Video Structuring: From Pixels to Visual Entities

Ruxandra Tapu, Zaharia Titus

VIDEO STRUCTURING: FROM PIXELS TO VISUAL ENTITIES

Ruxandra Tapu, Titus Zaharia

Institut Mines-Telecom / Telecom SudParis, ARTEMIS Department, UMR CNRS 8145 MAP5
5 Rue Charles Fourier, 91000, Evry, France
phone: +33 (0)1 60 76 46 74, email: {ruxandra.tapu, titus.zaharia}@it-sudparis.eu

ABSTRACT

In this paper we propose a method for automatic structuring of video documents. The video is firstly segmented into shots based on a scale space filtering graph partition method. For each detected shot the associated static summary is developed using a leap key-frame extraction method. Based on the representative images obtained, we introduce next a combined spatial and temporal video attention model that is able to recognize moving salient objects. The proposed approach extends the state-of-the-art image region based contrast saliency with a temporal attention model. Different types of motion presented in the current shot are determined using a set of homographic transforms, estimated by recursively applying the RANSAC algorithm on the interest point correspondence. Finally, a decision is taken based on the combined spatial and temporal attention models. The experimental results validate the proposed framework and demonstrate that our approach is effective for various types of videos, including noisy and low resolution data.

Index Terms— Saliency maps, temporal attention model, homography transform, RANSAC algorithm.

1. INTRODUCTION

Recent advances in the field of image/video acquisition and storage devices have determined a spectacular increase of the amount of audio-visual content transmitted, exchanged and shared over the Internet. In the past years, the only method of searching information in multimedia databases was based on textual annotation, which consists of associating a set of keywords to each individual item. Such a procedure requires a huge amount of human interaction and is intractable in the case of large multimedia databases. On the contrary, content-based indexing approaches aim at automatically describing the content. In the case of complex video documents, the richness of the associated information requires a preliminary phase, consisting of structuring the video into pertinent elements that can be described accurately. Such elements are most of the time scenes, shots, key-frames and objects of interest.

The human brain and visual system actively seek for regions of interest by paying more attention to some specific parts of the image/ video. Visual saliency [1] can be defined as the perceptual feature that allows an object to stand out from his neighbors by capturing our attention. Humans can easily understand a scene based on the selective visual attention which makes it possible to detect the region of interest in images or interesting actions in videos sequences [2].

Fig. 1 presents the proposed analysis framework. First, the video is temporally segmented into shots. For each determined shot, a set of representative key-frames is detected. Then, for each key-frame a salient region is obtained by combining spatial and motion information. The main contribution introduced in this paper concerns a novel bottom-up approach for modeling the spatio-temporal attention in videos. The temporal attention model is determined using a set of homographic transforms, while the spatial attention model exploits contrast-based saliency maps.

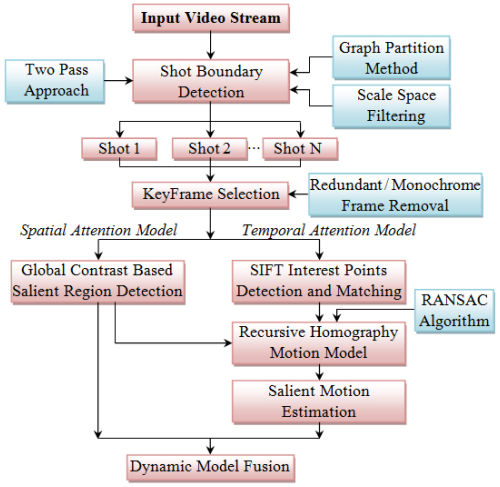


Fig.1. Spatio-temporal salient object detection framework

The rest of this paper is organized as follows. Section 2 presents and analyzes the related work. The shot detection algorithm together with the keyframe selection procedure is recalled in Section 3. Section 4 introduces the novel spatio-temporal attention system proposed and details the main steps involved: region based contrast computation, SIFT interest point matching and homographic motion modeling. The experimental results obtained are presented and discussed in details in Section 5. Finally, Section 6 concludes the paper and opens perspectives of future work.

2. RELATED WORK

The image attention models can be divided into two categories: bottom-up and top-down approaches. The last category is task driven, using prior knowledge of the video flow or its content. Their major drawback is the lack of generality, since the same context is not available in every video

document. To solve this problem, various bottom-up approaches have been introduced [3], [4], [5], usually referred to as saliency detection or stimuli-driven techniques. Most of them model the human reaction to external stimuli, and exploit low level features, in order to build a saliency map which emphasizes relevant regions. Determining salient regions which can be consistent with the human visual attention represents a difficult challenge in computer vision.

One of the first techniques proposed in the literature [6] uses several feature attributes such as color, intensity and orientation. In this case, the image saliency is defined based on the surrounded differences obtained across multi-scale image features. A modified version of the technique introduced in [6] is presented in [7]. In [8], authors determine center-surround contrast by using a Difference of Gaussian (DoG).Various methods [9], [10] are based solely on computational rules and not on the biological vision principles. Here, the authors propose to determine the saliency based on center-surround feature distances.

In [11], the saliency is established by applying heuristic measures (histogram threshold) on the estimated interest zone. More recent techniques [12] model simultaneously low-level and high-level features and visual organization rules in order to determine salient objects along with their corresponding context. The major drawback of the above-cited methods is caused by the use of local contrast measures that tend to produce high saliency values for pixels situated near an edge instead of uniformly highlighting objects of interest [13].

In order to overcome such limitations, other algorithms exploit global contrast measures. Thus, in [14], authors compute the pixel saliency based on its relative difference established by iterative comparison with all the image pixels. Only the luminance information is here taken into account. In [2], a frequency-based saliency detection system is introduced, that defines the pixel saliency value as the difference between the current color considered and the average color of the whole image. Both bottom-up and top-down approaches are useful to detect salient object in still images. However, when applied to video sequences, the results are often unsatisfactory, since no motion information is taken into account.

Existing video-dedicated approaches [15], [16], [17] combine relevant motion models with spatial attention in order to build efficient saliency models. In [16], the representative region is established by using the spatio-temporal energy accumulation of coherent moving objects, while in [17] the authors exploit the motion vectors magnitude and phase histograms.

However, in practice the video motion can be caused by the salient objects, but also by background objects or camera movement. In this case different types of motion need to be analyzed and appropriately taken into account. Naturally, such methods can be applied on arbitrary frames in the video sequences. However, in order to reduce the computational burden, we solely consider a set of representative key-frames, determined as described in the next section.

3. SHOT BOUNDARY DETECTION AND KEYFRAME EXTRACTION SYSTEM

Prior to detecting key-frames, we first apply a shot detection procedure, with the help of the method introduced in [18]. Based on a graph partition model combined with a non-linear scale-space filtering mechanism, the method achieves high precision (superior to 90%)) and recall rates (superior to 95%) whatever the movie quality and genre and for both abrupt and gradual transitions.

We develop next an automatic static storyboarding system, which extracts a variable number of keyframes from each detected shot, depending on the visual content variation. The first keyframe extracted is selected by definition, *N* frames away after a detected transition. Parameter *N* should be compatible with the size of the analysis window used for the shot detection in order to guarantee that the selected image does not belong to a gradual effect. Next, we introduced a leap-extraction method that considers for analysis only the images located at integer multipliers of *N*.

The current frames are compared with the existing shot keyframes set already extracted. If the visual dissimilarity (chi-square distance of HSV color histograms) between the analyzed frames is significant (above a pre-establish threshold), the current image is added to the keyframe set.

The resulting storyboard still may contain useless blank or test card frames. In order to eliminate them, an additional post-processing step eliminates all images from the selected set of keyframes assuring that the static story board captures all informational content of the original movie while discarding irrelevant images. To check if a selected keyframe is useless we computed its interest points based on SIFT descriptor. A test card or blank frame is detected and removed if the resulted number of keypoints is close to zero.

4. SPATIO-TEMPORAL VISUAL SALIENCY

The spatio-temporal saliency model introduced in this paper is based on the stationary saliency technique so-called region-based contrast (RC) [13], which is enhanced with a novel temporal attention model.

4.1. Stationary attention model

A powerful method of computing bottom-up visual cues is proposed in [13]. First, the input image is segmented into regions based on a graph partition strategy [19].

The saliency value of a region (r_k) is defined based on the color contrast to all other regions in the image:

$$S(r_k) = \sum_{r \neq r_k} w(r_i) \cdot d_r(r_k, r_i) \,, \qquad (1)$$

where $w(r_i)$ is the weight of region r_i, computed as the total number of pixels included in the region while $d_r(,)$ is the color distance metric between regions defined as:

$$d_r(r_1, r_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} p(c_{1,i}) \cdot p(c_{2,j}) \cdot \delta\left(c_{1,i}, c_{2,j}\right), \qquad (2)$$

where $p(c_{k,i})$ is the probability of the *i*-th color $c_{k,i}$ among all n_k colors in the *k* region ($k = 1,2$).

In order to increase the effects of closer regions and decrease the impact of farther regions a spatial weighting term is introduced in equation 1:

$$S(r_k) = \sum_{r \neq r_k} \exp\left(\frac{d_s(r_k, r_i)}{\sigma^2}\right) w(r_i) \cdot d_r(r_k, r_i). \quad (3)$$

Here, $d_s(r_k, r_i)$ is the spatial distance between regions and σ^2 is a parameter controlling the strength of the spatial weighting mechanism.

4.2. Temporal attention model

Intuitively, the salient motion is the movement that attracts the attention of a human subject. Most of the previously developed methods [15], [16] are based only on the temporal difference of adjacent frames and cannot effectively identify the salient motion.

In this paper, we propose a novel temporal attention technique that combines the above-described spatial (stationary) saliency model, on the interest point (I_p) correspondences between successive video keyframes and the homography transforms that model the motion of moving regions. The algorithm consists of the following steps:

Step 1: *Interest point detection and matching* – The Scale Invariant Feature Transform (SIFT) [20] is applied on two successive frames (by taking as starting frame each detected key-frame). The correspondence between the interest points is established using KD-tree matching technique [21].

Let $p_{1i}(x_{1i}, y_{1i})$ be the *i*-th key point in the first image and $p_{2i}(x_{2i}, y_{2i})$ be the correspondence in the second image. The associated motion vectors (v_{ix}, v_{iy}) and magnitude $(D_{i(1,2)})$ are also computed in this step:

$$v_{ix} = x_{2i} - x_{1i}\ ; v_{iy} = y_{2i} - y_{1i}, \quad (4)$$

$$D_{i(1,2)} = \sqrt{v_{ix}^2 + v_{iy}^2}\ ,\ i = \overline{1,n}, \quad (5)$$

where *n* is the total number of correspondences.

Step 2: *Interest points saliency initialization* – For the current key-frames the interest point's saliency values are determined based on the technique described in Section 4.1.

Step 3: Background / Camera motion detection –We start our analysis by identifying a subset of *m* keypoints located in the background (Fig 2) and identified based on their saliency value. More precisely, an interest point $p_{1,i}$ is defined as a background point if.

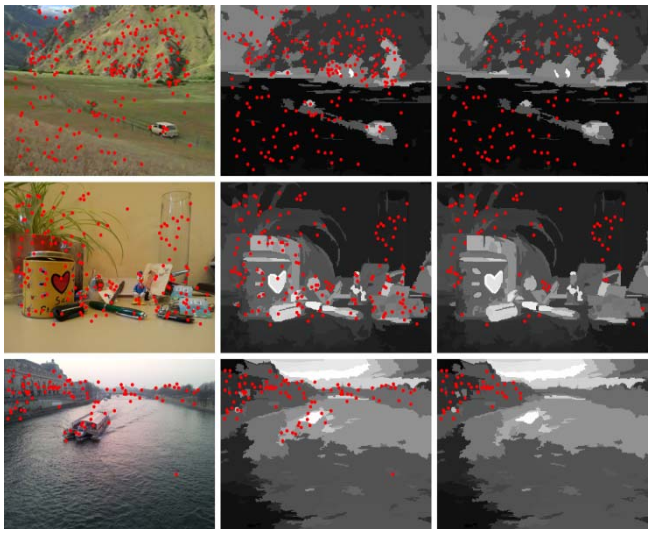
$$Sal(p_{1,i}) \leq T_h\ , \quad (6)$$

where $Sal(p_{1,i})$ is the saliency value of point $p_{1,i}$ while T_h is the average saliency value over the considered keyframe.

The subset of *m* background interest points is used to determine the geometric transformation between the selected images, by considering a homographic motion model, determined with the help of a RANSAC (*Random Sample Consensus*) [22] algorithm.

The RANSAC technique can be summarized as follows. Starting from a random sample of 4 interest point correspondences, a homographic matrix $\boldsymbol{H}$ is computed. Then, each other pair of points is classified as an inlier or outlier depending of its concurrence with $\boldsymbol{H}$.

After all of the interest points are considered for the estimation of matrix $\boldsymbol{H}$, the iteration that yields the largest number of inliers is selected.



(a) (b) (c)

Fig.2 Interest points selection process. (a) SIFT interest points extraction; (b) spatial saliency map; (c) subset of keypoints used for camera/background motion estimation.

Based on the matrix $\boldsymbol{H}$, for a current point $p_{1i} = [x_{1i}, y_{1i}, 1]^T$ expressed in homogeneous coordinates, its estimated correspondence position $p_{2i}^{est} = [x_{2i}^{est}, y_{2i}^{est}, 1]^T$ is determined as:

$$\begin{bmatrix} x_{2i}^{est} \\ y_{2i}^{est} \\ w \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \cdot \begin{bmatrix} x_{1i} \\ y_{1i} \\ 1 \end{bmatrix}, \quad (7)$$

where:

$$w = 1/(h_{20} \cdot x_{2i}^{est} + h_{21} \cdot y_{2i}^{est} + h_{22}). \quad (8)$$

The estimation error is defined as the difference between estimated and actual position of the considered interest point, as described in equation (9):

$$\epsilon(p_{1i}, \boldsymbol{H}) = \|p_{2i}^{est} - p_{2i}\|\,. \quad (9)$$

Ideally $p_{2i}^{est} = [x_{2i}^{est}, y_{2i}^{est}, 1]^T$ should be as close as possible to $p_{2i} = [x_{2i}, y_{2i}, 1]^T$.

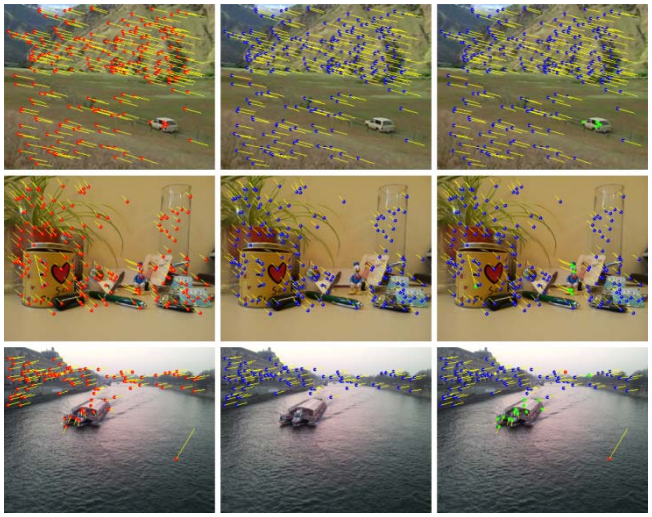
In the case where the estimation error $\epsilon(p_{1i}, \boldsymbol{H})$ is inferior to a predefined threshold *E*, the corresponding pixels are marked as belonging to background. The outliers, *i.e.*, pixels with estimation error $\epsilon(p_{1i}, \boldsymbol{H})$ exceeding the considered threshold, are considered to belong to foreground objects.

In our experiments, the background/foreground separation threshold *E* has been set to 4 pixels.

Step 4: Different types of motion recognition - In practice, multiple moving objects are present in the scene. In this case we determine a new subset of points formed by all the outliers and all the points not considered in previous step (obtained after subtracting from all the interest points the subset *m*).

For the current subset we computed a novel homography and determine again the inliers and outliers based on the projection error. The inliers thus determined form a new motion class. For the remaining outliers, the process is applied recursively until all points belong to a motion class (Fig. 3).

After all interest points are clustered into motion classes, we determine the salient movement.



(a) (b) (c)

Fig.3. Interest point classification. (a) Motion vectors; (b) Camera / background motion detection; (c) Motion classes (green–foreground and blue-background points).

Step 5: *Salient motion detection* – For all the motion classes determined at *Step 3* and *4* we compute their spatial saliency value as:

$$SalClass(M_i) = \frac{\sum_{j=1}^{m_i} Sal(p_{1,i})}{m_i}, i = \overline{1,N}, \quad (10)$$

where m_i is the total number of points included in motion class M_i and N is the total number of classes. In this case, the salient motion is determined as:

$$SalientMotion = \max_{i=1,N}\{SalClass(M_i)\}. \quad (11)$$

Step 6: Interest point refinement – For all the interest points included in the salient motion class we computed the distance to the closest neighbor. If the computed distance is two times bigger than the average class distance that point is eliminated. In this case the point classification to the current cluster is probably erroneous, caused by SIFT mismatching or false homography estimation.

Step 7: Salient regions detection – Based on the interest points included in the salient motion class we determine the relevant regions. A region is considered as salient if it contains an interest point. For interest points situated on a border, the region with highest initial saliency value is preferred.

The spatial and temporal attention models are now combined in order to produce the final video saliency map. In scenes with no independently moving objects, the system detects a single motion class, corresponding to the camera motion. In this case, the segmentation is performed based solely on the spatial attention model. On the other hand, if strong motion contrast is presented in the scene the saliency map is constructed based on the temporal attention model.

Step 8: Object detection – The object is extracted based on the GrabCut algorithm [23], which is automatically initialized with the ternary saliency map detected at step 7. In our case the pixels marked as certainly foreground are the pixels from the salient regions, the pixels labeled as probably foreground are situated between the salient regions (Fig. 4d), while pixels marked as certain background are encountered outside the salient regions.

5. EXPERIMENTAL RESULTS

We tested the proposed methodology on a set of 20 general purpose videos [24][25]. The videos are mostly documentaries, noisy, and vary in style and date of production. The resolution is of 341 x 256 pixels. Each video contains a single salient object. In terms of content, the objects correspond to humans performing various activities, and animals in the wild, ground vehicles, aircrafts...

The test database includes dark, cluttered and highly dynamic scenes make them very challenging for an automatic object extraction system. In addition, various types of both camera and object motions, are present. Some object detection results are presented in Fig. 4. Let us first note that for videos with rich texture or including multiple objects, the result of a spatial attention model is, in most of the cases, unrepresentative. For example; if we consider the case of the second video in Fig. 4, the salient object is a moving car of small dimensions with similar colour features as the background. As it can be observed, the spatial attention model detects as salient the sky, but after incorporated the motion information the method is able correct identify the car. The impact of motion information is even more important for scenes with rich texture (videos 2 and 3). In this case, the output of a spatial saliency system is useless because the technique is not able to distinguish between different types of regions. However, with the help of the dynamic model the temporal attentions become dominant and we are able to identify the objects of interest.

The proposed spatio-temporal visual saliency (STVS) method is compared with the state of the art graph-based visual saliency (GBVS) technique [7]. As it can be noticed from Fig. 4, the GBVS method is not able to correctly identify salient objects of large dimensions (video 1) neither to strictly localize only the salient object in textured movies (videos 2 and 3). On the contrary, the proposed STVS method successfully recovers the objects of interest.

6. CONCLUSIONS AND PERSPECTIVES

In this paper we have proposed an automatic salient object extraction system based on a spatiotemporal attention detection framework. The spatial model is developed starting from the region-based contrast applied now on video streams, while the temporal model rely on the interest points correspondence and geometric transforms between key-frames.

The technique is robust to complex background distracting motions and does not require any initial knowledge about the object size or shape. The various experimental results, obtained on various video sequences demonstrate the effectiveness of the proposed technique.

In our future work, we plan to extend the proposed method by taking into account not merely successive frames, but the whole content of a video shot in order to (1) increase the robustness of the algorithm and (2) obtain a video object tracking method.

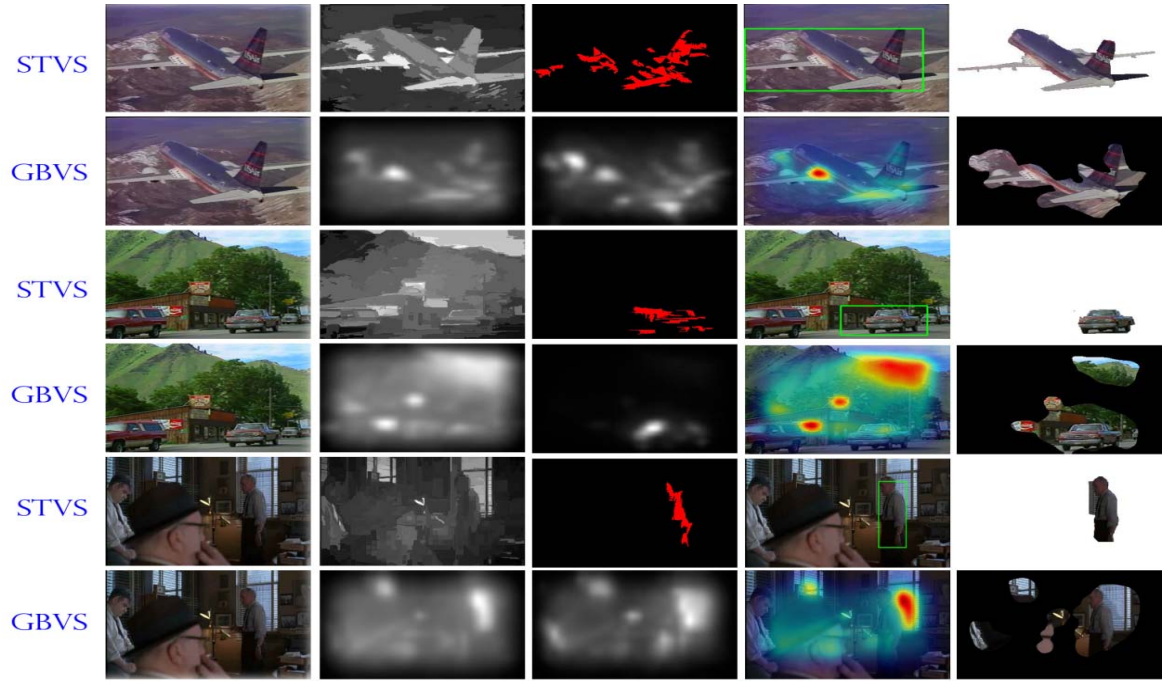


Fig.4. Salient map extraction process. (a) Key-frame selected from a video shot; (b) Spatial saliency map; (c) Temporal saliency map; (d) Candidate regions selection; (e) Detected object.

ACKNOLEDGEMENTS

Part of this work has been supported by the French FUI7 project 3D-LIVE.

REFERENCES

[1]. W. Kim, C. Jung, and C. Kim, "Spatiotemporal saliency detection and its applications in static and dynamic scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 4, pp. 446–456, 2011.

[2]. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection", *In CVPR*, pp. 1597–1604, 2009.

[3]. H. Li and K. N. Ngan, "Saliency model-based face segmentation and tracking in head-and-shoulder video sequences," *J. Vis. Commun. Image Representation*, vol. 19, pp. 320–333, 2008.

[4]. Y.-T. Chen and C.-S. Chen, "Fast human detection using a novel boosted cascading structure with meta stages," *IEEE Trans. Image Process.*, vol. 17, no. 8, pp. 1452–1464, Aug. 2008.

[5]. F. Liu and M. Gleicher, "Region enhanced scale-invariant saliency detection," in *Proc. IEEE ICME*, pp. 1477–1480, 2006.

[6]. L. Itti, C. Koch, and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis", *IEEE TPAMI*, 20(11), pp. 1254–1259, 1998.

[7]. J. Harel, C. Koch, P. Perona. "Graph-based visual saliency", *Advances in Neural Information Processing Systems*, pp. 545- 554, 2007.

[8]. T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object", in *Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[9]. R. Achanta, F. Estrada, P. Wils, and S. Susstrunk, "Salient region detection and segmentation", *International Conference on Computer Vision Systems*, 2008.

[10]. Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing", In *ACM International Conference on Multimedia*, pp.374-381, 2003.

[11]. Y. Hu, X. Xie, W.-Y. Ma, L.-T. Chia, and D. Rajan, "Salient region detection using weighted feature maps based on the human visual attention model", *Pacific Rim Conference on Multimedia*, pp. 993-1000, 2004.

[12]. S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection", In Proc. *Computer Vision and Pattern Recognition*, pp. 2376–2383, 2010.

[13]. M. Cheng, N. Zhang, G. Mitra, X. Huang, S. Hu, "Global contrast based salient region detection," In *IEEE CVPR*, pp. 409–416, 2011.

[14]. Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues", *In ACM Multimedia*, pages 815–824, 2006.

[15]. Y. L. Tian and A. Hampapur, "Robust Salient Motion Detection with Complex Background for Real Time Video Surveillance. In Proceedings of the IEEE Workshop on Motion and Video Computing - Vol. 2, 2005.

[16]. A. Belardinelli, F. Pirri, and A. Carbone, "Motion Saliency Maps from Spatiotemporal Filtering. In Attention in Cognitive Systems", Lecture Notes in Artificial Intelligence, pp. 112-123, 2009.

[17]. F.F.E. Guraya, F.A. Cheikh, A. Tremeau, Y. Tong, H. Konik, "Predictive Saliency Maps for Surveillance Videos", In Conference of Distributed Computing and Applications to Business Engineering and Science, pp.508-513, 2010.

[18]. R. Tapu, T. Zaharia, F. Preteux, "A scale-space filtering-based shot detection algorithm", IEEE 26-th Convention of Electrical and Electronics Engineers in Israel, 2010, pp.919-923, 2010.

[19]. P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, 2004

[20]. Lowe, D., "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, pp. 1-28, 2004.

[21]. R. Panigrahy, "An improved algorithm finding nearest neighbor using kd-trees", In Proceedings of the 8th Latin American conference on Theoretical informatics, LATIN'08, pp. 387–398, 2008.

[22]. J. J. Lee and G. Y. Kim. "Robust estimation of camera homography using fuzzy RANSAC", International Conference on Computational Science and Its Applications, 2007.

[23]. C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts", Proc. ACM SIGGRAPH, pp. 309-314, 2004.

[24]. www.di.ens.fr/willow/research/videoseg

[25]. www.archive.org