*Journal of Information Hiding and Multimedia Signal Processing* ©2019 ISSN 2073-4212
**Ubiquitous International** Volume **10**, Number **2**, March **2019**

# Graph-based Clustering with Spatiotemporal Contour Energy for Video Salient Object Detection

Bing Liu*, Mingzhu Xu and Ping Fu

School of Electronics and Information Engineering,
Harbin Institute of Technology, Harbin, China
liubing66@hit.edu.cn

Received December 2018; revised January 2019

Abstract. *In this paper, we propose a novel graph-based clustering model with spatiotemporal contour energy for video salient object detection, which can preserve the salient object and suppress the irrelevant surrounding background regions effectively. In order to estimate the salient object robustly in spatiotemporal domain, a novel spatiotemporal contour energy is modeled by exploiting optical flow, spatial contour and gradient flow field, which can enhance the energy inside the salient object and weaken it outside the salient object. Then, we estimate the saliency degree of superpixels by computing the geodesic distance of spatiotemporal contour energy between each superpixel node and border background nodes on a superpixel level graph model. The comprehensive experiments show that the proposed model outperforms the state of the art models, which are evaluated on two challenging datasets by three widely used performance metrics. We also applied the saliency map of the proposed method as a prior knowledge to unsupervised video object segmentation, showing that the proposed method can improve the segmentation performance of unsupervised video object segmentation.*
**Keywords:** Video saliency, Spatiotemporal Contour Energy, Graph-based Clustering.

1. **Introduction.** Human observer can be attracted by the most salient and attention-grabbing object in a visual scene[1]. Salient object detection (SOD) aims to distinguish the salient object from the complex visual scene. SOD is motivated by biologically plausible human visual attention mechanisms including the center-surround contrast [2] and feature integration theory (FIT) [3]. It also serve as an initial preprocessing step to select a certain subset of visual information for further enhanced processing. Then, the limited computational resources may be directed toward processing the salient object and improve overall performance. SOD can be applied to many visual tasks, including image/video compression[4], content-aware image/video retargeting[5], content-based image retrieval[6], object recognition and tracking [7],[8], unsupervised video object segmentation [9],[10] and video summarization[11].

A salient object in videos is defined as the one that consistently receives the highest fixation densities[12]. The perceptual vision research shows that the contrast priors are the most important factors in low-level visual saliency: the appearance contrast as spatial saliency cues and the motion contrast as temporal saliency cues. How to locate the salient object in a dynamic visual scene and separate the entire salient object from the background are two key issues for SOD in videos. Several existing video-based SOD models [13],[14] model the motion information of the salient object based on the statistical

*Corresponding author

motion histogram of the optical flow and compute the contrast between the salient object and the background. A novel spatiotemporal contour (STC) is estimated from the contour of the optical flow and the contour of the color frame in [15],[16]. This STC is used to distinguish the salient object from the background. However, although existing video-based SOD models have achieved adequate performance for some easy videos, they may hardly handle more challenging videos, such as motion blur, a dramatic changing background.

In this paper, we propose a novel spatiotemporal contour energy to represent the salient object in a dynamic visual scene and highlight the entire salient object uniformly on a superpixel level graph model. We also applied the saliency map of the proposed method as a prior knowledge to unsupervised video object segmentation, the experiment results show that the proposed method can improve the segmentation performance of unsupervised video object segmentation.

The remainder of this paper is organized as follows. Section 2 gives in detail of the proposed salient object detection for video. In Section 3, experiments are conducted to validate the effectiveness and superiority of the proposed method. An application of our proposed method to unsupervised video object segmentation is given in section 4 and the article is concluded in section 5.

2. **The Proposed Method.** As shown in Fig.1, we firstly model a robust spatiotemporal contour energy by exploiting optical flow, gradient of color frame and gradient flow field, which can highlight the dominant motion regions. Then, an undirected weighted graph is constructed on each single frame, where the nodes are superpixels and two spatially adjacent nodes are connected. Finally, the saliency degree of the each superpixel can be estimated by measuring the geodesic distance between the node and the background prior nodes. In the following part, we will describe the spatiotemporal contour energy and graph-based clustering, respectively.

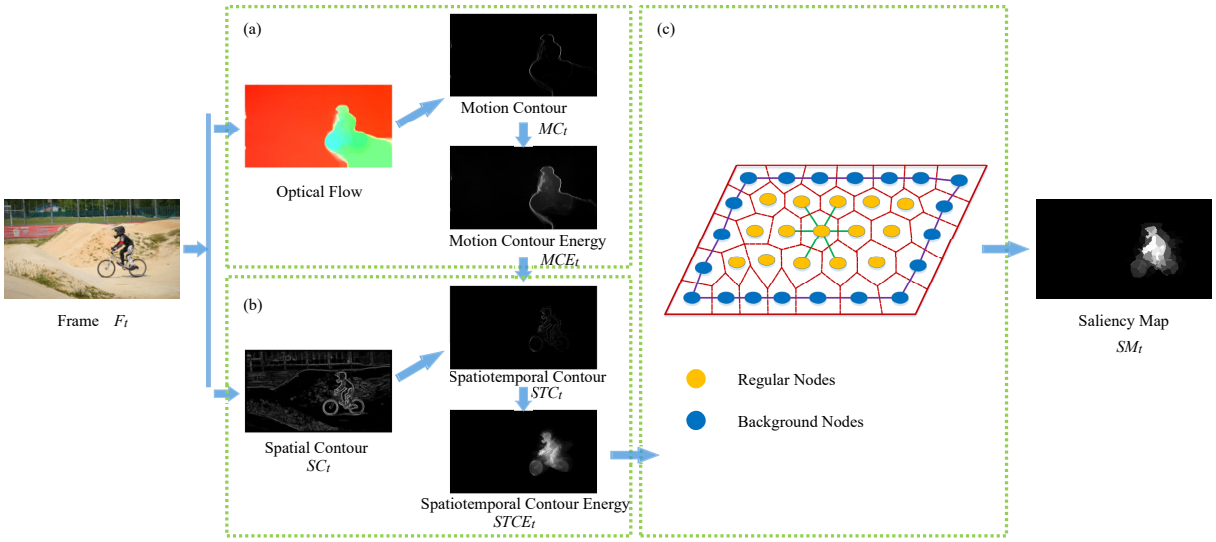


FIGURE 1. Illustration of the proposed framework. We integrate a new spatiotemporal contour energy in a graph model to cluster the salient object.

2.1. **Spatiotemporal contour energy.** For each adjacent frame-pair $\{F_t, F_{t+1}\}$, we can obtain the spatial contour $SC_t$ of each color frame (e.g., $F_t$) by using the contour detection method in [17]. To model the motion information, we extract the motion vector field $V_t$ by computing optical flow (LDOF [18]) between frame-pair $F_{t-1}$ and $F_{t+1}$. The motion

contour $MC_t$ represent the displacement range of the moving salient object and can be obtained by a gradient operation in (1).

$$MC_t(x, y) = \|\nabla V_t(x, y)\| \tag{1}$$

The existing work [15] multiplies $SC_t$ and $TC_t$ to generate a spatiotemporal contour, which represents the spatial contour of the moving salient object. However, the motion contour is larger than the spatial contour and multiplying them directly may yields an inaccurate STC. The inaccuracy may be propagated to the subsequent processing. Different from existing work, we first compute a new motion contour energy (MCE) by using the gradient flow field [15]. The formulation (2) is used to compute the contour energy based on gradient flow field ($GradFF$) operator.

$$MCE_t(x, y) = GradFF(MC_t(x, y)) \tag{2}$$

The $GradFF$ represents the gradient flow operation. The pixel inside a closed boundary region gain a high value and those pixels outside the region gain a low value. The details can be refered to [15].

$MCE_t$ is a mask showing the moving object, which can be seen in Fig.1 (a). Then, our new spatiotemporal contour $STC_t$ can be obtained by multiplying $MCE_t$ with spatial contour $SC_t$ via (3) and it can preserve the spatial contour of the moving salient object inside $MCE_t$ and exclude the irrelevant surrounding background noise outside $MCE_t$ effectively, as shown in Fig.1 (b).

$$STC_t = \log\{SC_t \cdot (1 - \exp(-\lambda \cdot MCE_t)) + 1\} \tag{3}$$

The log is a base-2 logarithm and the coefficient $\lambda$ is set to 1.0. We compute our new spatiotemporal contour energy $STCE_t$ by the $GradFF$ operator. It can be formulated as (4)

$$STCE_t(x, y) = GradFF(STC_t(x, y)) \tag{4}$$

2.2. **Graph-based clustering.** To achieve a reliable saliency estimation, we integrate our new spatiotemporal contour energy in a graph model to cluster the salient object. Specifically, we perform SLIC [19] to segment each frame into a set of superpixels. Then an undirected weighted graph $G_t = (V_t, E_t)$ on each frame is constructed: the nodes $V_t$ are superpixels and any spatial adjacent nodes are connected by the edge $E_t$. Finally, the saliency degree can be estimated by geodesic distance measurement.

Firstly, we select the superpixels touched four image borders as background nodes for that they are likely belong to the background. Any border superpixels are connected each other to form a closed-form graph model. As shown in Fig.1 (c), all blue nodes (border superpixels) are all connected.

Secondly, we design the edge weight based on the spatiotemporal contour energy (STCE) difference between two connected superpixels. The STCE difference between two adjacent superpixels can be defined as follows:

$$W_t(v_i, v_j) = \begin{cases} |STCE_t^i - STCE_t^j|, & if\ i\ ,\ j\ are\ connected \\ 0, & otherwise \end{cases} \tag{5}$$

The $STCE_t^i$ denotes the average STCE of all pixels in the $i$ th superpixel.

Finally, we employ the geodesic distance measurement to estimate the saliency of every superpixel. Specifically, the saliency degree of a superpixel is computed by accumulating the edge weights along the shortest path from $sp_i$ to the borders background nodes of the

frame. It can be formulated as follows:

$$sal_t(v_i) = \min \sum_{m=i,n;m,n\in\aleph}^{n\in B} W_t(v_m, v_n) \tag{6}$$

where $m, n$ are indexes of two connected superpixels and $\aleph$ denotes the set of neighboring nodes.

3. **Experimental Evalution.** We perform an extensive experiments to demonstrate the superiority of our proposed method. In the following part, we give a brief introduction of the datasets and the evaluation criteria, then the results of the proposed method will be compared to the 10 state-of-the-art models. A detail discussion is also followed.

3.1. **Experimental datasets and evaluation criteria.** The experiments are conducted on two widely used benchmark datasets: the UVSD dataset [20] and the DAVIS dataset [21]. UVSD consists of 18 unconstrained videos with complicated motion and complex scenes. A total of 3184 frames in this dataset, and each frame is pixel-wisely annotated on the salient object within each video. DAVIS is a challenging dataset for video object segmentation and it is also popular for video-based SOD. A total of 50 video sequences in this dataset and it has two different resolutions 480p and 1080p, and all 3455 frames are pixel-wisely annotated as the ground truth. We test our proposed method on the 480p for the computation efficiency.

Three widely used performance metrics, which are PR curves, Fmeasure curves and mean absolute error(MAE), are adopted to evaluate the performance of the proposed model.

Here, PR curves refers to the Precision-Recall curves. The precision value is defined as the fraction of salient pixels correctly assigned to all pixels of the extracted regions, while the recall value is defined as the ratio of detected salient pixels with respect to the ground truth foreground pixels. For a saliency map, we generate a set of binary images by using different threshold values from 0 to 255. The precision/recall pairs of all the binary maps are computed to plot the precision-recall curve in the rectangular coordinate system.

Fmeasure is served as the overall performance measure, which is defined as:

$$Fmeasure = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{7}$$

where $\beta^2$ is set to 0.3 to assign a higher importance to precision. We also compute the Fmeasure by different pairs of precision and recall acquired from the above calculation and plot the Fmeasure results in the rectangular coordinate system.

The MAE computes the average difference between the saliency map and the ground truth.

3.2. **Comparison to the state-of-the-art SOD models.** In this section, we compare the proposed method with 10 state-of-the-art models: SAG [22][16], CG [15], RWRV [23], SCUW [24], CBCS [25], MC [26], DRFI [27], SORBD [28], MR [29], SF [30]. The first five models are designed for video-based SOD, while the last five models are designed for image-based SOD. The MC is a new deep models for salient object detection in still images. The results of all these baseline models are generated by using the publicly available codes with the default parameters.

The quantitative comparison results can be seen in Fig.2, from which the top row shows the comparison results on the DAVIS dataset and the bottom row shows the results on the UVSD dataset. In Fig.2, we can observe that our proposed method outperforms all

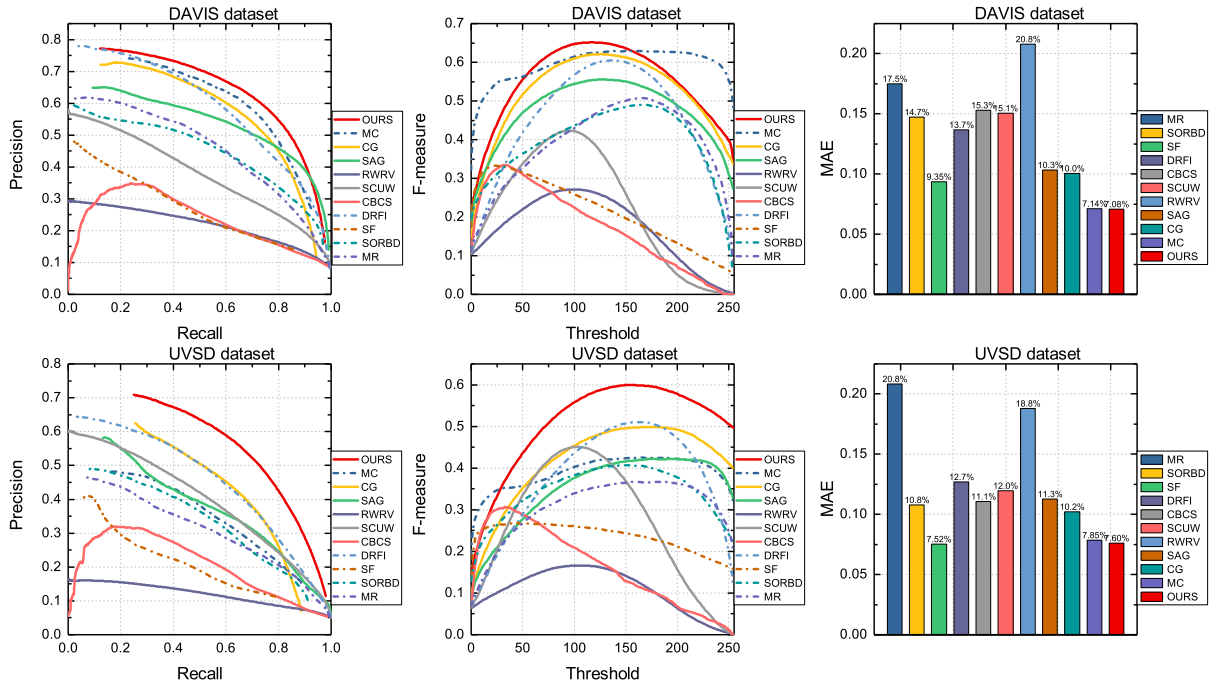


FIGURE 2. PR curves, Fmeasure curves and MAEs of different models on two wildly used benchmark datasets. The solid lines show the performance of five video-based SOD models, while the dashed lines show the performance of five image-based SOD models.

10 state-of-the-art methods, including the deep models MC on both datasets in all three metrics.

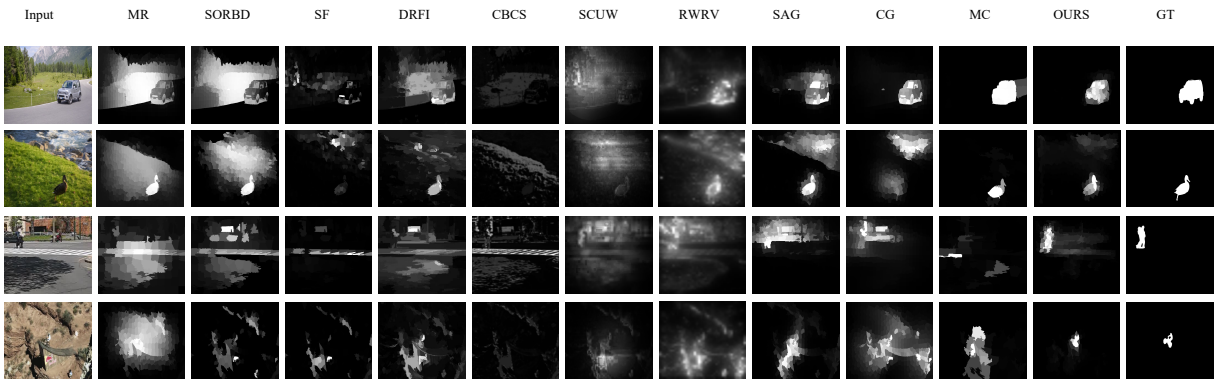


FIGURE 3. Examples of saliency maps generated by the 10 state-of-the-art models and our proposed method on DAVIS and UVSD datasets. The first column shows the input frames, the 2nd-5th columns show the saliency maps of four image-based salient object detection models, the 6th-10th columns show the saliency maps of five video-based salient object detection models, the 11th column shows the saliency maps of the deep learning models, the 12th column shows the saliency results of our proposed model and the final column shows the ground-truth maps.

From Fig.2, it also can be observed that the models designed for videos (represented by the solid lines) achieve a higher performance than the models designed for still images (represented by the dash lines) in terms of the all three metrics generally. The reason

behind this may lie in that the motion cues have been extensively exploited in video-based models and the motion cues play a key role in the salient object detection for videos.

We also compare the saliency results generated by the proposed method with 10 state-of-the-art models visually. In Fig.3, the first two rows show the samples selected from DAVIS dataset and the last two rows show the samples selected from UVSD dataset. The 2-11th columns show the saliency results generated from the 10 baseline models, the 12th column shows the results of our proposed method and the last column shows the manually annotated groundtruth map. These video examples contain complex visual scenes, such as low contrast between the salient object and background (e.g., the 4th row), motion blur caused by camera jitter (e.g., the 3th row), dynamic background noise (e.g., the 2th row).

From Fig.3, we can see that our proposed method achieve a superior performance than other 10 baseline models (including one deep learning method). The five video-based saliency models can better highlight the salient object than the 4 conventional image-based saliency models, by exploiting dominant motion cues. The deep learning method MC also achieve good performance, which benefits from the large scale training data. Our proposed model achieves the state-of-the-art performance and is independent of the large-scale training data, benefiting from the spatiotemporal contour energy.

4. **Application To Unsupervised Video Object Segmentation.** Unsupervised video object segmentation (VOS) is formulated as a binary labeling optimization problem, which aims to separate foreground objects from background in a video and outputs a binary map [10],[31],[21]. On the other hand, salient object detection (SOD) aims to detect the salient object in a video and outputs a probability map (non-binary map) where the value of each pixel represents its probability of belonging to salient objects [32],[33]. Although unsupervised VOS and SOD are different tasks, SOD are also beneficial to unsupervised VOS when the primary objects are the salient objects. Many unsupervised VOS methods use saliency map as an initial foreground likelihood estimation and a set of significant post-processing are also incorporated to improve the final segmentation map [9],[22],[10]. Faktor [9] used saliency map as an initial foreground likelihood votes and correct these votes by consensus voting across entire sequence. Wang [22] estimated the saliency map based on geodesic distance and segmented the salient object by exploiting saliency term, appearance term and location term. Jang [10] constructed a hybrid energy function in terms of a saliency map and optimized the foreground and background distributions in alternate convex optimization way, which achieves state-of-the-art performance.

Next, we validate the impact of our proposed SOD method for improving the segmentation performance of unsupervised VOS method and demonstrate the applicability of our proposed SOD method to unsupervised VOS. Specifically, we use the saliency maps acquired from our proposed method as a substitute for saliency maps in NLC [9]. In other words, we start a crude saliency votes from our saliency map, instead of the saliency map in NLC, and then correct these votes in a consensus voting way across entire sequence to get the final segmentation map iteratively. We use OURS+NLC and NLC [9] to represent them respectively. Additionally, we also give the results of several most recent unsupervised VOS methods, such as ACO [10], FST [31].

We tested all the methods on DAVIS and UVSD datasets using the publicly available source codes with the default parameters or the results in the DAVIS benchmark [21]. In order to evaluate the results properly, we adopted 2 widely used metrics (IOU and Recall) in DAVIS benchmarks and a newly enhanced-alignment metric (E-measure) [34]. IOU measure the intersection over union (IOU) between a segmented map and its ground-truth. Recall measure the fraction of all frames scoring IOU higher than 0.5. E-measure

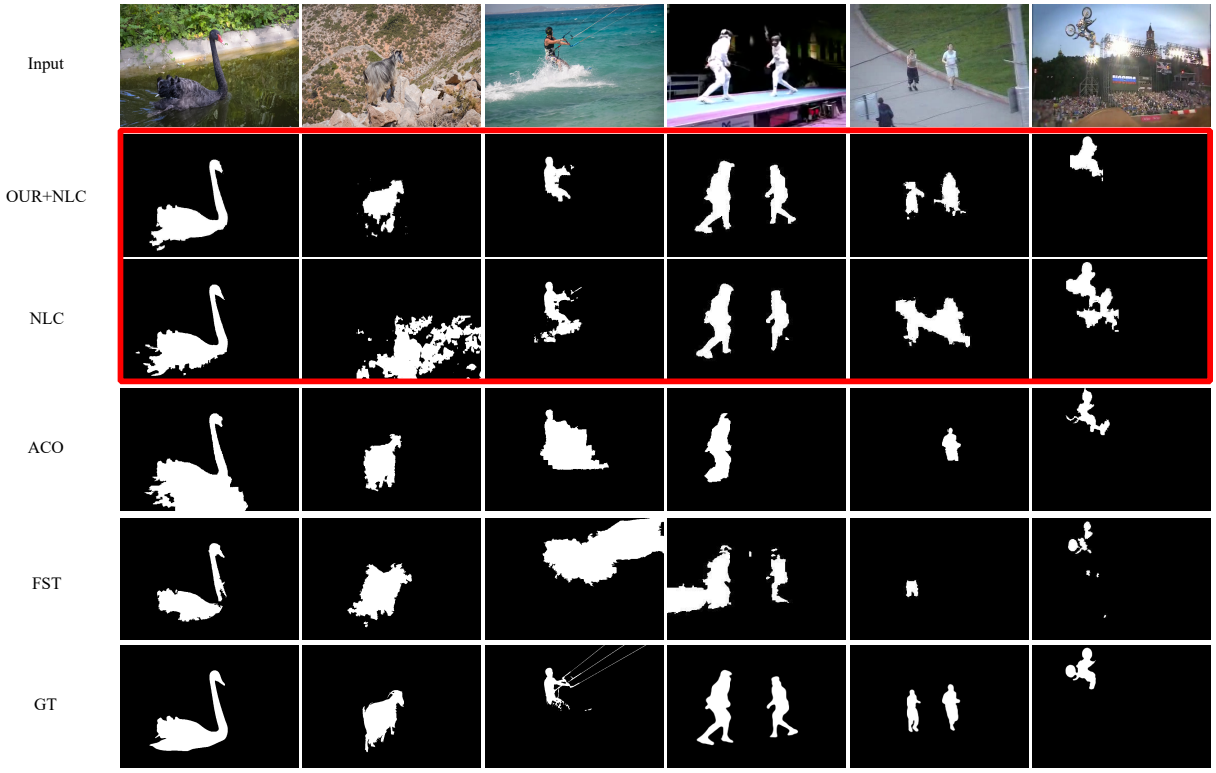


FIGURE 4. Visual comparison between the improved NLC method by our proposed SOD method and the original NLC method on both datasets, which can be seen clearly in the red bounding box. The results of the other 2 most recent unsupervised VOS models (ACO and FST) are shown in 5-5th rows. The 1-3th columns show some examples selected from the DAVIS dataset and the 4-6th columns are selected from the UVSD dataset.

evaluate the object structure similarity (between segmented map and ground-truth) by considering both global statistics and local pixel matching in a compact term. Note that, the higher the metric value, the higher the performance, for all metrics.

TABLE 1. Quantitative Comparison Between Improved NLC Method By Our Proposed SOD Method And The Original NLC Method. The Best is Remarked In Boldface. The Results Of The Other 2 Most Recent Unsupervised VOS Models (ACO and FST) Are Shown In 5-6th Rows.

| Methods | DAVIS | | | UVSD | | |
|---|---|---|---|---|---|---|
| | IOU | Recall | E-measure | IOU | Recall | E-measure |
| OURS+NLC | **0.581** | **0.659** | **0.871** | **0.478** | **0.529** | **0.832** |
| NLC [9] | 0.551 | 0.558 | 0.850 | 0.463 | 0.497 | 0.799 |
| ACO [10] | 0.518 | 0.589 | 0.818 | 0.446 | 0.458 | 0.802 |
| FST [31] | 0.558 | 0.649 | 0.852 | 0.401 | 0.329 | 0.778 |

From TABLE 1, we can see that our proposed SOD method improves the segmentation performance of NLC by 0.030, 0.101 and 0.021 on the DAVIS dataset and 0.015, 0.032 and 0.033 on the UVSD dataset, in terms of IOU, Recall and E-measure respectively. It demonstrates that a good saliency map can improve the performance of unsupervised VOS. This results also proves the superiority of our proposed method.

OURS+NLC performs better than the other two unsupervised VOS methods (ACO, FST) on both datasets in all metrics, which can be seen in TABLE 1 clearly.

Some examples of the segmentation results are shown in Fig.4. We can see that our proposed SOD method improved the segmentation performance of NLC clearly, which are enclosed in a red bounding box.

5. **Conclusions.** In this work, we present a novel graph-based clustering with a new spatiotemporal contour energy for salient object detection in videos. We provide detailed derivations of our new spatiotemporal contour energy and present a superpixel-based graph model that incorporates our new spatiotemporal contour energy to cluster the salient object. We also validate the performance improvement of unsupervised VOS by applying our proposed method to unsupervised VOS.

**Acknowledgment.** This work is supported by National Science Foundation of China under Grant No. 61671170, Natural Science Foundation of Heilongjiang under Grant No. F2015003, the Open Projects Program of National Laboratory of Pattern Recognition under Grant No.201700019.

## REFERENCES

[1] A. Borji, D. N. Sihite, and L. Itti, Salient Object Detection: A Benchmark, *Springer Berlin Heidelberg*, 2012.

[2] C. Koch and S. Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, *Hum Neurobiol*, vol. 4, no. 4, pp. 219-227, 1987.

[3] A. M. Treisman and G. Gelade, Gelade g. a feature integration theory of attention. cog. psychol, *Cogn Psychol*, vol. 12, no. 1, pp. 97-136, 1980.

[4] C. Guo and L. Zhang, A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression, *Trans. Img. Proc.*, vol. 19, no. 1, pp. 185-198, 2010.

[5] Y. Fang, Z. Chen, W. Lin, and C. W. Lin, Saliency detection in the compressed domain for adaptive image retargeting, *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 21, no. 9, p. 3888, 2012.

[6] L. Li, S. Jiang, Z. J. Zha, Z.Wu, and Q. Huang, Partial-duplicate image retrieval via saliency-guided visual matching, *IEEE Multimedia*, vol. 20, no. 3, pp. 13-23, 2013.

[7] Z. Ren, S. Gao, L. T. Chia, and W. H. Tsang, Region-based saliency detection and its application in object recognition, *IEEE Transactions on Circuits Systems for Video Technology*, vol. 24, no. 5, pp. 769-779, 2014.

[8] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, Adaptive object tracking by learning background context, *in Computer Vision and Pattern Recognition Workshops*, pp. 23-30, 2012.

[9] A. Faktor and M. Irani, Video segmentation by non-local consensus voting, 2014.

[10] W. D. Jang, C. Lee, and C. S. Kim, Primary object segmentation in videos via alternate convex optimization of foreground and background distributions, *in Computer Vision and Pattern Recognition*, pp. 696-704, 2016.

[11] Q. G. Ji, Z. D. Fang, Z. H. Xie, and Z. M. Lu, Video abstraction based on the visual attention model and online clustering, *Signal Processing Image Communication*, vol. 28, no. 3, pp. 241-253, 2013.

[12] J. Li, C. Xia, and X. Chen, A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection, *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 349-364, 2017.

[13] Z. Liu, X. Zhang, S. Luo, and O. L. Meur, Superpixel-based spatiotemporal saliency detection, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 9, pp. 1522-1540, 2014.

[14] J. Li, Z. Liu, X. Zhang, O. L. Meur, and L. Shen, Spatiotemporal saliency detection based on superpixel-level trajectory, *Signal Processing: Image Communication*, vol. 38, no. C, pp. 100-114, 2015.

[15] W. Wang, J. Shen, and L. Shao, Consistent video saliency using local gradient flow optimization and global refinement, *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185-4196, 2015.

[16] W. Wang, J. Shen, R. Yang, and F. Porikli, Saliency-aware video object segmentation, *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 40, no. 1, pp. 20-33, 2018.

[17] M. Leordeanu, R. Sukthankar, and C. Sminchisescu, Efficient closed-form solution to generalized boundary detection, *in European Conference on Computer Vision*, pp. 516-529, 2012.
[18] T. Brox and J. Malik, Large displacement optical flow: Descriptor matching in variational motion estimation, *IEEE Trans Pattern Anal Mach Intell*, vol. 33, no. 3, pp. 500-513, 2011.
[19] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, Slic superpixels compared to state-of-the-art superpixel methods, *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 34, no. 11, pp. 2274-2282, 2012.
[20] Z. Liu, J. Li, L. Ye, G. Sun, and L. Shen, Saliency detection for unconstrained videos using superpixel-level graph and spatiotemporal propagation, *IEEE Transactions on Circuits Systems for Video Technology*, vol. PP, no. 99, pp. 1-1, 2016.
[21] F. Perazzi, J. Pont-Tuset, B. Mcwilliams, L. V. Gool, M. Gross, and A. Sorkine-Hornung, A benchmark dataset and evaluation methodology for video object segmentation, *in Computer Vision and Pattern Recognition*, pp. 724-732, 2016.
[22] W.Wang, J. Shen, and F. Porikli, Saliency-aware geodesic video object segmentation, *in Computer Vision and Pattern Recognition*, pp. 3395-3402, 2015.
[23] H. Kim, Y. Kim, J. Y. Sim, and C. S. Kim, Spatiotemporal saliency detection for video sequences based on random walk with restart, *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2552-2564, 2015.
[24] 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, 2014.
[25] H. Fu, X. Cao, and Z. Tu, Cluster-based co-saliency detection, *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 22, no. 10, pp. 3766-3778, 2013.
[26] R. Zhao, W. Ouyang, H. Li, and X. Wang, Saliency detection by multi-context deep learning, *in Computer Vision and Pattern Recognition*, pp. 1265-1274, 2015.
[27] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, Salient object detection: A discriminative regional feature integration approach, *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2083-2090, 2013.
[28] W. Zhu, S. Liang, Y. Wei, and J. Sun, Saliency optimization from robust background detection, *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2814-2821, 2014.
[29] C. Yang, L. Zhang, H. Lu, R. Xiang, and M. H. Yang, Saliency detection via graph-based manifold ranking, *in IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3166-3173, 2013.
[30] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, Saliency filters: Contrast based filtering for salient region detection, *in Computer Vision and Pattern Recognition*, pp. 733-740, 2012.
[31] A. Papazoglou and V. Ferrari, Fast object segmentation in unconstrained video, *in IEEE International Conference on Computer Vision*, pp. 1777-1784, 2014.
[32] A. Borji, M. M. Cheng, H. Jiang, and J. Li, Salient object detection: a benchmark, *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706-5722, 2015.
[33] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, Structure-measure: a new way to evaluate foreground maps, *in ICCV*, 2017.
[34] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, Enhanced-alignment measure for binary foreground map evaluation, *in IJCAI*, 2018.