SPATIO-TEMPORAL FUSION OF VISUAL ATTENTION MODEL

Anis RAHMAN, Guanghan SONG, Denis PELLERIN, Dominique HOUZET

Department Images and Signal, GIPSA-lab
961 rue de la Houille Blanche, BP 46, 38402 Grenoble Cedex, France

ABSTRACT

The motivation behind a spatio-temporal visual saliency model is to extract salient information from two distinct pathways: static (intensity) and dynamic (motion). Consequently, the information from these pathways is combined to get the final visual saliency map. Since the response of the pathways is different, the step of combination of the maps is important. As a consequence, we study six recent fusion techniques against two video databases using human eye positions from an eye tracker. A criterion is used to evaluate and underline the significance of these fusion methods.

1. INTRODUCTION

Visual attention is a process to attend to regions in a visual scene that appears to be salient from their surroundings. The map to represent this spotlight of focus in the field of computer vision is called visual saliency map. In the human vision system, the raw information from the visual stimuli is decomposed into several paths that process this information for certain features. At the end of all the processing, these feature maps are combined together into a final visual saliency map that represents the regions of attention. It is important to understand this function of the human vision system, and to create models for computing that can be used to extract relevant information. This capability is potentially applicable in the domains of video compression, video synthesis and analysis, robotics, and many more.

The objective of the paper is to have a better understanding of the potentialities of different fusion methods, and to make the best choice for our application. Furthermore, the fusion method must be adaptive to the changing environment and quick to process.

The rest of the paper is organized as follows: Section 2 introduces the spatio-temporal visual saliency model employed to compute the visual saliency. Section 3 presents the methods used for information fusion for the two-pathway visual saliency model. Section 4 describes the video databases, eye tracker experiment, and criterion used for the evaluation of the results. Section 5 presents the findings on the videos, and demonstrates the performance of the different fusion methods evaluated. Section 6 concludes the paper.

The research is supported by Rhône-Alpes region (France) under the CIBLE project No. 2136.

2. SPATIO-TEMPORAL SALIENCY MODEL

Different spatio-temporal models exist to extract the visual saliency using spatial and temporal information, and to process them separately. Here, the model [5] presented is based on the human visual system, where the layout of the different steps of the model makes it biologically inspired. Likewise, the information is decomposed into two distinct pathways: static and dynamic pathways to produce static and dynamic maps, as shown in Figure 1. We use a faster GPU implementation [8] to evaluate the model for the test video databases.

The static pathway starts with retinal preprocessing to enhance the details of the input visual scene by increasing the higher luminance frequencies. This preprocessed data is passed through a 2D bank of Gabor filters arranged to model the receptive fields in the visual cortex. This bank uses six orientations and four frequencies, and results in 24 partial maps. The resulting maps interact with each other that mimic the lateral connections in the neuronal environment. Finally, all the partial maps are concatenated into a static visual saliency map M_s. Besides the static pathway, the dynamic pathway estimates the region of motion in the visual scene. The pathway starts with camera motion compensation to estimate the dominant motion of the salient regions. This preprocessing is followed by 2D motion estimation of local motion against the background. This estimation utilizes speed, orientation, and direction in the moving scene. In the end, temporal filtering is performed to minimize the effects of bad estimations that result in the dynamic visual saliency map M_d. Ultimately, the final visual saliency map M_{sd} is produced by the fusion of both the static and dynamic maps treated with a center effect using a 2D Gaussian window. Here, the inclusion of center effect is important because in real videos the objects of interest are often placed in the center of the screen.

3. DIFFERENT FUSION METHODS

The work evaluates six recent fusion techniques for the fusion of static and dynamic saliency maps from the spatio-temporal model. All these intermediate maps have unequal influences in the final visual saliency map, due to the varying input for the separate pathways. Therefore, the motivation behind the evaluation is to find an efficient and robust fusion method that not only extracts all useful information, but also reduces the effects of false findings.

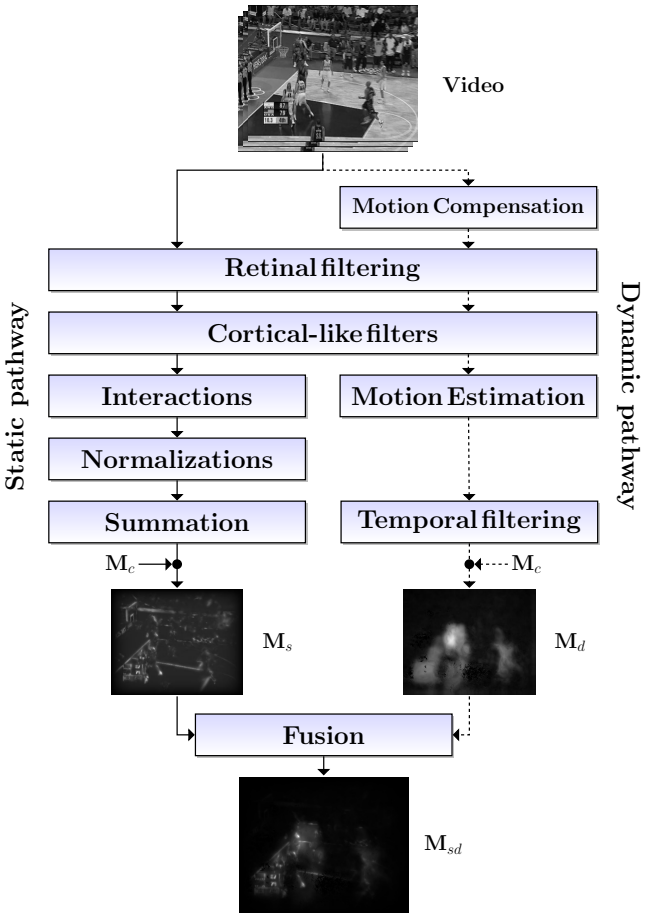


Figure 1: Block diagram of spatio-temporal visual saliency model, where M_s, M_d and M_{sd} are static, dynamic and fused visual saliency maps treated with center effect M_c.

3.1 Fusion using Shannon's information theory [2]

Using the Shannon information theory, the conspicuous spots are taken as events. Hence, the information conveyed by each event is calculated by counting the values above a threshold. This probability is used to yield the information conveyed by each conspicuity map.

$$P(M) = \frac{M > \tau}{M}$$
$$\tau = 0.6 \cdot MAX(M_s \cup M_d)$$

The weights for the static and dynamic map are obtained using:

$$I(M) = -log(P(M))$$
$$W(M) = I(M)MAX(M)$$

and, we get the final map using equation:

$$M_{sd} = W(M_s)I(M_s)M_s + W(M_d)I(M_d)M_d$$

3.2 Motion priority fusion model [3]

The work uses the notion of motion priority, as the human vision system pays more attention to the regions in motion against the static background. Here, strong motion contrast will increase the weight for the dynamic map, whereas the fusion weight of the spatial information causes it to decrease. The dynamic weights for the two pathways are calculated as:

$$W_d = \alpha\ exp(1-\alpha)$$
$$W_s = 1 - W_d$$
$$\alpha = MAX(M_d) - MEAN(M_d)$$

and, then the final saliency map is computed using:

$$M_{sd} = W_s M_s + W_d M_d$$

3.3 Binary threshold mask fusion model [4]

The fusion method uses a mask for the dynamic map, which enhances the robustness when the motion parameters are not estimated correctly. It is useful to exclude the inconsistent regions, and requires no selection of a weighting factor for the spatial and temporal information. Furthermore, the use of MAX operator avoids the suppression of insignificant salient regions.

$$M_{sd} = MAX(M_s, M_d \cap M_{st})$$

Here, $M_{st}(\tau = \bar{M_s})$ is the thresholded static saliency map.

3.4 Max skewness fusion model [5]

The fusion model modulates the static and dynamic saliency maps using the maximum and skewness respectively using:

$$M_{sd} = \alpha M_s + \beta M_d + \gamma M_s M_d$$
$$where, \begin{cases} \alpha = MAX(M_s) \\ \beta = SKEWNESS(M_d) \\ \gamma = \alpha\beta \end{cases}$$

3.5 Key memory fusion model [7]

The fusion model uses temporal changes to improve the mean μ and variance S that are calculated as:

$$\mu_s^k = (1-\alpha)\mu_s^{k-1} + \alpha\mu_s^k$$
$$\mu_d^k = (1-\alpha)\mu_d^{k-1} + \alpha\mu_d^k$$
$$S_s^k = (1-\alpha)S_s^{k-1} + \alpha S_s^k$$
$$S_d^k = (1-\alpha)S_d^{k-1} + \alpha S_d^k$$
$$\alpha = \begin{cases} 1/k & 1 \le k \le K \\ 1/K & k > K \end{cases}$$

where K depicts the rate of illumination changes that is set to 2. Whereas, the weight is calculated as:

$$W_k = \frac{(\mu_s^k - \mu_d^k)}{(\delta_s^k + \delta_d^k)}$$

Finally, the fused saliency map is computed as:

$$M_{sd} = W_k M_s + M_d$$

3.6 Dynamic weight fusion model [10]

The fusion method uses a dynamic weight calculated from the ratio of the means of static and dynamic maps ($\bar{M}_s$ and $\bar{M}_d$) from the model.

$$M_{sd} = \alpha M_d + (1-\alpha) M_s$$

$$\alpha = \frac{\bar{M}_d}{\bar{M}_s + \bar{M}_d}$$

4. MATERIALS AND METHOD

4.1 Video databases

The different fusion methods detailed in Section 3 are tested against two video databases named GS [9] and SM [5] that are assembled using the approach followed by Carmi and Itti [1]. Here, each database is composed of videos with varying content from films, documentaries, sports etc . All the videos are decomposed into small clip snippets of several seconds that are randomly joined together into a set of clips of ∼30s. This random fusion of the clip snippets is interesting to study the influences of the early attention rather than the participant anticipating the transitions among the visual frames. The general information regarding the video databases used is illustrated in Table 1. The main differences between the two video databases are frame size, content and video quality.

4.2 Ground truth

The eye tracker experiment involves about 20 participants instructed to do free viewing of the video stimuli; the two video databases. The respective eye positions for each participant are recorded using Eyelink II eye tracker, which afterwards are combined to build density maps for each video frame. These resulting density maps are compared against the visual saliency maps to find the relevance of the model.

4.3 Criterion

Normalized Scanpath Saliency (NSS) [6] is the criterion employed to evaluate the results of the model with their corresponding eye fixations. It is a kind of Z-score to compare a saliency map from the model to eye position density map of the participants.

$$NSS(k) = \frac{\overline{M_h(x,y,k) \times M_m(x,y,k)} - \overline{M_m(x,y,k)}}{\sigma_{M_m(x,y,k)}}$$

where $\overline{M_m(x,y,k)}$ is the average of the map $M_m(x,y,k)$, $\sigma_{M_m(x,y,k)}$ is the standard deviation of the map $M_m(x,y,k)$ and $M_h(x,y,k)$ is the density map of normalized eye positions. Here, m and h notations correspond to the model and human respectively.

4.4 Dispersion

Dispersion is the measure used to analyze how eye positions change overtime, we consider the dispersion of these positions among the participants. It is defined by the equation:

$$D = \frac{1}{N^2} \sum_{i,j<i} d_{i,j}^2$$

where N is the number of participants and $d_{i,j}$ is the distance between the eye positions of participants i and j. A lower value shows the eye positions to be closer for the participants.

5. RESULTS

The static and dynamic saliency maps are combined into a visual saliency map using different fusion methods described in Section 3. These resulting maps are compared against the experimental eye position density maps using NSS as the criteria.

Here, Figures 2 and 3 illustrate the evolution of mean NSS over time for the two video databases, where time is represented by the frame position of the clip snippet (one image = 40ms). The curve is plotted by averaging the NSS values of the 1st frame of each clip snippet, likewise, the same process is repeated for the first 70 frames of every clip snippet. It is observed that the evolution curve starts off with a low mean NSS value at the beginning of each frame, and it quickly reaches to a peak value at about 13th frame (520ms). This phenomenon is explained by the fact that at the change of every clip snippet the real eye positions correspond to the salient regions from the previous clip snippet. The mean NSS curve reaches its maximum value after the involvement of bottom-up influences on the visual stimulus, and then decreases with time. Similarly, Figure 4 shows that the value of dispersion is high at the beginning of the videos, and it drops to a lowest value as all the participants find a common region of interest. It is significant that about the same time the value of NSS is at a peak value.

Table 2 shows the NSS mean values and gains after fusion, and Figure 5 illustrates resulting saliency maps for the two test video databases. Firstly, the fusion methods for $M_{sd}han$ and $M_{sd}lu$ consider a threshold to extract only the useful information from the partial maps. Secondly, we know that in a human visual system attention is often influenced by motion, that is incorporated in the fusion method used for $M_{sd}jiang$. Likewise, $M_{sd}marat$ uses skewness as a the motion

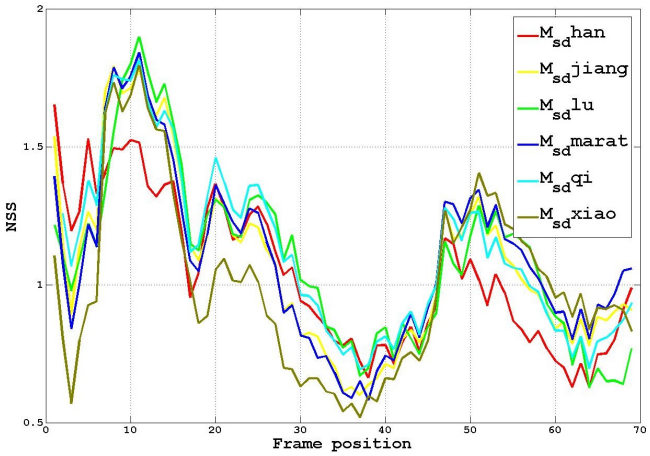


Figure 2: Evolution of NSS for GS video database using various fusion techniques

					Experimental video databases	
Name	Number of participants (M/F)	Total clips	Number of clip snippets per clip	Clip snippet duration	Total frames	Frame size
GS	12/3	10	6	5-8s	10000	608×272
SM	20/10	20	15	1-3s	14000	720×576

Table 1: General information about the video databases used.

							Different fusion methods		
Video database	Criterion	M_s	M_d	$M_{sd}han$ [2]	$M_{sd}jiang$ [3]	$M_{sd}lu$ [4]	$M_{sd}marat$ [5]	$M_{sd}qi$ [7]	$M_{sd}xiao$ [10]
GS	NSS	0.57	1.02	1.02	1.26	1.14	1.19	1.17	1.25
	$Gain\ for\ NSS\ (\cdot/M_d)$	-	-	0%	23%	12%	17%	15%	22%
SM	NSS	0.88	1.19	1.33	1.40	1.37	1.28	1.43	1.35
	$Gain\ for\ NSS\ (\cdot/M_d)$	-	-	12%	18%	15%	7%	20%	13%

Table 2: Mean NSS for various fusion methods evaluated against two video databases

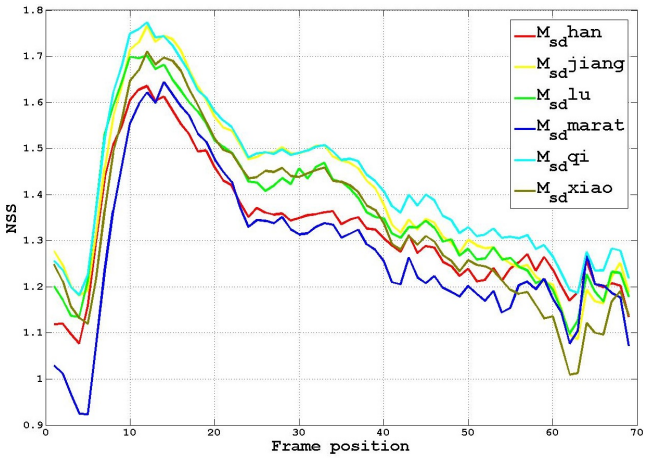


Figure 3: Evolution of NSS for SM video database using various fusion techniques

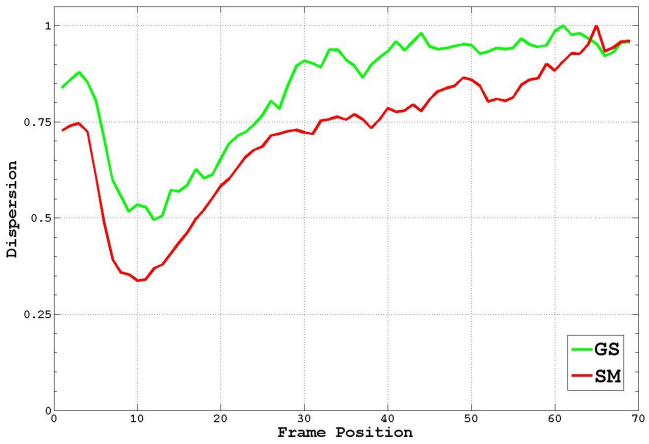
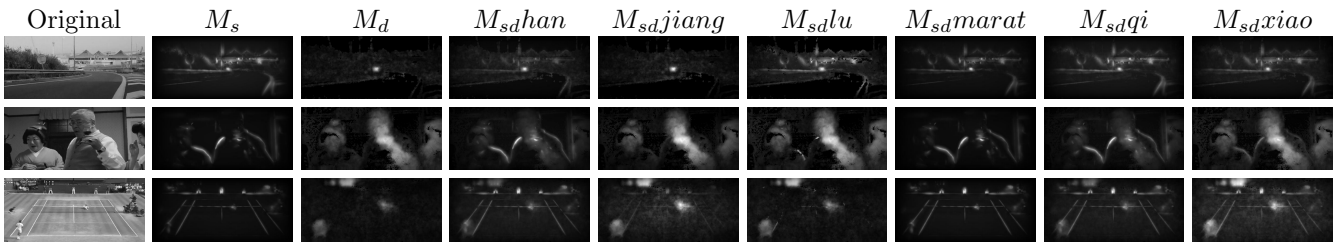


Figure 4: Dispersion D for eye position as a function of frame for the two video databases

priority parameter. Thirdly, in $M_{sd}qi$ the fusion used is additive, but the weight from the approximation determines the validity of the maps. Lastly, $M_{sd}xiao$ uses dynamic weight computed as the ratio of the means of static and dynamic maps.

In Figure 5, the first row for each database represent a scene with high motion, whereas the other samples represent indoor, outdoor and sport scenes. The resulting visual saliency maps show that $M_{sd}jiang$ and $M_{sd}xiao$ give priority to the salient objects in the dynamic saliency map M_d. In the case of indoor and sport scenes, the contours from the static maps M_s and motion from the dynamic maps M_d are fused into final saliency maps depending on the fusion method used.

In Table 2, the NSS values for partial maps M_s and M_d from the two separate pathways of the model show that globally results for SM video database are better than the GS video database. Besides this, the results show that the difference of the amount of salient information in the partial maps contributes unequally in the final visual saliency map. This unequal influence is achieved by computing dynamic weights for the partial maps. Resultantly, we get better results for fused maps $M_{sd}jiang$ and $M_{sd}xiao$ for GS video database. Whereas, in case of SM video database, the fusion maps $M_{sd}qi$ takes into account the quality of both static and dynamic maps. Additionally, the use of a memory effect further enforces the fusion results, and hence we obtain a gain of 20%. Furthermore, we know that dynamic information is important in human visual system, and hence a priority will improve the results. This is observed for GS video database, where the fusion maps $M_{sd}jiang$ has the best results.



(a) Examples of saliency maps using different fusion methods for GS video database

(b) Examples of saliency maps using different fusion methods for SM video database

Figure 5: Some results for the two video databases

6. CONCLUSION

The study evaluates six fusion methods for a spatio-temporal model against two test video databases with varying features. In a nutshell, each of the fusion methods has their separate advantages, which could be chosen in function of the application or combined intelligently to result in a method that is more robust.

REFERENCES

[1] R. Carmi and L. Itti. Visual causes versus correlates of attentional selection in dynamic scenes. *Vision Res.*, 46(26):4333 – 4345, 2006.

[2] B. Han and B. Zhou. High speed visual saliency computation on gpu. In *IEEE Int. Conf. Image Process.*, pages 361–364, 2007.

[3] P. Jiang and X. Qin. Keyframe-based video summary using visual attention clues. *IEEE Multimedia*, 17(2):64 –73, 2010.

[4] T. Lu, Z. Yuan, Y. Huang, D. Wu, and H. Yu. Video retargeting with nonlinear spatial-temporal saliency fusion. In *IEEE Int. Conf. on Image Process.*, 2010.

[5] 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. *Int. J. Comput. Vision*, 82(3):231–243, 2009.

[6] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005.

[7] F. Qi, X. Song, and G. Shi. Lda based color information fusion for visual objects tracking. In *IEEE Int. Conf. on Image Process.*, pages 2201 –2204, 2009.

[8] A. Rahman, D. Houzet, D. Pellerin, S. Marat, and N. Guyader. Parallel implementation of a spatio-temporal visual saliency model. *Journal of Real-Time Image Processing*, 6:3–14, 2011.

[9] G. Song, D. Pellerin, and L. Granjon. Sound effect on visual gaze when looking at videos. In *16th European Signal Process. Conf.*, 2011.

[10] X. Xiao, C. Xu, and Y. Rui. Video based 3d reconstruction using spatio-temporal attention analysis. In *Proc. IEEE Int. Conf. on Multimedia and Expo*, pages 1091–1096, 2010.