A Performance Evaluation of Fusion Techniques for Spatio-Temporal Saliency Detection in Dynamic Scenes

Satya Muddamsetty, Désiré Sidibé, Alain Trémeau, Fabrice Mériaudeau

A PERFORMANCE EVALUATION OF FUSION TECHNIQUES FOR SPATIO-TEMPORAL SALIENCY DETECTION IN DYNAMIC SCENES

Satya M. Muddamsetty^a*, Désiré Sidibé*^{a}*, Alain Trémeau*^{b} *and Fabrice Mériaudeau*^{a}

^{a}Université de Bourgogne, Le2i UMR CNRS 6306,12 rue de la fonderie, 71200 Le Creusot France
^{b}Université Jean Monnet, Laboratoire Hubert Curien UMR CNRS 5116

ABSTRACT

Visual saliency is an important research topic in computer vision applications, which helps to focus on regions of interest instead of processing the whole image. Detecting visual saliency in still images has been widely addressed in literature. However, visual saliency detection in videos is more complicated due to additional temporal information. A spatio-temporal saliency map is usually obtained by the fusion of a static saliency map and a dynamic saliency map. The way both maps are fused plays a critical role in the accuracy of the spatio-temporal saliency map. In this paper, we evaluate the performances of different fusion techniques on a large and diverse dataset and the results show that a fusion method must be selected depending on the characteristics, in terms of color and motion contrasts, of a sequence. Overall, fusion techniques which take the best of each saliency map (static and dynamic) in the final spatio-temporal map achieve best results.

***Index Terms*—** Spatio-temporal saliency, context information, fusion, performance evaluation

1. INTRODUCTION

Visual saliency is a selective mechanism which drives our attention and limits the processing of incoming information. It has been applied to different application domains including object detection [1], predicting human eye fixations [2], segmentation [3], image/video compression [4], video surveillance [5], image retargeting [6] and mobile robot navigation [7].

According to psychological studies [8], visual attention follows two basic principles: bottom-up and top-down factors. In bottom-up approach, saliency (or attention) is based on center-surround contrast or rarity which suppresses frequently occurring features. In top-down approach, attention is based on context and high level factors of the images such as human faces. There have been many theories and models of visual attention, the most influential being the *feature integration theory* (FIT) of Treisman and Gelade [9] and the *guided search* model of Wolfe [10]. In FIT the author claims that simple features like color, intensity orientation, spatial frequency and motion, are processed rapidly in parallel over the entire visual field which is known as pre-attentive mode and in the second stage objects are identified separately in the attentive mode which requires focused of attention [9]. In the *guided search* model, the goal is to explain and predict the results of visual search experiments. This model considers top-down process along with bottom-up saliency to distinguish the target from the distractors.

Many methods have been proposed for visual saliency detection in images and videos over the past decade and a survey of state of the art methods can be found in [11]. However, most attention has been given to visual saliency detection in static images [3, 12, 13]. Almost all these methods are based on the bottom-up approach and divergence analysis, and they use low-level features such as color, intensity, spatial frequency and orientation to detect salient regions in an image [14].

To deal with video sequences, the temporal information should be considered. Therefore, a static saliency map is often computed for each frame of a sequence and combined with a dynamic map to get the final spatio-temporal saliency map. The accuracy of the saliency model depends on the quality of both the static and dynamic saliency maps and also on the fusion method.

Very few methods deal with videos. In [15] the spatio-temporal saliency map is computed by discriminant center-surround saliency with dynamic textures. [1] proposed a information theoretic saliency which is computed from spatio-temporal volumes and fused by dynamic weight method.

In this paper, we focus mainly on the fusion step and evaluate the performances of nine different fusion techniques on a large dataset of complex dynamic scenes. The paper is organized as follows. In section 2, we describe the spatio-temporal saliency computation approach. Section 3 summarizes the different fusion techniques used for evaluation, and experiments and results are illustrated in Section 4. Finally, Section 5 gives concluding remarks.

2. SPATIO-TEMPORAL VISUAL SALIENCY

Most of the spatio-temporal saliency models are obtained by the fusion of a static saliency map with a dynamic saliency map. Both maps have to be estimated with accuracy in order to get a correct spatio-temporal saliency map. The following subsections briefly describe the method for static and dynamic saliency estimation.

2.1. Static saliency

Many methods have been proposed in literature for visual saliency detection in still images [3, 12, 13], and a review of current approaches can be found in [11]. In our work, we used a saliency detection method based on context information [13] since this method was shown to perform best in a recent evaluation [16].

The context-aware saliency detection method is based on the distinctiveness of a region with respect to its local and global surroundings [13]. The method follows four principles of human visual attention such as low level considerations (contrast and color), global considerations, visual organization rules and high level factors. First, a local single-scale saliency is computed for each pixel in the image. The dissimilarity measure between a pair of patches is defined by:

$$d\left(p_i, q_k\right) = \frac{d_{color}\left(p_i, q_k\right)}{1 + c.d_{position}\left(p_i, q_k\right)}, \qquad (1)$$

where $d_{color}(p_i, q_k)$ is the Euclidean distance between image patches p_i and q_k of size 7×7 centered at pixel i and pixel k in CIELAB color space, and $d_{position}(p_i, q_k)$ is the Euclidean distance between the positions of patches p_i and q_k. c is a constant scalar value set to $c = 3$ in our experiments (changing the value of c does not significantly affect the final result). The single-scale saliency value of pixel i at scale r is then defined as:

$$S_i^r = 1 - e^{-\frac{1}{K}\sum_{k=1}^{K} d(p_i^r, q_k^r)}. \qquad (2)$$

In the second step, a pixel is considered salient if its K most similar patches $\{q_k\}_{k=1}^{K}$ at different scales are significantly different from it. The global saliency of a pixel i is taken as the mean of its saliency at different scales.

The final step includes the immediate context of the salient object. The visual contextual effect is simulated by extracting the most attended localized areas at each scale. A pixel i is considered as a focus of attention at scale r which is normalized to the range [0, 1], if the dissimilarity measure of Eq. 1 exceeds a given threshold ($S_i^r > 0.8$). Then, each pixel which is outside of attended areas is weighted according to its Euclidean distance to the closest focus of attention pixel. The final saliency map which includes the context information is computed as:

$$\hat{S}_i = \frac{1}{M}\sum S_i^r(1 - d_{foci}^r(i)), \qquad (3)$$

where M is the total number of scales and $d_{foci}(i)$ is the Euclidean positional distance between pixel i and the closest focus of attention pixel at scale r.

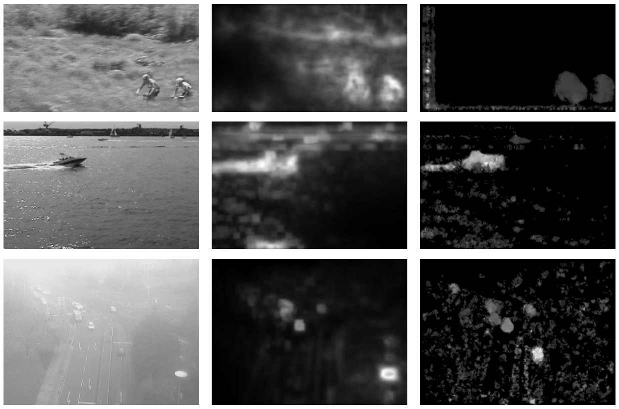

Fig. 1. Examples of static and dynamic saliency detection. From top to bottom row: *Cyclists*, *Boats* and *Traffic* sequences.

2.2. Dynamic saliency

Dynamic saliency is obtained based on the relative motion between successive frames. In order to consider only the motion of objects in the scene, it is necessary to compensate for the background or camera motion. This background motion is computed using a 2D parametric affine motion estimation algorithm developed in [17]. The algorithm provides dominant motion compensation between two successive frames using a robust multi-resolution estimation approach.

After compensation of the dominant motion, the local motion of objects which are present in the frame is estimated by the polynomial expansion technique which accurately computes the displacement field between two frames [18].

Finally, a temporal median filtering is applied to remove noise. If a pixel has a high motion vector in one frame but not in the previous ones then it is probably due to noise resulting from the motion estimation algorithm. This temporal median filter is applied on five successive estimated motion vectors. After temporal median filtering, a normalization step is applied and salient motion information is found which is different from its surroundings.

Some examples of static and dynamic saliency maps obtained by the contex-aware method described in Section 2.1 and the motion estimation technique described in Section 2.2 are shown in Fig. 1.

3. FUSION TECHNIQUES

In the bottom-up visual attention process, low-level features are processed separately to produce feature maps, which are then fused into a master saliency map that shows the most salient regions among all feature maps spatially and temporally. The fusion step is an important component in bottom-up spatio-temporal saliency modeling. Different fusion methods have been used by authors in literature and we briefly described the most common ones below. In the following, the static saliency map, the dynamic saliency map and the fused spatio-temporal saliency map are referred to as M_S, M_D and M_F respectively.

Mean fusion [12]: this fusion method takes the pixel average of both static and dynamic saliency maps.

$$M_F = (M_S + M_D)/2. \tag{4}$$

Max fusion [2]: this is a winer takes all (WTA) strategy in which the maximum value between the two saliency maps is taken for each pixel.

$$M_F = max(M_S, M_D). \tag{5}$$

Multiplication fusion [2]: a pixel by pixel multiplication is done, corresponding to a logical *AND*.

$$M_F = M_S \times M_D. \tag{6}$$

Maximum skewness fusion [2]: this fusion technique takes advantage of the characteristics of the static and the dynamic saliency maps. The static pathway is modulated by its maximum value α. The dynamic saliency map is modulated by its skewness value β. The reinforcement term γ gives more importance to the areas that are salient both in a static and dynamic way.

$$M_F = \alpha M_S + \beta M_D + \gamma(M_S \times M_D), \tag{7}$$

with $\alpha = max(M_S)$, β= $skewness(M_D)$ and $\gamma = \alpha\beta$.

Binary thresholded fusion [6]: first, a binary mask M_B is generated by thresholding the static saliency map (the mean value of M_S is used as threshold). The binary mask is used to exclude spatiotemporal inconsistent areas and to enhance the robustness of the final saliency map when the global motion parameters are not estimated properly.

$$M_F = max(M_S, M_D \cap M_B). \tag{8}$$

Motion priority fusion [19]: this fusion technique is based on *motion priority* which states that a viewer might pay more attention to the motion caused by a moving object even when the static background is more attractive [19]. The perception of moving objects saliency increases nonlinearly with motion contrast and shows significant saturation and threshold effects.

$$M_F = (1 - \alpha)M_S + \alpha M_D, \tag{9}$$

with $\alpha = \lambda e^{1-\lambda}$ and $\lambda = max(M_D) - mean(M_D)$.

Dynamic weight fusion [20]: in this fusion method, the weights of the static and dynamic saliency maps are determined by the ratio between the means of both maps for each frame.

$$M_F = \alpha M_D + (1 - \alpha)M_S, \tag{10}$$

where $\alpha = mean(M_D)/(mean(M_S) + mean(M_D))$.

Information theory fusion [21]: this fusion technique is based on information theory.

$$M_F = \alpha_S I(M_S)M_S + \alpha_D I(M_D)M_D, \tag{11}$$

where the weights α_S and α_D are given by $\alpha_i = max(M_i)I(M_i)$, $I(M_i)$ being the importance of the saliency map M_i.

Scale invariant fusion [22]: in this fusion technique, the input images are analyzed at three different scales from 32×32 to 128×128 to original image size. Three fused maps are obtained which are finally combined linearly into the final spatio-temporal saliency map.

$$M_F = \sum_{l=1}^{3} w_l M_F^l, \tag{12}$$

where $M_F^l = \alpha M_S + (1 - \alpha)M_D$ with $\alpha = 0.5$ is the fused map at scale l and the coefficients of the linear combination are $w_1 = 0.1$, $w_2 = 0.3$ and $w_3 = 0.6$.

4. EXPERIMENTS AND RESULTS

In this section, we evaluate the performances of different fusion approaches described in Section 3 to compute spatio-temporal saliency maps. For a quantitative evaluation, we use a large dataset of complex dynamic scenes [15]. The dataset contains twelve video sequences captured with different challenges such as dynamic background scenes with moving trees, snow, smoke, fog, pedestrians, waves in the sea and moving cameras.

For each sequence, a manual segmentation of the salient objects is available for every frame and served as ground truth. We can therefore evaluate the different fusion techniques by generating Receiver Operating Characteristic (ROC) curves and evaluating the Area Under ROC Curve (AUC). For each fusion technique, the obtained spatio-temporal saliency map is first normalized to the range $[0, 1]$, an then binarized using a varying threshold $t \in [0, 1]$. With the binarized maps, we compute the true positive rate and false positive rate with respect to the ground truth data.

Table 1 summarizes the results obtained with all sequences by the different fusion techniques. We observe that the best performances are obtained by the *Mean* [12], *Scale Invarian*t [22], *Max* [2] and *Dynamic Weight* [20] fusion methods respectively. In particular, the *Mean* fusion technique achieves an average AUC value of 0.9325 for all twelve sequences. Those fusion methods take the best of each saliency map (static and dynamic) in the final spatio-temporal map: the static saliency value is given more importance if it is higher at a given position and vice-versa. On the contrary, the *Motion Priority* [19] and the *Multiplication* [2] fusion techniques give the least performances. In particular, the *Motion Priority* method gives an average AUC value of 0.7943 for

Sequence	Mean	Max	AND	MSF	BTF	DWF	MPF	ITF	SIF	Avg AUC
Birds	0.9713	0.9794	0.9023	0.9563	0.9852	0.9669	0.7639	0.9097	0.9245	0.9288
Boats	0.9891	0.9745	0.9867	0.9881	0.9695	0.9827	0.9808	0.9889	0.9829	**0.9826**
Cyclists	0.9628	0.9497	0.8862	0.9418	0.9533	0.9602	0.8394	0.9248	0.9498	0.9298
Chopper	0.9784	0.9847	0.6891	0.6956	0.9852	0.9850	0.6791	0.9628	0.9711	0.8812
Freeway	0.7128	0.6633	0.7023	0.7614	0.5087	0.5456	0.7581	0.6218	0.7452	**0.6688**
Peds	0.9608	0.9435	0.8984	0.9380	0.9441	0.9512	0.8852	0.9400	0.9558	0.9352
Jump	0.9395	0.9314	0.8949	0.9212	0.9459	0.9479	0.8535	0.8804	0.9197	0.9149
Ocean	0.8273	0.7465	0.8108	0.8126	0.7535	0.7810	0.8032	0.8063	0.8412	0.7980
Surfers	0.9453	0.9782	0.7993	0.9208	0.9844	0.9545	0.6251	0.9334	0.8757	0.8907
Skiing	0.9678	0.9784	0.5195	0.6491	0.9807	0.9796	0.4905	0.9394	0.9365	0.8268
Landing	0.9701	0.9524	0.9718	0.9703	0.9521	0.9579	0.9047	0.9353	0.9720	0.9541
Traffic	0.9645	0.9566	0.8860	0.9540	0.8736	0.9615	0.9477	0.9640	0.9593	0.9408
Avg AUC value	**0.9325**	0.9199	0.8289	0.8758	0.9030	0.9145	**0.7943**	0.9006	0.9200	

Table 1. Fusion techniques evaluation results. Mean (Mean fusion), Max (Max fusion), AND (Multiplication fusion), MSF (Maximum skewness fusion), BTF (Binary thresholded fusion), DWF (Dynamic weight fusion), MPF (Motion priority), ITF (Information theory fusion), SIF (Scale invariant fusion).

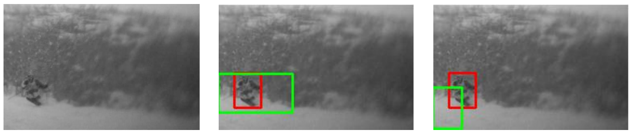

Fig. 2. Example of salient region segmentation with the *Skiing* sequence. From left to right: input frame; detection with *Binary Threshold* and with *Motion Priority* fusion techniques. Red box indicates ground truth and green box indicates the detected salient region.

all sequences, which is 17% less than the value obtained by the *Mean* fusion technique. This can be explained by the fact that this fusion approach gives more importance to motion information. Therefore, when the motion contrast is not estimated properly, the final saliency map is not accurate. This problem can be observed with the *Skiing* sequence for which the *Binary Threshold* and the *Motion Priority* fusion methods achieve AUC values of 0.9807 and 0.4905 respectively. The salient object segmentation results for those two fusion methods are shown in Fig. 2. As can be seen, for this sequence with low motion contrast *Motion Priority* fusion method fails to localize the target due to the incorrect estimation of the dynamic saliency map. The red box shows the ground truth location of the salient object while the green bow is the output of the estimated spatio-temporal saliency detection method.

Analyzing the sequences individually, we see that the best and least performances are obtained with the *Boats* and *Freeway* sequences, respectively, with average AUC values of 0.9826 and 0.6688 for all fusion techniques. The *Boats* sequence shows good color and motion contrasts, so both static and dynamic maps are estimated correctly (as shown in Fig. 1). As a consequence, all fusion techniques perform well. On the other hand, the color contrast of the *Freeway* sequence is very limited. So fusion methods such as *Binary Threshold* (BTF) and *Dynamic Weight* which give high importance to the static map perform poorly, with AUC values of 0.5087 and 0.5456 respectively. For instance, in the BTF technique, the mean value of the static map is used to generated a binary mask which is then combined with the dynamic map. It is clear that if the static map is not accurate, the final spatio-temporal saliency map will be inaccurate as well.

5. CONCLUSION

In this paper a performance evaluation of fusion techniques for spatio-temporal saliency detection in dynamic scenes is presented. The nine fusion techniques are evaluated on a large dataset of twelve complex dynamic scenes. The results show the consistency of fusion approaches that base decision on the scene's characteristics as the final spatio-temporal saliency map takes the best of each individual saliency map (static and dynamic). This include *Mean*, *Scale Invariant*, *Max* and *Dynamic Weights* fusion methods. On the other hand, fusion techniques which are based on a strong a priori such as *Motion Priority* fusion achieve good results only when the underlying assumption is satisfied. Thus, they performances vary depending on the sequence.

It is clear that the accuracy of a spatio-temporal saliency map depends on the quality of both static and dynamic maps, which are based on the scene's contents. Therefore, it would be useful to derive the weights (fusion technique) based on the images contents. We are currently investigating in this direction.

6. REFERENCES

[1] L. Chang, Pong C. Yuen, and Guoping Qiu, "Object motion detection using information theoretic spatio-temporal saliency," *Pattern Recogn. 2009*, vol. 42, no. 11, pp. 2897–2906.

[2] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *IJCV,2009*, vol. 82, no. 3, pp. 231–243.

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

[4] C. L. Guo and L. M. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing,*, vol. 19, no. 1, pp. 185 –198, 2010.

[5] F. F. E. Guraya H . Konik T. Yubing, F. A. Cheikh and A. Trémeau, "A spatiotemporal saliency model for video surveillance," *Cognitive Compuation,2011*, vol. Volume 3, Issue 1, pp. pp 241–263.

[6] T. Lu, Z. Yuan, Y. Huang, D. Wu, and H. Yu, "Video retargeting with nonlinear spatial-temporal saliency fusion," in *ICIP,2010*.

[7] C. Siagian and L. Itti, "Biologically inspired mobile robot vision localization," *IEEE Transactions on Robotics*, vol. 25, no. 4, pp. 861–873, July 2009.

[8] S. Frintrop, *Computational Visual Attention*, Springer, 2011.

[9] A. M. Treisman and G. Gelade, "A feature-integration theory of attention.," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[10] J M Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[11] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on PAMI,*, vol. 35, no. 1, pp. 185–207, 2013.

[12] C. Koch L. ltti and E. Neibur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on PAMI,1998*, vol. vol 20, pp. 1254–1259.

[13] S. Goferman, L. Zelnik-manor, and A. Tal, "Context-aware saliency detection," in *IEEE Conf. on CVPR,2010*.

[14] J.B. Huang and N. Ahuja, "Saliency detection via divergence analysis: An unified perspective," in *In proc of ICPR,2012*. IAPR.

[15] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Transactions on PAMI,*, vol. 32, no. 1, pp. 171 –177, 2010.

[16] A. Borji, N. Dicky. Sihite, and L. Itti, "Salient object detection: A benchmark," in *ECCV (2)*, 2012, pp. 414–429.

[17] P. Anandan M. J. Black, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *CVIU*, vol. Vol 63, No. 1,, pp. 74–104, 1996.

[18] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis, 13th Scandinavian Conference, SCIA 2003, Halmstad, Sweden, June 29 - July 2, 2003, Proceedings*, Josef Bigün and Tomas Gustavsson, Eds. 2003, vol. 2749 of *Lecture Notes in Computer Science*, pp. 363–370, Springer.

[19] J. Peng and Q. Xiaolin, "Keyframe-based video summary using visual attention clues," *IEEE on MultiMedia,*, vol. 17, no. 2, pp. 64 –73, 2010.

[20] X. Xiao, C. Xu, and Y. Rui, "Video based 3d reconstruction using spatio-temporal attention analysis," in *Multimedia and Expo (ICME), 2010*.

[21] B. Han and B. Zhou, "High speed visual saliency computation on gpu," in *InProc of ICIP,2007*.

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