcelos, “Generalized Stauffer-Grimson background subtraction for dynamic scenes,” *Machine Vision and Applications*, vol. 22, no. 5, pp. 751–766, 2011.

[42] V. Mahadevan and N. Vasconcelos, “Background subtraction in highly dynamic scenes,” in *Proc. Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–6, June 2008.

[43] L. Li, W. Huang, I. Y. -H. Gu, and Q. Tian, “Statistical modeling of complex backgrounds for foreground object detection,” *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.

[44] C. Stauffer and W. E. L. Grimson, “Adaptive background mixture models for real-time tracking,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 246–252, June 1999.

[45] G. Dalley, J. Migdal, and W. Grimson, “Background subtraction for temporally irregular dynamic textures,” in *Proc. IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 1–7, Jan. 2008.

[46] S. Zhang, H. Yao, and S. Liu, “Dynamic background modeling and subtraction using spatio-temporal local binary patterns,” in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 1556–1559, Oct. 2008.

Spatiotemporal Saliency Detection Using Textural Contrast and Its Applications

Wonjun Kim, *Student Member, IEEE,* and Changick Kim, *Senior Member, IEEE*.

***Abstract*—Saliency detection has been extensively studied due to its great possibilities for various computer vision applications. However, most existing methods are easily biased toward edges or corners, which are statistically significant, but not necessarily salient. Moreover, they often fail to find salient regions in complex scenes due to ambiguities between salient regions and highly textured backgrounds. In this paper, we present a novel unified framework for spatiotemporal saliency detection based on *textural contrast*. Our method is simple, robust, yet biologically plausible and it can thus be easily extended to various applications such as image retargeting, object segmentation, and video surveillance. Based on various data sets, we conduct comparative evaluations of 12 representative saliency detection models presented in literature, and the results show that the proposed scheme outperforms other previously developed methods in detecting salient regions of the static and dynamic scenes.**

***Index Terms*—Saliency detection, computer vision applications, human visual attention, textural contrast, comparative evaluations.**

I. Introduction

THE human visual system (HVS) has an outstanding ability to quickly grasp the most relevant regions at a glance without any prior knowledge. Therefore, we can easily understand contextual information of a given scene based on this selective visual attention in an efficient manner. There are numerous factors contributing such visual saliency. Among them, as reported by many biological experiments, the most important factor is contrast [1]. That is, the relevant element is not the absolute amplitude of visual signals (e.g., intensity, color, etc.) but rather contrast between these amplitudes at a given point and its surroundings. The importance of contrast has been strongly supported by a meaningful result in cognitive neuroscience, showing that the receptive field of the retina is performed based on the center-surround cell network (i.e., center-surround contrast) in which cone synaptic input is fed into the center of the receptive field via the dendritic tree and the second input is provided by excitatory surrounds using gap junction between cells [2]. Therefore, computational modeling of this biological system enables various applications (e.g., content-aware image resizing [3], [4], [27], object detection and segmentation [6], [7], [8], adaptive image and video compression [9], video summarization [10], image quality assessment [11], [12], [13], video surveillance [27], and so on), requiring only limited processing resources. For this reason, saliency detection has been extensively studied by researchers in psychology, cognitive neuroscience, and computer vision.

Wonjun Kim and Changick Kim are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea. E-mail: {jazznova, changick}@kaist.ac.kr

In the field of computer vision, many computational models have been proposed to accomplish this task automatically, and the comprehensive survey on recent developments is also found in [14]. According to the literature review, most of previous methods can be divided into two major groups, i.e., top-down and bottom-up approaches. First of all, the top-down approaches are task-driven and can thus be regarded as solving the problem of visual recognition [15], [16]. In this category, salient visual attributes are defined as descriptors delineating specific objects, such as face, text, etc., for the given task. These approaches mostly require prior knowledge, which is not available in every image, and it thus leads to hard generalization. The majority of saliency detection methods are driven by the biological plausibility of the bottom-up mechanisms. More specifically, most of bottom-up approaches have been proposed based on a set of simple low level features such as luminance, color, and orientation, followed by some center-surround operations. Again, this is because local image features become stimuli of interest when they are best distinguishable from its surroundings that may be of possible interest (it is referred to as the discriminant center-surround hypothesis) [17].

In this paper, we introduce a novel unified framework for detecting salient regions in both images and videos. The key idea behind our approach is to mimic the biological system by formulating two main contrast mechanisms occurring in the retina and the visual cortex. Specifically, we propose to use *textural contrast* defined as the combination of luminance contrast (for retina level) and directional coherence contrast (for visual cortex level), and extend its concept to the spatiotemporal domain with temporal gradients. By incorporating the responses of *textural contrast* into a multiscale framework, we can generate reliable saliency maps for images and videos. Compared to traditional bottom-up models, one important advantage of the proposed method is that it greatly eliminates unwanted fine details whereas highlighting salient regions quite uniformly owing to the ability of providing the contextual information regarding underlying image structures. Note that this work is extended from our previous one [18] and differs in the following respects: 1) we provide more technical details about the appropriateness of *textural contrast* for saliency detection; 2) we extend the concept of the directional coherence presented in [18] to detect saliency from videos as well; 3) we provide comparative evaluations of 12 representative saliency detection models both qualitatively and quantitatively. Moreover, various saliency-inspired applications are also demonstrated.

The rest of this paper is organized as follows. A systematic review of previous bottom-up methods is presented in Section

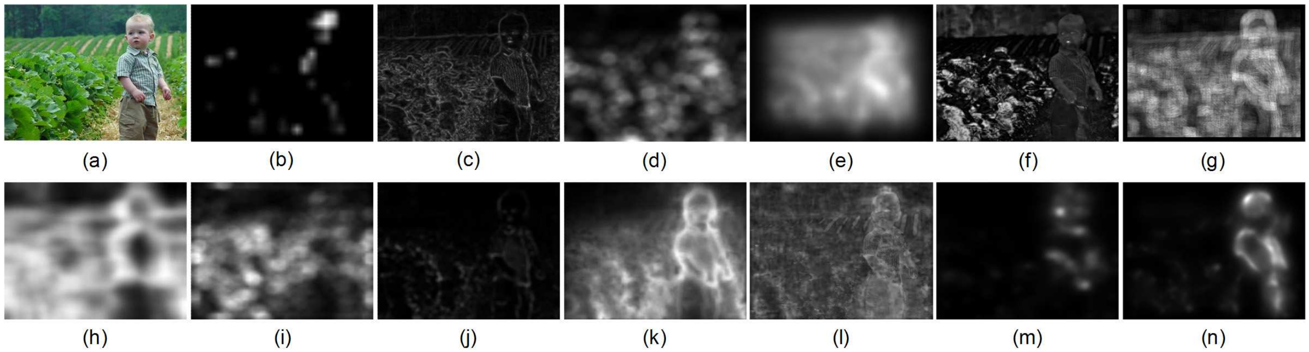


Fig. 1. (a) Input image. Saliency maps generated by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method. Note that pixels in high intensities are highly likely to be salient. For simplicity, we refer to the first author of each method.

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 images and videos are introduced in Section V. Conclusion follows in Section VI.

II. A Review of Bottom-up Saliency Detection Models

In this section, we briefly review several bottom-up models, which are representative in literature, and discuss about their limitations. The main advantage of these models lies in data-driven nature, i.e., do not require any prior knowledge. Most of bottom-up approaches can be further divided into two categories: statistical and spectral methods. Statistical methods are fundamentally based on the center-surround hypothesis. Specifically, they mostly adopt the difference between feature statistics obtained from center and surrounding regions as a measure of saliency. The first computational and statistical model is developed in a center-surround framework by Itti *et al.* [19]. Their saliency map is generated based on the linear combination of normalized feature maps obtained from three basic components: intensity, color, and orientation. Inspired by their success in predicting human fixations, several models, more or less based on different mathematical tools, have been investigated in literature. Ma and Zhang [20] compute the distance between Lab color features obtained from center and surrounding regions on the quantized block image. Harel *et al.* [21] define the graph using pixel positions and weight values proportional to their dissimilarity obtained from orientation, intensity, and its variation. The resulting graphs are treated as Markov chains and their equilibrium distribution is adopted as saliency maps. Achanta *et al.* [22] formulate the problem of detecting salient regions as conducting the band pass filtering, which is simply implemented by computing the difference between mean of colors over the whole image and Gaussian blurred version of the original image. Bruce and Tsotsos [23] propose to employ Shannon's self-information measure and adopt the independent component analysis (ICA) to efficiently estimate the one-dimensional probability density function. In [24], authors compute statistical likelihood of the feature response from each pixel to those of surrounding regions as a measure of saliency. To consider the local structure more precisely, they propose to use the local steering kernel estimated from a collection of spatial gradient vectors. Xu *et al.* [25] propose to utilize the spatial-frequency information to be robust to the complex background. They compute the residual of the Renyi entropy via the pseudo-Wigner-Ville distribution for finding salient regions. Goferman *et al.* [4] incorporate positional information into color contrast between image patches for detecting salient regions. Liu *et al.* [26] propose a set of features including multiscale contrast, center-surround histogram, and color spatial distribution to describe a salient object locally, regionally, and globally. A conditional random field is employed to efficiently combine these features. Authors of [27] exploit ordinal signatures of the feature distribution. The rationale behind this method is that ordinal signatures are robust to small variations occurring in the feature distribution and thus the difference of them between center and surrounding regions indicates salient locations even under highly textured backgrounds. They also propose a framework for spatiotemporal saliency detection by involving the sum of difference obtained from temporal gradients.

On the other hand, spectral methods also have been constantly proposed. These methods attempt to efficiently eliminate background by analyzing filter responses in the frequency domain based on the assumption that less periodicity makes the rare event (i.e., saliency) on the corresponding location in the reconstruction of the original image. It should be emphasized that spectral methods are highly correlated with the human vision mechanism, which is able to grasp salient regions at a glance, since they promptly work in a global view. Hou and Zhang [28] firstly introduce global contrast in the frequency domain to detect salient regions by using spectral residual, which is simply defined by subtracting a smoothed version of the log magnitude spectrum from the original one. However, it is well known that what actually locates saliency is the phase information, rather than the magnitude information. In this sense, Bian and Zhang [29] normalize Fourier coefficients with respect to their magnitudes and only use the phase information to find salient regions. Similarly, Guo and Zhang [9] build the spatiotemporal saliency map using the phase spectrum of the quaternion Fourier transform (PQFT), which is composed of color, intensity, and temporal gradients.

Even though such bottom-up models have been extensively studied, they still suffer from two main limitations: 1) a bias toward edges or corners and 2) vulnerability to cluttered and highly textured background. These limitations are illustrated in Fig. 1. Specifically, previous methods tend to emphasize only high contrast edges and thus easily fail to capture the whole regions of saliency. They also tend to highlight cluttered background rather than salient regions and thus the saliency maps by previous models are expected to be highly unreliable. To tackle these limitations, we propose to use *textural contrast* for finding salient regions. Moreover, we provide a novel unified framework for spatiotemporal saliency detection by involving directional coherence contrast of temporal gradients. In the following, we will explain the proposed spatiotemporal saliency detection scheme and its excellence in detail.

III. Proposed Method

In general, the brain and the vision systems work together to identify relevant regions in a given scene. We aim to model this biological system focusing on the main visual stream from the retina to the visual cortex. The first type of information captured by our visual system in the retina is luminance contrast. At higher levels of processing in the visual cortex, orientation contrast is involved to understand the context. It is important to note that the conjunction of luminance contrast and orientation contrast makes the corresponding region to be more salient than using either of them separately [1]. Motivated by this fact, we attempt to model such biological mechanisms using *textural contrast* defined by allowing for both luminance contrast and directional coherence contrast. In particular, we propose to exploit the directional coherence to estimate orientation contrast since the use of gradient information in a pixel-wise manner often leads to failure in describing the underlying image structure, especially in cluttered and highly textured regions. We also extend the concept of the directional coherence to the temporal domain for spatiotemporal saliency detection.

A. Spatial saliency by textural contrast

First of all, we define luminance contrast by considering how distinctive the intensity attribute of each pixel is compared to the global one. For the improved dynamic ranges useful for effectively suppressing high contrast in the background, the n-th order statistics are applied as follows:

$$S_L^k(i) = \left| \bar{I}^k - \frac{1}{N} \sum_{j \in B_i} I^k(j) \right|^n , \quad (1)$$

where k denotes the frame index. $\bar{I}^k$ denotes the mean of luminance values over the whole image (i.e., the largest surrounding region). B_i and N represent the neighbor region (5×5 pixels in our implementation) centered at the i^{th} pixel position and its size, respectively. The luminance contrast maps generated by using various n values are shown in Fig. 2. From exhaustive experiments, it is carefully observed that the second-order moment (i.e., $n = 2$) yields the best results in

Fig. 2. (a) Input image, (b) first-order model, (c) second-order model, and (d) fourth-order model.

Fig. 3. More examples for luminance contrast maps. (a) Input image, (b) first-order model, (c) second-order model, and (d) fourth-order model.

suppressing irrelevant regions while sufficiently emphasizing the salient region. More examples are shown in Fig. 3.

Along with luminance contrast, we also attempt to depict the local image structure based on directional coherence contrast obtained from center and surrounding regions. It is important to note that we focus on directional coherence rather than directly using gradient information, which is unreliable in cluttered and highly textured regions. In detail, directional patterns in center and surrounding regions provide a good approximation to the underlying image structure, which is indeed coherent with the visual attention. To do this, we allow for the structure tensor, which efficiently summarizes the dominant orientation and the energy along this direction based on the local gradient field, defined as follows:

$$\mathbf{T}_s^k(i) = \begin{bmatrix} \sum_{j \in B_i} I_x^k(j)^2 & \sum_{j \in B_i} I_x^k(j) I_y^k(j) \\ \sum_{j \in B_i} I_x^k(j) I_y^k(j) & \sum_{j \in B_i} I_y^k(j)^2 \end{bmatrix} , \quad (2)$$

where I_x^k and I_y^k denote the gradient in horizontal and vertical directions at the k^{th} frame, respectively. The usefulness of the structure tensor defined in (2) for our task stems from the fact that the relative discrepancy between two eigenvalues (i.e., $\lambda_1 \geq \lambda_2 \geq 0$) of $\mathbf{T}_s^k(i)$ indicates how intensively gradients in the local region are distributed along the dominant direction (i.e., the degree to which those directions are consistent). For better understanding, we illustrate the distributions of gradients obtained from selected image patches as shown in Fig. 4. As can be seen, the gradients belonging to the textural boundary attracting the visual attention (①) are intensively distributed along the dominant direction compared to those of the highly

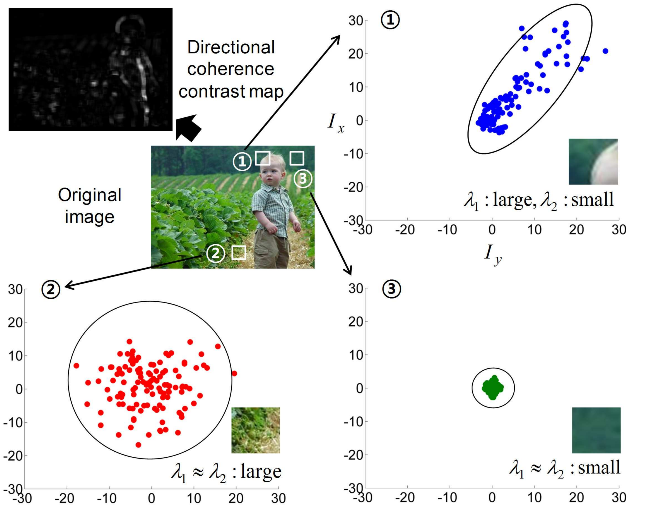


Fig. 4. Gradients obtained from selected image patches are illustrated. Note that λ_1 and λ_2 represent the energy along the dominant orientation of the gradient field and its perpendicular direction, respectively.

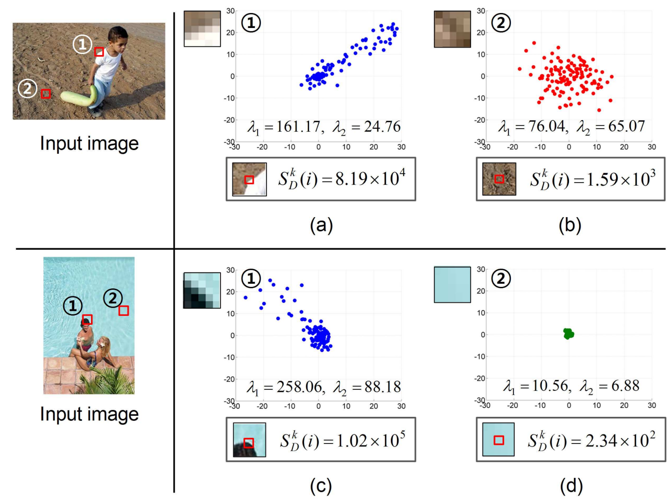


Fig. 5. Directional coherence contrast obtained from (a)(c) salient region, (b) highly textured region, and (d) flat region. Note that $S_D^k(i)$ values from (a) and (c) are much larger than those of (b) and (d) (i.e., irrelevant regions).

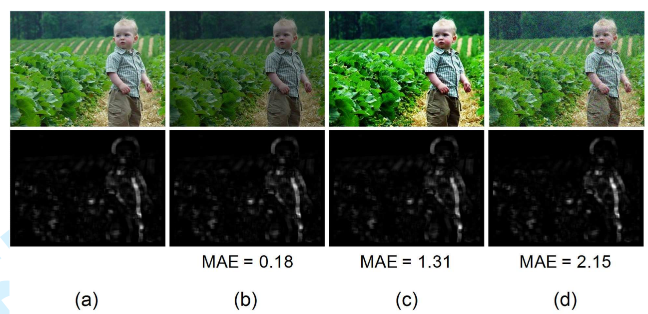


Fig. 6. Directional coherence contrast maps in challenging conditions. (a) Original image, (b) brightness change, (c) contrast change, and (d) white Gaussian noise (0, 0.01). Note that the number marked below each sub-figure denotes the average MAE value, which is obtained from the comparison with the directional coherence contrast map of (a).

textured region (②) or the uniformly textured (i.e., flat) region (③). Thus, we define our directional coherence at each pixel position as follows:

$$\xi = (\lambda_1 - \lambda_2)^2. \tag{3}$$

Here the larger the value ξ is, the higher the directional coherence is. Note that the average of gradients does not guarantee the reliable measure since aligned but oppositely oriented gradients would cancel out in this average. In what follows, directional coherence contrast between center and surrounding regions can be formulated as follows:

$$S_D^k(i) = \sum_{j \in W_i} |\xi^k(j) - \xi^k(i)|, \tag{4}$$

where W_i is a set of neighboring pixels centered at the i^{th} pixel position. Note that the size of W_i is set to 7×7 pixels in our implementation. An example of the directional coherence contrast map (i.e., the gray-scale representation of $S_D^k(i)$) is shown in Fig. 4. More examples for the directional coherence contrast are shown in Fig. 5. We confirm that salient regions yield quite large values compared to irrelevant regions. We also demonstrate some examples of the directional coherence maps in various challenging conditions in Fig. 6. More specifically, those maps provide the reliable image structure even with the drastic change of brightness and contrast (see Fig. 6(b) and (c)). In addition to this, directional coherence contrast is quite robust to the presence of noise (see Fig. 6(d)). For generality, we compute the average MAE (mean absolute error) values obtained from over 500 natural images (see the number marked below each sub-figure in Fig. 6). Note that small MAE values show that directional coherence contrast between center and surrounding regions is invariant to a wide range of variations. Therefore, it is thought that directional coherence contrast is highly desirable to measure visual saliency.

Since salient regions are assumed to contain both luminance contrast and directional coherence contrast as mentioned, our spatial saliency map at the k^{th} frame is thus computed by combining such two responses as follows:

$$S^k(i) = S_L^k(i) \times S_D^k(i). \tag{5}$$

Here each response is smoothed by Gaussian filtering as in [28] and $S^k(i)$ is normalized to [0,255] for gray-scale representation. It is worth noting that the combination strategy defined in (5) produces more desirable saliency maps while effectively suppressing false positives in the background. This is because either of two responses may be high in the irrelevant region. The example of our spatial saliency map at the single scale is shown in Fig. 7.

B. Combining with temporal saliency

For the spatiotemporal saliency detection, we need to involve motion stimuli, which can be defined by spatiotemporal orientation (equivalent to the velocity [17]). To compute motion contrast (i.e., motion energy associated with different velocities) strongly attracting the visual attention in videos, we propose to apply the concept of the directional coherence to temporal gradients. First of all, the structure tensor of temporal

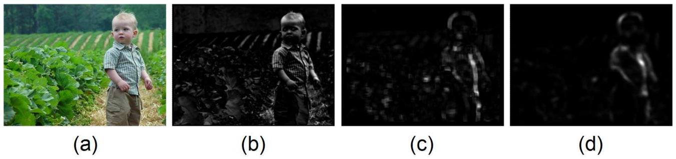


Fig. 7. (a) Original image, (b) luminance contrast map, (c) directional coherence contrast map, and (d) our spatial saliency map (single scale).

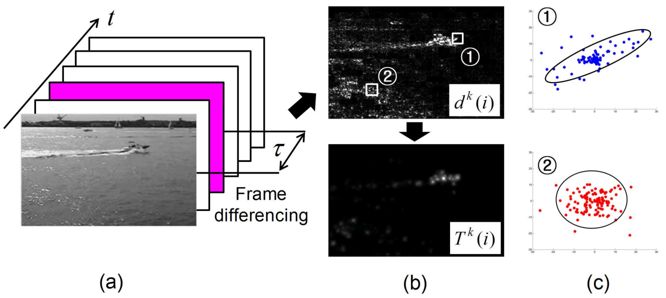


Fig. 8. (a) Input image sequence, (b) results of frame differencing (i.e., temporal gradient, $d^k(i)$) (top) and the proposed temporal saliency map (bottom), and (c) gradient distributions for selected image patches from $d^k(i)$.

gradients can be represented similarly with (2) as follows:

$$\mathbf{T}_t^k(i) = \begin{bmatrix} \sum_{j\in B_i} d_x^k(j)^2 & \sum_{j\in B_i} d_x^k(j)d_y^k(j) \\ \sum_{j\in B_i} d_x^k(j)d_y^k(j) & \sum_{j\in B_i} d_y^k(j)^2 \end{bmatrix}, \quad (6)$$

where $d^k(i) = I^k(i) - I^{k-\tau}(i)$ ($\tau = 3$ in our work). Based on this, the temporally directional coherence can be defined by using the difference between two eigenvalues, λ_1 and λ_2, of $\mathbf{T}_t^k(i)$, i.e., $\phi = (\lambda_1 - \lambda_2)^2$. In what follows, we adopt contrast of the temporally directional coherence as the measure of temporal saliency as follows:

$$T^k(i) = \sum_{j\in W_i} |\phi^k(j) - \phi^k(i)|, \quad (7)$$

where W_i is a set of neighboring pixels centered at the i^{th} pixel position as mentioned before. The overall procedure for generating the temporal saliency map is shown in Fig. 8. It is worth noting that our approach for temporal saliency detection has a great ability to suppress irrelevant motions (e.g., rippling water) occurring in the background while still highlighting a region of interest (e.g., a moving boat). This is because temporal gradients by irrelevant motions are generally unstructured in a local regions (see Fig. 8(b)) and they thus yield low contrast of the temporally directional coherence. Moreover, we compare ours with the center-surround temporal gradient patterns, i.e., $\sum_{j\in W_i} |d^k(j) - d^k(i)|$, and results are shown in Fig. 9. We can see that the center-surround temporal gradient patterns often fail to suppress a snowfall in the background whereas the tensor-based analysis allows the proposed temporal saliency to be more closely correlated with visual attention.

Finally, the proposed spatiotemporal saliency map at the single scale can be defined by the combination of the spatial and temporal saliencies, i.e., $S^k(i)$ and $T^k(i)$, as follows:

$$V^k(i) = \alpha \cdot S^k(i) + (1-\alpha) \cdot T^k(i), \quad (8)$$

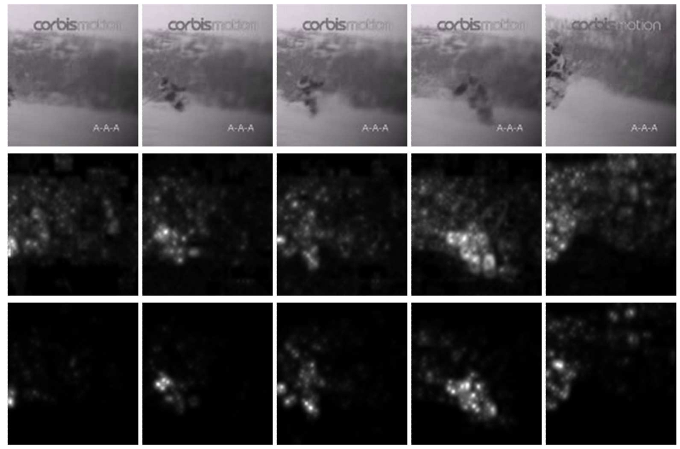


Fig. 9. Performance comparison between the center-surround temporal gradient patterns and the proposed temporal saliency. Ski sequences obtained from [41] (top), results of the center-surround temporal gradient patterns (middle), and our temporal saliency (bottom).

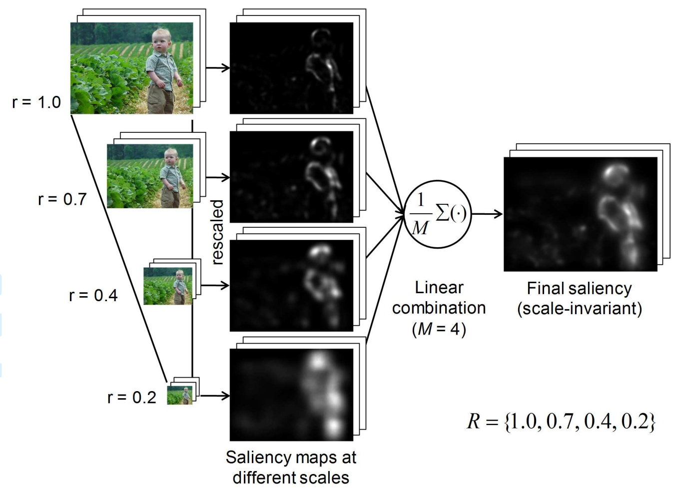


Fig. 10. Scale-invariant spatiotemporal saliency map. Note that the saliency map computed at each scale is resized to the size of the original image.

where $\alpha \in [0, 1]$ denotes the weighting factor for balancing between the spatial and temporal saliency. In general view, since moving objects are more attractable than static objects and backgrounds in videos [31], we set α to 0.3 (i.e., weigh temporal saliency more than spatial one) in our implementation. For the grey-scale representation, the spatiotemporal saliency values defined in (8) are normalized from 0 to 255.

C. Scale-invariant spatiotemporal saliency map

Since the size of salient regions is unknown, saliency maps are usually built on the combination of outputs from different scales [19], [21], [9], [4], [27]. Specifically, let $R = \{r_1, r_2, \cdots, r_M\}$ denote the set of scales to be considered to conduct the multiscale analysis. Note that we treat all image levels equally by taking them into account in a unified solution since no level is more important than others in HVS. Therefore, the scale-invariant spatiotemporal saliency map is finally computed by the linear combination of outputs obtained

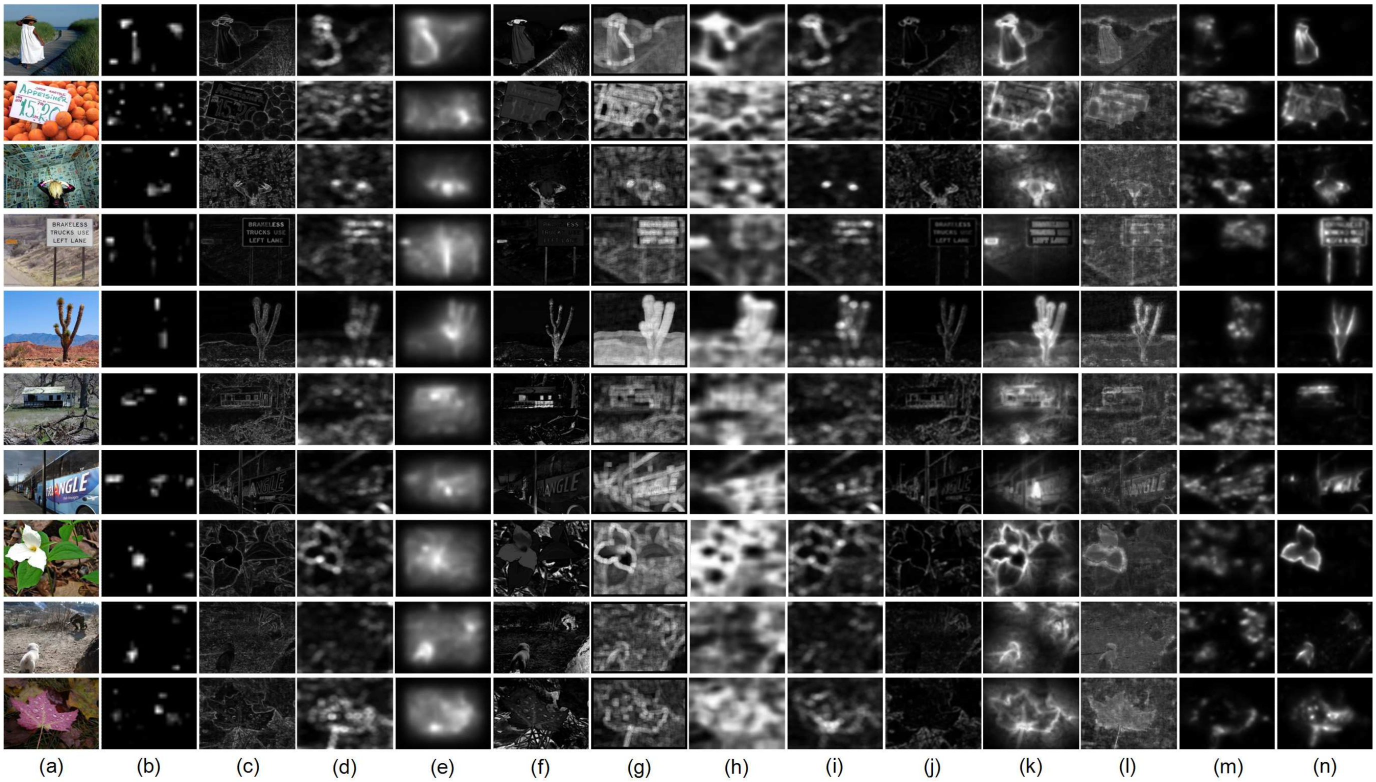


Fig. 11. (a) Input image. Saliency maps generated by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC).

from each scale with the same weight as follows:

$$\tilde{V}^k(i) = \frac{1}{M} \sum_{r \in R} V_r^k(i), \tag{9}$$

where V_r^k denotes the spatiotemporal saliency map computed by using the scale factor r, which is subsequently rescaled to the size of the original image frame (i.e., finest scale). The scale-invariant spatiotemporal saliency map with $R = \{1.0, 0.7, 0.4, 0.2\}$ is shown in Fig. 10. Specifically, the whole body of the child (i.e., large scale feature) is mostly detected at the coarse scale ($r = 0.2$) while all the details (i.e., small scale features) are captured at the fine scale ($r = 1.0$). By combining outputs from each scale, we can highlight the whole region of salient objects accurately regardless of their sizes through this multi-scale analysis. Thus, we confirm that the proposed method provides well-discriminative representation for visual saliency while suppressing the non-salient regions (e.g., cluttered and highly textured background).

IV. Experimental Results

A. Performance evaluation in images

In this subsection, we demonstrate the performance of the proposed algorithm for static images. Our experiments were conducted on total 800 images collected from the most popularly used data sets in saliency detection tests, namely MSRA data set [26] and PASCAL VOC data set [32]. Images from both data sets are taken in indoor and outdoor environments and contain a wide range of salient objects

TABLE I
Performance variation with different block sizes

	3×3 pixels	5×5 pixels	7×7 pixels	9×9 pixels
F_β	0.693	0.699	0.703	0.709
sec	0.45	0.58	0.81	1.07

such as human, car, train, building, sign, animal, and so on. We used 5×5 pixels of the block (i.e., B_i) for computing luminance contrast and structure tensor, and four scale factors, i.e., $R = \{1.0, 0.7, 0.4, 0.2\}$ as mentioned. The performance variation according to the size of B_i is shown in Table I. Note that the computation of F-measure values will be explained in detail in the latter part of this subsection. By considering both accuracy and processing time (estimated using the image whose size is 400×300 pixels), it is thought that our basic setting (i.e., 5×5 pixels for B_i) is reasonable.

To show the superiority of our proposed method, we compared our approach (we refer to it as TC) with 12 representative models presented in literature, which are proposed by Itti [19], Ma [20], Hou [28], Harel [21], Achanta [22], Bruce [23], Seo [24], Guo [9], Xu [25], Goferman [4], Liu [26], and Kim [27]. Note that we refer to the first author of each method for simplicity. Some experimental results for saliency detection are shown in Fig. 11 and 12. From the results of previous methods, it is easy to see that a lot false positives are generated in highly textured and cluttered backgrounds. In particular, false positives near high contrast edges in the background are hard to be eliminated

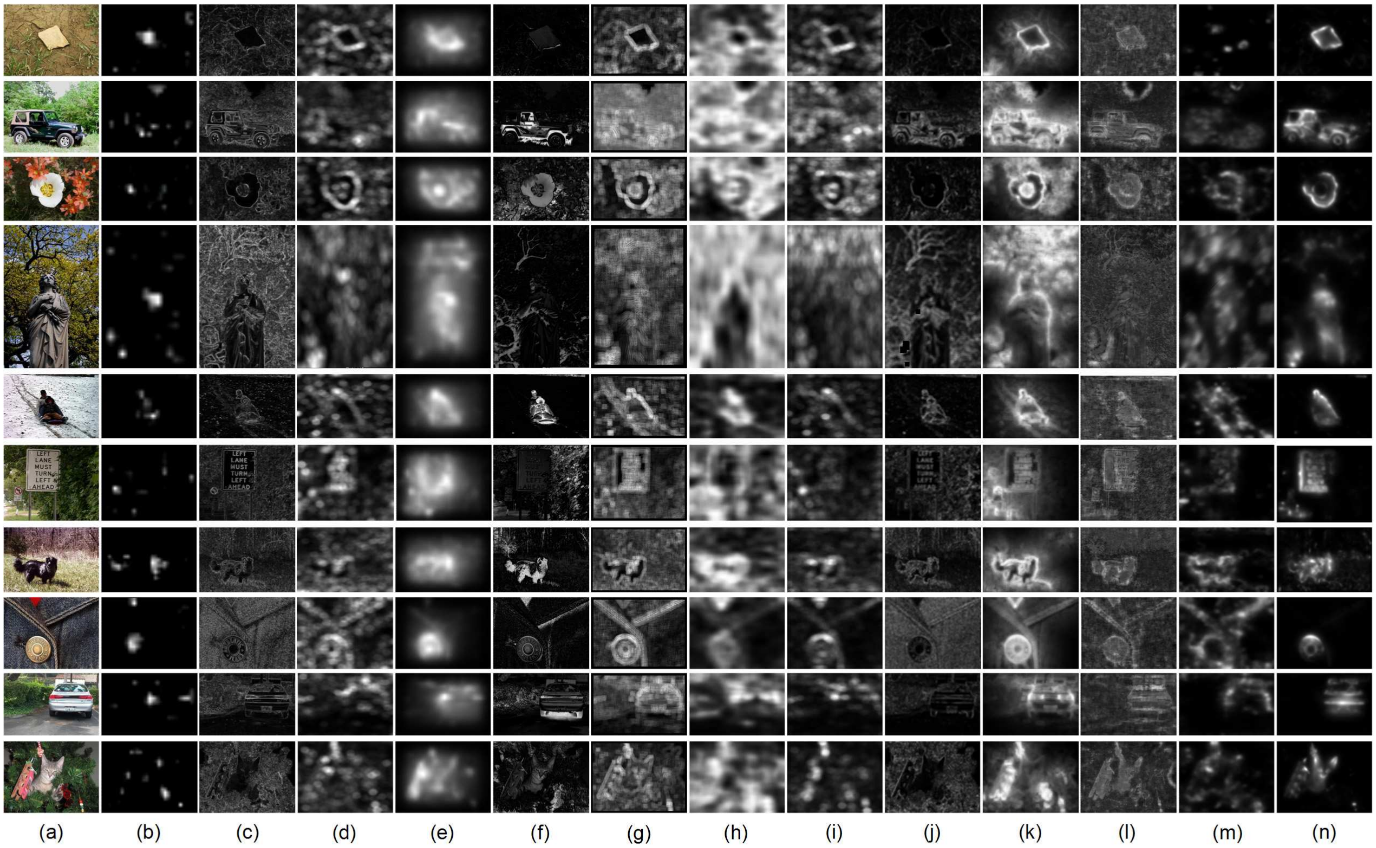


Fig. 12. (a) Input image. Saliency maps generated by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC).

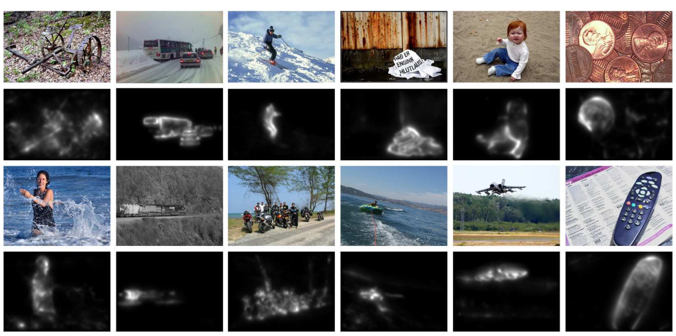


Fig. 13. More examples of saliency maps (odd rows: input images, even rows: saliency maps by the proposed method).

by previous models. Even worse, the complicated cluttered backgrounds are more emphasized rather than the salient objects in several results. Also, those methods often fail to capture the whole region of salient objects due to complex colors and textures, which makes the further applications (e.g., image retargeting and object segmentation) unreliable. In contrast to these results, our proposed approach efficiently deals with such challenging conditions, providing visually acceptable saliency strongly coherent with the human visual attention (see Fig. 11(n) and 12(n)). More examples of saliency maps generated by the proposed method are shown in Fig. 13.

For the quantitative evaluation, we compared the binary mask for salient regions, which are obtained by thresholding our saliency map, with that of other methods. Note that the ground truth images for our image data set are manually generated. The detection accuracy is evaluated by using two quantities, i.e., recall and precision, defined as follows:

$$Recall = \frac{TP}{GT}, \quad Precision = \frac{TP}{TP + FP}, \tag{10}$$

where GT denotes the total number of the ground truth pixels in the data set. TP and FP denote the number of true positives and false positives, respectively. Based on these quantities, we plot the ROC curve varying the thresholding value with respect to the whole image data set as shown in Fig. 14(a) and (b). This curve is useful to investigate how reliably each method highlights salient regions while suppressing non-salient ones in various images. More specifically, most of previous methods are vulnerable to highly textured background, thus yielding relatively low precision values at the same recall rate as shown in Fig. 14(a) and (b). Among them, models proposed by Harel [21], Bruce [23], Guo [9], Goferman [4], Liu [26], and Kim [27] perform quite well even though they still provide less reliable visual saliency compared to the proposed method. In contrast to that, it is easy to see that our saliency map clearly outperforms state-of-the-art algorithms. Moreover, we also computed the F-measure defined as follows:

$$F_\beta = \frac{(1+\beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Preicsion + Recall}. \tag{11}$$

Here we use $\beta^2 = 0.3$ in our work to emphasize the precision

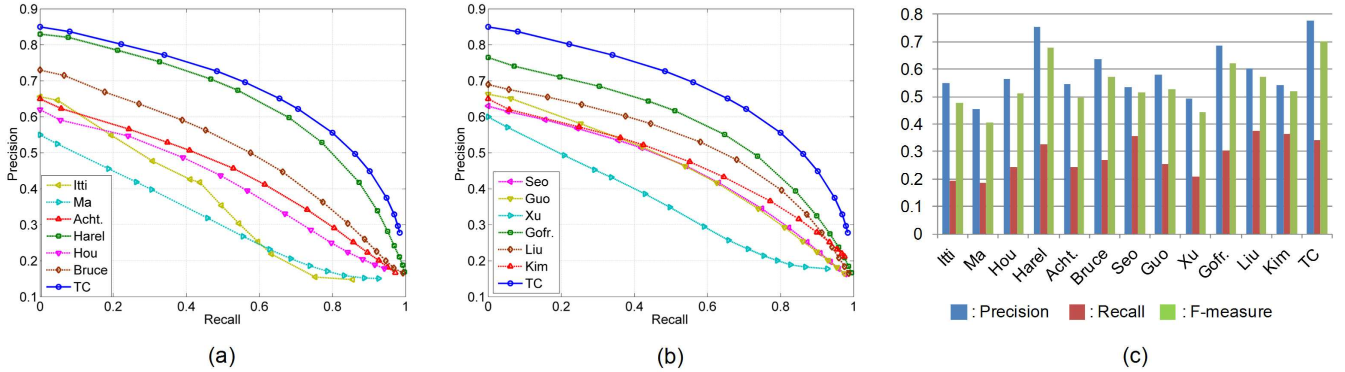


Fig. 14. . (a)(b) ROC curves. (c) Precision-recall bars with F-measures. Note that the proposed method shows the highest F_β values, meaning that our model accurately indicates salient regions without severe false positives.

TABLE II
PERFORMANCE COMPARISON OF PROCESSING SPEED

Method	Itti	Ma	Hou	Harel	Achanta	Bruce	Seo	Guo	Xu	Goferman	Liu	Kim	Proposed
speed (sec/frame)	1.22	0.62	0.01	0.93	0.03	4.07	0.75	0.01	16.32	26.88	2.15	0.06	0.58

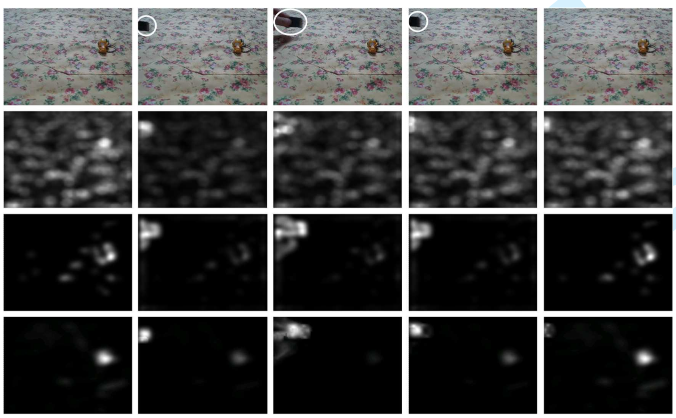


Fig. 15. Spatiotemporal saliency maps of the indoor video (the first row) generated by selected models of Guo [9] (the second row), Kim [27] (the third row), and the proposed method (the fourth row).

more than recall as in [22]. It is important to note that this F-measure effectively represents the ability to suppress false positives while preserving the salient region. Based on this quantity, we efficiently compared our approach with other methods proposed in literature as shown in Fig. 14(c). As can be seen, our proposed method shows the best detection performance with the highest average values of F_β. Note that the proposed method has the slightly lower recall but has the highest precision, indicating that it is better suitable for further computer vision applications such as image retargeting, object segmentation, etc. In summary, the proposed saliency detection method has the best overall performance (on F-measure) among all the methods.

The framework of the proposed method has been implemented by using Visual Studio 2005 (C++) on the low-end PC (Core2Duo 3.0GHz). We compared the processing speed of our model with that of above-mentioned 12 competitive methods as shown in Table II. Note that the processing speed is averaged over a number of images of size 400×300 pixels in our database. Specifically, the processing speed of the proposed method is slightly slower than several methods such as models by Hou [28], Achanta [22], Guo [9], and Kim [27], but it provides much better detection performance. With particular regard to methods of Xu [25] and Goferman [4], the processing speed of the proposed method is a lot faster than that of their method with still higher detection accuracy. From experimental results in images, it is clearly thought that our proposed approach can provide an efficient way of building a reliable saliency map.

B. Performance evaluation in videos

To evaluate the performance of the proposed method in videos, we used various image sequences captured in both indoor and outdoor scenes respectively, which are resized to 256×196 pixels. For computing our spatiotemporal saliency map, we employ three scales, which are 128×128, 64×64, and 32×32 pixels. First, to justify the efficiency of our spatiotemporal saliency, we compared the proposed method with two spatiotemporal schemes proposed by Guo [9] and Kim [27] using a simple video obtained from [27] as shown in Fig. 15. Specifically, since there are no moving objects in the beginning part of the given video, Kim [27]'s model and our model successfully select a small doll as salient areas whereas Guo [9]'s model pays attention to highly textured backgrounds, which are less salient as shown in the first column of Fig. 15. After that, all the methods capture the moving object as salient areas successfully, but highly textured backgrounds are rarely suppressed in Guo [9]'s model. It should be emphasized that the proposed method better captures the salient region (i.e., a small doll) compared to Kim [27]'s saliency model.

In addition, we also demonstrate the spatiotemporal saliency maps for more complicated videos obtained from [33] as shown in Fig. 16. Note that we highlight relevant regions with regard to values of our spatiotemporal saliency map in this

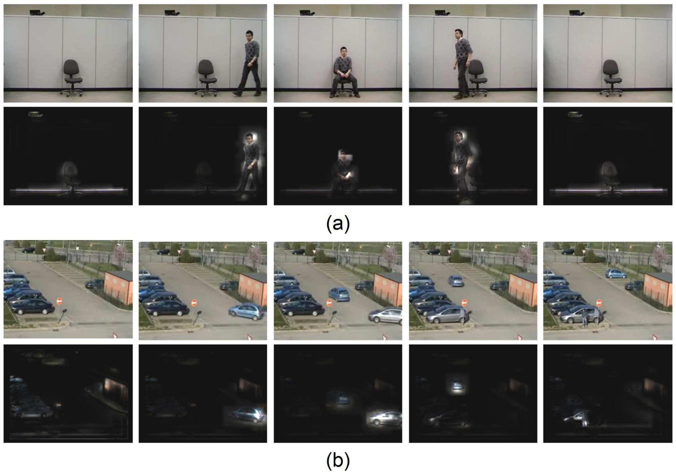


Fig. 16. Spatiotemporal saliency maps generated by the proposed method. Note that the top and the bottom row of each sub-figure show the input image sequences and the corresponding saliency map, respectively.

example. In Fig. 16(a), a man is walking toward a chair and sits for a while. Then, he leaves his seat. In the beginning part of this video, we can say that the chair and a black bag on the cabinet attract the visual attention. When the man comes on the scene, the proposed method efficiently emphasizes the moving object until he disappears. On the other hand, our model firstly selects several parked cars, a road sign, and a building as salient regions in Fig. 16(b). When new cars enter in the parking lot, the proposed method successfully selects moving cars as the most salient areas while retaining static salient areas with relative small importance. The processing speed of the proposed method achieves averagely about 15 fps on our test videos, and it can thus be sufficiently applied for real-time applications. Based on this, it is thought that our spatiotemporal saliency map can efficiently provide reduced search regions for object segmentation, recognition, and tracking tasks in various videos, leading to reduction of computational complexity.

V. Applications in Images and Videos

Owing to the outstanding ability of the proposed method in detecting salient regions as proven in the previous section, we apply our saliency map to three representative applications, i.e., image retargeting (or content-aware image resizing), object segmentation, and background subtraction in dynamic texture scenes, to show its plentiful possibilities in the field of computer vision.

A. Image retargeting

In this subsection, we introduce the most popularly adopted application, i.e., image retargeting, in which the saliency map can be employed. Image retargeting is the process of adaptively resizing a given image to fit the size of arbitrary displays based on the image importance model. For the success of image retargeting, the image importance model needs to be carefully defined since it guides further resizing procedures. To this end, L_1-norm and L_2-norm of gradient magnitude have been popularly used to measure the image importance at each pixel position [34], [35]. However, these often lead to severe distortion when the scene is cluttered or the background is complex. In order to overcome this limitation, the proposed gray-scale saliency map can be employed as a measure of the image importance since it successfully preserves the underlying image structure while suppressing the background. In Fig. 17, we apply our saliency map to various resizing operators, i.e., dynamic programming [34], importance diffusion [36], and fisheye-view warping [37]. It is easy to see that the proposed saliency map performs well regardless of types of resizing operators. Note that we employ the importance diffusion [36] as a resizing operator to achieve the target sizes in our experiments.

To confirm the appropriateness of the proposed saliency map for image retargeting, we compared ours with other saliency maps generated by above-mentioned 12 methods as shown in Fig. 18 and 19. Retargeting results in the vertical direction are shown in Fig. 18. In this example, original image whose height is 400 pixels is reduced in the vertical direction by 100 pixels. Figure 19 shows the horizontal retargeting results of various images. More specifically, the image importance obtained by previous algorithms provide quite reasonable viewing effects up to a certain target height or width. However, as the reduction ratio further increases, the resizing operator starts to remove a connected path of pixels across salient objects, leading to drastic distortions in the resized image. From Fig. 18 and 19, it is easy to see that our saliency

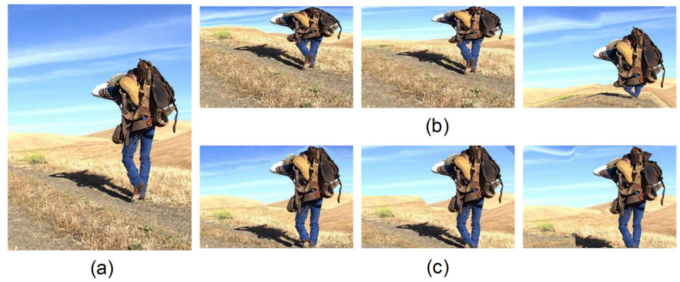


Fig. 17. (a) Input image. (b) Retargeting results by dynamic programming [34], importance diffusion [36], and fisheye-view warping [37] (from left to right) with their own importance measures. (c) Retargeting results by the proposed saliency map with resizing operators used in (b).

Fig. 18. (a) Input image. Results of image retargeting based on saliency models by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC). Note that the original image is reduced in the vertical direction by 100 pixels.

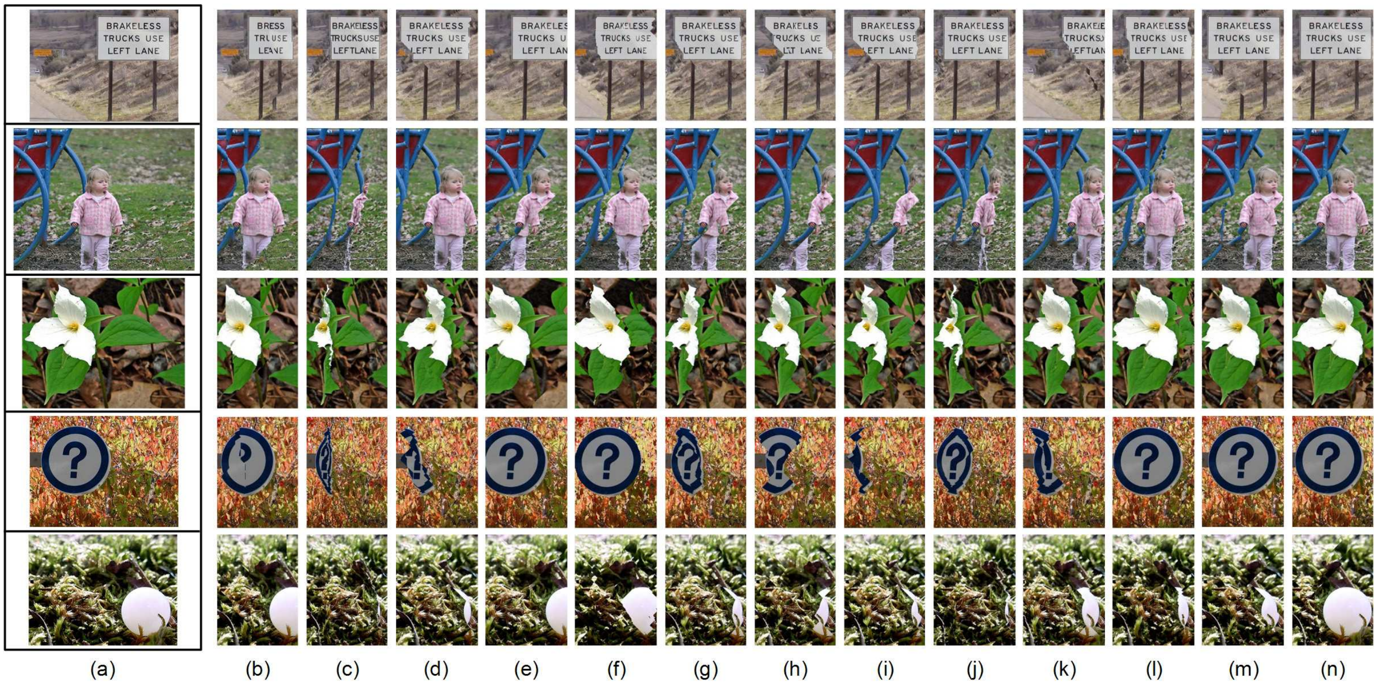


Fig. 19. (a) Input image. Results of image retargeting based on saliency models by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC). Note that each sub-figure is reduced in horizontal direction by 220, 180, 200, 180, and 180 pixels (from top to bottom).

TABLE III
PERFORMANCE COMPARISON BY USING THE AVERAGE TARGETING RATE (TR) VALUES

Method	Itti	Ma	Hou	Harel	Achanta	Bruce	Seo	Guo	Xu	Goferman	Liu	Kim	Proposed
TR	0.729	0.581	0.731	0.903	0.786	0.762	0.684	0.687	0.662	0.834	0.888	0.845	**0.953**

model provides visually acceptable retargeting results while preserving viewers' experience compared to other models. For the quantitative analysis, we compute the targeting rate, which is defined as $TR = TP/GT$ [38] where GT denotes the total number of ground truth pixels belonging to salient objects that should be preserved during the retargeting procedure and TP denotes the amount of preserved ground truth pixels. Table III shows the comparison results of the average TR values on total 100 images randomly collected from our image data set and we can see that our proposed method outperforms other previous ones in image retargeting.

B. Object segmentation

Even though there have been notable research advances for object segmentation, previous methods (e.g., graph-cut [39] and grab-cut [40] schemes) still require the user interaction for seed selection. To be a fully automatic object segmentation, the saliency map can be straightforwardly employed as a seed map since it successfully highlights pixels relevant to the foreground objects in a given scene as shown in Section IV. In this subsection, we show the utility of the proposed saliency map for object segmentation. To do this, we employ the graph-cut scheme [39], which is most widely used for object segmentation. Note that we set the threshold value to satisfy the condition that pixels having higher intensity than three times of the global mean in the saliency map are determined as seeds for foreground objects. Several segmentation results driven by various saliency maps are shown in Fig. 20. As can be seen, saliency maps generated by previous methods often lead to high-level false negative (i.e., objects are misclassified as background regions) as well as false positive (i.e., background regions are misclassified as objects) rates. In contrast to that, the proposed saliency map provides desirable segmentation results even with the complex background due to its ability of suppressing cluttered and highly textured components. In particular, our saliency map leads to the successful segmentation whereas most of previous models fail to segment the foreground (i.e., white statue) due to the cluttered background in the sixth row of Fig. 20. More segmentation results based on our saliency map are shown in Fig. 21.

We also provide the quantitative performance comparison in Table IV. In this analysis, we adopt the previously introduced quantities, i.e., recall, precision, and F-measure (see (10) and (11)). Note that we set $\beta = 1.0$ to assign the balanced weights between recall and precision for the F-measure. From Table IV, we can see that the proposed saliency map based segmentation achieves the highest recall, precision, and F-measure. It should be emphasized again that such favorable segmentation results can be obtained since our proposed scheme performs robustly even in cluttered scenes, which fits with the human visual perception.

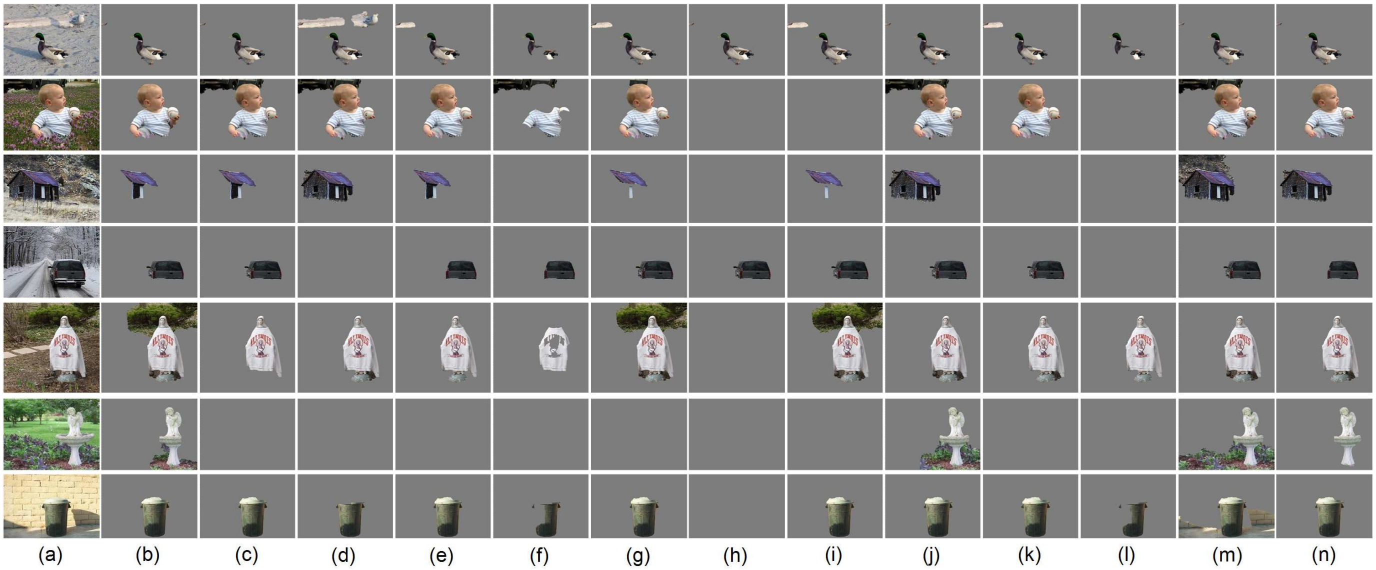


Fig. 20. (a) Input image. Results of object segmentation based on saliency models by (b) Itti [19], (c) Ma [20], (d) Hou [28], (e) Harel [21], (f) Achanta [22], (g) Bruce [23], (h) Seo [24], (i) Guo [9], (j) Xu [25], (k) Goferman [4], (l) Liu [26], (m) Kim [27], and (n) the proposed method (TC).

TABLE IV
PERFORMANCE COMPARISON OF OBJECT SEGMENTATION

Method	Itti	Ma	Hou	Harel	Achanta	Bruce	Seo	Guo	Xu	Goferman	Liu	Kim	Proposed
$Recall$	0.807	0.773	0.829	0.818	0.576	0.633	0.314	0.801	0.792	0.746	0.392	0.848	**0.882**
$Precision$	0.708	0.672	0.664	0.721	0.638	0.583	0.264	0.662	0.629	0.649	0.455	0.664	**0.738**
F_β	0.754	0.721	0.737	0.766	0.606	0.608	0.287	0.725	0.702	0.694	0.421	0.745	**0.803**

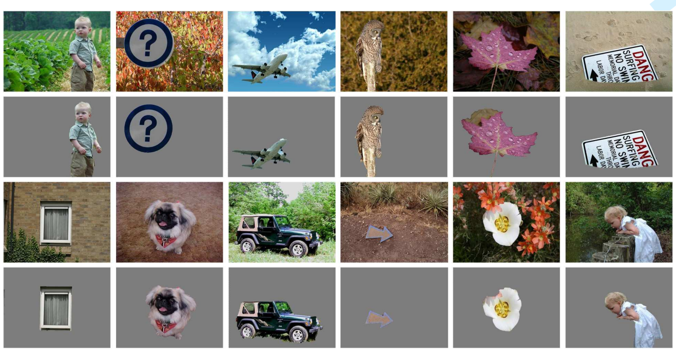


Fig. 21. More segmentation results based on our saliency maps (odd rows: input images, even rows: segmentation results by the proposed method).

C. Background subtraction in dynamic texture scenes

With increasing interest in high-level safety and security, video surveillance systems have been in critical demand. For the success of such systems, background subtraction, one of essential tasks in video surveillance, has been studied in various environments. However, motion patterns of the background (e.g., waving leaves, spouting fountain, rippling water, etc.), which are not static over short time periods, often cause to high-level false positives and it thus is still hard to handle them in the traditional background subtraction frameworks. In this subsection, we address this limitation through our temporal saliency detection scheme. From the visual saliency point of view, the problem of robust background subtraction narrows down to suppressing the locations having non-salient motions. The proposed temporal saliency map has several advantages over the traditional background subtraction models as follows: 1) since our scheme is based on the contrast of the temporally directional coherence, it has a great ability to suppress island-like false positives occurring in irrelevant regions and 2) the proposed method is completely unsupervised and thus does not require training and initializing the background model.

For the performance evaluation of background subtraction in dynamic texture scenes, we first tested our method and several background subtraction algorithms on three short videos, i.e., boat (180×100 pixels) [41], bottle (244×180 pixels) [42], and campus (160×128 pixels) [43] sequences, and the corresponding detection results are shown in Fig. 22. As can be seen, traditional background subtraction models (i.e., traditional mixture of Gaussian (t-MoG) model [44] and generalized mixture of Gaussian (g-MoG) model [45]) yield many false positives due to rippling water and strongly waving leaves whereas the proposed temporal saliency map is quite robust to such background motions. Note that, to make fair comparisons, we report experimental results at a true positive (TP) rate of 0.8, which is good enough to correctly extract moving objects for further applications. We also plot the ROC curve using the campus sequence in Fig. 23. Note that we include the method based on spatiotemporal local binary patterns (STLBP) [46] in the ROC curve. For the quantitative evaluation, the false positive (FP) rates, which are obtained from 20 frames randomly selected from each video, are shown in Table V. From Fig. 22, 23, and Table V, we confirm that

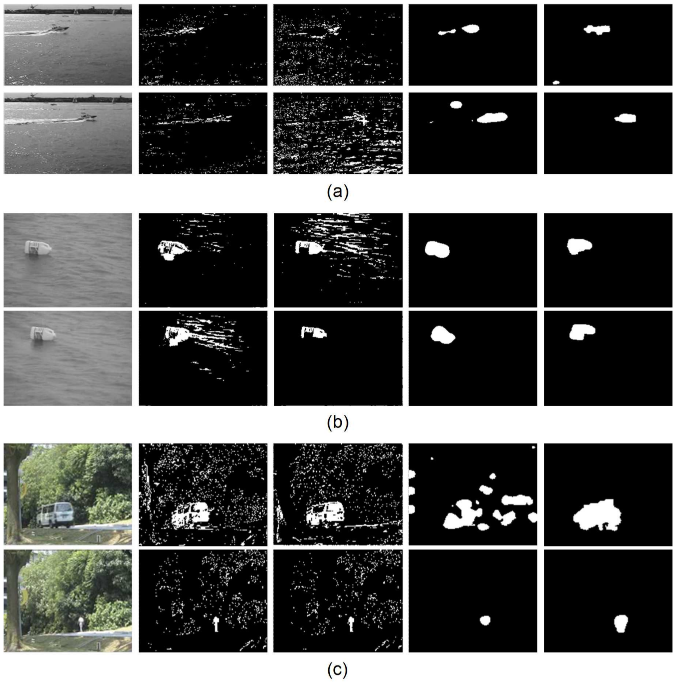


Fig. 22. 1st column: input frames of each video, i.e., (a) boat, (b) bottle, and (c) campus. Background subtraction results by t-MoG [44], g-MoG [45], Guo [9], and the proposed method (from the 2nd column to the 5th column).

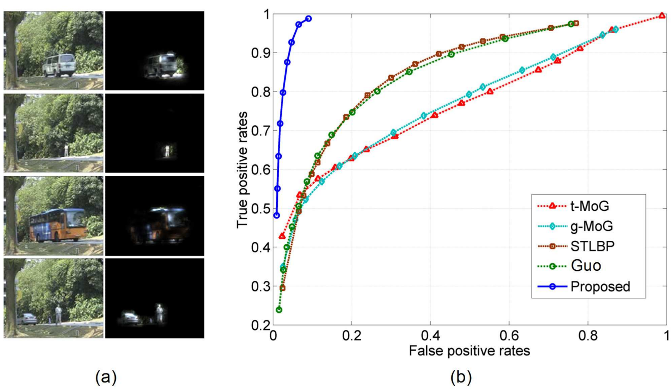


Fig. 23. (a) Input sequences (left) and some examples of the corresponding temporal saliency maps by the proposed method (right). (b) ROC curves determined by t-MoG [44], g-MoG [45], STLBP [46], Guo [9], and the proposed method.

the proposed saliency-based scheme is capable of correctly extracting moving objects with low false positive rates in dynamic texture scenes.

VI. Conclusion

A novel method for detecting salient regions in the spatiotemporal domain has been proposed in this paper. The key idea of the proposed method is that the biological mechanism of the bottom-up visual attention can be approximated by exploiting two main contrasts captured in the retina and the visual cortex. To this end, we propose to use textural contrast defined based on combination of luminance contrast and directional coherence contrast, and extend its concept to

TABLE V
Performance comparison by FP rates measured at TP = 0.8

Method	t-MoG [44]	g-MoG [45]	Guo [9]	Proposed
boat [41]	0.057	0.092	0.015	0.007
bottle [42]	0.027	0.058	0.004	0.006
campus [43]	0.552	0.511	0.264	0.023

the spatiotemporal domain with temporal gradients. By incorporating the responses of the proposed contrast mechanism into a multiscale framework, we can generate reliable saliency maps. Based on extensive experimental results, we confirm that the proposed method efficiently highlights relevant regions in images and videos even with the cluttered background. Furthermore, to show the plentiful possibilities of the visual saliency, we apply the proposed saliency map to various real-world applications such as image retargeting, automatic object segmentation, background subtraction in dynamic texture scenes. From the comparison results, it is thought that the proposed method is effective enough to be applicable to various vision-based intelligent applications.

References

[1] R. VanRullen, "Visual saliency and spike timing in the ventral visual pathway," *Journal of Physiology*, vol. 97, pp. 365–377, 2003.

[2] O. S. Packer and D. M. Dacey, "Synergistic center-surround receptive field model of monkey H1 horizontal cells," *Journal of Vision*, vol. 5, pp. 1038–1054, 2005.

[3] R. Achanta and S. Susstrunk, "Saliency detection for content-aware image resizing," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 1005–1008, Nov. 2009.

[4] S. Goferman, L. Z. Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2376–2383, June 2010.

[5] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 446–456, Apr. 2011.

[6] Y. Fu, J. Cheng, Z. Li, and H. Lu, "Saliency cuts: An automatic approach to object segmentation," in *Proc. IEEE International Conference on Pattern Recognition (ICPR)*, pp. 1–4, Dec. 2008.

[7] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): meaningful attention using stochastic image modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 693–708, Apr. 2010.

[8] O. L. Meur and J. -C. Chevet, "Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2801–2813, Nov. 2010.

[9] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[10] S. Marat, M. Guironnet, and D. Pellerin, "Video summarization using a visual attention model," in *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 1784–1788, 2007.

[11] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality ? Applying to image quality metric," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 2, pp. 169–172, Oct. 2007.

[12] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency inspired full-reference quality metric for packet-loss-impaired video," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 81–88, Mar. 2011.

[13] H. Liu and I. Heynderickx, "Visual attention in object image quality assessment: based on eyetracking data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, July 2011.

[14] A, Toet, "Computational versus psychophysical bottom-up image saliency: a comparative evaluation study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2131–2146, Nov. 2011.

[15] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 1, pp. 253–256, Sept. 2003.
[16] H. Li and K. N. Ngan, "Saliency model-based face segmentation and 646 tracking in head-and-shoulder video sequences," *Journal of Visual Communication and Image Representation*, vol. 19, no. 5, pp. 320–333, July 2008.
[17] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, no. 7, pp. 1–18, 2008.
[18] W. Kim and C. Kim, "Saliency detection via textural contrast," *Optics Letters*, vol. 37, no. 9, pp. 1550–1552, May 2012.
[19] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no, 11, pp. 1254–1259, Nov. 1998.
[20] Y. F. Ma and H. J . Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. ACM International Conference on Multimedia*, pp. 374–381, 2003.
[21] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 19, pp. 545–552, 2007.
[22] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1597–1604, June 2009.
[23] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: an information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 1–24, 2009.
[24] H. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of Vision*, vol. 8, no. 12, pp. 1–27, 2009.
[25] Y. Xu, Y. Zhao, C. Jin, J. Qu, L. Liu, and X. Sun, "Salient target detection based on pseudo-Wigner-Ville distribution and Renyi entropy," *Optics Letters*, vol. 35, pp. 475–477, 2010.
[26] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. -Y. Shum, "Learning to detect a salient object," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[27] W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 446–456, Apr. 2011.
[28] X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8, June 2007.
[29] P. Bian and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency advances in neuro-information processing," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 251–258. 2009.
[30] A. B. Chan, V. Mahadevan, and N. Vasconcelos, "Generalized Stauffer-Grimson background subtraction for dynamic scenes," *Machine Vision and Applications*, vol. 22, no. 5, pp. 751–766, 2011.
[31] L. Wolf, M. Guttmann, and D. Cohen-Or, "Non-homogeneous content-driven video retargeting," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 1–6. Oct. 2007.
[32] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A, Zisserman, "The PASCAL visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
[33] http://imagelab.ing.unimore.it/visor/index.asp.
[34] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Transactions on Graphics*, vol. 26, no. 3, pp. 267–276, 2007.
[35] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Transactions on Graphics*, vol. 27, no. 3, pp. 1–8, 2008.
[36] S. Cho, H. Choi, Y. Matsushita, S. Lee, "Image retargeting using importance diffusion," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 997–980, Nov. 2009.
[37] F. Liu and M. Gleicher, "Automatic image retargeting with fisheye-view warping," in *ACM Annual Symp. on User Interface Software and Technology*, pp. 153–162, 2005.
[38] W. Kim and C. Kim, "A texture-aware salient edge model for image retargeting," *IEEE Signal Processing Letters*, vol. 18, no. 11, pp. 631–634, Nov. 2011.
[39] Y. Boykov and G. Funka-Lea, "Graph cuts and efficient N-D image segmentation," *International Journal of Computer Vision*, vol. 70, no. 2, pp. 109–131, 2006.
[40] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut-Interactive foreground extraction using iterated graph cuts," in *Proc. ACM SIGGRAPH*, pp. 309–314, 2004.
[41] A. B. Chan, V. Mahadevan, and N. Vasconcelos, "Generalized Stauffer-Grimson background subtraction for dynamic scenes," *Machine Vision and Applications*, vol. 22, no. 5, pp. 751–766, 2011.
[42] V. Mahadevan and N. Vasconcelos, "Background subtraction in highly dynamic scenes," in *Proc. Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–6, June 2008.
[43] L. Li, W. Huang, I. Y. -H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Transactions on Image Processing*, vol. 13, no. 11, pp. 1459–1472, Nov. 2004.
[44] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 246–252, June 1999.
[45] G. Dalley, J. Migdal, and W. Grimson, "Background subtraction for temporally irregular dynamic textures," in *Proc. IEEE Workshop on Applications of Computer Vision (WACV)*, pp. 1–7, Jan. 2008.
[46] S. Zhang, H. Yao, and S. Liu, "Dynamic background modeling and subtraction using spatio-temporal local binary patterns," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 1556–1559, Oct. 2008.