Motion Compensated Temporal Image Signature Approach

Haroon Qureshi and Markus Ludwig
Institut für Rundfunktechnik GmbH (IRT), D-80939 Munich, Germany

Keywords: Saliency, Visual Attention Modeling, DCT Based Object Detection.

Abstract: Detecting salient regions in a temporal domain is indeed a challenging problem. The problem gets trickier when there is a moving object in a scene and becomes even more complex in the presence of camera motion. The camera motion can influence saliency detection as on one side, it can provide important information about the location of moving object. On the other side, camera motion can also lead to wrong estimation of salient regions. Therefore it is very important to handle this issue more sensible. This paper provides a solution to this issue by combining a saliency detection approach with motion estimation approach. This further extends the Temporal Image Signature (TIS) (Qureshi, 2013) approach to the more complex level where not only object motion is considered but also camera motion influence is compensated.

1 INTRODUCTION

Motion of pixels in 2D images can be perceived as a result of projection of object motion and camera motion in a 3D world coordinate system. In the case there is no camera motion, anything that moves is salient. But in case of camera motion, it is important to identify salient regions in images that correspond to moving objects only. Therefore it is of interest to separate the two motions from each other and in other words to detect and compensate the camera motion.

In Temporal Image Signature (TIS) approach (Qureshi, 2013) a combination of two saliency approaches (Image Signature based approach (IS) (Hou et al., 2012) and Temporal Spectral Residual (TSR) approach (Cui et al., 2009)) was shown and camera motion was ignored. It was also suggested that compensating the camera motion effect while detecting salient regions might certainly improve the results. The camera motion artifacts were evident in the results. In this paper, a camera motion along with the moving object is considered.

With the advancement in modern digital technologies, the demand for better quality video assessment in term of efficient transmission of media contents has also increased. Therefore, there is a need for intelligent video compression. One possible way is to compress the video equally without considering the contents which requires a specific amount of bandwidth. Another but intelligent way is to distribute the amount of compression of parts based on contents. This can be done by dividing regions that require higher compression quality, thus higher bitrate, from other regions that do not require such high quality. The first regions should be compressed less and second regions could be compressed higher. The information to divide the regions can be perceived from saliency detection methods. The proposed method is an initial step towards developing the intelligent video coding approach by detecting the salient regions in a scene.

The proposed approach compensates the effect of camera motion by using motion estimation and combines that information with the salient regions. Several options exist in order to compensate camera motion. One way is to measure camera movement physically. Another automatic but general way is to estimate camera motion using local and global motion information of objects in a scene. This can be done by computing motion vectors fields locally or globally. In this paper, global motion (i.e. background motion) is considered as the camera motion.

The paper is organized as follows: In Section 2, some existing state-of-the-art approaches are explained. Section 3 describes the proposed work in detail along with the computation and fusion of saliency map with the motion information. Finally, the evaluation of the proposed model with other state-of-the-art approaches is presented in section 4.

2 RELATED WORK

The idea of visual attention modelling is strongly in-

Qureshi H. and Ludwig M..
Motion Compensated Temporal Image Signature Approach.
DOI: 10.5220/0005303305120516
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 512-516
ISBN: 978-989-758-089-5

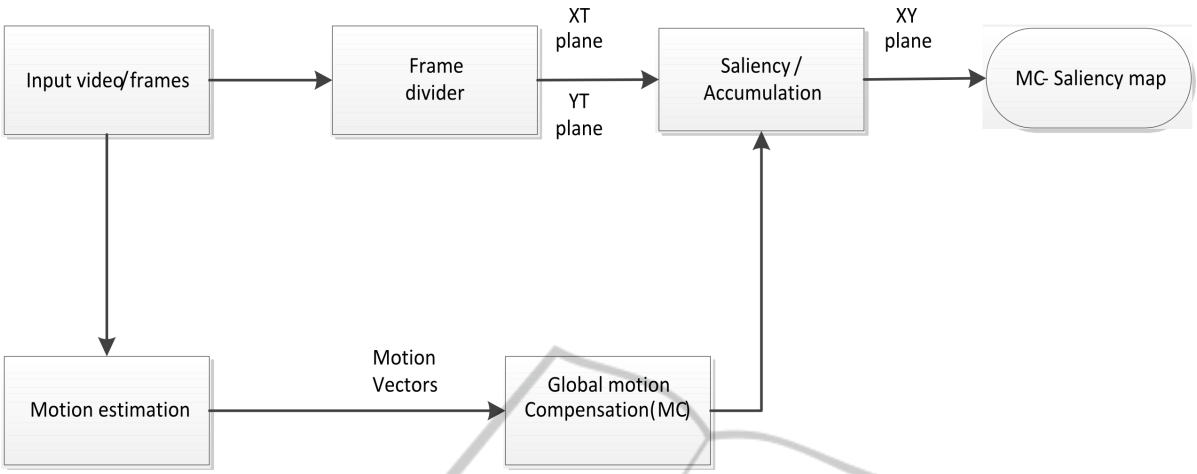

Figure 1: The proposed system.

spired by two models that are based on the human visual perception. The first model "Feature Integration Theory" (FIT) (Treisman and Gelade, 1980) explains the behavior of a human when it looks at the visual scene. The second model Guided Search (GS) (Treisman, 1986) can be seen as an extension to FIT. Generally saliency detection approaches can be categorized into two parts, image based saliency and video based saliency approaches.

The main goal of image based saliency methods is to figure out salient regions from the background. Numerous processes have been suggested in this area. Itti et al. simulated the process of human visual search in order to detect salient regions in still images (Itti and Koch, 2001). Huo et al. proposed Spectral Residual approach (SR) (Hou and Zhang, 2007) by considering irregularity clue from the smooth spectrum for saliency detection. Whereas Achanta et al. and Cheng et al. estimate saliency using frequency-tuned saliency and contrast based concept (Achanta et al., 2009)(Cheng et al., 2011). More recently DCT and its variation (Qureshi, 2013) (Schauerte and Stiefelhagen, 2012) is also utilized in detecting salient regions.

The main aim of video based saliency is to separate salient motions from the background. In the context of video many approaches have been proposed. In fact, many similar approaches from still image domain were developed also in the video or temporal domain as well. For example, Temporal Spectral Residual (TSR) (Cui et al., 2009) approach is an extension of Spectral Residual (SR) (Hou and Zhang, 2007) and Temporal Image Signature (TIS) (Qureshi, 2013) approach is an extension of the Image Signature (IS) (Hou et al., 2012) approach. For video content, in order to detect salient region as accurate as possible, different clues (e.g., camera motion, face, text and speech) are used by the researchers. For example many researchers proposed to use human face as an important clue while detecting salient regions because of its importance of attracting human's attention (Qureshi and Ludwig, 2013) (M. Cerf and Koch, 2009). Motion (e.g., camera, object) can also play an important clue to detect automatic regions of interest.

Studies of different visual models have shown that motion may indeed play an important role and provide useful information about salient areas. A tool that often used in motion analysis is the Motion Vector Field that can be retrieved by different algorithms. Liang et al. suggests a method called phase-correlation Motion Estimation (Guo et al., 2008). Based on the motion vector field researchers describe a method that estimates global motion in a scene and use it to compensate camera movement (Deigmoeller, 2010) (Hadi Hadizadeh, 2014). Some researchers used camera motion information to find salient regions of interest (Abdollahian and Edward J, 2007). A very comprehensive survey, analysis of scores, datasets, and model of state of the art technologies in visual attention modeling is presented in the papers (Ali Borji, 2013) (Borji. A and Itti, 2013). In this paper, a method proposed in (Chen and Bajić, 2010) is used for the proof-of-concept for motion estimation.

There are many applications of saliency models which have been developed over the years and thus further increased interest in attention modeling. For example, object based segmentation (Han et al., 2006), image retargeting (Achanta and Süsstrunk, 2009), face detection (Qureshi and Ludwig, 2013) (M. Cerf and Koch, 2009) , compression (Hadi Hadizadeh, 2014) and video summarization (Y.-F. Ma and Zhang, 2005).

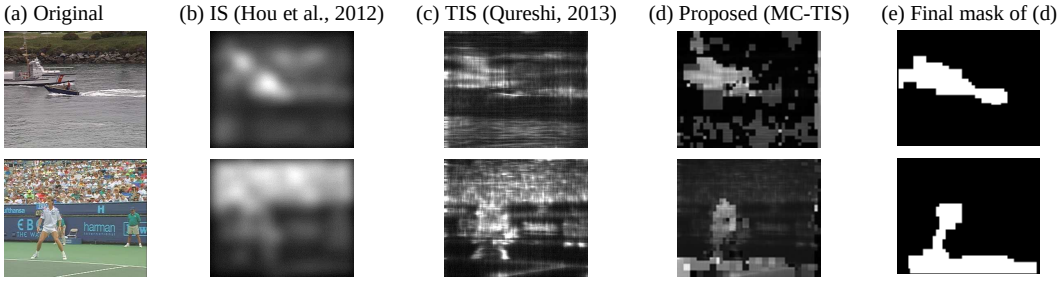

Figure 2: Visual comparison of Proposed approach with motion compensation Vs IS (Hou et al., 2012) and TIS (Qureshi, 2013) with no motion compensation.

3 PROPOSED APPROACH

This paper proposes a combination of motion detection algorithms (Chen and Bajić, 2010)[1] with Temporal Image Signature (TIS), a saliency detection approach (Qureshi, 2013). The system is shown in Figure 1. Temporal Image Signature (TIS) (Qureshi, 2013) approach completely ignores the effect of camera motion. The proposed approach extends Temporal Image Signature (TIS) while compensating the effect of camera motion.

It is now well established that local object motion is an important clue to grab the human attraction (Qureshi and Ludwig, 2013) (Itti et al., 1998). In Temporal Image Signature (TIS) approach (Qureshi, 2013), it was observed that the accuracy of TIS approach degrades whenever there is a camera motion present in the scenes. Although, the model was able to detect salient regions successfully. But in some cases due to severe camera motion where background motion competes with the foreground object motion background motion or camera motion could confuse the salient object motion.

The processing of the proposed approach is as follows: At first, a number of frames are sliced into the horizontal (XT) and vertical planes (YT). XT and YT are the planes of image lines in a temporal domain. Then the IS (Hou et al., 2012) approach is applied separately on all planes. IS approach defines the saliency using the inverse Discrete Cosine Transform (DCT) of the signs in the cosine spectrum. Secondly, the global motion estimation algorithm (Chen and Bajić, 2010), followed by global motion compensation (Chen and Bajić, 2010) is applied on each frame separately in XY domain. In the final step, first salient information in the XT and YT plane is accumulated by transformation back into the XY domain and it is then it is fused with the motion compensated frames using coherent-normalization-based fusion method (C. Chamaret and Meur, 2010). This results in a final map which is a combination of saliency map and motion estimation. Figure 2 provides a visual comparison of the proposed approach (MC-TIS) after compensating camera motion with the results of TIS approach (Qureshi, 2013) and Image Signature (IS) approach (Hou et al., 2012) with no compensation. The binary mask of the proposed approach is also shown.

[1]matlab implementation for motion estimation is available on http://www.sfu.ca/ ibajic/software.html

Fusion Process

Given a video clip (I) with size m * n * t where m*n is the image size and t is the number of frames being processed, TIS approach as proposed in (Qureshi, 2013) can be modified as follows.

$$MapXT_m = sign(DCT(I_{XT_m})) \quad (1)$$

$$MapYT_n = sign(DCT(I_{YT_n})) \quad (2)$$

$$(MapXT_m) \xrightarrow[Transform]{} hMapXY_t \quad (3)$$

$$(MapYT_n) \xrightarrow[Transform]{} vMapXY_t \quad (4)$$

Adding eq (3) and (4)

$$SMap(t) = hMapXY(t) + vMapXY(t) \quad (5)$$

Transformation of frames to global motion compensated (GMC) frames can be written as.

$$GMC_t \xrightarrow[GMC-MVs]{} XY_{t_{MC}} \quad (6)$$

Saliency map can be fused with the information of global motion compensation information by adding equation (5) and (6) using coherent-normalization-based fusion method (C. Chamaret and Meur, 2010).

$$fSMap(t) = (1-\alpha)SMap(t) + \alpha XY_{t_{MC}} + \beta SMap(t)XY_{t_{MC}} \quad (7)$$

MapXT and MapYT represent horizontal and vertical maps, SMap is the saliency map using TIS approach, fSMap is the final map after the combination,

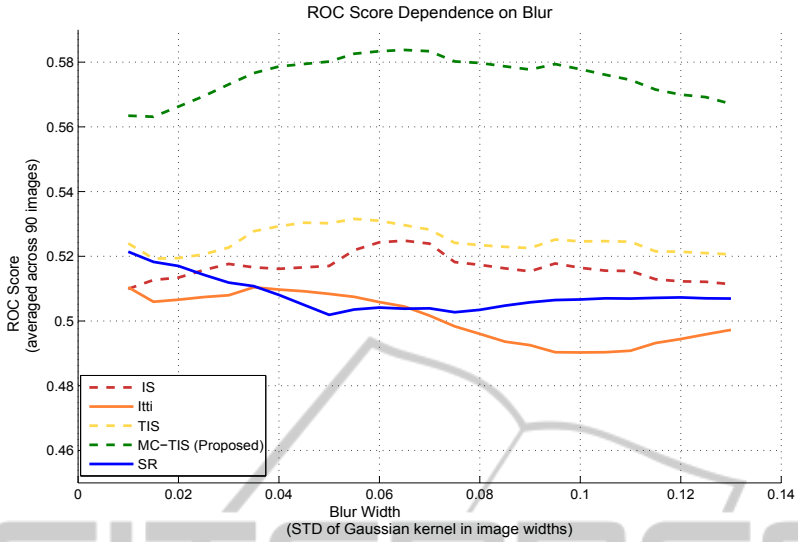

Figure 3: The ROC metric comparison.

alpha and beta are positive constant of values in the range between 0 to 1. MC represent global motion compensated frames, MVs are the motion vectors, I_{XT} and I_{YT} are the slices of the images in horizontal and vertical axis. 'sign (DCT(I))' is the IS process applied on the image. The algorithm is graphically depicted in Figure 1.

4 EVALUATION

To validate the saliency maps generated by our algorithm, the data set of human eye tracking database for standard video sequences introduced by Hadizadeh, Enriquez, and Bajić (H. Hadizadeh, 2012) is used to compare the various saliency map algorithms. This dataset includes a database of gaze locations by 15 subjects free-viewing 90 color images (288 * 352 pixels). In order to evaluate[2] the reliability between a particular saliency map and a set of fixations of the image, Receiver Operating Characteristics (ROC) Area Under the Curve (AUC) score is computed for each image. It is computed in a way as described in Image Signature (Hou et al., 2012). Area under the ROC curve, focus on saliency location at gaze positions.

We compare our saliency maps generated from our proposed approach (MC-TIS) to the following published saliency algorithms: the original Itti-Koch (Itti) saliency model (Itti et al., 1998), Image Signature (IS) (Hou et al., 2012), Temporal Image Signature (TIS) (Qureshi, 2013) and Spectral Residual (SR) (Hou and Zhang, 2007). Figure 3 shows dependence of ROC score of five algorithms on blur when applied to the final saliency maps. The AUC score of all five algorithms under its optimal mean is also shown in Table 1. From Figure 3, it can be seen that the performance of MC-TIS (proposed approach) is better than others saliency algorithms. Hence the regions emphasized by the proposed saliency algorithm matches to a large extent with those image regions seen by humans in free viewing conditions.

Table 1: AUC score of all 5 algorithms.

Algorithm (name)	AUC (mean)
Proposed (MC-TIS)	**0.5838**
Temporal Image Signature (TIS)	0.5315
Image Signature (IS)	0.5248
Spectral Residual (SR)	0.5214
Itti	0.5104

5 CONCLUSION

The proposed approach showed the advantage of using motion estimation information in combination with the saliency information. The addition of motion information can indeed play an important filter in removing unknown artifacts that can origin from the camera motion. The presence of camera motion can influence a saliency model between actual object or salient motion and the background or camera motion. Therefore, it becomes very important to compensate

[2]matlab script for benchmarking developed in California Institute of Technology is available on http://goo.gl/bu1tkc

the camera motion as it is more visible in Figure 2. in case of the TIS approach (c). It was also shown that the detected salient regions by the proposed approach (MC-TIS) have a large overlap with the locations of human eye movement fixations as compared to other saliency algorithms.

Our proposed system may fail in some difficult situations, such as in case of more severe camera motions or camera motion is wrongly estimated. So the success depends strongly on the quality and accuracy of the used motion estimation method as well. Of course the proposed approach can also prove effective in other computer vision problems, e.g. in object categorization or object recognition, video encoding where compression plays an important role.

Interesting avenues for future research are to investigate the combination of motion estimation and saliency algorithms for the application of intelligent video compression.

ACKNOWLEDGEMENTS

This work was supported by the ROMEO project (grant number: 287896), funded by the EC FP7 ICT collaborative research programme.

REFERENCES

Abdollahian, G. and Edward J, D. (2007). Finding regions of interest in home videos based on camera motion. In *IEEE International Conference on Image Processing (ICIP)*, volume 4.

Achanta, R., Hemami, S. S., Estrada, F. J., and Süsstrunk, S. (2009). Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604. IEEE.

Achanta, R. and Süsstrunk, S. (2009). Saliency detection for content-aware image resizing. In *IEEE Intl. Conf. on Image Processing*.

Ali Borji, L. I. (2013). State-of-the-art in visual attention modeling. In *IEEE transactions on Pattern Analysis and Machine Intelligence*, volume 35, pages 185–207.

Borji. A, Tavakoli. H, S. D. and Itti, L. (2013). Analysis of scores, datasets, and models in visual saliency prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 921–928.

C. Chamaret, J. C. C. and Meur, O. L. (2010). Spatio-temporal combination of saliency maps and eye-tracking assessment of different strategies. In *Proc. IEEE Int. Conf. Image Process*, pages 1077–1080.

Chen, Y.-M. and Bajić, I. V. (2010). Motion vector outlier rejection cascade for global motion estimation. *IEEE Signal Process. Lett*, 17(2):197–200.

Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X., and Hu, S.-M. (2011). Global contrast based salient region detection. In *CVPR*, pages 409–416.

Cui, X., Liu, Q., and Metaxas, D. (2009). Temporal spectral residual: fast motion saliency detection. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 617–620, New York, NY, USA. ACM.

Deigmoeller, J. (2010). Intelligent image cropping and scaling. In *PhD thesis, Brunel University*.

Guo, C., Ma, Q., and Zhang, L. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *CVPR'08*.

H. Hadizadeh, M. J. Enriquez, I. V. B. (2012). Eye-tracking database for a set of standard video sequences. *IEEE Trans. on Image Processing*, 21(2):898–903.

Hadi Hadizadeh, I. V. B. (2014). Saliency-aware video compression. *IEEE Trans. on Image Processing*, 23(1):19–33.

Han, J., Ngan, K. N., Li, M., and Zhang, H. (2006). Unsupervised extraction of visual attention objects in color images. *IEEE Trans. Circuits Syst. Video Techn.*, 16(1):141–145.

Hou, X., Harel, J., and Koch, C. (2012). Image signature: Highlighting sparse salient regions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(1):194–201.

Hou, X. and Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR07). IEEE Computer Society*, pages 1–8.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Review Neuroscience*, 2(3):194–203.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259.

M. Cerf, E. P. F. and Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. In *Journal of vision*, volume 9.

Qureshi, H. (2013). Dct based temporal image signature approach. *Proceedings of the 8th International Conference on Computer Vision Theory and Applications (VISAPP '13)*, 1:208–212.

Qureshi, H. and Ludwig, M. (2013). Improving temporal image signature approach by adding face conspicuity map. *Proceedings of the 2nd ROMEO Workshop*.

Schauerte, B. and Stiefelhagen, R. (2012). Predicting human gaze using quaternion dct image signature saliency and face detection. In *Proceedings of the IEEE Workshop on the Applications of Computer Vision (WACV)*. IEEE.

Treisman, A. (1986). Features and objects in visual processing. *Sci. Am.*, 255(5):114–125.

Treisman, A. M. and Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136.

Y.-F. Ma, X.-S. Hua, L. L. and Zhang, H.-J. (2005). A generic framework of user attention model and its application in video summarization. In *Trans. Multi*, volume 7, pages 907–919.