SPATIOTEMPORAL UTILIZATION OF DEEP FEATURES FOR VIDEO SALIENCY DETECTION

Trung-Nghia Le

Department of Informatics,
The Graduate University for Advanced Studies, Japan
ltnghia@nii.ac.jp

Akihiro Sugimoto

National Institute of Informatics, Japan
sugimoto@nii.ac.jp

ABSTRACT

This paper presents a method for detecting salient objects in a video where temporal information in addition to spatial information is fully taken into account. Following recent reports on the advantage of deep features over conventional hand-crafted features, we propose the SpatioTemporal deep Feature (STF feature) that utilizes local and global contexts over frames. With this feature, we compute the saliency map for each frame through supervised learning of the Random Forest. We then refine the saliency maps using our proposed SpatioTemporal Conditional Random Field (STCRF). STCRF is our extension of CRF toward the temporal domain and formulates relationship between neighboring regions both in a frame and over frames. STCRF leads to temporally consistent saliency maps over frames, contributing to detect boundaries of salient objects accurately and to reduce noise. Our intensive experiments using publicly available benchmark datasets confirm that our proposed method significantly outperforms state-of-the-art methods.

Index Terms— Video saliency, SpatioTemporal CRF, spatiotemporal deep feature, salient object detection

1. INTRODUCTION

Salient object detection from videos plays an important role for many applications such as video re-targeting or visual tracking. Saliency computation methods for videos are usually developed from bottom-up saliency models for still images by incorporating motion features to deal with moving objects[1][2]. Top-down methods were also developed to integrate different features[3] for video saliency computation. These existing saliency computation methods for videos are based on hand-crafted features, which are not sufficiently robust for challenging cases, especially when the salient object is presented in low-contrast and cluttered background; thus they often fail in complex scenes.

Recent advances in deep learning using Convolutional Neural Network (CNN) enable us to extract directly from raw images/videos deep features, which are more powerful for discrimination and, furthermore, more robust than hand-crafted features[4][5]. Indeed, saliency models for videos using deep features[6][7][8] have demonstrated superior results over existing work utilizing only hand-crafted features. They, however, extract deep features from each frame independently and employ frame-by-frame processing to compute saliency, resulting in not working well on dynamically moving objects. This is because temporal information over frames is not taken into account in computing either deep features or saliency maps.

Fig. 1. Examples of results obtained by our proposed method. Top row images are original video frames, followed by the corresponding saliency maps obtained using our method. The second row images are before the refinement and the third row images are our final results.

Computed saliency maps do not always reflect the shapes of salient objects in videos. In order to segment salient objects, in particular, object boundaries, as accurately as possible while reducing noise, the refinement is usually applied to the saliency maps as post-processing. Dense Conditional Random Field (CRF) has been used to refine the saliency map to improve spatial coherence and contour localization[6]. However, CRF is applied to each frame of a video separately, meaning that only spatial contextual information is captured. Again, temporal information over frames is not taken into account.

Motivated by the above observation, we propose a novel framework using spatiotemporal information as fully as possible for salient object detection in videos. Our method con-

978-1-5386-0560-8/17/$31.00 ©2017 IEEE

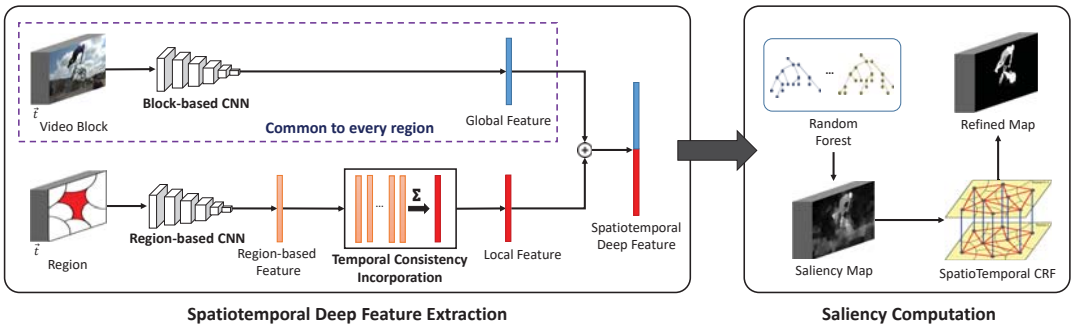


Fig. 2. Pipeline of the proposed method at a single scale.

sists of the feature extraction and the saliency computation; each of which utilizes spatiotemporal information as much as possible. We introduce the SpatioTemporal deep Feature (STF feature) that utilizes both local and global contexts over frames. Our STF feature consists of local and global features. The local feature aggregates over frames, deep features extracted from each frame using a region-based CNN, and global features is computed from temporal-segments using a block-based CNN. The saliency computation, on the other hand, has two steps: map computation and refinement. In the map computation, we supervisedly train the Random Forest (RF) using STF features. The refinement is executed by using our extension of CRF, SpatioTemporal CRF (STCRF), in which temporal consistency of regions over frames as well as spatial relationship between regions in a frame is formulated. With this refinement, boundaries of salient objects are detected accurately with reduced noise (cf. Fig.1). Intensive experiments demonstrate the superiority of our method against the state-of-the-art methods.

We note that though an extension of dense CRF into both spatial and temporal domains, called Dynamic CRF (DCRF), has been used for object segmentation[9] and saliency computation[3] in videos, our STCRF shares with DCRF only the very general-level idea of utilizing spatial and temporal information. The way to construct STCRF is totally different from DCRF. DCRF is formulated at pixel-level while STCRF is at region-level. Accordingly, STCRF is capable of exploiting spatial and temporal information more semantically, which is more suitable to detect salient objects in videos. To facilitate such semantic level expansion, the energy function in STCRF is defined using only deep features, while DCRF is totally based on the combination of classical hand-crafted features such as color and optical flow.

2. PROPOSED METHOD

Our goal is to compute the saliency map to accurately segment salient regions in every frame from an input video with keeping in mind that temporal information is as fully used as possible.

We segment an input video at multiple scales and compute a saliency map at each scale at each frame, and then aggregate all saliency maps at different scales at each frame into the final saliency map. This follows our intuition that objects in a video contain various salient scale patterns and an object at a coarser scale may be composed of multiple parts at a finer scale. In this work, we employ the temporal superpixel segmentation method[10] to segment a video at three scale levels.

Figure 2 illustrates the pipeline of our proposed method at a single scale. The final saliency map is computed by aggregating all saliency maps (output of Fig. 2) at different scales. In the following subsections, we explain how to compute a saliency map at a scale.

2.1. Spatiotemporal Deep Feature Extraction

Our proposed STF feature is the concatenation of local and global features (cf. Fig.2). The local feature is extracted using a region-based CNN followed by aggregation over frames, while the global feature is computed using a block-based CNN whose input is a temporal segment of the video.

A segmented region, namely, a superpixel, at each frame is fed into the region-based CNN to extract its region-based feature. As our region-based CNN, we use the publicly available pre-trained R-CNN model[4]. The region-based feature contains the local context of the region but does not contain temporal information because it is computed frame-wisely. In order to incorporate temporal information, we aggregate region-based features over frames, resulting in the consistent local feature over frames. Just uniformly averaging region-based features over frames is not wise because of pixel fluctuation occurring over time due to the lossy compression, degrades accuracy of corresponding regions over frames. We thus linearly combine region-based features at neighboring frames, similarly to [2], using weights modeled by the Gaussian distribution centered at the frame to compute its local feature. With these weights, region-based features at frames having larger distance to a frame of interest, less contribute to

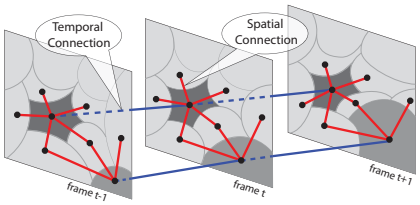


Fig. 3. Graphical representation of our introduced STCRF.

the local feature of the frame.

To compute global features, we feed a temporal segment (sequential frames) of a video into a block-based CNN. The global feature obtained in this way takes its temporal consistency into account in its nature. As our block-based CNN, we employ the pre-trained C3D model[5], which is known to be effective for extracting spatiotemporal features for action recognition. As temporal segments, each frame is expanded into both directions in the temporal domain to obtain a 16-frame block. For each input block, we feed it into the pre-trained C3D model only once and assign the extracted global feature identically to all the regions in the block. This distributes the global context to each region and, at the same time, reduces the computational cost.

2.2. Saliency Map Computation Using RF

In computing the saliency map using STF features, we employ Random Forest (RF), similarly to [11]. This is because saliency computation methods using a neural network require a large dataset for training while methods using RF require much less training dataset, and large video datasets with annotated ground truth for salient object detection are not publicly available. Differently from feeding hand-crafted features to RF as in [11], we feed our STF features to RF.

In our experiments, we employed RF with 500 decision trees. In the training phase, to build the decision tree in the forest, at each split node, we randomly chose 15 elements in the feature vector to compute the best pair of a feature index and a threshold.

2.3. Refinement Using SpatioTempral CRF

CRF is used to enhance accuracy (particularly in object boundaries) of the saliency map while reducing noise, because CRF captures the spatial relationship between regions in a frame. We extend CRF toward the temporal domain to have ability of capturing temporal consistency of regions over frames as well. We call our extended CRF, SpatioTemporal CRF (STCRF in short).

STCRF graph construction: For a segmented temporal region in a block (temporal-segment) of the video, i.e., temporal superpixels, at a scale, we construct a STCRF graph. Each vertex of the graph represents a region, and each edge represents the neighboring relationship between regions in space or in time. Considering all the neighboring relationships, however, leads to a dense graph especially when the video volume is large, and the constructed graph becomes practically useless in the sense of memory consumption and processing time in the inference. We therefore restrict such edges only that represents the adjacency relationship (cf. Fig. 3). Furthermore, we partition the video into chunks of consecutive blocks so that inference in each block is performed separately.

In the experiments, an input video is decomposed into overlapping blocks with a fixed size where the overlapping rate is 50%. We note that each block length is equal to about two seconds. The saliency score of a region is refined by uniformly averaging saliency scores of the region over all the blocks having the region. This averaging reduces processing time while keeping accuracy.

Energy function for STCRF: We define the energy function of our STCRF graph so that probabilistic inference is realized by minimizing the function, like CRF. The energy function E has, as its input for training, a block $\boldsymbol{x}$ and labels $\boldsymbol{l} = \{l_i \mid l_i \in \mathcal{V}\}$ where l_i is the label for region i and $\mathcal{V}$ is the set of vertices, i.e., regions in block $\boldsymbol{x}$. E has the unary term and the binary term:

$$E(\boldsymbol{l}, \boldsymbol{x}; \boldsymbol{\theta}) = \sum_{i \in \mathcal{V}} \psi_{\mathrm{u}}(l_i, \boldsymbol{x}; \theta_{\mathrm{u}}) + \sum_{(i,j) \in \mathcal{E}} \psi_{\mathrm{b}}(l_i, l_j, \boldsymbol{x}; \theta_{\mathrm{b}}),$$

where ψ_{u} and ψ_{b} are the unary and binary potentials, both of which depend on observation $\boldsymbol{x}$. $\mathcal{E}$ is the set of edges of the STCRF graph. $\boldsymbol{\theta} = (\theta_{\mathrm{u}}, \theta_{\mathrm{b}})$ is the model parameters to be found through training.

The unary potential is defined by each region i independently from the saliency score S_i of the region:

$$\psi_{\mathrm{u}}\left(l_i, \boldsymbol{x}; \theta_{\mathrm{u}}\right) = \theta_{\mathrm{u}} S_i(\boldsymbol{x}).$$

The binary potential provides the deep feature based smoothing-term that encourages assigning similar labels to regions with similar deep features. Depending on spatial adjacency or temporal adjacency, the potential is differently formulated: with further separation of θ_{b} into θ_{bs} and θ_{bt},

$$\psi_{\mathrm{b}}\left(l_i, l_j, \boldsymbol{x}; \theta_{\mathrm{b}}\right) = \begin{cases} \theta_{\mathrm{bs}} \exp\left(-\frac{\|F_i(\boldsymbol{x}) - F_j(\boldsymbol{x})\|^2}{2\sigma^2}\right) & (i,j) \in \mathcal{E}_{\mathrm{s}} \\ \theta_{\mathrm{bt}} & (i,j) \in \mathcal{E}_{\mathrm{t}} \end{cases},$$

where $\mathcal{E}_{\mathrm{s}}$ and $\mathcal{E}_{\mathrm{t}}$ denotes the set of edges representing spatial adjacency and that representing temporal adjacency. Namely, $\mathcal{E} = \mathcal{E}_{\mathrm{s}} \cup \mathcal{E}_{\mathrm{t}}$ and $\mathcal{E}_{\mathrm{s}} \cap \mathcal{E}_{\mathrm{t}} = \emptyset$. $F_i(\boldsymbol{x})$ is the STF feature of region i, and σ is the parameter for the distance function. Differently from other works, our model takes advantage of contextual deep feature smoothness to improve the final result.

In order to minimize our energy function $E(\boldsymbol{l}, \boldsymbol{x}; \boldsymbol{\theta})$, we employ the Loopy Belief Propagation (LBP) inference [12], which is a generalization of forward-backward procedure. We

Table 1. Compared state-of-the-art methods and classification.

target	hand-crafted feature	deep feature
video	LD[3], LGFOGR[13], RST[2], SAG[14], STS[1]	None
image	None	DCL[6], DHSNet[7], ELD[8]

Table 2. Quantitative comparison with state-of-the-art methods on three datasets, using F-measure (F-Adap) (higher is better) and Mean Absolute Errors (MAE) (smaller is better). The best and the second best results are shown in blue and red, respectively. Our method, denoted by STCRF, is marked in **bold**.

Dataset	10-Clips		SegTrack2		DAVIS	
Metric	F-Adap ⇑	MAE ⇓	F-Adap ⇑	MAE ⇓	F-Adap ⇑	MAE ⇓
STCRF	0.927	0.021	0.817	0.024	0.794	0.030
LD[3]	0.637	0.197	0.286	0.281	0.252	0.302
LGFOGR[13]	0.629	0.207	0.500	0.117	0.537	0.102
RST[2]	0.827	0.055	0.510	0.125	0.627	0.077
SAG[14]	0.755	0.117	0.504	0.106	0.494	0.103
STS[1]	0.591	0.177	0.471	0.147	0.379	0.183
DCL[6]	0.935	0.031	0.734	0.060	0.664	0.067
DHSNet[7]	0.923	0.022	0.733	0.050	0.715	0.048
ELD[8]	0.893	0.023	0.611	0.065	0.572	0.081

used the UGM toolbox[1] for both training and inference processes in the experiments.

3. EXPERIMENTAL RESULTS

3.1. Benchmark Datasets and Evaluation Criteria

We evaluated the performance of our method on three public benchmark datasets: 10-Clips dataset[15], SegTrack2 dataset[16], and DAVIS dataset[17], which are with 10, 14, and 50 video sequences, respectively. SegTrack2 and DAVIS are challenging datasets and good platforms to evaluate the robustness of methods because of frequent occurrences of occlusions, motion blur, and appearance changes though they are for video object segmentation but not for salient object detection. All the datasets contain manually annotated pixel-wise ground-truth for every frame.

In training RF and STCRF models, we took an approach where we use all three datasets together rather than training RF and STCRF for each dataset. This is because each dataset is too small to train reliable models. Our approach also enables the trained model not to over-fit to a specific dataset.We mixed all three datasets into one (larger) dataset and sampled only 10% frames for the training set and the remaining frames for the testing set. The parameters required in our method are two standard deviations appearing in the linear combination of region-based features and in the binary potential of the STCRF energy function. We set them to be, respectively, 2.0 and 5.0.

We evaluated the performance using **F-measure**, and Mean Absolute Error (**MAE**). F-measure is a balanced measurement between $Precision$ and $Recall$ as follows: $F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}$. We remark that we set $\beta^2 = 0.3$ for F-measure so that precision is more considered. MAE, on the other hand, is the average over the frame of pixel-wise absolute differences between the ground truth and obtained saliency scores.

For a threshold, we binarize the saliency map to compute Precision and Recall at each frame in a video and then take the average over the video. After that, the mean of the averages over videos in a dataset is computed. F-measure is computed from the final Precision and Recall. When binarizing results for the comparison with the ground truth, we also used **F-Adap**[18], an adaptive threshold $\theta = \mu + \eta$ where μ and η are the mean value and the standard deviation of the saliency scores of the obtained map. MAE are computed in the same way.

3.2. Comparison with the State-of-the-Arts

We compared the performance of our proposed method (denoted by STCRF) with several state-of-the-art methods for salient object detection, which are classified in Table 1. We remark that we run original codes provided by the authors with recommended parameter settings for obtaining results. We also note that we frame-wisely applied the methods developed for the still image to videos.

Figure 4 shows examples of obtained results. Qualitative evaluation confirms that our method produces the best results on each dataset. Our method can handle complex foreground and background with different details, giving accurate and uniform saliency assignment. In particular, object boundaries are clearly kept with less noise, compared with the other methods.

In order to quantitatively evaluate the obtained results, we first computed F-measure curves (cf. Fig.5). F-measure indicates that our method significantly outperforms (is at least comparable with) the other methods at every threshold on all the datasets. Since 10-Clips dataset is easiest among the three, any methods can achieve good results while the other two datasets are challenging, meaning that the effectiveness of methods becomes discriminative. Indeed, compared with the second best and third best methods, DCL[6] and DHSNet[7], our method is comparable on 10-Clips dataset and significantly better on the other datasets.

Table 2 illustrates the evaluations in terms of F-Adap and MAE. Our method achieves the best performance under any metric on all the datasets. There is only one case that STCRF is the second best method while DCL[6] is the best method on 10-Clips dataset, which is the easiest dataset among the three datasets.

3.3. Detailed Analysis of the Proposed Method

To demonstrate the effectiveness of utilizing STF features and STCRF, we performed experiments under four different controlled settings. We compared the proposed method, denoted

[1] http://www.cs.ubc.ca/ schmidtm/Software/UGM.html

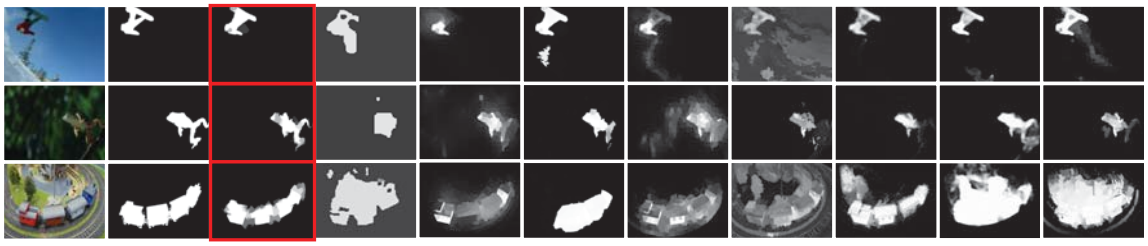
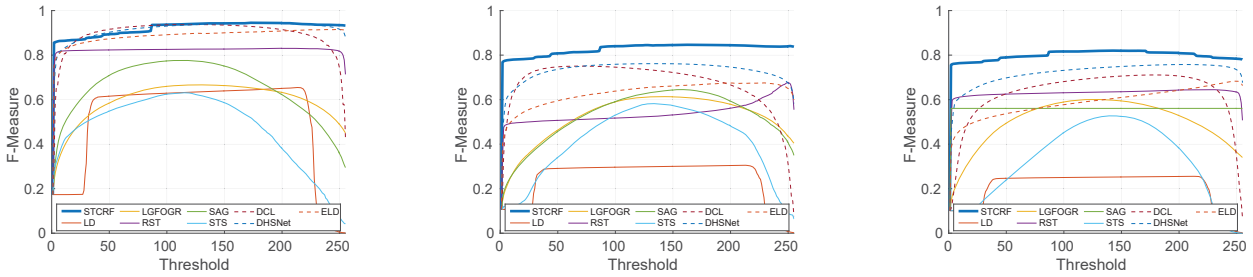


Fig. 4. Visual comparison of our method against the state-of-the-art methods. From left to right, original image and ground-truth are followed by outputs obtained using our method (STCRF), LD[3], LGFOGR[13], RST[2], SAG[14], STS[1], DCL[6], DHSNet[7], ELD[8], in this order. Our method surrounded with red rectangles achieves the best results.



(a) 10-Clips Dataset (b) SegTrack2 Dataset (c) DAVIS Dataset

Fig. 5. Quantitative comparison with state-of-the-art methods on three benchmark datasets, using F-measure with different thresholds. Our method is denoted by STCRF (thick blue).

by STF+STCRF, with three baseline methods as in Table 3. We note that LF and STF are for the evaluation of STF against local features alone and that STF+CRF is for the evaluation of STCRF against CRF.

Figure 6 shows F-measure under different thresholds. We see that STF+STCRF achieves better (at least comparable) results than all the baseline methods on the three datasets. The t-test with significant level 0.15 at every threshold confirmed that for 10-Clips and DAVIS datasets, F-measure of STF+STCRF is significantly better than that of STF+CRF (the best one among the baselines) for the thresholds in $[60, 90]$ while two methods have the same performance for the other thresholds. For SegTrack2 dataset, the two methods are confirmed to have the same performance for any threshold.

Evaluation results using F-Adap and MAE are demonstrated in Table 3, indicating that STF+STCRF exhibits the best performance on all the three datasets. We also see that (1) combing global and local features improves accuracy against using local features alone and that (2) the refinement with STCRF effectively works and brings more gain than that with CRF (cf. Fig. 1). In conclusion, our (complete) method captures local and global contexts over frames to produce accurate final saliency maps.

4. CONCLUSION

Differently from the still image, the video has temporal information and how to incorporate temporal information as effectively as possible is the essential issue for dealing with the video. This paper focused on detecting salient objects from a video and proposed a method using STF features and STCRF. Our method takes into account temporal information in a video as much as possible in different ways, namely, feature extraction and the saliency computation. Our proposed STF feature utilizes local and global context in both spatial and temporal domains. STCRF is capable of capturing temporal consistency of regions over frames and spatial relationship between regions.

5. REFERENCES

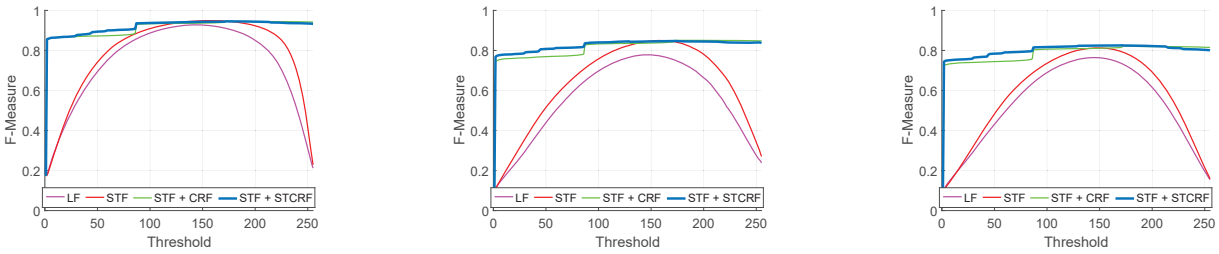
[1] F. Zhou, S. B. Kang, and M.F. Cohen, "Time-mapping using space-time saliency," in *CVPR*, June 2014, pp. 3358–3365. 1, 4, 5

[2] T.-N. Le and A. Sugimoto, "Contrast based hierarchical spatial-temporal saliency for video," in *PSIVT*. 2015, vol. 9431 of *LNCS*, pp. 734–748, Springer. 1, 2, 4, 5

[3] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *TPAMI*, vol. 33, no. 2, pp. 353–367, Feb 2011. 1, 2, 4, 5

Table 3. The detail of settings of the experiments, using F-measure (F-Adap) and Mean Absolute Errors (MAE) (compared with the baseline methods). The best results are shown in blue (higher is better for F-Adap and lower is better for MAE). Our complete method, denoted by STF-STCRF, is marked in **bold**.

setting description	local feature	global feature	saliency refinement	10-Clips		SegTrack2		DAVIS	
				F-Adap ⇑	MAE ⇓	F-Adap ⇑	MAE ⇓	F-Adap ⇑	MAE ⇓
LF	with	without	without	0.868	0.095	0.590	0.125	0.600	0.122
STF	with	with	without	0.887	0.075	0.658	0.099	0.650	0.107
STF+CRF	with	with	CRF	0.916	0.021	0.789	0.026	0.763	0.031
STF+STCRF	with	with	Spatiotemporal CRF	0.927	0.021	0.817	0.024	0.794	0.030



(a) 10-Clips Dataset (b) SegTrack2 Dataset (c) DAVIS Dataset

Fig. 6. Comparison with the baseline methods using F-measure under different thresholds. Our (complete) method is denoted by STF+STCRF (thick blue.)

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, June 2014, pp. 580–587. 1, 2

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *ICCV*, Dec 2015, pp. 4489–4497. 1, 3

[6] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *CVPR*, June 2016, pp. 478–487. 1, 4, 5

[7] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *CVPR*, June 2016, pp. 678–686. 1, 4, 5

[8] G. Lee, Y. W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, June 2016, pp. 660–668. 1, 4, 5

[9] Y. Wang, K.-F. Loe, and J.-K. Wu, "A dynamic conditional random field model for foreground and shadow segmentation," *TPAMI*, vol. 28, no. 2, pp. 279–289, Feb 2006. 2

[10] J. Chang, D. Wei, and J.W. Fisher, "A video representation using temporal superpixels," in *CVPR*, June 2013, pp. 2051–2058. 2

[11] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, June 2013, pp. 2083–2090. 3

[12] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999, pp. 467–475. 3

[13] 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, Nov 2015. 4, 5

[14] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *CVPR*, June 2015, pp. 3395–3402. 4, 5

[15] K. Fukuchi, K. Miyazato, A. Kimura, S. Takagi, and J. Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *ICME*, June 2009, pp. 638–641. 4

[16] F. Li, T. Kim, A. Humayun, D. Tsai, and J. M. Rehg, "Video segmentation by tracking many figure-ground segments," in *ICCV*, Dec 2013, pp. 2192–2199. 4

[17] 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 *CVPR*, June 2016, pp. 724–732. 4

[18] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *ICCV*, Dec 2013, pp. 1761–1768. 4