Video Saliency Detection Using Adaptive Feature Combination and Localized Saliency Computation

Eunpil Park, Byeong-Ju Han, Seungjoon Yang, and Jae-Young Sim
School of Electrical and Computer Engineering, UNIST, Ulsan, Korea
E-mail : {cosmos,bjhan,syang,jysim}@unist.ac.kr

***Abstract*—A novel saliency detection algorithm for videos is proposed in this paper. We adaptively determine the weights of color and motion features to extract combined global feature contrast by adopting the compactness prior of salient object. We localize a saliency searching area in a current frame using the saliency distribution computed at the previous frame. We estimate the saliency by computing a relative feature distance with respect to the salient object and local background, which is weighted by global feature contrast. Experimental results show that the proposed algorithm captures salient objects faithfully on various videos, and outperforms the state-of-the art video saliency detection methods qualitatively and quantitatively.**

I. INTRODUCTION

Saliency detection is automatic extraction of visually meaningful regions from images and videos. Saliency detection has been studied to facilitate various applications such as image retrieval [1], segmentation [2], and action recognition [3]. During the past decade, intensive research has been performed on saliency detection of still images. Most of the image saliency detection techniques compute saliency values based on the center-surround contrast: a salient region yields a distinct feature compared to its surrounding area. Moreover, the constraints on the locations of salient foreground objects and background are also employed. The center prior assumes a salient object is highly probable to be located near the image center [4], and the boundary prior regards the image boundaries are usually included to the background [5]. Recently, machine learning techniques were adopted to detect saliency of still images [6], [7], [8], [9].

While only the spatial features are employed to estimate saliency of still images, spatial and temporal features should be considered together for saliency detection of videos. Therefore, the conventional image saliency detection methods often fail to capture the video saliency successfully since they do not use temporal information. Video saliency detection techniques have been devised in two directions. First, spatial and temporal saliency maps are obtained separately, and a final saliency map is generated as a weighted summation of the two saliency maps [10], [11], [12], [13], [14]. Huang et al. [10] empirically determined the weights for combining spatial and temporal saliency maps. Fang et al. [11] measured uncertainty of spatial and temporal saliency values for each pixel, and used a larger weight to the saliency value with lower uncertainty. Liu et al. [12] determined the weights by checking the consistency between spatial and temporal maps. Muthuswamy et al. [13] assigned a high weight to the motion saliency map when motion contrast is high. Li et al. [14] employed a higher weight for the saliency map which highlights moving objects more faithfully. These methods suffer from the limitation that inaccurately assigned saliency values in either of the two saliency maps degrade the quality of a final saliency map. On the other hand, spatial and temporal features are combined first, and the combined feature is used to generate a final saliency map [15], [16], [17], [18], [19], [20], [21], [22]. Seo et al. [15] computed the gradient of a pixel which is compared to that of spatially and temporally surrounding pixels. Rahtu et al. [16] constructed the histogram of spatial and temporal features for video saliency detection. Xue et al. [17] separated salient foreground objects from the background using a low-rank matrix technique. Lee et al. [18] employed a support vector machine to combine spatial and temporal features. Kim et al. [19] computed a final saliency map by designing a spatial transition matrix and a temporal restarting term based on the random walk with restart. Wang et al. [20] integrated the gradients of the spatial and temporal feature maps. Wang et al. [21], [22] employed the spatiotemporal feature map and the results of the previous frame to separate the foreground and the background regions. However, these methods do not utilize the spatial and temporal features adaptively according to their changeable characteristics in various videos.

In this paper, we propose a novel video saliency detection algorithm which combines the color and motion features adaptively and estimates the saliency on localized searching areas. We first extract global feature contrast by combining the color and motion features adaptively at each frame based on the compactness prior of salient object. We localize a searching area of saliency detection in a current frame using the saliency distribution in the previous frame, since the salient regions detected in previous frames are highly probable to be also salient in a current frame in typical videos. We also measure the relative feature distances with respect to a salient object and non-salient local background, respectively, which are then weighted by the global feature contrast to estimate the final saliency. Experimental results demonstrate that the proposed algorithm yields a reliable performance of saliency detection for video sequences with diverse characteristics of color and motion, and outperforms the existing state-of-the-art methods.

II. FEATURE CONTRAST COMPUTATION

We partition each frame in an input video sequence into superpixels using SLIC [23]. We compute the contrast of color

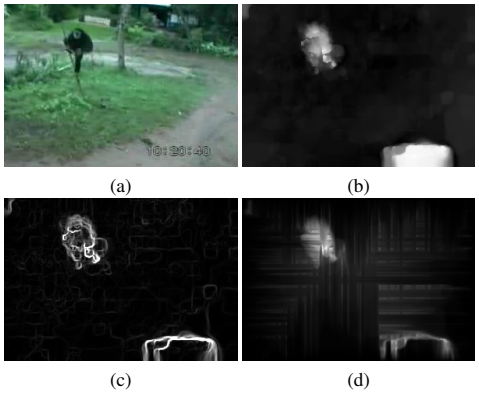


Fig. 1: Relative motion feature contrast. (a) An input frame. (b) The initial map of relative motion features and (c) its gradient magnitude map. (d) The refined map of motion feature contrast.

and motion features at each superpixel based on the boundary prior, which are then combined adaptively according to the compactness prior of salient object.

A. Color Feature Contrast

We extract a color feature $c(\mathbf{p}_i) \in \mathbb{R}^3$ for the *i*-th superpixel $\mathbf{p}_i$ as the average color of the pixels belonging to $\mathbf{p}_i$ based on the broadly tuned color space [24]. We compute color feature contrast using the boundary prior, where a superpixel yielding large feature distances to the boundary superpixels is assigned a high contrast value. In order to handle the exceptional cases that salient objects are placed touching the image boundaries [25], we also partition the boundary superpixels into the four sets $\mathcal{B}_{\text{top}}$, $\mathcal{B}_{\text{bottom}}$, $\mathcal{B}_{\text{left}}$, and $\mathcal{B}_{\text{right}}$, which are from the top, bottom, left, and right boundaries, respectively. We measure the color dissimilarity at $\mathbf{p}_i$ with respect to each of the four boundary sets. For example, the color dissimilarity $d_{\text{top}}(\mathbf{p}_i)$ between $\mathbf{p}_i$ and $\mathcal{B}_{\text{top}}$ is computed as the minimum distance from $c(\mathbf{p}_i)$ to the color features of all the superpixels in $\mathcal{B}_{\text{top}}$, given by

$$d_{\text{top}}(\mathbf{p}_i) = \min_{\mathbf{p}_k \in \mathcal{B}_{\text{top}}} \|c(\mathbf{p}_i) - c(\mathbf{p}_k)\|. \quad (1)$$

In the same way, $d_{\text{bottom}}(\mathbf{p}_i)$, $d_{\text{left}}(\mathbf{p}_i)$, and $d_{\text{right}}(\mathbf{p}_i)$ are measured by using $\mathcal{B}_{\text{bottom}}$, $\mathcal{B}_{\text{left}}$, and $\mathcal{B}_{\text{right}}$, respectively. Finally, we estimate the color feature contrast $f_{\text{c}}(\mathbf{p}_i)$ of $\mathbf{p}_i$ as the maximum among the four dissimilarity values.

$$f_{\text{c}}(\mathbf{p}_i) = \max\left\{d_{\text{top}}(\mathbf{p}_i), d_{\text{bottom}}(\mathbf{p}_i), d_{\text{left}}(\mathbf{p}_i), d_{\text{right}}(\mathbf{p}_i)\right\}. \quad (2)$$

B. Motion Feature Contrast

In a video sequence, relative motion of a foreground object with respect to the background is more recognizable than absolute motion. At each frame of an input video sequence, we first estimate the background motion as the average optical flow vector [26] of the pixels in the boundary superpixels based on the boundary prior assumption. Then we obtain $m^t(\mathbf{x})$, a relative foreground motion at a pixel $\mathbf{x}$ in the t th frame I^t, given by

$$m^t(\mathbf{x}) = \frac{1}{2N+1} \sum_{k=t-N}^{t+N} \|o^k(\mathbf{x}) - \boldsymbol{\mu}^k\| \quad (3)$$

where $o^k(\mathbf{x})$ is the optical flow vector of $\mathbf{x}$ in I^k, and $\boldsymbol{\mu}^k$ denotes the background motion vector in I^k, respectively. Note that we consider $2N$ neighboring frames to estimate the relative motion at a current frame, in order to avoid salient foreground objects which are static at the current time instance from being assigned low motion feature values. We empirically set $N = 2$. Fig. 1(b) shows an example of the initial map of relative motion features obtained from an input frame in Fig. 1(a).

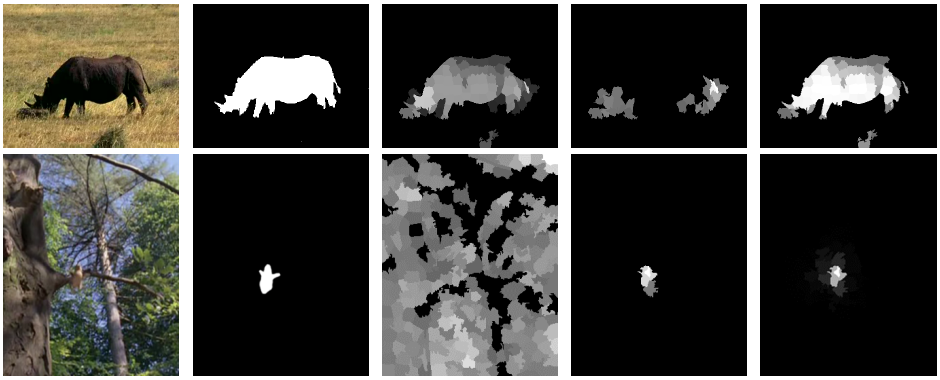
We also refine the initially obtained relative motion features to suppress high values assigned to some background regions using [20]. We compute the gradient of initial motion features, and accumulate the gradient magnitudes at each pixel along the four directions. Then we define a motion feature contrast of each pixel by taking the minimum among the four values of the accumulated gradient magnitudes. Fig. 1(c) shows the gradient map of Fig. 1(b), and Fig. 1(d) shows the refined motion feature contrast map where we see that the relatively high contrast values in the bottom background region observed in the initial map are effectively alleviated. Finally, we define a motion feature contrast $f_{\text{m}}^t(\mathbf{p}_i)$ of a superpixel $\mathbf{p}_i$ in I^t by taking the average over all pixels in $\mathbf{p}_i$.

C. Adaptive Combination of Color and Motion Features

Fig. 2 shows the resulting maps of superpixel-wise feature contrast on two video sequences with different characteristics. In the first row, the rhinoceros does not move and shakes its head and tail slightly, and therefore the resulting motion feature contrast in Fig. 2(d) fails to capture the whole salient object region due to the lack of sufficient motion information. However, the color feature contrast in Fig. 2(c) detects most of the object region faithfully. In contrary, the bird in the second row is falling fast, and the motion feature contrast in Fig. 2(d) successfully indicates the bird. However, lots of the background regions are assigned high contrast values of color feature in Fig. 2(c), since the bird has a similar color to that of the cluttered background.

In order to exploit the color and motion features adaptively according to their confidence, we combine the two features based on the compactness prior assumption that a salient object yields a compact shape [24]. To this end, at the t-th frame, we first normalize the color feature contrast $f_{\text{c}}^t(\mathbf{p}_i)$ and the motion feature contrast $f_{\text{m}}^t(\mathbf{p}_i)$ into the range of [0, 1], respectively. Then we combine $f_{\text{c}}^t(\mathbf{p}_i)$ and $f_{\text{m}}^t(\mathbf{p}_i)$ by assigning a higher weight to the feature which exhibits more compact distribution of superpixels with high feature contrast. Specifically, at each superpixel $\mathbf{p}_i$ in I^t, we compute a global feature contrast $f_{\text{cm}}^t(\mathbf{p}_i)$ given by

$$f_{\text{cm}}^t(\mathbf{p}_i) = \left(\frac{\sigma_{\text{m}}^t}{\sigma_{\text{c}}^t + \sigma_{\text{m}}^t}\right) \cdot f_{\text{c}}^t(\mathbf{p}_i) + \left(\frac{\sigma_{\text{c}}^t}{\sigma_{\text{c}}^t + \sigma_{\text{m}}^t}\right) \cdot f_{\text{m}}^t(\mathbf{p}_i) \quad (4)$$



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

Fig. 2: Color and motion features. (a) Input frames. (b) The ground truth saliency maps. (c) Color feature contrast maps. (d) Motion feature contrast maps. (e) Combined feature contrast maps.

where σ_{c}^t and σ_{m}^t are the standard deviations weighted by the color and motion feature contrast, respectively.

$$\sigma^t = \sqrt{\frac{\sum_{\mathbf{x}\in\Omega^t} f^t(\mathbf{x}) \cdot \parallel \mathbf{x} - \bar{\mathbf{x}}^t \parallel^2}{\sum_{\mathbf{x}\in\Omega^t} f^t(\mathbf{x})}} \tag{5}$$

where $f^t(\mathbf{x})$ is the feature contrast of the superpixel including a pixel $\mathbf{x}$, and Ω^t is the set of the pixels belong to the superpixels with feature contrast values larger than 40% of the maximum value in I^t. $\bar{\mathbf{x}}^t$ is the average position of Ω^t weighted by the feature contrast, given by

$$\bar{\mathbf{x}}^t = \frac{\sum_{\mathbf{x}\in\Omega^t} f^t(\mathbf{x}) \cdot \mathbf{x}}{\sum_{\mathbf{x}\in\Omega^t} f^t(\mathbf{x})}. \tag{6}$$

Note that the weights in (4) are not fixed and rather adaptively determined at each frame in a video sequence, respectively, which reflect the relative confidence or contribution of color and motion features for reliable extraction of global feature contrast. As shown in Fig. 2(e), the combined feature contrast detects the salient objects reliably, even though either of the color and motion features is not extracted faithfully.

III. Saliency Evaluation

The human visual system tends to recognize the visual contents of a current frame together with that of the previous frames. Moreover, a salient object appears in similar locations between adjacent frames in a typical video sequence. Therefore, we first localize a set of candidate superpixels to search for a salient object in a current frame by using the saliency distribution computed at the previous frame. Then, motivated by the object tracking method [27], we estimate the saliency for each candidate superpixel by using the relative feature distances with respect to a salient object and its local background region. Finally, the saliency values are spatially and temporally refined based on the energy minimization framework.

A. Localized Searching Area

We define Φ^t a local searching area for a salient object in I^t, which is composed of the superpixels geometrically close to the salient region found at the previous frame I^{t-1}.

$$\Phi^t = \left\{ \mathbf{p} \mid \delta(\mathbf{p}) < \frac{\tau\sqrt{W^2+H^2}}{\lambda_\alpha(\mathbf{p})} \right\} \tag{7}$$

where $\delta(\mathbf{p})$ is the shortest distance among the distances from a superpixel $\mathbf{p} \in I^t$ to the superpixels in I^{t-1} with saliency values larger than 0.5. W and H denote the width and height of image frame, and τ is set to be 0.05 empirically. Note that the distance threshold is weighted by

$$\lambda_\alpha(\mathbf{p}) = \exp\left(-\alpha \cdot f_{\mathrm{cm}}^t(\mathbf{p})\right), \tag{8}$$

which encourages the superpixels with high feature contrast of $f_{\mathrm{cm}}^t(\mathbf{p})$ to be included to Φ^t. We set $\alpha = 3$ empirically. At the first frame, we determine Φ^1 as the set of superpixels satisfying the condition $\lambda_\alpha(\mathbf{p}) < 0.4$, since there is no previous frame. Fig. 3(c) shows the local searching area Φ^t for an input frame in Fig. 3(a), where we see that the salient objects as well as some non-salient background regions are included.

Next, we find $\mathcal{R}_{\mathrm{s}}^{t-1}$ the set of superpixels with the top 30% of saliency values in I^{t-1}. Fig. 3(d) shows the associated region of $\mathcal{R}_{\mathrm{s}}^{t-1}$ obtained from the previous frame in Fig. 3(b). Moreover, we find $\mathcal{R}_{\mathrm{ns}}^t$ the region of the three layers of superpixels in I^t surrounding Φ^t. Fig. 3(e) shows $\mathcal{R}_{\mathrm{ns}}^t$ of Φ^t in Fig. 3(c). We see that $\mathcal{R}_{\mathrm{s}}^{t-1}$ and $\mathcal{R}_{\mathrm{ns}}^t$ exhibit color appearances similar to that of the salient foreground object and its local background, respectively.

B. Saliency Value Computation

Note that the salient superpixels in Φ^t have similar color features to that of $\mathcal{R}_{\mathrm{s}}^{t-1}$, while the non-salient superpixels in Φ^t have similar colors to that of $\mathcal{R}_{\mathrm{ns}}^t$. Based on this property, we evaluate saliency values of the superpixel in Φ^t using the color feature distances from $\mathcal{R}_{\mathrm{s}}^{t-1}$ and $\mathcal{R}_{\mathrm{ns}}^t$. Specifically, we

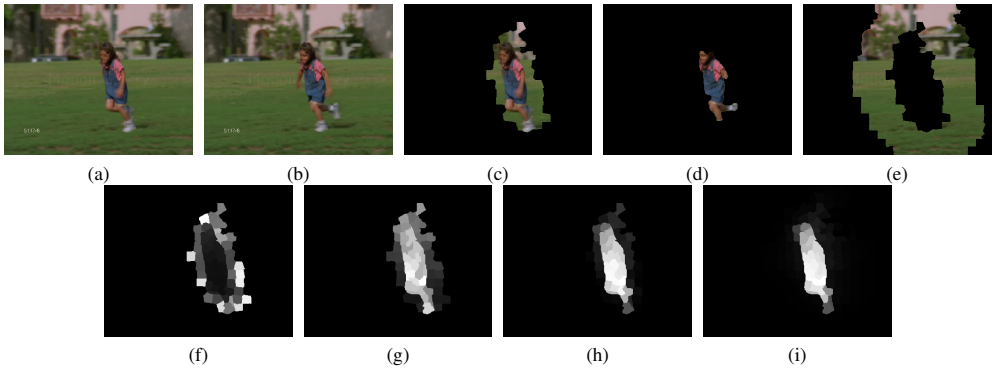


Fig. 3: Localized searching area. (a) Input frame I^t. (b) Previous frame I^{t-1}. (c) Local searching area Φ^t. (d) $\mathcal{R}_{\rm s}^{t-1}$ in previous frame. (e) $\mathcal{R}_{\rm ns}^{t}$ in current frame. (f) Feature distance $d_{\rm s}^t$ on Φ^t. (g) Feature distance $d_{\rm ns}^t$ on Φ^t. (h) Initial saliency map $\tilde{S}^t$. (i) Final saliency map S^t.

measure the weighted dissimilarity of a color feature at each superpixel $\mathbf{p}_i$ in Φ^t from that of $\mathcal{R}_{\rm s}^{t-1}$ and $\mathcal{R}_{\rm ns}^{t}$, respectively, given by

$$d_{\rm s}^t(\mathbf{p}_i) = \frac{\lambda_\gamma(\mathbf{p}_i)}{|\mathcal{R}_{\rm s}^{t-1}|} \sum_{\mathbf{p}_j \in \mathcal{R}_{\rm s}^{t-1}} \|c(\mathbf{p}_i) - c(\mathbf{p}_j)\|, \quad (9)$$

$$d_{\rm ns}^t(\mathbf{p}_i) = \frac{1}{\lambda_\gamma(\mathbf{p}_i) \cdot |\mathcal{R}_{\rm ns}^{t}|} \sum_{\mathbf{p}_j \in \mathcal{R}_{\rm ns}^{t}} \|c(\mathbf{p}_i) - c(\mathbf{p}_j)\|. \quad (10)$$

Note that the feature distances are also weighted by $\lambda_\gamma(\mathbf{p}_i)$, such that a superpixel with a high feature contrast is assigned a decreased feature distance to $\mathcal{R}_{\rm s}^{t-1}$ and at the same time an increased distance to $\mathcal{R}_{\rm ns}^{t}$. We set $\gamma = 1.5$ empirically. At the first frame, $\mathcal{R}_{\rm s}^{t-1}$ is not available, and thus we set $d_{\rm s}^1(\mathbf{p}_i)$ as the maximum feature distance.

Figs. 3(f) and (g) show the resulting feature distances of $d_{\rm s}^t(\mathbf{p}_i)$ and $d_{\rm ns}^t(\mathbf{p}_i)$, respectively, for the superpixels of Φ^t in Fig. 3(c). In Fig. 3(f), we see that relatively large distances of $d_{\rm s}^t(\mathbf{p}_i)$ are associated with the background superpixels which yield quite dissimilar colors from the foreground objects, while small distances are assigned to the foreground superpixels. In contrary, as shown in Fig. 3(g), relatively large distances of $d_{\rm ns}^t(\mathbf{p}_i)$ are assigned to the foreground superpixels, but the background superpixels have small distances. Based on this property, we compute an initial saliency value $\tilde{S}^t(\mathbf{p}_i)$ for each superpixel $\mathbf{p}_i \in \Phi^t$ using a relative feature distance of $d_{\rm s}^t(\mathbf{p}_i)$ and $d_{\rm ns}^t(\mathbf{p}_i)$ [27].

$$\tilde{S}^t(\mathbf{p}_i) = \exp\left\{ \xi \left(\frac{d_{\rm ns}^t(\mathbf{p}_i)}{d_{\rm s}^t(\mathbf{p}_i) + d_{\rm ns}^t(\mathbf{p}_i)} \right)^2 \right\} \quad (11)$$

where we set $\xi = 2$ empirically. Fig. 3(h) shows the resulting initial saliency values, where high saliency values are assigned to the foreground superpixels which yield small distances of $d_{\rm s}^t(\mathbf{p}_i)$ and large distances of $d_{\rm ns}^t(\mathbf{p}_i)$. We initialize the zero saliency value for all the superpixels outside of Φ^t. The initial saliency values are normalized into the range of [0, 1].

Then we obtain $\mathbf{S}^t$ a set of the final saliency values $S^t(\mathbf{p}_i)$'s of all the superpixels in I^t by minimizing the following cost function, a modified version of the cost function in [20].

$$\begin{aligned} E(\mathbf{S}^t) = & \sum_{\mathbf{p}_i \in I^t} \left(S^t(\mathbf{p}_i) - \tilde{S}^t(\mathbf{p}_i) \right)^2 \\ & + \sum_{\mathbf{p}_i \in I^t} \sum_{\mathbf{p}_j \in \mathcal{N}(\mathbf{p}_i)} w_{\mathbf{p}_i,\mathbf{p}_j} \left(S^t(\mathbf{p}_i) - S^t(\mathbf{p}_j) \right)^2 \\ & + \sum_{\mathbf{p}_i \in I^t} w_{\mathbf{p}_i,\mathbf{q}_i} \left(S^t(\mathbf{p}_i) - S^{t-1}(\mathbf{q}_i) \right)^2. \end{aligned} \quad (12)$$

The first term denotes a data cost. The second term computes a spatial smoothness cost which encourages adjacent superpixels with similar color features to have similar saliency values to each other. $\mathcal{N}(\mathbf{p}_i)$ is the set of the adjacent superpixels to $\mathbf{p}_i$. The weight is defined as $w_{\mathbf{p}_i,\mathbf{p}_j} = \exp\left(-20\|c(\mathbf{p}_i) - c(\mathbf{p}_j)\|\right)$. The last term represents a temporal smoothness cost for temporal coherency of saliency distribution, where temporal neighboring superpixels with similar colors are assigned similar saliency values. $\mathbf{q}_i$ denotes the superpixel in I^{t-1} closest to $\mathbf{p}_i \in I^t$. Fig. 3(i) shows the final saliency distribution where the initially computed saliency values of the background superpixels in Φ^t are further suppressed.

IV. Experimental Results

We evaluate the performance of the proposed algorithm using the two datasets: VideoSeg [29] and SegTrack [28]. The VideoSeg is composed of 10 videos which include large foreground objects with high feature contrast, and the SegTrack is composed of 6 videos exhibiting relatively small objects with various motion characteristics. As did in [14], we do not include the 'penguin' in the SegTrack for experiments, since the provided ground truth saliency maps are not plausible. Also, we remove the frames at the boundaries of the 'cheetah' and 'monkeydog' in SegTrack. We compare the performance of the proposed algorithm qualitatively and quantitatively with that of the six video saliency detection methods: LSD [17], DCMR [10], RSF [16], UW [11], SP [12] and RWRV [19], using the source codes from the authors' websites.

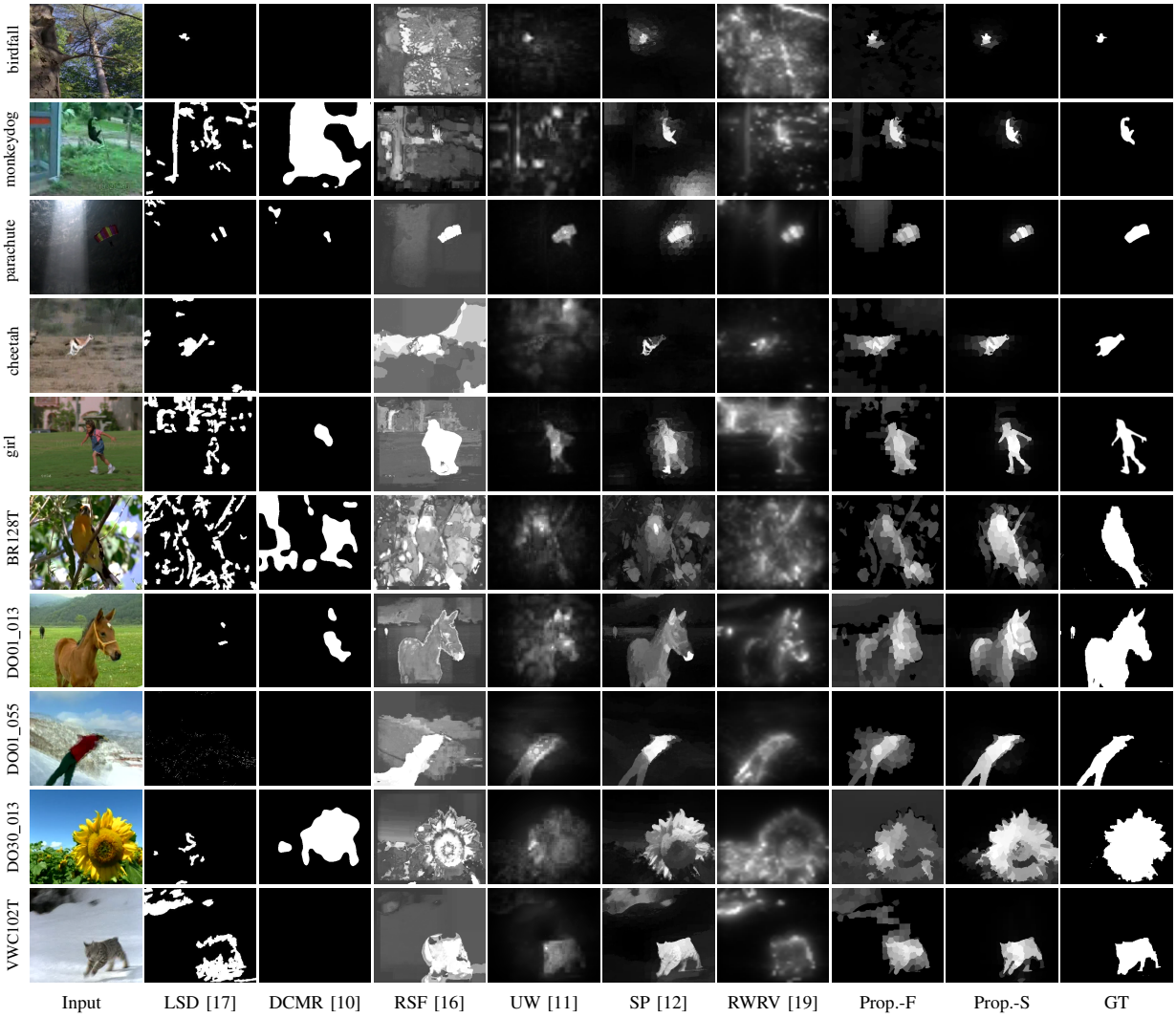


Fig. 4: Comparison of the resulting saliency maps. The first five videos are from the SegTrack [28], and the other five videos are from the VideoSeg [29]. The last column shows the ground truth (GT) saliency maps.

Fig. 4 compares the resulting saliency maps obtained by the proposed algorithm and the existing methods. LSD and DCMR rarely detect the salient objects. RSF tends to highlight the background regions significantly. UW finds overall areas of salient objects but fails to extract the whole salient objects completely. SP integrates the spatial and temporal saliency maps adaptively, and provides relatively good performance as shown in 'parachute' and 'girl' sequences. However, it often results in smoke effects around the salient objects, and it also detects some background regions as salient as shown in 'monkeydog' and 'VWC102T' sequences. RWRV tends to highlight the boundaries of salient objects. It also blurs the saliency maps and captures lots of background regions to be salient, as shown in 'birdfall,' 'girl,' and 'DO30_013' sequences. Prop.-F and Prop.-S show the maps of the combined feature contrast in (4) and the final saliency maps obtained by the proposed algorithm, respectively. We see that the proposed feature extraction catches the salient objects reliably on most sequences, even when a salient object exhibits various colors as shown in 'cheetah,' 'girl,' and 'DO01_055' sequences. Prop.-F also highlights local background regions surrounding the salient objects. However, Prop.-S successfully suppresses these artifacts by local saliency evaluation. In particular, Prop.-S clearly highlights the saliency objects in the 'BR128T' and 'DO01_013' which exhibit little motion features of salient objects. Fig. 5 provides the quantitative comparison results in terms of precision, recall and F-measure score that are evaluated by comparing the ground truth saliency maps and

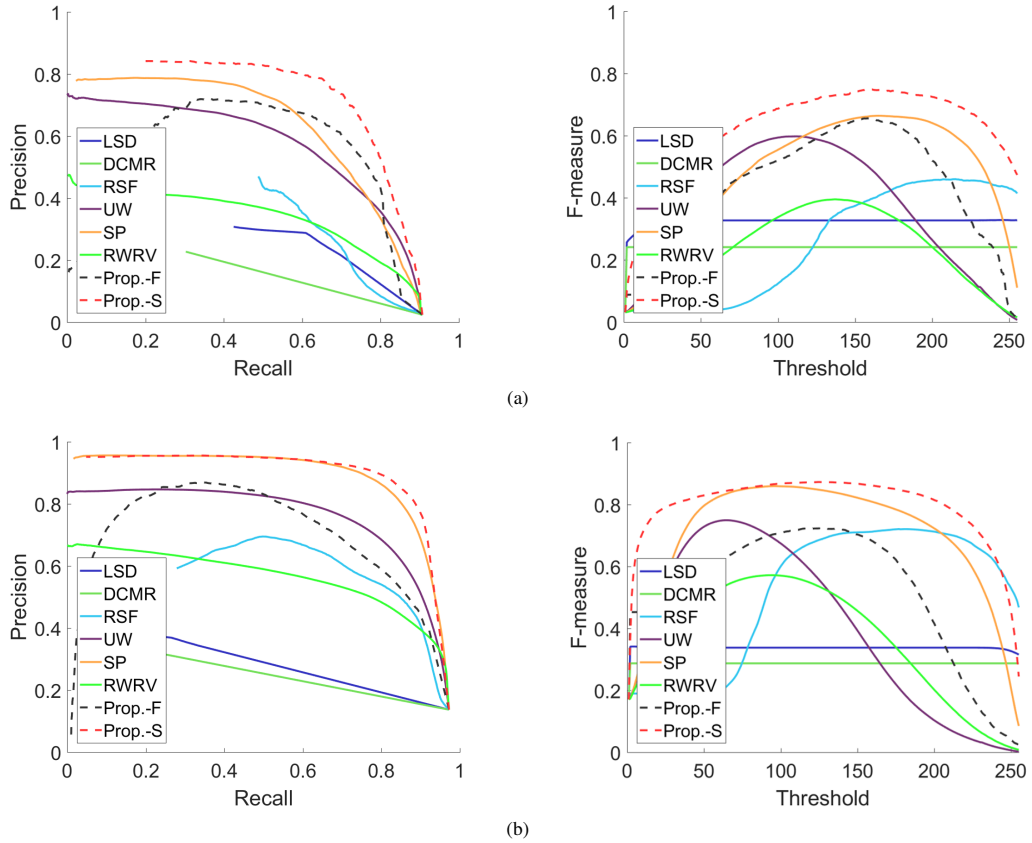


Fig. 5: Quantitative comparison of saliency detection algorithms in terms of the precision, recall, and F-measure score. (a) SegTrack [28] and (b) VideoSeg [29].

the resulting saliency maps which are binarized with various thresholds from 0 to 255 [30]. We see that the proposed algorithm yields the best performance compared with the six existing methods in most of the thresholds.

V. Conclusion

In this paper, we proposed a novel saliency detection algorithm for videos. We combined the color and motion feature contrast adaptively according to their confidence at each frame. We locally constrained the searching area for saliency detection, and evaluated saliency values using a relative feature distance with respect to the salient object and its local background. Experimental results demonstrated that the proposed algorithm detects the video saliency faithfully and yields a better performance than the existing state-of-the-art methods.

Acknowledgement

This work was supported by Institute for Information & Communications Technology Promotion(IITP) grant funded by the KOREA government (MSIT) (No. 2017000670021001, Information-Coordination Technique Enabling Augmented Reality with Mobile Objects).

References

[1] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Fast image retrieval: Query pruning and early termination," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 648–659, May 2013.

[2] L. Wang, J. Xue, N. Zheng, and G. Hua, "Automatic salient object extraction with contextual cue," in *Proceedings of the IEEE International Conference on Computer Vision*, Nov. 2011, pp. 105–112.

[3] X. Wang and C. Qi, "Saliency-based dense trajectories for action recognition using low-rank matrix decomposition," *J. Vis. Commun. Image Represent.*, vol. 41, pp. 361–374, Nov. 2016.

[4] Y. Luo, J. Yuan, P. Xue, and Q. Tian, "Saliency density maximization for efficient visual objects discovery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 12, pp. 1822–1834, Dec. 2011.

[5] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 29–42.

[6] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[7] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep cnn features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, 2016.

[8] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 598–606.

[9] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[10] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1336–1349, Aug. 2014.

[11] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3910–3921, Sept. 2014.

[12] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sept. 2014.

[13] K. Muthuswamy and D. Rajan, "Particle filter framework for salient object detection in videos," *IET Computer Vision*, vol. 9, no. 3, pp. 428–438, 2015.

[14] J. Li, Z. Liu, X. Zhang, O. Le Meur, and L. Shen, "Spatiotemporal saliency detection based on superpixel-level trajectory," *Signal Process.: Image Commun.*, vol. 38, pp. 100–114, Oct. 2015.

[15] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 15–15, 2009.

[16] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *Proceedings of the European Conference on Computer Vision*, 2010, pp. 366–379.

[17] Y. Xue, X. Guo, and X. Cao, "Motion saliency detection using low-rank and sparse decomposition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 1485–1488.

[18] S.-H. Lee, J.-H. Kim, K. P. Choi, J.-Y. Sim, and C.-S. Kim, "Video saliency detection based on spatiotemporal feature learning," in *Proc. IEEE ICIP*, 2014, pp. 1120–1124.

[19] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2552–2564, Aug. 2015.

[20] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.

[21] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3395–3402.

[22] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, 2018.

[23] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[24] J.-S. Kim, J.-Y. Sim, and C.-S. Kim, "Multiscale saliency detection using random walk with restart," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 2, pp. 198–210, Feb. 2014.

[25] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3166–3173.

[26] C. Liu, "Beyond pixels: Exploring new representations and applications for motion analysis," Ph.D. dissertation, MIT, 2009.

[27] D.-Y. Lee, J.-Y. Sim, and C.-S. Kim, "Visual tracking using pertinent patch selection and masking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3486–3493.

[28] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *The IEEE International Conference on Multimedia and Expo*, Jul. 2009, pp. 638–641.

[29] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 190–202, 2012.

[30] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2009, pp. 1597–1604.