SPATIO-TEMPORAL SALIENCY BASED ON RARE MODEL

Marc Décombas[ab], *Nicolas Riche*[c], *Frédéric Dufaux* [b],*Béatrice Pesquet-Popescu* [b],*Matei Mancas*[c], *Bernard Gosselin*[c], *Thierry Dutoit*[c]

[a] Thales Communications & Security – Laboratoire MMP, 92230 Gennevilliers, France
[b]Télécom ParisTech – Dept. Traitement du Signal et des Images, 75014 Paris, France
{marc.decombas, frederic.dufaux, beatrice.pesquet}@telecom-paristech.fr
[c]University of Mons (UMONS) – Faculty of Engineering (FPMs) , Mons, Belgique
{Nicolas.Riche, Matei.Mancas, Bernard.Gosselin, Thierry.Dutoit}@umons.ac.be

ABSTRACT

In this paper, a new spatio-temporal saliency model is presented. Based on the idea that both spatial and temporal features are needed to determine the saliency of a video, this model builds upon the fact that locally contrasted and globally rare features are salient. The features used in the model are both spatial (color and orientations) and temporal (motion amplitude and direction) at several scales. To be more robust to moving camera a module computes the global motion and to be more consistent in time, the saliency maps are combined together after a temporal filtering. The model is evaluated on a dataset of 24 videos split into 5 categories (Abnormal, Surveillance, Crowds, Moving camera, and Noisy). This model achieves better performance when compared to several state-of-the-art saliency models.

***Index Terms*—** Visual attention, Saliency, Rarity Mechanism, Optical Flow

1. INTRODUCTION

The aim of visual saliency models is to automatically predict human attention. The term *attention* refers to the process that allows one to focus on some stimuli at the expense of others and has been introduced in [1], [2]. Human attention mainly consists of bottom-up and top-down processes. Bottom-up attention uses features extracted from the signal to find the most salient objects. Top-down attention uses a priori knowledge about the scene or task-oriented knowledge in order to modify the bottom-up saliency. This domain is a very active area due to several important applications such as gaze prediction [3], content aware compression [4], video retargeting [5], and video summary [6]. The general idea of saliency models is that rare, novel or surprising information is salient. The objective of those models is to identify unusual features in a given spatio-temporal context like in [7], [8], [9], [10], [11], [12], [13], [14], [15], [16]. In this paper, the models [12], [13], [14], [15], [16] has been chosen for the comparison.

Seo and Milanfar proposed a framework for static and space-time saliency detection [12]. Their saliency model is based on the comparison of a local window centered on a pixel location and its neighboring windows. The temporal aspect is taken into account by defining windows as spatio-temporal cubes. In [13], Culibrk *et al.* introduced a model based on motion and simple static cues. More precisely, a multi-scale background modeling and foreground segmentation is carried out. This model employs the principles of multi-scale processing, cross-scale motion consistency, outlier detection and temporal coherence. Zhang *et al.* proposed a Bayesian framework for saliency detection called SUN [14]. Their approach is based on the assumption that visual saliency is the probability of a target at every location given the visual features observed. Mancas *et al.* considered in [15] the use of dynamic features, while static cues such as color and textures are not taken into account. More specifically, the approach is based on motion features extraction, spatio-temporal filtering and rare motion extraction. The RARE saliency model is introduced in [16]. The idea is to find areas in the frame exhibiting features, which are rare and contrasted with the others, and to assign them a higher saliency value. However, in [16], the RARE algorithm is implemented using only static features, such as colors and orientations, but ignoring dynamic features.

The proposed method is built upon [15],[16]. As both temporal and spatial features are important, we propose here to integrate dynamic features to the static model presented in [16]. This new model is referred to as Spatio-Temporal RARE (ST-RARE). More precisely, motion amplitude and direction, which can efficiently represent temporal information, are added to color and orientation to have a more accurate saliency map and better temporal robustness. These information are also used to perform tracking and temporal filtering of the saliency maps.

The paper is structured as follows. In Sec. 2, the proposed ST-RARE model is described in detail. Sec. 3 provides an evaluation of the proposed model on a wide variety of videos against eye-tracking data. Finally, Sec. 4 includes a discussion and conclusion.

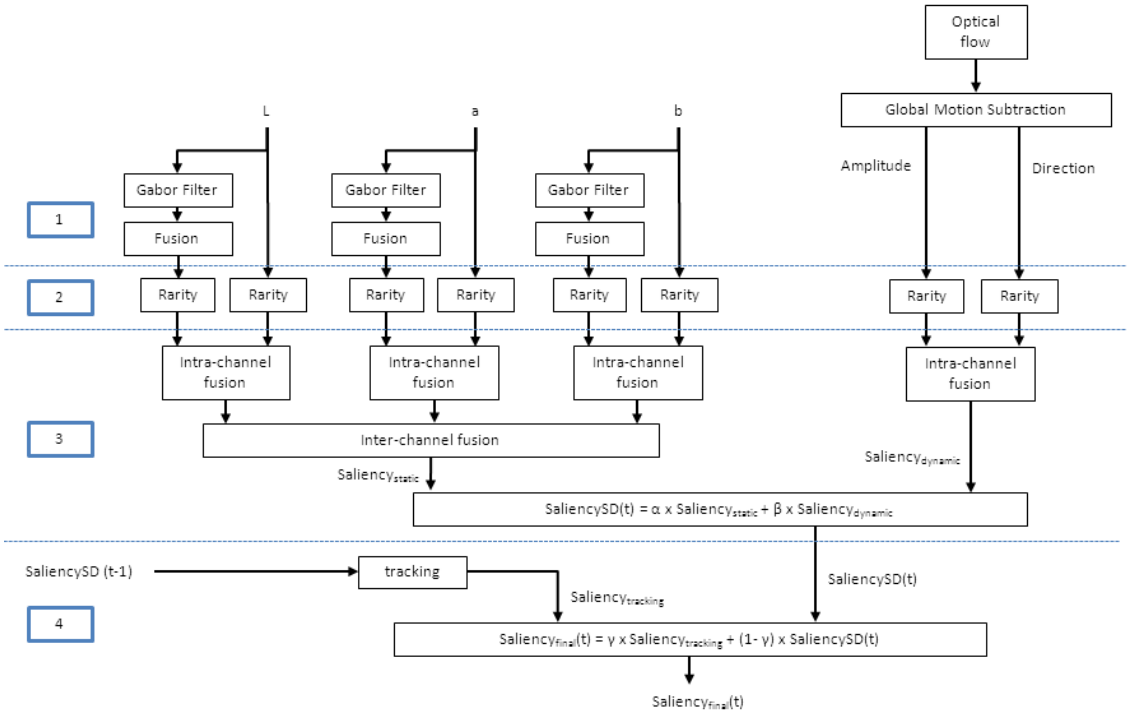

Fig.1. *Overview of the ST-RARE saliency model. From top to down: (1) feature extraction, (2) multi-scale rarity mechanism, (3) fusion steps, and (4) tracking and temporal filtering (the static features are on the left while the dynamic features are on the right).*

2. SPATIO TEMPORAL RARE MODEL

Fig.1 represents the proposed ST-RARE model. The proposed ST-RARE model combines spatial ($Saliency_{Static}$) and temporal ($Saliency_{dynamic}$) information to provide the map Saliency SD. The spatial module is similar to RARE [16] and the temporal module is one of the contributions of this paper. To have a better temporal robustness, tracking is used to combine SaliencySD at time *t*-1 and *t* to generate the final saliency map.

2.1. Feature extraction

Spatial features are computed in the CIE Lab colour space. This perceptually motivated colour space has the advantage to better decorrelate colour information. At this stage, a first pathway directly uses the CIE Lab colour transformation and computes the areas in the image containing the rarest colour. In parallel, a second pathway extracts orientation features by using Gabor filters with 8 orientations and 3 scales. The decomposition at several scales is recombined in a single map for each orientation. The orientation maps are also fused together into a single map [16]. These two operations are denoted as fusion, just after the "Gabor filter" module in Fig.1.

Temporal features are computed using the optical flow from [17] extracted from the luminance component. As the optical flow is computed pairwise, temporal coherence is not guaranteed. To reduce noise, a mean filter is temporally applied on both the horizontal and vertical directions of the optical flow. In the case of a moving camera, the background has a global motion, whereas other objects follow their own local motion. In order to better identify the salient moving objects, global motion is computed (as the average horizontal and vertical movements) and subtracted from the motion intensity obtained by optical flow. This pre-processing is illustrated in Fig. 1. From the local motion two basic temporal features are extracted: the motion amplitude *A* and direction *D*, defined as:

$$A = \sqrt{\Delta x^2 + \Delta y^2}$$

$$D = arctan2\ \Delta y, \Delta x$$

where Δx and Δy are the vector components obtained by the optical flow. These two temporal features are denoted in Fig. 1 as "Amplitude" and "Direction".

In summary, we have six spatial feature maps: three low-level (which are the colours from the first path) and three medium-level (the orientation and texture information coming from the Gabor filters) and two temporal features maps: motion amplitude and direction.

2.2. Multi-scale rarity mechanism

This mechanism is the one used in [16]. A feature is not necessary salient alone, but only in a specific context. The

mechanism of multi-scale rarity allows detecting both locally contrasted and globally rare regions in the image.

First, for each feature map, a Gaussian Pyramid decomposition is built at four different scales. For each scale (for pixel neighbourhoods with increasing sizes), the occurrence probability p of the pixels is computed using histograms. Then, saliency is obtained by computing $-\log(p)$ where p is the occurrence probability of a feature map at a given scale. The $-\log(p)$ increases the saliency inside the rare regions for each feature,. Saliency will be higher for rare regions in the frame.

Finally, the rarity maps of each scale are summed up and normalized to obtain a multi-scale contrast and rarity map per feature. As the input of this stage is a set of eight features, the output will consist in a set of eight rarity maps.

2.3. Spatial and temporal combination

The fusion is illustrated in the third part of Fig. 1. It is achieved in two main steps for the static pathway: an intra-channel fusion followed by an inter-channel one. Indeed, first an intra-channel fusion is computed between colour and orientation rarity maps by providing a higher weight to the maps which have important peaks compared to their mean [18]. This process is named efficiency. This leads to 3 final maps, one per colour channel. Second, an inter-channel fusion between these three maps uses the same principle of efficiency for computing weights for each map.

As the temporal aspect in the saliency maps is one of our contributions, we propose for the dynamic pathway to do an intra-channel fusion between the amplitude and direction feature conspicuity maps.

Next, a linear combination is applied between static and dynamic maps to obtain the following map:

$$\text{SaliencySD}\ t = \alpha\ \text{Saliency}_{\text{static}} + \beta\ \text{Saliency}_{\text{dynamic}}$$

where

$$\alpha = \left(max_{static} - mean_{static}\right)^2$$

$$\beta = \left(max_{dynamic} - mean_{dynamic}\right)^2$$

Where max_{static} is the maximum of the saliency static and it is the same principle for the other parameters. Generally, if a frame contains slow motion, Saliency_{static} will have a higher weight. Conversely, in the presence of fast motion, $\text{Saliency}_{dynamic}$ will become dominant.

2.4. Temporal tracking

The last step is the temporal tracking framework in order to improve temporal coherence and robustness, as shown in Fig. 1. A prediction of the saliency at time t, $\text{Saliency}_{\text{tracking}}$, is obtained by motion compensation of the saliency at time t-1. The final saliency map is then obtained by a linear combination of the SaliencySD at time t and $\text{Saliency}_{\text{tracking}}$,

$$\text{Saliency}_{\text{final}}\ t = \gamma\ \text{Saliency}_{\text{tracking}} + \left(1-\gamma\right)\ \text{SaliencySD}(t)$$

with a weighting factor γ empirically set to *0.3*. This approach provides a higher saliency value in temporally consistent regions and filters out noisy estimates, improving overall robustness.

3. PERFORMANCE EVALUATION

3.1. Dataset

The ASCMN (Abnormal, Surveillance, Crowd, Moving, and Noise) video benchmark [19] is used for evaluation. It is composed of 24 videos separated into 5 categories: Abnormal with surprising motion objects, Surveillance with normal motion objects, Crowd with several crowd densities, Moving with moving camera, and Noise with long period of noise and sudden salient object. Ground truth has been computed for ASCMN with eye tracking data from 13 viewers, acquired using a commercial FaceLab eye tracking system [20]. This system allows small head movements and is thus less intrusive than other eye tracking systems, making the viewer feel more comfortable. The viewers are PhD students and researchers ranging from 23 to 35 years old, both males and females. The eye gaze positions are recorded and superimposed on the initial video for all the viewers, as shown in the first column of Fig. 2.

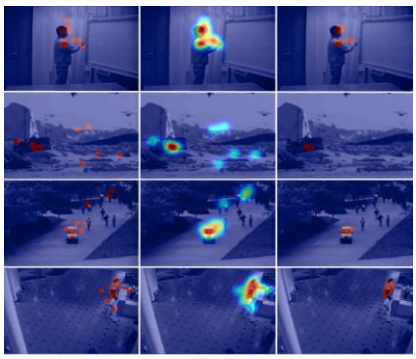

Fig. 2 *Eye tracking results. First column: gaze positions, Second column: heatmap, Third column: thresholded heatmap*

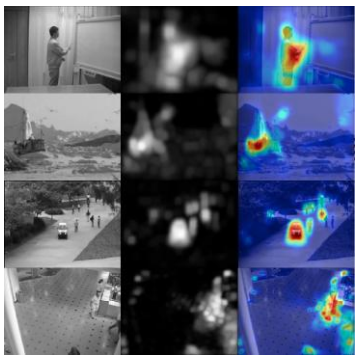

Fig. 3 *Visual results. First column: Original image, Second column: Saliency map (high saliency=high intensity), Third column: Heatmaps (superposition of the original image with the saliency map. high saliency=red).*

A Gaussian convolution is applied on subjects' gaze positions to obtain a "heatmap" which can also be superimposed on corresponding video frame (Fig. 2, second column). This post-processing step is useful in estimating the mean gaze density, eliminating the outliers and giving more importance to the focus points common to several users. Finally, depending on the metric used to assess the correspondence between the eye tracking results and an automatic saliency model, a thresholded version of the heatmap can be obtained (Fig. 2, third column).

3.2. Metrics

To compare the results of our approach with different models, three different metrics are used. The Area Under the ROC curves (AUROC) [21] focuses on saliency location at gaze positions. The Normalized Scanpath Saliency (NSS) [22] focuses on saliency values at gaze positions. KL-Divergence [23] focuses on the discrepancy of saliency and gaze distributions. For AUROC and NSS, high scores indicate better performance. Conversely, low scores are better for KL-Divergence.

3.3. Experimental results

Fig. 3 presents experimental results for the same frames as in Fig. 2. On the first row, the man is salient and the object is well detected, although we observe a small position shift when compared to the eye tracking reference (Fig. 2). On the second row, the boat is well detected and our approach is not disturbed by the birds. Compared to the eye tracking data, our approach detects a bigger salient region which corresponds to the whole object. On the third row, people are well detected and fast objects are more salient, as in the eye tracking reference. On the fourth row, the main moving man is well detected but our approach gives additional salient regions corresponding to other moving people. Globally, we can see in these examples that our approach detects well salient regions.

In Fig. 4, the sequences are evaluated with the metric proposed in Sec. 3.2. We can see that the proposed ST-RARE has globally better results than the other algorithms with the AUROC and the NSS metric. With the KL-Divergence, Seo [12] and Culibrk [13], obtain a better score due to a more precise distribution, but ST-RARE still reaches good performances. We can also observe that the proposed ST-RARE always outperforms the previous dynamic-only or static-only versions of RARE in [15] and [8] respectively.

The proposed model clearly adapts efficiently to very different types of video with fast and slow motion. It reaches good scores for all the categories except for crowd. In this case, the ST-RARE algorithm detects the rarity of multiple local motions and gives importance to these parts.

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new spatio-temporal saliency algorithm. It builds upon the RARE algorithm [8] by adding dynamic features. More specifically, a fusion algorithm takes into account both spatial and temporal maps. It also integrates a tracking module to improve accuracy and robustness. Experimental results show the relative efficiency of the proposed saliency approach when compared to five state-of-the-art models.

The model adapts efficiently to various classes of video containing very different types of motion. It is possible to add other modules to our flexible framework in order to improve the motion model.

5. ACKNOWLEDGEMENTS

N. Riche is supported by the Fonds pour la formation a la Recherche dans Industrie et dans Agriculture (FRIA).

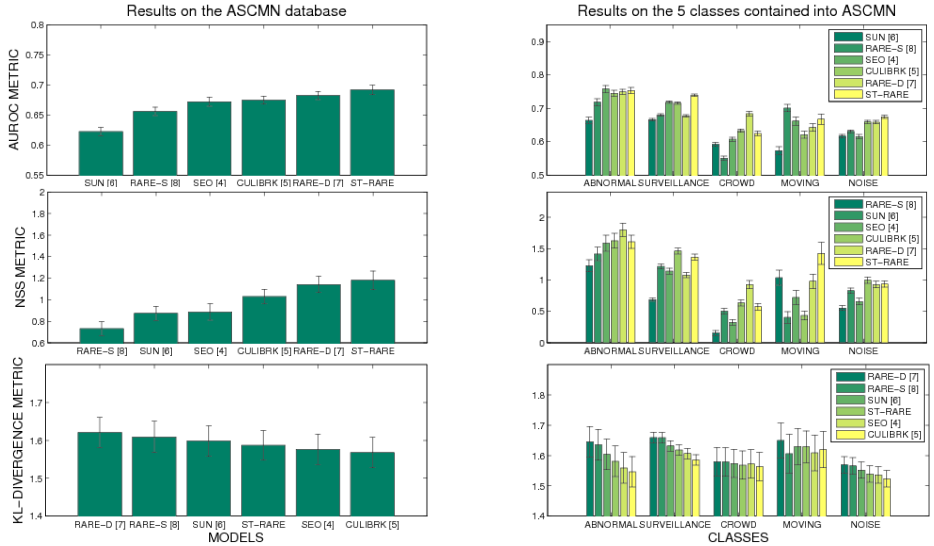

Fig. 4 *Evaluation of ST RARE using AUROC (first row), NSS (second row) and KL-Divergence (third row) metrics.*

7. REFERENCES

[1] C. Koch, S. Ullman, “Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry,” Human Neurobiology, vol. 4, pp. 219-227, 1985.

[2] J.K. Tsotsos, S.M. Culhane, W.Y.K. Wai, Y.H. Lai, N. Davis, F. Nuflo, “Modelling Visual Attention via Selective Tuning,” Artificial Intelligence, vol. 78, no. 1-2, pp. 507-545, Oct. 1995.

[3] S. Lu, J.H. Lim “Saliency Modeling from Image Histograms”, European Conference on Computer Vision (ECCV), pp. 321-332, Florence, Italy, 2012:

[4] M. Décombas, F. Dufaux, E. Renan, B. Pesquet-Popescu, F. Capman, “Improved seam carving for semantic video coding,”, in Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP 2012), Banff, Canada, Sept. 2012

[5] M. Rubinstein, A. Shamir, S. Avidan “Improved seam carving for video retargeting,” ACM Trans. Graphics, vol. 27, no. 3 , pp. 1-16, 2008

[6] Z. Li, P. Ishwar, J. Konrad, “Video condensation by ribbon carving,” IEEE Trans. Image Processing, vol. 18, no. 11, pp. 2572-2583, Nov. 2009

[7] L. Itti, P. Baldi, "A principled approach to detecting surprising events in video", in Proc. IEEE Computer Vision and Pattern Recognition (CVPR), San Diego, CA, June 2005

[8] C. L. Guo , Q. Ma, L. M. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform", in Proc. IEEE Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, June 2008

[9] D. Gao , V. Mahadevan, N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency", Neural Information Processing Systems (NIPS), Dec. 2007

[10] E. Rahtu, J. Kannala, M. Salo and J. Heikkilä, "Segmenting Salient Objects from Images and Videos", European Conference on Computer Vision (ECCV), Heraklion, Greece, Sept. 2010

[11] V. Mahadevan and N. Vasconcelos, "Spatiotemporal Saliency in Dynamic Scenes," IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) , vol.32, no.1, pp.171-177, Jan. 2010

[12] H. Seo, P. Milanfar, “ Static and space time visual saliency detection by self-resemblance,” Journal of vision, vol. 9, no. 12, pp. 1-12, 2009

[13] D. Culibrk, M. Mirkovic, V. Zlokolica, P. Pokric, V. Crnojevic, D. Kukolj, “Salient motion features for video quality assessment,” in Proc. IEEE International Conference on Image Processing (ICIP), Hong Kong, Hong Kong, Oct. 2010

[14] L. Zhang, M. Tong, T. Marks, H. Shan, G. Cottrell, “SUN : A Bayesian framework for saliency using natural statistics,” Journal of vision, vol. 9, no. 7, pp. 1-20, 2008

[15] M. Mancas, N. Riche, J. Leroy, B. Gosselin, “ Abnormal motion selection in crowds using bottom-up saliency,” in Proc. IEEE International Conference on Image Processing (ICIP), Bruxelles, Belgium, 2011

[16] N. Riche, M. Mancas, B. Gosselin, T. Dutoit, “ Rare: A new bottom-up saliency model” in Proc. IEEE International Conference on Image Processing (ICIP), Orlando, FL, Oct. 2012

[17] A. Chambolle, T. Pock, “A first-order primal-dual algorithm for convex problems with application to imaging,” Technical Report, 2010

[18] L. Itti, C. Koch, “Comparaison of feature combination strategies for saliency-based visual attention systems,” in Proc. SPIE, Human Vision and Electronic Imaging (HVEI), San Jose, CA, 1999

[19] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, T. Dutoit, “Dynamic saliency models and human vision: a comparative study on videos,” in Proc. Asian Conference on Computer Vision (ACCV), Daejon, South Korea, 2012

[20] Seeing Machines: Facelab commercial eye tracking system, http://www.seeingmachines.com/product/facelab/

[21] B. Lau, B.: Evaluation measures for saliency maps: AUROC,http://www.subcortex.net/research/code/area_under_roc_curve

[22] A. Borji, A.: Evaluation measures for saliency maps: CC_and_NSS,https://sites.google.com/site/saliencyevaluation/evaluation-measures

[23] M. Mancas, N. Riche, Computational attention website http://tcts.fpms.ac.be/attention