SALIENT OBJECT DETECTION IN IMAGE SEQUENCES VIA SPATIAL-TEMPORAL CUE

Chuang Gan[1,2]*, Zengchang Qin*[2]*, Jia Xu*[1]*, Tao Wan*[2,3]

[1] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2] Intelligent Computing and Machine Learning Lab, Beihang University, Beijing, China
[3] School of Biological Science and Medical Engineering, Beihang University, Beijing, China

ABSTRACT

There are large amounts of videos available online with the increasing popularity of Internet world wide. It is a non-trivial task for accurately searching and categorizing these videos due to the variety of contents contained in the videos. Visual saliency models provide a possible way to solve this problem by locating and extracting salient objects from the background within images, which can reduce the search effort and assist object detection and recognition tasks. Compared to static images, videos contain motion information which might be more likely to attract human attention. In this paper, we present a new region contrast based saliency detection model using spatial-temporal cues (RCST). We extend the general static image saliency computation model to handle videos by incorporating both spatial and temporal features. Four general saliency principles and three methods are introduced to evaluate the saliency detection performances of the RCST based method in terms of qualitative and quantitative evaluations using a publicly available video segmentation database. The experimental results demonstrate that our algorithm outperforms the existing state-of-the-arts methods.

Index Terms— object detection, saliency, spatial-temporal cue.

1. INTRODUCTION

Rapid development of computer infrastructure including increased speed of processors, less expensive but increasing capacity of storage device and easily accessible Internet have brought in a vast number of videos in past decades. It is a great challenge for us to search or categorize these videos. Salient areas in an image or a video are generally regarded as the focus in human eyes. Visual saliency models can help us to locate salient objects from the background for the purpose of effective search. Saliency detection can help audience to locate the most attractive and important content from extensive images and videos.

This work was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61061130540. Emails: zcqin@buaa.edu.cn, tao.wan.wan@gmail.com

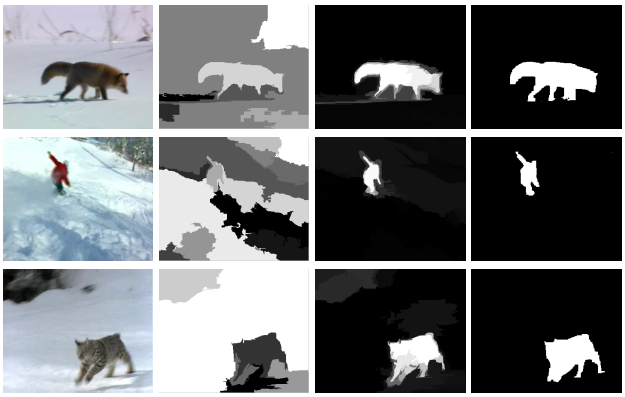


Fig. 1. Three examples illustrating the saliency detection results compared to ground truth. From left to right: (a) input frames, (b) graph based segmentation, (c) salient objects detected by our method, and (d) ground truth.

Visual saliency is originally a task of predicting the eye-fixations on images, and recently has been extended to locate a region containing the salient object. There are various applications including the salient object detection and recognition [1, 2], image compression [3], image cropping [4], image retrieval [5], photo collage [6, 7] and so on. The study on human visual systems suggests that the saliency is related to uniqueness, rarity and surprise of a scene, characterized by primitive features like color, texture, shape and etc. For example, Fig. 1 illustrates the procedure of graph segmentation based saliency detection for an input frame. Recently, a lot of efforts have been made to design various algorithms to compute the saliency for static images [8, 9, 10, 11, 12, 13]. However, there is not much literature to consider extending the saliency computation models to videos related tasks.

In this paper, we propose a novel saliency model combining both color and motion feature to create a saliency map. Different from previous work such as [12, 14], our method can be summarized by the following four steps:

1. Initial graph based image segmentation.
2. Local contrast based region refinement.

3. Saliency map computation by combining color and motion information weighted by regions spatial distance and area ratio.

4. Fusing spatial weight of each region to generate final saliency maps.

Comprehensive experiments have been done on a public dataset[1], and comparisons are made to state-of-the-art models. The rest of paper is organized as follows. In Section 2, we review the related work of saliency models development. In Section 3, we introduce our region contrast based saliency detection model via spatiotemporal cues (RCST). In Section 4, extensive experiments have been done and compared with seven state-of-the-art methods. Section 5, the conclusion and related application have been discussed.

2. PREVIOUS WORK

The basis of most recently bottom-up computation saliency models are inspired by the concept of the Feature Integration Theory (FIT) by Treisman and Gelade [15], which posits that different kinds of attention are responsible for binding various features into consciously experienced wholes. In [16], saliency models can be roughly divided into two kinds: local contrast and global contrast.

Local contrast based method estimates saliency of a particular patch based on their dissimilarity with neighbors. Itti *et al.*[17] proposed central-surrounded differences based on a set of pre-attentive image feature. Ma and Zhang [18] proposed a saliency map by using a fuzzy growth model. Liu *et al.* [11] find multi-scale contrast by linearly combining contrast in a Gaussian image pyramid. Goferman *et al.*[19] simultaneously model local low-level clues, global considerations, visual organization rules, and high-level features to highlight salient objects along with their contexts. Jiang *et al.*[13] computes differences between the color histogram of a region and its immediately neighboring regions are used to evaluate the saliency score. The saliency maps, computed from multi-scale image segmentation to capture non-local contrast.

Global contrast based method considers the contrast relationship over the whole image. Zhai and Shah [12] define pixel-level saliency based on a pixel's contrast to all other pixels. However, for efficiency they use only luminance information, thus ignoring distinctiveness clues in other channels. Hou and Zhang [10] detect saliency in frequency domain by tuning the amplitude of spectrum of image. Achanta *et al.* [8] propose a frequency tuned method that directly defines pixel saliency using a pixel's color difference from the average image color. Cheng *et al.* [9] uses color histogram to compute color difference between pixels and the whole image, aims to directly compute the global uniqueness. Based on the regional contrast, element color uniqueness and spatial distribution are introduced to evaluate the saliency scores of regions [20].

However, these methods are limited to static images. If given a video sequence, the detection of saliency may vary significantly. For example, we may focus more on the dog that runs across the road compared with other cars that go along in the same direction, so the dog should be considered as the salient object. It may seem difficult to distinguish the salient dog based on the usual saliency model in static images. Thus it is necessary to incorporate some temporal cues to aid on salient object detection in video. Some previous work has added motion information into the saliency model. For example, Zhai and Shah [12] propose a method based on spatial-temporal cue, and Guo *et al.*[14] proposed a method based on phase spectrum quaternion fourier transform to efficiently compute the spatiotemporal salient map. However, the experiments results are not satisfactory.

To solve the saliency region detection task in videos, we first define four general saliency principles specific for videos, and then proposed the region contrast method based on the principles. To better illustrate the proposed method, extensive experiments have been done and evaluated based on three different objective comparison measures.

3. THE RCST-BASED SALIENCY DETECTION METHOD

In order to solve the problem of detecting salient regions in image sequences, we first introduce four basic principles of human visual attention. Based on these principles, a region contrast based method is proposed to incorporate color and motion cues for video tasks. The framework of processing is illustrated in Fig. 2 and outcomes of the saliency detection are in Fig. 3. Compared to previous works, it can been seen that our method is more robust and has promising performances for videos related application.

3.1. Principles of salient region in image sequences

Based on human cognition of attention, we summarize four principles for salient regions in video.

- The salient region always stands out from surrounding context in a certain aspect, such as color and motion.
- The color and motion within the same salient region is coherent.
- The spatial distribution of salient regions always more centralized than the background.
- The salient region is most probably placed near the center of the image.

[1] http://www.brl.ntt.co.jp/people/akisato/saliency3.html

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

Fig. 2. The series of images showing the procedure of the RCST-based salient object detection. From left to right: (a) input image, (b) graph based segmentation, (c) segments after the region merging process, (d) region contrast based saliency map, (e) saliency map with spatial weight, and (f) ground truth.

The first principle, based on bottom-up salient stimuli, has been noted in previous work [13]. However, what makes a little different is that we also incorporate motion cue into consideration, and then make our saliency computation model more suitable for salient objects detection in videos. The second principle is based on the assumption that the pixels within the same objects always have coherent motion, even when the camera is moving. The third principle is based on the assumption that the distribution of color and motion in the background is always more centralized. The last principle is also known as golden ratio. The camera always lets the salient objects locate near the center of scene, so that the viewer can easily realize the core of images.

3.2. Local contrast based super pixels generation

As we know, the initial segmentation results always have great impacts on the final saliency detection performances. In this section, we propose a method based on principle 2 to help get a robust segmentation result. Further experiments show that this step can help the saliency maps be less sensitive to the segmentation parameters. The detailed discussions of the parameters selection will be shown in Section 4.

3.2.1. Feature Extraction

As above mentioned, one major contribution of our $RCST$ model is to take both color and motion cues into consideration. We will briefly describe the feature extracted in these two levels, respectively in the following.

For color perspective, we extract the color histogram in $CIE\ L*a*b$ and hue space as region descriptors. And for motion perspective, we compute horizontal and vertical moving information of each pixel in the input frames based on optical flow approach proposed by Liu [21], which can be represented by (dx, dy). Then the moving orientation and magnitude will be computed based on Eq. (1) and Eq. (2).

$$ori = \arctan(dy/dx) \tag{1}$$

$$mag = \sqrt{dx^2 + dy^2} \tag{2}$$

Thus, each pixel in the image can be represented by six dimensions feature (L, a, b, h, ori, mag), then they will be quantized into several bins. To be noted that the histogram of color(L, a, b, h) and histogram of motion(ori, mag) are extracted respectively. Then the histogram distance of color and motion will be computed separately, and fused to evaluate the similarity between the neighbor regions.

3.2.2. Region segmentation

The first step of our method is to generate sub region $(superpixels)$. The algorithm will be introduced in the following.

We apply the graph based image segmentation approach [22] to initially decompose the image into N several regions and then merge the similar regions based on principle 2. The distance between the region i and region j can be formulated as Eq. (3):

$$D(i,j) = \beta d_{color}(i,j) + \lambda d_{motion}(i,j) \tag{3}$$

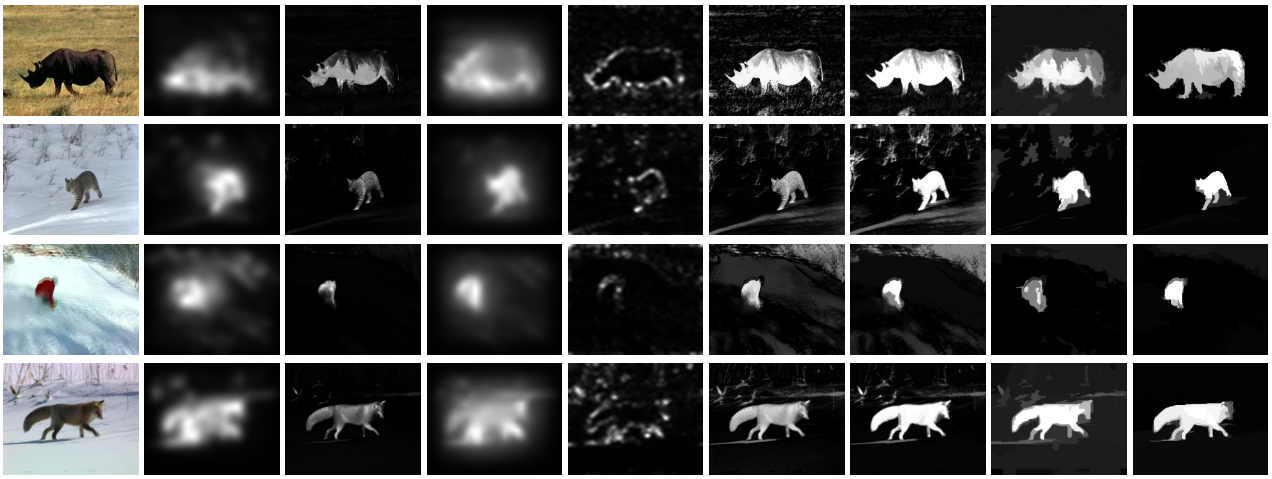
β and λ are the weight of color histogram distance and motion histogram between two regions. If $D(i,j)$ below threshold th, then the two regions will be merged. As usual, this merging process can help reduce half of the number of the regions, and it can make the whole object group into the same region, which not only contributes a lot to final saliency maps generation, but also improves region contrast computation efficiency.

3.3. Region contrast based saliency computation model

After initial pre-processing, each frame in the videos can be decomposed into several sub region. The following will describe how the principles 1, 3, and 4 that we proposed above can be formulated to the $RCST$ saliency computation model.

3.3.1. Saliency map based on region contrast

Based on principle 1, we compute the saliency score for each region. The saliency score for each region i is mainly based on the color and motion contrast with remaining regions. We should follow two rules for different remaining regions in contrast computation.



(a) input (b) IT [17] (c) FT [8] (d) GB [23] (e) SR [10] (f) LC[12] (g) HC [9] (h) RC[9] (i) RCST

Fig. 3. Visual comparison of saliency maps obtained from different methods. The RCST-based method achieved superior detection performance with distinct object boundary and clean background compared to the existing state-of-the-art detection methods.

- The nearer region should be assigned more weight for contrast computation.
- The larger region should be assigned more weight for contrast computation.

Then the formulation of the proposed model can be represented by Eq. (4) and Eq. (5).

$$S(i) = -log(1 - \sum_{k=1}^{n} \alpha_{ik} w_k \times D(i,k)) \tag{4}$$

$$\alpha_{ik} = \frac{1}{2\sigma_1} \exp(-((\frac{x_i - x_k}{W})^2 + (\frac{y_i - y_k}{H})^2)) \tag{5}$$

w_k is the ratio of area remaining region k to the area of the whole input image. $D(i,j)$ evaluates the contrast between region i and region k. α_{ik} is the spatial weight of each remaining region k. σ_1 is a weighted parameter. n is the number of remaining regions. W and H represent the width and height of the input image, respectively. x_i, y_i represent the average x and y position belonging to region i.

3.3.2. Spatial weight

Based on principles 3 and 4, we will assign the region nearer to center and more centralized a higher spatial weight. The formulation is below as Eq. (6).

$$E(i) = \frac{1}{2\sigma_2} \exp(-((\frac{x_i - x_0}{W} \times \frac{\mathrm{var}(x_i)}{W})^2 + (\frac{y_i - y_0}{H} \times \frac{\mathrm{var}(y_i)}{H})^2)) \tag{6}$$

σ_2 is a weighted parameter, $\mathrm{var}(x_i)$ and $\mathrm{var}(y_i)$ represent variance of x and y coordinate positions in region i. x_0 and y_0 represent the center position of the input frame.

3.3.3. Final saliency map

For each region i can be represented by two saliency scores $S(i)$ and $E(i)$. Then final saliency map is then fusing region contrast saliency and spatial weight map as Eq. (7), then normalized to [0, 255].

$$F(i) = S(i) \times E(i) \tag{7}$$

In the experiments, we find that the weight between the region contrast saliency map and spatial weight map may fluctuate in different situations. To ensure fairness, we assign them to the same weight.

4. EXPERIMENTAL RESULTS

In this section, we have evaluated the results of our approach on the publicly available database provided by Kimura [24]. We compare the proposed method with seven state-of-art saliency detection methods, according to: the number of citation [17, 10], recently [9] and variety[8, 12, 23, 23] . We apply our method and others to compute saliency maps for the image sequences in the database. In order to comprehensive evaluate the accuracy of our method for salient object segmentation, we perform three different objective comparison measures. Visual comparison of saliency map can be seen in Fig. 3.

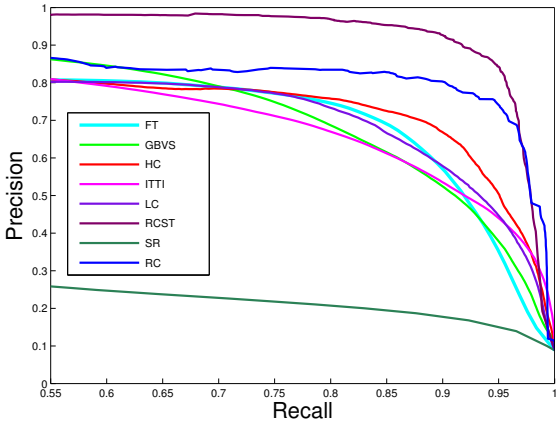


Fig. 4. Precision and recall curve. The RCST-based detection method yielded better detection precisions as well as higher recall values.

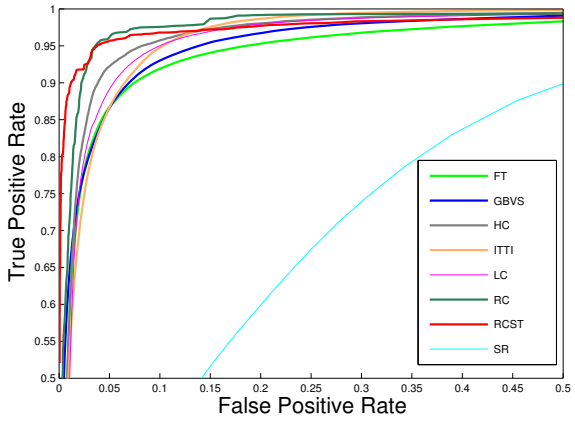


Fig. 5. ROC curve. The RCST-based detection method yielded the best performance when there is less false positive rate.

4.1. Experimental design

This database contains 10 uncompressed video clips of natural scenes with 12 frames a second, including at least one target objects or something others. Length varies between 5-10 seconds. And it also provided corresponding ground truth in the form of accurate human-marked labels for salient regions excluding the first 15 frames. In the experiment, we select six videos from the video segmentation database (nearly 600 images), which have obvious motion information, as our video saliency object detection database.

4.1.1. Implementation details

In the experiments, the segmentation and threshold parameters are a little bit sensitive. More region generated will cause better result but may bring the computation burden. Therefore, there should be a balance between the accuracy and efficiency. The parameter of graph based segmentation we set is (0.4, 350, 1200). The threshold of region merging we set is 0.5. The weight of color contrast and motion contrast is set 0.6 and 0.4, respectively. Gaussian Smoothing parameters we set is 0.5.

Both qualitative and quantitative evaluations were utilized to assess performances of new developed method. Obtained results were compared to seven reference detection methods.

We first provide the visual comparison of different methods in Fig. 3. It can be seen that our method can deal with better in different cases where the background is cluttered. For example, other approaches may be distracted by the textures on the background while our method almost successfully highlights the whole salient object. Besides, our method has less false positives than other approaches, which can be very useful in real problems application.

4.1.2. Precision and Recall

To quantitatively evaluate the object segmentation results, the performances of our algorithm is measured by its precision and recall rate. Precision corresponds to the percentage of detected saliency pixel correctly assigned, while recall corresponds to the fraction of detected salient pixels in relation to the ground truth. High recall can be achieved at the expense of reducing precision. So it is necessary and important to measure them together. In the experiment, we use the most directly way to evaluate saliency map threshold at fixed number. We vary the threshold from 0 to 255, which is shown in Fig. 4. It can be seen that our method performs better detection precisions given higher recall.

Receiver operating characteristic (ROC) curves show the trade-off between misses and false positives. The axes for an ROC curve are fallout and recall. Recall is the same as above. Fallout, or false alarm rate, is the probability that a true negative was labeled a false positive. We vary the threshold as above, and it can be seen in Fig. 5 that our method has better performances when there is less false positive rate.

5. CONCLUSION AND FUTURE WORKS

We presented a novel salient object detection method for videos to combine motion information and color contrast features. In this work, four principles of salient regions were introduced to identify salient objects within the image sequences. The method is simple to implement and fast to compute. Further, it is robust when the background of image is noisy. The method was validated using a popular database, which is publicly available online. The qualitative and quantitative results have confirmed that the integration of motion and color features improve the detection quality and accuracy compared to the usage of color feature alone.

The future work will focus on development of a detection method for multiple salient objects in video. The presented method can also be extended to process surveillance videos in order to perform abnormal detection tasks. We believe that the RCST-based method can further enhance retrieval performance for traditional video search problems.

6. REFERENCES

[1] Christopher Kanan and Garrison W. Cottrell, "Robust classification of objects, faces, and flowers using natural image statistics," in *CVPR*, 2010, pp. 2472–2479.

[2] Dirk Walther and Christof Koch, "Modeling attention to salient proto-objects," *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[3] Laurent Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1304–1318, 2004.

[4] Luca Marchesotti, Claudio Cifarelli, and Gabriela Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *ICCV*, 2009, pp. 2232–2239.

[5] Tao Chen, Ming-Ming Cheng, Ping Tan, Ariel Shamir, and Shi-Min Hu, "Sketch2photo: internet image montage," *ACM Trans. Graph.*, vol. 28, no. 5, 2009.

[6] Stas Goferman, Ayellet Tal, and Lihi Zelnik-Manor, "Puzzle-like collage," *Comput. Graph. Forum*, vol. 29, no. 2, pp. 459–468, 2010.

[7] Jingdong Wang, Long Quan, Jian Sun, Xiaoou Tang, and Heung-Yeung Shum, "Picture collage," in *CVPR (1)*, 2006, pp. 347–354.

[8] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009, pp. 1597–1604.

[9] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu, "Global contrast based salient region detection," in *CVPR*, 2011, pp. 409–416.

[10] Xiaodi Hou and Liqing Zhang, "Saliency detection: A spectral residual approach," in *CVPR*, 2007.

[11] Tie Liu, Jian Sun, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum, "Learning to detect a salient object," in *CVPR*, 2007.

[12] Yun Zhai and Mubarak Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *ACM Multimedia*, 2006, pp. 815–824.

[13] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Tie Liu, and Nanning Zheng, "Automatic salient object segmentation based on context and shape prior," in *Proceedings of the British Machine Vision Conference*. 2011, pp. 110.1–110.12, BMVA Press.

[14] Chenlei Guo, Qi Ma, and Liming Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *CVPR*, 2008.

[15] A. M. Treisman and G. Gelade, "A feature-integration theory of attention.," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[16] Zhendong Mao, Yongdong Zhang, Ke Gao, and Dongming Zhang, "A method for detecting salient regions using integrated features," in *ACM Multimedia*, 2012, pp. 745–748.

[17] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[18] Yu-Fei Ma and HongJiang Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *ACM Multimedia*, 2003, pp. 374–381.

[19] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal, "Context-aware saliency detection," in *CVPR*, 2010, pp. 2376–2383.

[20] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012, pp. 733–740.

[21] Ce Liu, *Beyond pixels: exploring new representations and applications for motion analysis*, Ph.D. thesis, Massachusetts Institute of Technology, 2009.

[22] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, no. 2, pp. 167–181, 2004.

[23] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545–552.

[24] Ken Fukuchi, Kouji Miyazato, Akisato Kimura, Shigeru Takagi, and Junji Yamato, "Saliency-based video segmentation with graph cuts and sequentially updated priors," in *ICME*, 2009, pp. 638–641.