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

Contemporary video search and categorization are non-trivial tasks due to the massively increasing amount and content variety of videos. We put forward the study of visual saliency models in video. Such a model is employed to identify salient objects from the image background. Starting from the observation that motion information in video often attracts more human attention compared to static images, we devise a region contrast based saliency detection model using spatial-temporal cues (RCST). We introduce and study four saliency principles to realize the RCST. This generalizes the previous static image for saliency computational model to video. We conduct experiments on a publicly available video segmentation database where our method significantly outperforms seven state-of-the-art methods with respect to PR curve, ROC curve and visual comparison.

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

1. INTRODUCTION

A vast number of videos are made available by rapid development of computer infrastructure such as the speedup of processors, the increase of the storage capacity and the easier access to Internet. It is a great challenge to search or to categorize these videos. Salient areas in an image or a video are generally regarded as the focal areas of human eyes. Thus, saliency detection can locate and isolate the most attractive and important content from extensive images and videos, to assist video related applications.

Visual saliency is originally regarded as a task of predicting the eye-fixations on images, and has been recently extended to locate a region containing the salient object. It has 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 perception system suggests that the saliency is referred to uniqueness, rarity and attractive of a scene, which can be represented by visual features such as color, motion, texture, shape and so on [8]. Recently, a lot of research effort have been made to design various algorithms to compute the saliency for static images [9, 10, 11, 12, 13, 14]. However, extending the saliency computation models to videos related tasks is far from being well understood.

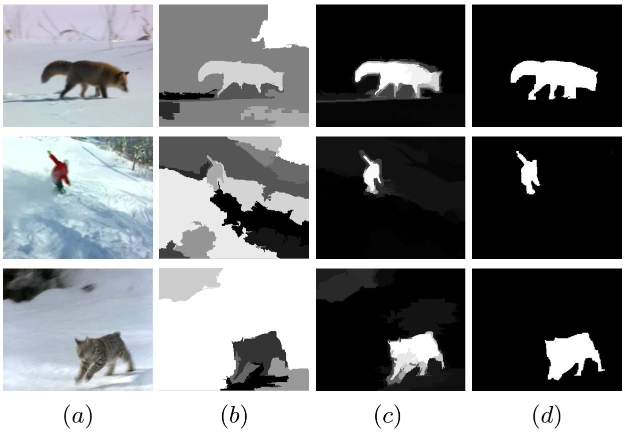

Fig. 1. Three examples illustrating the saliency detection results compared to ground truth. From left to right: (a) input frames, (b) graph based segmentation, (c) salient objects detected by our method, and (d) ground truth.

In this paper, we propose a novel saliency model combining both color and motion feature to create a saliency map. For example, Fig. 1 illustrates the procedure of graph segmentation based saliency detection for an input frame. We first define four general saliency principles specific for videos, and then introduce the region contrast method based on our principles. Comprehensive experiments on a public database[2] show that our RCST model outperforms seven state-of-the-art

---

[*] **Jia Xu is currently a visiting researcher at Natural Language Computing Group, Microsoft Research Asia, Beijing, China, under the "Star-Track" Visiting Young Faculty Program.**

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, 61250110577 and Danish National Research Foundation for the Sino-Danish Center for the Theory of Interactive Computation. Emails: ganchuang1990@gmail.com, zcqin@buaa.edu.cn, xu@tsinghua.edu.cn, tao.wan.wan@gmail.com.

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

methods with the respect to PR and ROC curve as well as visual comparison.

The core contribution of this paper is incorporating the spatial-temporal cues in the saliency detection in video. We introduce principles for salient regions in video and realize it in the following four steps:

- Initialize graph based image segmentation.
- Refine region based on local contrast (Section 3.2).
- Compute saliency map by combining color and motion information weighted by region spatial distance and area ratio (Section 3.3.1).
- Fuse spatial weight of each region to generate final saliency maps (Section 3.3.2 and 3.3.3).

The rest of paper is organized as follows. In Section 2, we review the related work of saliency models. In Section 3, we describe our region contrast based saliency detection model via spatiotemporal cues (RCST). In Section 4, we present extensive experiments compared to seven state-of-the-art methods. In Section 5, we conclude the results and discuss the future work.

2. PREVIOUS WORK

The crux of most recent bottom-up computation saliency models are inspired by the concept of the Feature Integration Theory (FIT) by Treisman and Gelade [15], which asserts that various visual feature are responsible for different kinds of attention system. As [16], saliency models can be roughly divided into two kinds: local contrast and global contrast.

Local contrast based methods estimate saliency of a particular patch based on their dissimilarity with its 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.* [12] find multi-scale contrast by linearly combining contrast in a Gaussian image pyramid. Goferman *et al.*[19] simultaneously model local low-level cues, global considerations, visual organization rules and high-level features to extract salient objects along with their contexts. Jiang *et al.*[14] compute color histogram of a region, and the differences with its neighboring regions are then used to evaluate the saliency score. Besides, saliency maps computed from multi-scale image segmentation to captures non-local contrast.

Global contrast based methods consider the contrast relationship over the whole image. Zhai and Shah [13] define pixel-level saliency based on a pixel's contrast to all other pixels. However, for efficiency they only use luminance information, thus do not take distinctiveness clues in other channels into consideration. Hou and Zhang [11] detect saliency in frequency domain by tuning the amplitude of image spectrum. Achanta *et al.* [9] propose a frequency tuned method that directly defines pixel saliency using each pixel's color difference from the average image color. Cheng *et al.* [10] apply color histogram to compute color difference between pixels and the whole image, in order to directly compute the color global uniqueness. Based on the regional contrast, element color uniqueness and spatial distribution are introduced to evaluate the saliency scores of regions [20].

Nonetheless, all these methods are limited to static images. 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 to other cars that go along in the same direction, so the dog should be considered as the salient object. It may be difficult to identify the dog as a saliency based on the usual saliency model in static images. Thus, it is necessary to incorporate some temporal cues to aid salient object detection in video. Some previous work has added motion information into the saliency model. For example, Zhai and Shah [13] propose a method based on spatial-temporal cue, and Guo *et al.*[21] proposed a method based on phase spectrum quaternion Fourier transform to efficiently compute the spatio-temporal salient map. However, their experimental results are not satisfactory. We will introduce our RCST-based model in the following section as well as its experimental results in Section 4.

3. RCST-BASED SALIENCY DETECTION

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 depicted in Fig. 3. Compared to previous works, it can been seen that our method is more robust and achieves promising performance for video related applications.

3.1. Principles of salient region in image sequences

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

1. The salient region always stands out from surrounding context in a certain aspect, such as color and motion.
2. The color and motion is coherent in one salient region.
3. The spatial distribution of salient regions is always more centralized than the background.
4. The salient region is usually close to the image center.

The first principle, based on bottom-up salient stimuli, has been noted in previous work [14]. Different than their approach, we incorporate motion cue into consideration. This

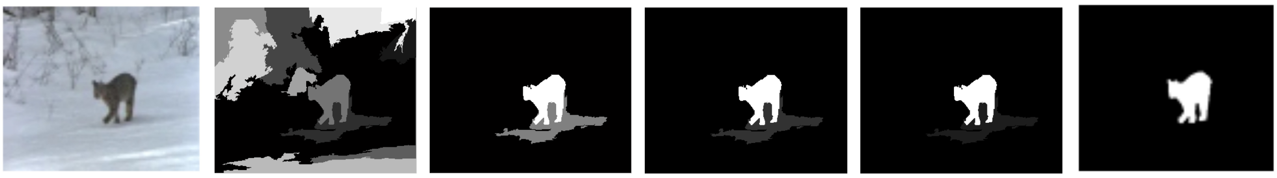
(a) (b) (c) (d) (e) (f)

Fig. 2. The series of images show 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.

makes 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 the "golden ratio": the camera always lets the salient objects locate close to the center of scene so that the core of images can be easily caught.

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 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 mentioned above, 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. 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 [22], 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 a six-dimensional 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 as follows.

We apply the graph based image segmentation approach [23] 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 $D(i,j)$ is 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 regions. The following describes 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 i, which is mainly based on the color and motion contrast with remaining regions. We observe that salients often appear in the *nearer* or *larger* region, hence assign a higher weight to these regions in the contrast computation. This model is formulated as

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

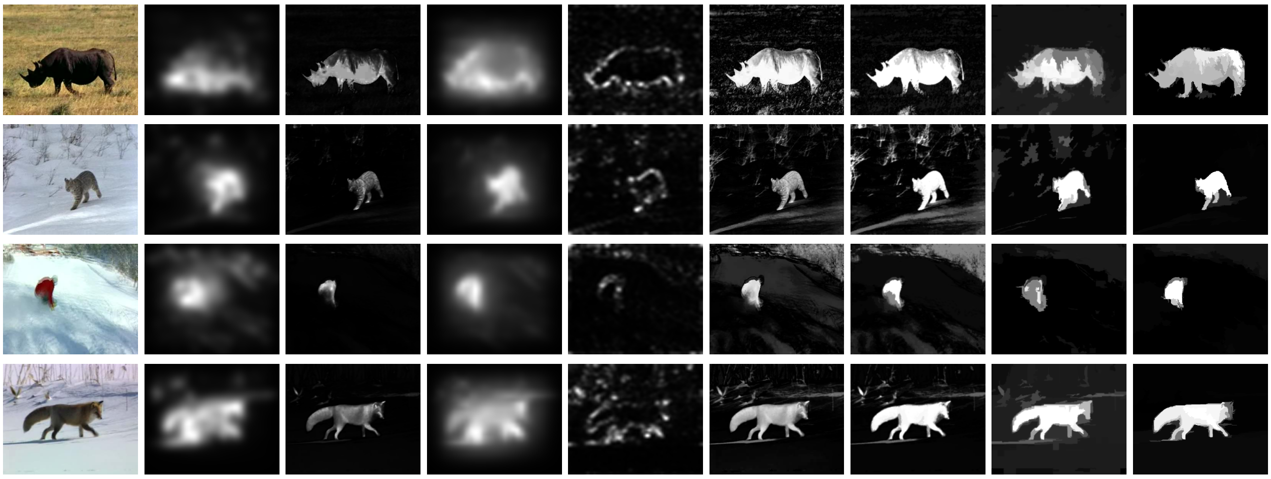
(a) input (b) IT [17] (c) FT [9] (d) GB [24] (e) SR [11] (f) LC[13] (g) HC [10] (h) RC[10] (i) RCST

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 state-of-the-art detection methods.

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

where 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 principle 3 and 4, we assign the region close to center or more centralized with a higher spatial weight:

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

where 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*

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), and subsequently 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 the same weight to them.

4. EXPERIMENTAL RESULTS

In this section, we present the results of our approach on the publicly available database provided by Kimura [25]. We compare the proposed method with seven state-of-the-art saliency detection methods, according to: the number of citation [17, 11], recent [10] and variety[9, 13, 24, 24] . We apply our method and others to compute saliency maps for the image sequences in the database. In order to comprehensively evaluate the accuracy of our method for salient object segmentation, we perform evaluation using the PR and ROC curve as well as the visual comparision. The visual comparison of saliency map can be seen in Fig. 3.

4.1. Experimental setting

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.

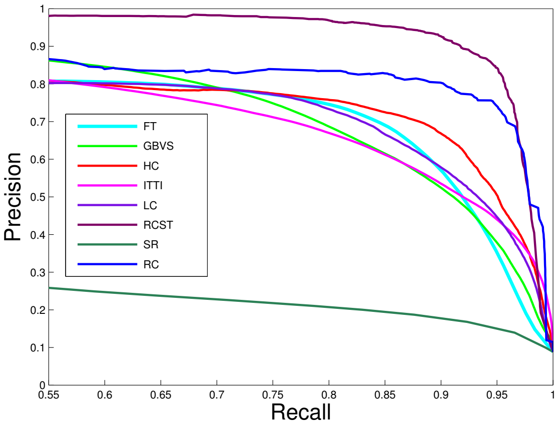

Fig. 4. Precision and recall (PR) curve. The RCST-based method yielded higher detection precision and recall values.

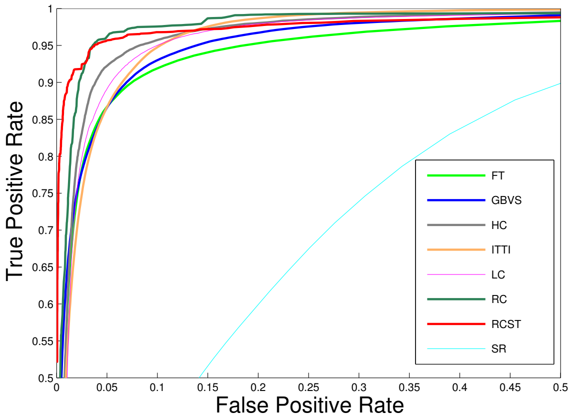

Fig. 5. ROC curve. The RCST-based method yielded the best performance when the false positive rate is low.

4.1.1. Implementation details

In the experiments, the segmentation and threshold parameters are somewhat sensitive. More generated regions will incur a 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 are both 0.5.

Both qualitative and quantitative evaluations were utilized to assess the performance of this 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. PR and ROC curve

To quantitatively evaluate the object segmentation results, the performance of our algorithm is measured by its precision and recall rate(PR) curve. 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. CONCLUSIONS AND FUTURE WORK

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] Ali Borji, Dicky N. Sihite, and Laurent Itti, “Salient object detection: A benchmark,” in *ECCV (2)*, 2012, pp. 414–429.

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

[10] 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.

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

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

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

[14] 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.

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

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

[23] 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.

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

[25] 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.