VIDEO SALIENCY BASED ON RARITY PREDICTION: HYPERAPTOR

Ioannis Cassagne [a], Nicolas Riche [a], Marc Décombas [a],
Matei Mancas [a], Bernard.Gosselin [a], Thierry Dutoit [a], Robert Laganiere [b]

[a]University of Mons (UMONS) – Faculty of Engineering (FPMs) , Mons, Belgique
{Nicolas.Riche, Matei.Mancas, Bernard.Gosselin, Thierry.Dutoit}@umons.ac.be
{marc.decombas, ioannis.cassagne}@gmail.com
[b]University of Ottawa (UOTTAWA) –Ottawa, Canada
laganier@uottawa.ca

ABSTRACT

Saliency models are able to provide heatmaps highlighting areas in images which attract human gaze. Most of them are designed for still images but an increasing trend goes towards an extension to videos by adding dynamic features to the models. Nevertheless, only few are specifically designed to manage the temporal aspect.

We propose a new model which quantifies the rarity natively in a spatiotemporal way. Based on a sliding temporal window, static and dynamic features are summarized by a time evolving "surface" of different features statistics, that we call the "hyperhistogram". The rarity-maps obtained for each feature are combined with the result of a superpixel algorithm to have a more object-based orientation. The proposed model, Hyperaptor stands for hyperhistogram-based rarity prediction. The model is evaluated on a dataset of 12 videos with 2 different references along 3 different metrics. It is shown to achieve better performance compared to state-of-the-art models.

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

1. INTRODUCTION

The visual saliency models aim to automatically predict human attention. In [1][3], the human attention has been introduced and can be defined as the process that allows one to focus on some important stimuli at the expense of others. Two main processes can be defined in human attention called bottom-up and top-down. The most salient objects are found using features extracted from the signal with the bottom-up approach while the top-down attention uses a priori task-oriented or scene knowledge to modify the bottom-up saliency. Even if at a first glance there are lots of attention models, the philosophy behind is the same: identify unusual features in a given spatio-temporal context by searching rare, novel or surprising information. Attention models application are very numerous. Among the existing applications, one can find gaze prediction [3], content aware compression [4], video retargeting [5] or video summary [6].

Itti et al. [7] proposed a static model based on three features: color, luminance and orientation. Harel et al. [8] improve this model to create features maps at multiple spatial scales and propose a Graph-Based Visual Saliency model (GBVS). This approach builds a fully connected graph over all grid locations of each feature map. Weights are assigned between nodes that are inversely proportional to the similarity of feature values and their spatial distance. In [9], Marat et al. propose a model that is inspired by the biology of the visual system, and breaks down each frame of a video intro three maps: 1) a static saliency map emphasizes regions that differ from their context, 2) a dynamic saliency map emphasizes moving regions and 3) a face saliency map emphasizes areas where faces are detected. Finally, they fuse all these maps into a master saliency map.

The saliency model of Rahtu et al. [10] has the advantage to be multi-scale, does not require training and is computed in the CIE Lab perceptual color space. To take into account the movement in the scene, motion intensity is added as an input feature.

Build upon [11][12], a Spatio-Temporal saliency model based on Rarity (ST-RARE) has been proposed in [13] and integrate dynamic features like motion amplitude and direction. A temporal filtering is also used to be more robust in the time.

In this paper, we propose a new hyperhistogram-based rarity prediction model called Hyperaptor. The contributions of this paper are 1) a new way to extract features with more temporal information, 2) a new process to select important features based on a surface of rarity, 3) a final map enhancement using a SLIC algorithm [14], a center Gaussian and a tracker. These contributions lead to a better model in eye gaze and salient objects prediction. It is more stable through time and more object-oriented.

The paper is structured as follows. In Sec. 2, Hyperaptor is described in detail. Sec. 3 provides an evaluation of the proposed model on a wide variety of videos against eye-

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

tracking data and manually objects segmentation. Finally, Sec. 4 presents a discussion and conclusion.

2. HYPERAPTOR MODEL

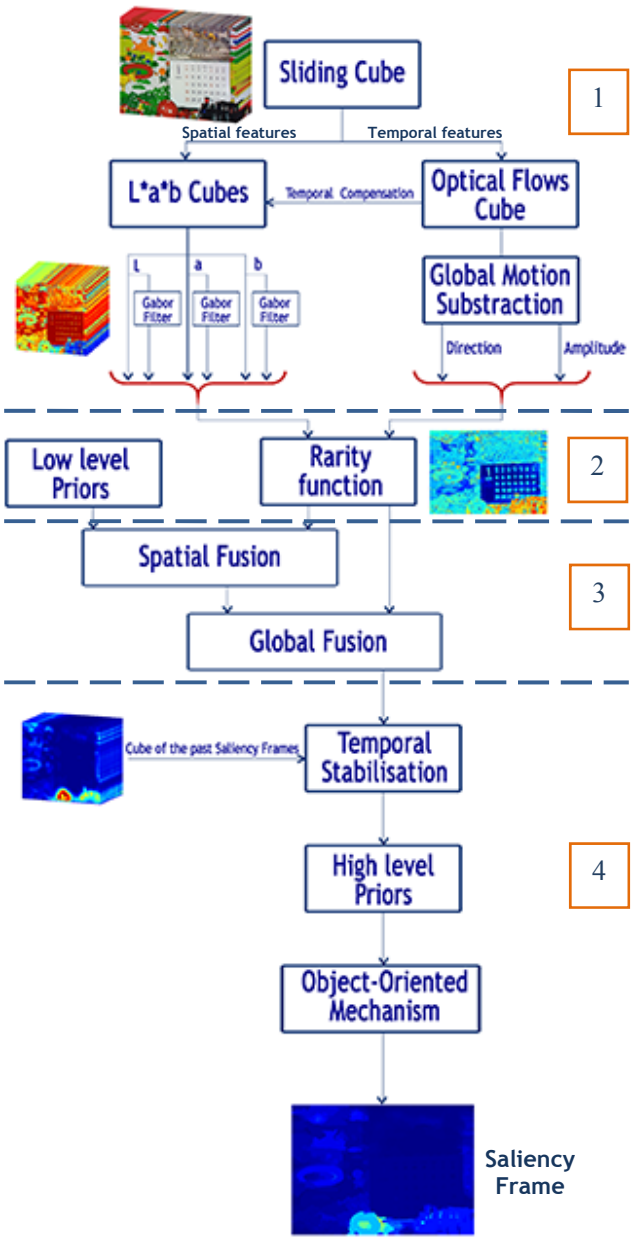

Figure 1 *Overview of Hyperaptor. From top to down: (1) features extraction on a sliding video cube, (2) low level priors and multi-scale rarity mechanism applied on features maps cube, (3) fusion steps, and (4) post-processing (temporal stabilization, high level priors and object oriented mechanism).*

Figure 1 represents the overall schema of Hyperaptor. A sliding "cube" (2D + *t*) is used to extract both spatial and temporal features in the video. After pre-processing of those features, their rarity based on the hyperhistograms is computed. These hyperhistograms are a temporal concatenation of all the histograms of features which are extracted in each frames in the cube. Low-level priors as specific behavior on colors are also added. A fusion of the different rarity feature maps is achieved and stabilized temporally using attention history. Finally, high level features (like face detection) and a superpixel algorithm are added to the model to provide an object-based approach. In the following subsections, we detail each of the algorithm steps.

2.1. Feature extraction

A video cube (x, y, t) is built with a temporal sliding window, that has been empirically fixed at 15 frames, to have static and dynamic information from the current frame but also the previous ones. Six static features are extracted from this video; three color feature cubes (luminance and two chrominances) are defined in the CIE Lab color space and eight orientation maps realized with a Gabor filter and combined together at three different scales allowing to have three texture feature cubes at three different scales.

The optical flow from [15] computed on the luminance component only is used to create two dynamic feature cubes (one for motion amplitude and the other one for motion direction). To manage the camera motion, a global motion subtraction, which is based on the mean value, is used. Two temporal features are extracted from the optical flow: the motion amplitude A and direction D, defined as:

$$A = \sqrt{\Delta x^2 + \Delta y^2}$$
$$D = arctan2(\Delta y, \Delta x)$$

Where $\Delta x$and Δy are the vector components obtained by the optical flow. These features are put together to build two cubes of dynamic features.

2.2. Low-level priors and multi-scale rarity mechanism

Two low-level prior maps are computed for spatial features maps from [16]: 1) the first is related to frequency. Indeed, the human behavior can be modeled by band-pass filtering. 2) The second is about colors. Some studies [16] find that warm colors, such as red and yellow, are more pronounced to the human visual system than cold colors.

A rarity mechanism is then applied on each feature map (temporal and spatial). The primary idea comes from [12][17] and is based on the fact that a feature is not necessary salient alone, but only in a specific context. Here it was extended to have a real temporal behavior. Indeed, the mono-dimensional feature histogram used on a frame becomes a hyperhistogram which is a 2D surface (Figure 2(b)).

The rarity mechanism is illustrated in Figure 2 on the luminance component in 3 steps: a) a Gaussian pyramid decomposition provides features maps cube at different scales, b) for each cube, a histogram surface (hyperhistogram) is processed, c) the self-information is computed on the entire hyperhistogram, but only the current frame is extracted.

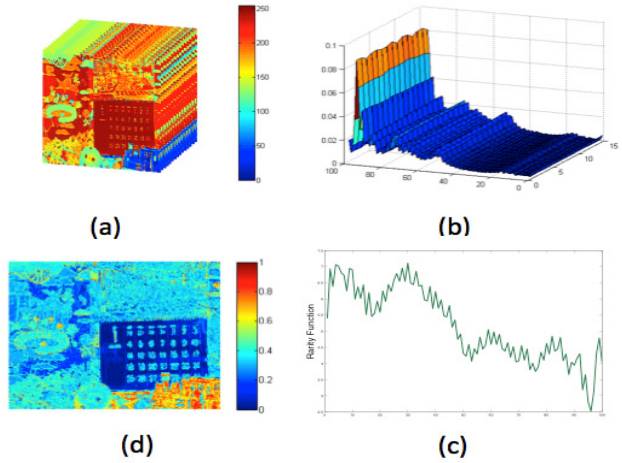

Figure 2 *Illustration of the rarity mechanism on a single scale of the luminance features maps cube (a). A rarity function (c) at the time t is computed from a histogram surface (b) to obtain a rarity luminance map (d).*

This mechanism provides higher scores for locally (in space and time) contrasted and globally (in space and time) rare regions.

2.3. Fusion

The fusion process has two main steps: 1) the spatial features mas are combined with the low-level priors map with a *max fusion*. The maximum value between the two maps is taken for each pixel. 2) These rarity spatial maps are then combined with the temporal features maps. Based on [18], the maps which have important peaks compared to their mean have a higher weight. A single saliency map is finally obtained.

2.4. Post-processing and enhancement

The saliency map obtained in the previous section is still enhanced using three different techniques.
Firstly, there is a post-processing step which performs temporal stabilization based on a mean of a short history of saliency map of previous frames.
Secondly, high-level priors are added. Previous studies [20] have shown that salient information is mainly located in the center of images for natural images. To model this prior, a centered Gaussian is used.
Finally, a SLIC algorithm [14] is used with DBSCAN [23] to extract superpixels in the frame. Those superpixels are groups of pixels with similar color levels. They extract shape information from the objects in the frame. The saliency map is averaged for each superpixel. In that way, the final map will be more object-oriented approach.

3. PERFORMANCE EVALUATION

3.1. Dataset and metrics

The STRAP video benchmark is based on [16] which provide 12 raw videos with eye tracking data and three manually segmented masks and nine other manually segmented binary masks have been made to cover the whole set of 12 videos. This new database is available on [20]. Figure 3 shows three different video sequences examples extracted from the database with the original frame on the left column, the manually segmented mask ground truth on the middle column and a heatmap of the eye tracking ground truth on the right column.

Figure 3 *Extract from the database with two references. First line: Hall, second line: Mobile, Third Line: Soccer. First column: Original videos, Second column: Manual binary map, Third column: Heatmap of the eye tracking ground truth*

To compare the results of Hyperaptor with different other video saliency models, three different metrics are used. Based on the eye tracking data, the Area Under the ROC curve (AUROC) [21] focuses on saliency location at gaze positions. The Normalized Scanpath Saliency (NSS) [22] focuses on saliency values at gaze positions. For AUROC and NSS, high scores indicate better performance on the eye-tracking ground truth. For the manually segmented objects ground truth, the F-measure metric is used. This metric is based on true positives (tp), true negatives (tn), false positives (fp) and false negatives (fn) that compare the predicted results with the reference results. It is defined as a combination of the Precision and the Recall where Precision is the number of relevant points compared with the total number of points found and Recall is the number of relevant points compared

with the total number of important points in the reference. This metric shows the capacity of the approach to predict the salient object and not only the eye gaze.

3.3. Experimental results

To validate Hyperaptor, qualitative and quantitative experimentation has been done. Figure 4 shows the qualitative results of three different video attention models as heatmaps superimposed on the current frame. Blue means areas which are not important while red, the regions of predicted interest. On the first column of Figure 4, we can see that our approach defines well the salient objects. For STRARE (middle column), the salient objects are well identified for *Hall* and *Soccer* video sequences (video frames in the 1st and 3rd lines) but the model highlights also part of the background which is not what it should be. For the frame extracted from the *Mobile* video sequence (the middle line), the ball is not detected as salient. For GBVS (3rd column), the approach works well only for *Hall* video sequence.
To compare the heatmaps with the ground truth, please refer to Figure 3.

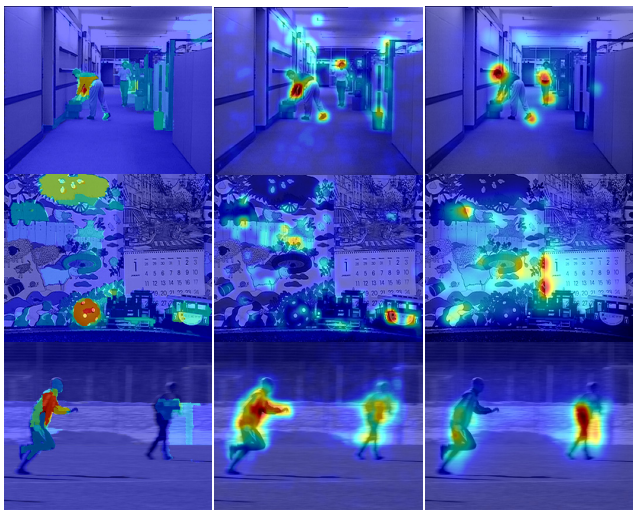

Figure 4 *Visual results as Heatmaps (superimposition of the original frame with the saliency map. High saliency=red). Columns from left to right; HYPERAPTOR– STRARE – GBVS.*

For the quantitative validation in Figure 5, the three previously described metrics are used to compare 4 state-of-the-art saliency algorithms and a constant centered Gaussian. It can be seen that with the AUROC metric, our model is third. This is due to the fact that this metric is strongly influenced by the centered Gaussian which is found in the Gaussian model and in GBVS.
The NSS metric is complementary to the AUROC. It can be seen that following this metric, our approach is statistically better than the state-of-the-art.
When we compare Hyperaptor with the other models on the manually segmented objects ground truth using the F-measure, Hyperaptor statistically overpasses all the other methods that , we remain, are not natively object-oriented.
Figure 5 shows that Hyperaptor is always better on than the state-of-the-art methods on two of the metrics (NSS and F-measure). The AUROC metric is very sensitive here to the centered Gaussian.

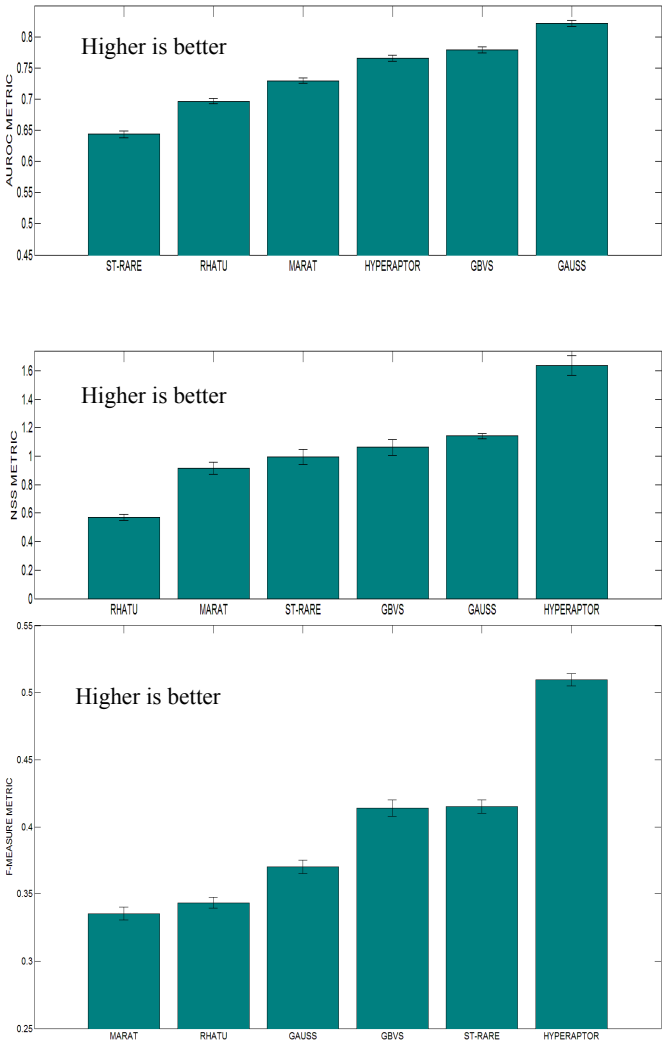

Figure 5 *Evaluation of Hyperaptor using AUROC (first row), NSS (second row) and F-Measure (third row) metrics compared with 4 different saliency models and a Gaussian*

4. CONCLUSION

In this paper, we propose a new video saliency approach which uses 2D histogram surfaces while computing the rarity. Both static and dynamic features are taken into account. Low and high level information are added along with a superpixels-based pre-segmentation. This novel approach is evaluated on 12 videos with both eye-tracking and manually segmented objects ground truth and three different

comparison metrics. The videos are very different in terms of content and motion (including cluttered background, moving background, moving camera, etc...). The 12 videos dataset is based on an existing work which is complemented with the unfinished manually segmented objects maps.

If we except the AUROC metric where Hyperaptor is 3rd, on NSS and F-measure, our model is way better than the competing approaches. The use of the superpixels in the proposed model lead to the fact that Hyperaptor is also very good for the detecting salient objects and not only in predicting eye gaze.

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, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. on pattern analysis and machine intelligence, vol. 20, no. 11, pp.1254-1259, 1998

[8] J. Harel, C. Koch, P. Perona, "Graph-based visual saliency," In Advances in neural information processing system, pp. 545-552, 2006

[9] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, A. Guérin-Dugué , "Modeling spatio-temporal saliency to predict gaze direction for short videos," International journal of computer vision, vol. 82, no. 3, pp. 231-243, 2009

[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] 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

[12] 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

[13] M. Décombas, N. Riche, F. Dufaux, B. Pesquet-Popescu, M. Mancas, B. Gosselin, T. Dutoit, Spatio-temporal saliency based on rare model, IEEE International Conference on Image Processing (ICIP), 2013.

[14] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," IEEE Transactions on Pattern Analysis and Machine Intelligence, , vol.34, no.11,pp. 2274-2282, 2012

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

[16] L. Zhang, Z. Gu, H. Li, "SDSP: A novel saliency detection method by combining simple priors, ". IEEE Trans. Image Processing, pp.171-175, 2013.

[17] M. Mancas. "Relative influence of bottom-up and top-down attention," Attention in Cognitive Systems, LNCS, Volume 5395/2009:pp. 212–226, 2009

[18] L. Itti, C. Koch, "Comparison of feature combination strategies for saliency-based visual attention systems". In International Society for Optics and Photonics ; Electronic Imaging'99, pp. 473-482., 1999.

[19] H. Hadizadeh, M. J. Enriquez, and I. V. Bajić, "Eye-tracking database for a set of standard video sequences," IEEE Trans. Image Processing, vol. 21, no. 2, pp. 898-903, Feb. 2012

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

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

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

[23] M. Ester, H.P. Kriegel, J. Sander, J., X. Xu, , "A density-based algorithm for discovering clusters in large spatial databases with noise," In Kdd, vol. 96, no. 34, pp. 226-231, 1996.

[24] V. Rijsbergen, C. Keith Joost, Information retrieval Butterworths, London, 1979.