# Spatio-temporal fusion of visual saliency model

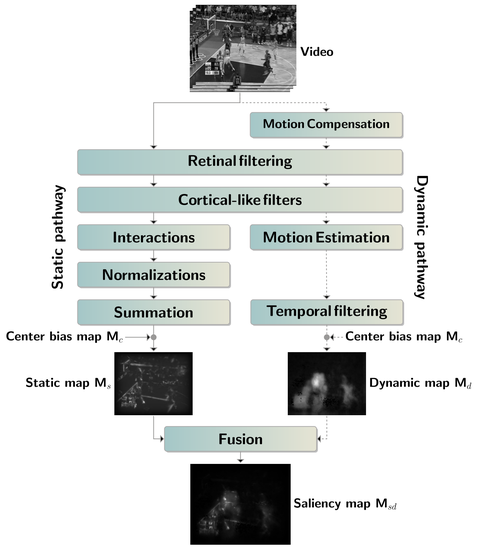
{Anis.Rahman, Guanghan.Song,Denis.Pellerin, Dominique.Houzet}@gipsa-lab.grenoble-inp.fr
Department Images and Signal, GIPSA-Lab, Grenoble, France

## Contribution

Spatio-temporal visual saliency model extracts salient information from two distinct pathways: static (intensity) and dynamic (motion). Fusion is important because both these pathways respond differently. Here, we study six fusion techniques against two video databases using human eye positions from an eye tracker.

## Model

Bottom-up visual saliency model determines source of attention, and its concentration to contribute or initiate other tasks. Such models are interesting for applications like robotics, image analysis, compression, video indexing.




Spatio-temporal saliency model proposed by Marat et al. [4].

## References

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

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

[3] 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.

[4] S. Marat, T. Ho 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.

[5] 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.

[6] 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.

## Fusion methods

**Shannon's information theory fusion [1]** considers conspicuous spots as events. Hence, information by each event is calculated using a threshold.

$$P(M) = \frac{M > \tau}{M}$$
$$\tau = 0.6 \cdot MAX(M_s \cup M_d)$$
$$I(M) = -log(P(M))$$
$$W(M) = I(M)MAX(M)$$
$$M_{sd} = W(M_s)I(M_s)M_s + W(M_d)I(M_d)M_d$$

**Motion priority fusion model [2]** uses notion that human vision system pays more attention to the regions in motion against the static background.

$$W_d = \alpha\ exp(1-\alpha)$$
$$W_s = 1 - W_d$$
$$\alpha = MAX(M_d) - MEAN(M_d)$$
$$M_{sd} = W_s M_s + W_d M_d$$

**Binary threshold mask fusion model [3]** uses masked dynamic map to enhance robustness of motion parameters. Also, MAX operator avoids suppression of insignificant salient regions after normalizations.

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

$M_{st}$ is the thresholded static saliency map (the threshold is $(\tau = \bar{M}_s)$.

**Max skewness fusion model [4]** modulates static and dynamic saliency maps using the maximum and skewness respectively.

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

**Key memory fusion model [5]** uses temporal changes to improve mean $\mu$ and variance $S$.

$$\mu_s^k = (1-\alpha)\mu_s^{k-1} + \alpha\mu_s^k$$
$$\mu_d^k = (1-\alpha)\mu_d^{k-1} + \alpha\mu_d^k$$
$$S_s^k = (1-\alpha)S_s^{k-1} + \alpha S_s^k$$
$$S_d^k = (1-\alpha)S_d^{k-1} + \alpha S_d^k$$
$$\alpha = \begin{cases} 1/k & 1 \le k \le K \\ 1/K & k > K \end{cases}$$
$$W_k = \frac{(\mu_s^k - \mu_d^k)}{(\delta_s^k + \delta_d^k)}$$
$$M_{sd} = W_k M_s + M_d$$

**Dynamic weight fusion model [6]** calculates dynamic weight from ratio of means of static and dynamic maps.

$$M_{sd} = \alpha M_d + (1-\alpha)M_s$$
$$\alpha = \frac{\bar{M}_d}{\bar{M}_s + \bar{M}_d}$$

## Test video databases

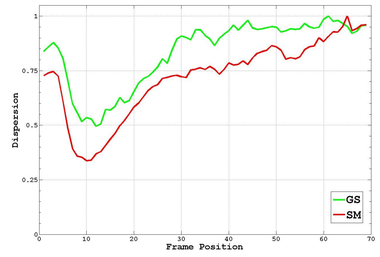
| | Experimental video databases | | | | | |
|---|---|---|---|---|---|---|
| Name | Participants (M/F) | Total clips | Clip snippets per clip | Clip snippet duration | Total frames | Frame size |
| GS | 12/3 | 10 | 6 | 5-8s | 10000 | $608 \times 272$ |
| SM | 20/10 | 20 | 15 | 1-3s | 14000 | $720 \times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

General information about the video databases used.

## Dispersion of eye positions

Dispersion is measure to analyze how eye positions change overtime, we consider dispersion of these positions among the participants. We observe the evolution of this dispersion along time.

$$D = \frac{1}{N^2} \sum_{i,j<i} d_{i,j}^2$$
$$where, \begin{cases} N: number\ of\ participants \\ d_{i,j}: distance\ between\ eye\ positions\ of\ participants\ i\ \&\ j \end{cases}$$

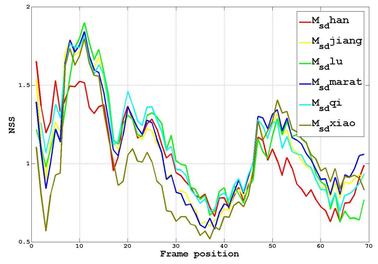



Dispersion $D$ for eye positions as function of frame position for the two video databases.

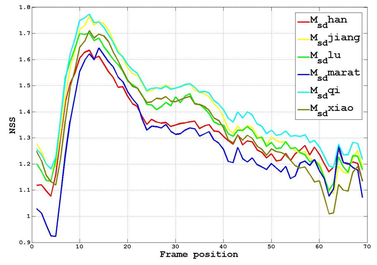
## Criterion of comparison

Normalized Scanpath Saliency (NSS) is a Z-score to compare saliency maps to eye position density maps from participants.

$$NSS(k) = \frac{\overline{M_h \times M_m} - \overline{M_h}\ \overline{M_m}}{\sigma_{M_m}}$$
$$where, \begin{cases} M_m: salience\ map\ of\ the\ model \\ M_h: density\ map\ of\ normalized\ eye\ positions \end{cases}$$

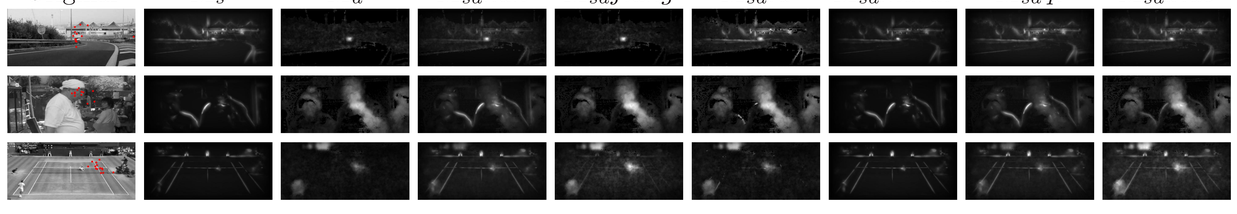



Evolution of NSS for GS video database.




Evolution of NSS for SM video database.

## Results: Saliency maps




Resulting saliency maps after fusion for SM video database.

## Results: NSS gain

| Video database | Criterion | $M_s$ | $M_d$ | $M_{sd}han$[1] | $M_{sd}jiang$[2] | $M_{sd}lu$[3] | $M_{sd}marat$[4] | $M_{sd}qi$[5] | $M_{sd}xiao$[6] |
|---|---|---|---|---|---|---|---|---|---|
| GS | $NSS$ | 0.57 | 1.02 | 1.02 | 1.26 | 1.14 | 1.19 | 1.17 | 1.25 |
| | $NSS\ Gain\ (\cdot/M_d)$ | - | - | 0% | 23% | 12% | 17% | 15% | 22% |
| SM | $NSS$ | 0.88 | 1.19 | 1.33 | 1.40 | 1.37 | 1.28 | 1.43 | 1.35 |
| | $NSS\ Gain\ (\cdot/M_d)$ | - | - | 12% | 18% | 15% | 7% | 20% | 13% |

Mean NSS for various fusion methods evaluated against two video databases.

## Conclusion

For the two used databases, NSS values are better for dynamic maps than for static maps. Hence, the best results are given by fusion methods based on motion priority. For future applications, the choice of one fusion method depends on the reliability of each pathway.