FUNNRAR: HYBRID RARITY/LEARNING VISUAL SALIENCY

P. Marighetto[1] *I. Hadj Abdelkader*[2] *S. Duzelier*[2] *M. Décombas*[4] *N. Riche*[1]
J. Jakubowicz[2] *M. Mancas*[1] *B. Gosselin*[1] *R. Laganière*[3]

[1] University of Mons - Mons, Belgium
[2] Telecom SudParis - Evry, France
[3] University of Ottawa - Ottawa, Canada
[4] Cooprev - Paris, France

ABSTRACT

Saliency models provide heatmaps highlighting the probability of each pixel to attract human gaze. To define image's important regions, features maps are extracted. The rarity, surprise or contrast are computed leading to conspicuity maps, showing important regions of each feature map. The final saliency map is obtained by merging these maps. The fusion process is usually a linear combination of the maps where the coefficients show their importance. We propose a novel generic fusion mechanism based on 1) using a rarity-based attention module and 2) using neural networks to achieve the fusion. The first layer of the NN merges the weighted feature maps into a saliency map. The second layer takes into account the spatial information. The approach is compared to 8 models using 4 different comparison metrics on open state-of-the-art databases.

***Index Terms*—** Visual attention, Saliency, Neural Network, Learning based saliency

1. INTRODUCTION

Visual saliency models aim to automatically predict human attention. Attention models application are very numerous. Among the existing applications, one can find gaze prediction [1], content aware compression [2], video retargeting [3] or video summarization [4].

In [1, 5, 6], 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. Attention is a competition between two components, one called top-down and the other called bottom-up. Bottom-up attention uses features extracted from the image to predict the salient part of the visual signal while top-down attention uses a priori task-oriented or scene knowledge to modify the bottom-up saliency.

In this paper and in most of the existing visual attention models mainly the bottom-up component is taken into account. Even if there are lots of attention models, the philosophy behind those models is the same: identify unusual features in a given spatio-temporal context by searching contrasted, rare, novel or surprising information.

Itti et al. [7] proposed one of the first computational models for images based on three features: color, luminance and orientation. Bruce et al. [8] made a model based on information maximization. Other approaches tend to use graph theory [9], or rarity mechanisms [10].

However, recently, learning, and especially deep learning were successfully used to predict human attention. Zhao et al. [11] reviewed recent advances in learning saliency-based visual attention. The learning process is used in two ways for visual attention models.

The first idea is to use deep learning for the entire algorithm: the nets will extract features, select the most salient ones and output a saliency map. [12] and [13] studied the use of convolutional networks in saliency prediction. SALICON model [14] is also based on deep neural networks. These last models use neural networks directly on images to compute saliency maps. A problem of those models is their high dependency on the dataset. Using images of websites or advertisements on models trained on natural images lead to bad results.

A second idea is to use deep learning only to select the interesting features in the image and than using a more classical approach. Li et Yu [15] used CNNs to extract features on three different scales and then apply a refinement method to obtain their saliency map. For these models, they also depend a lot on the dataset as the features extracted will stick to the data.

In this paper, we propose to keep a classical rarity-based attention brick and use learning only for the last step of the algorithm: conspicuity maps fusion. We propose a model called FuNNRar which aims in taking advantage of the power of learning for conspicuity maps fusion, but which remains generic enough to be applied to any dataset given its classical rarity-based core using [10]. This combination of a classical saliency pipeline and a learning algorithm is, at our best knowledge novel.

The contributions of this paper are 1) a new way to combine 6 features' conspicuity maps, obtained using Rare2012 model [10] into a saliency map based on a Neural Network (NN) where each input node is a pixel followed by 2) a second NN that takes into account the spatial correlation between pixels. To validate this approach, we propose 3) a new open dataset based on CAT2000 [16] and SALICON [17] which will be freely available on [18].

The paper is structured as follow. In Sec. 2, FuNNRar is described in detail. Sec. 3 provides an evaluation of the proposed model on the open images dataset. Finally, Sec. 4 presents a discussion and concludes the paper.

2. FUNNRAR MODEL

2.1. Initial saliency model

We used Rare2012 saliency model [10] to compute the rarity conspicuity maps which will be merged. There are three main steps in its architecture. First, low-level color and medium-level orientation features are extracted. Afterwards, a multi-scale rarity mechanism is applied. Finally, rarity maps are merged into a single final saliency map.

Here, we propose to replace the final fusion step, based on [19], by a learning mechanism. At the end of step 2, six rarity conspicuity maps are returned: three maps on color features and three on orientation in each color feature. Thus, 6 maps are in input of the proposed fusion mechanism.

2.2. Proposed fusion process

The fusion model relies on two types of neural networks (NN). The first NN are trained on the rarity conscpicuity maps, pixel by pixel, in order to provide suitable weights to merge the maps. The following neural networks are trained on the merged map, by processing pixel patches, in order to get a final map that takes into account the high correlation between pixels.

For both learning and testing phases, we resized the images to 50x50 pixels. The proposed fusion consists in computing 2500 NN, one for each pixel, to get the merged map. The neurons of the input layer are the 6 values from the 6 conspicuity maps for the same locus in the original image. They are connected to a hidden layer of 10 neurons, and one output neuron labeled by the eye-tracking map.

For the training phase, we implemented a supervised learning based on the back-propagation algorithm. The training dataset is composed of 6x10,000 inputs corresponding to the 6 conspicuity maps of the 10,000 images of our database and the output dataset of 10,000 eye-tracking maps.

This step already provides interesting results but does not take into account the correlation within a group of adjacent pixels.

In a second step, we improved the model by implementing a more complex NN that does the fusion considering the spatial correlation. We divided our 50x50 images dataset into a MxM patches of pixels, where M is a divisor of 50. We computed $(50/M)^2$ NN, each of them composed of $6 \times (M^2)$ inputs corresponding to the 6 respective patches from the conspicuity maps, as the patches will not overlap. The hidden layer implements X neurons, where X is the number higher than the number of inputs and with the same tenth power (for example, with 150 inputs, X = 200) and the labels used during the training phase come from the eye-tracking data.

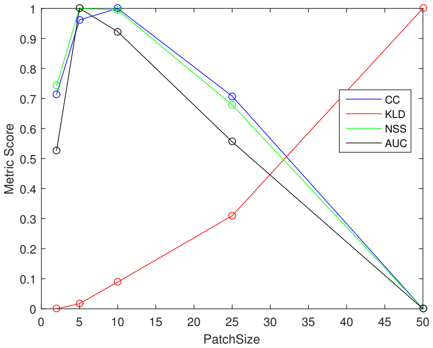

Fig. 1. Evaluation of different patches size.

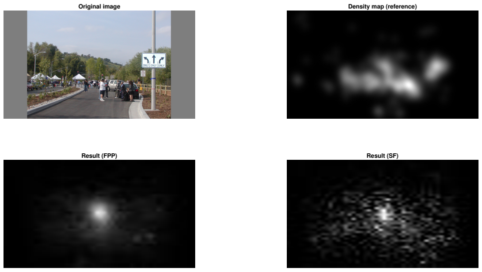

Fig. 2. Qualitative comparison of different fusion process. Top-left: initial image, top-right: density map (reference), bottom-left: FPP, bottom-right: SF.

We evaluated these 5 different patches size on the database and metrics that will be described in Sec. 3. Figure 1 shows the results. We can determine that the 5x5 patches have better results in general and will be used in the rest of the paper.

3. EVALUATION

3.1. Metrics

Based on [20], three metrics are sufficient to fairly evaluate saliency models, such as Kullback-Leibler Divergence, Normalized Scanpath Saliency and Area Under the ROC Curve. We will add to these metrics a fourth one, Pearson's Correlation Coefficient to confirm the results.

Kullback-Leibler Divergence (KLD) is a measure of the information lost when the saliency maps probability distribution is used to approximate the human eye fixation map probability distribution (density map). The lower the score, the better the saliency map approximates the density map.

Normalized Scanpath Saliency (NSS) quantifies the saliency map values at the eye fixation locations. Saliency map is normalized to zero mean and unit standard deviation. Larger score implies a better saliency map.

Area Under the ROC Curve (AUC) first selects all the fixated locations as positives and considers all the other locations as negatives. Multiple thresholds are then applied to the saliency map, and the numbers of true positives, and false positives are computed at each threshold. Finally, the ROC curve can be plotted according to the true positive rate and false positive rate at each threshold. Perfect saliency map leads to a score of 1, while random prediction has a score of 0.5.

Pearson's Correlation Coefficient (CC) describes the linear relationship between two variables. Its output range is [-1, 1]. Higher absolute score indicates higher similarity between saliency map and density map.

3.2. Dataset

To test our model, we merged two existing databases to create a new larger dataset for machine learning applied to saliency which is freely available [18]. The database has 3 different sections: 1) training is made of training data from SALICON [17] database's training part, composed of 10000 images, 2) validation is made of CAT2000 [16] database's training part, composed of 2000 images and 3) testing part is made of SALICON database's validation part, composed of 5000 images.

3.3. Results

From a qualitative point of view, Fig. 2 shows that compared to the eye-tracking density map reference, the pixel by pixel fusion process (FPP) saliency map is very focused on the center of the image. The spatial correlation (SF) saliency maps is a little wider, but it is still centered. These results show that the center bias is directly learned from the training set. At a first glance, a smoothed version of SF model seems closer to the reference.

We proceed to evaluate the impact of an averaging filter on SF, with square filters which size is between 2x2 and 20x20. Figure 3 shows results. We can see that in general, an averaging filter of size 5x5 improves the results.

Indeed, in the above section, two different comparisons are made. First, we compare the saliency maps obtained with the new fusion mechanisms to the former version of the fusion implemented in the classical algorithm [21]. In a second step, we compare the new model created with this mechanism to state-of-the-art models described in previous section.

The first comparison is held between FPP, SF, its filtered version (SFF) and the 'classic' mechanism (CF), which is our reference here. Figure 4 shows the average results on the training set.

We can see that all new fusion process have better scores on all the considered metrics compared to the classical fusion process, both increasing Rare2012 performance. FPP and SFF holds the best results on this database, which are quite similar. Thus, we selected the SFF method as the new fusion process used in FuNNRar model.

Then, 8 state-of-the-art models and FuNNRar model will be compared. The models used are Itti's model [7], AIM [8], GBVS [9], BMS [22], AWS [23], WMAP [24, 25], Context-aware saliency (CA) [26] and Rare2012 [21]. Figure 5 and figure 6 show the results.

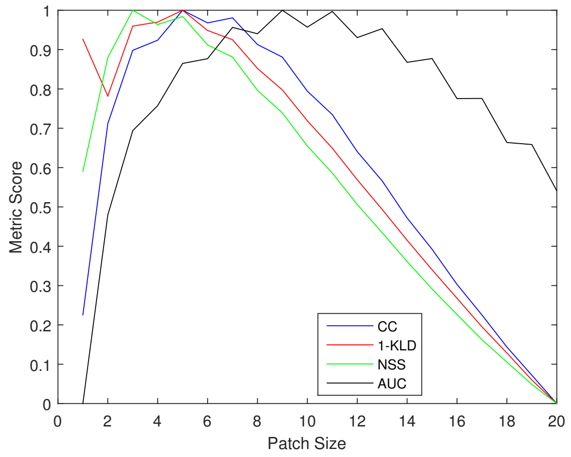

Fig. 3. Impact of an average filter on SF. Scores have been set between 0 and 1 for visualization purpose.

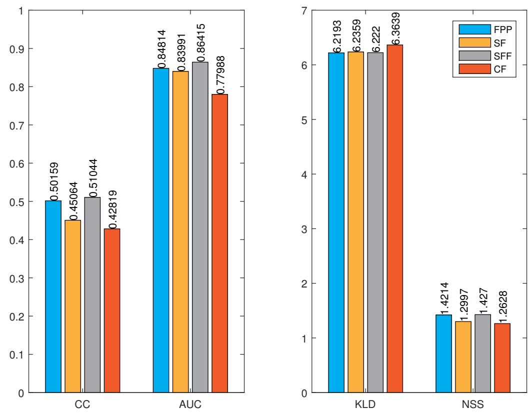

Fig. 4. Comparison of different fusion processes. From left to right: Fusion Pixel by Pixel (FPP, in blue) - Spatial Fusion (SF, in yellow) - Spatial Fusion Filtered (SFF, in gray) - Classic Fusion (CF, in red)

FuNNRar scores are better compared to all the other models on every metric. We can notice that GBVS equals FuN-

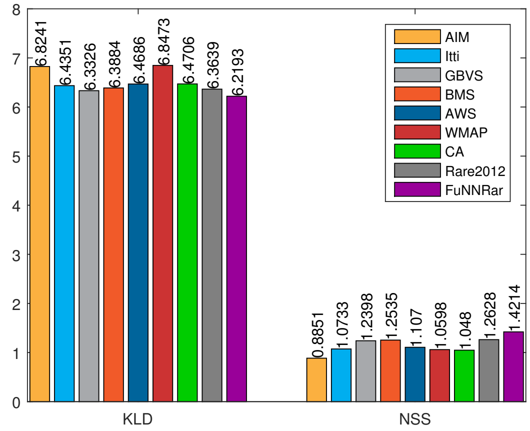

Fig. 5. KL-Divergence (left part) and NSS (right part) scores on saliency models.

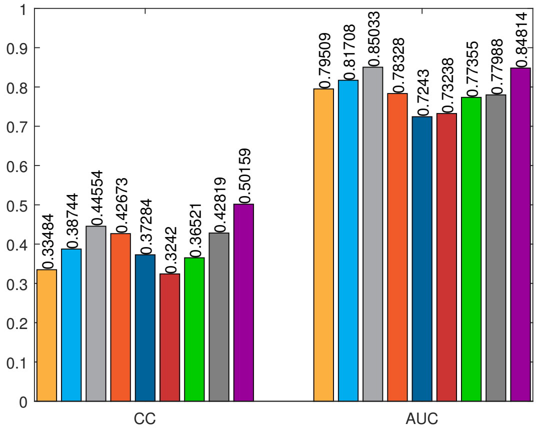

Fig. 6. CC (left part) and AUC (right part) scores on saliency models.

NRar on AUC score. This results show this new fusion process improve Rare2012 results to outperform state-of-the-art models.

4. CONCLUSION

This paper presents a new model to predict eye fixation based on the learning mechanism for merging the features' conspicuity maps and demonstrates its efficiency in saliency computation. The originality is to keep a classical structure and mix it with learning only for the last step of maps fusion.

FuNNRar is the first model to use a neural network that learns the merging weights assigned to each conspicuity map. With a few layers, the Neural Networks can already perform competitively on an open dataset based on CAT2000 and SALICON. It is by the way interesting to see that FPP which provides much more centered saliency maps than SF have better results, while SF seems more relevant from a qualitative point of view. This fact shows the huge importance of the centered bias in the database and the metrics usually used for saliency maps assessment.

As a future work, we plan to improve the model by training a sparse deep neural network that will merge the 6 maps and take into account the spatial correlation between pixel at the same time, what we have done separately with two distinct neural networks in FuNNRar in order to get both quantitative efficiency and qualitative improvement.

5. REFERENCES

[1] Shijian Lu and Joo-Hwee Lim, "Saliency modeling from image histograms," in *Computer Vision–ECCV 2012*, pp. 321–332. Springer, 2012.

[2] Marc Decombas, Frederic Dufaux, Erwann Renan, B Pequet-Popesu, and François Capman, "Improved seam carving for semantic video cod," in *Multimedia Signal Processing (MMSP), 2012 IEEE 14th International Workshop on*. IEEE, 2012, pp. 53–58.

[3] Michael Rubinstein, Ariel Shamir, and Shai Avidan, "Improved seam carving for video retargeting," in *ACM transactions on graphics (TOG)*. ACM, 2008, vol. 27, p. 16.

[4] Zhuang Li, Prakash Ishwar, and Janusz Konrad, "Video condensation by ribbon carving," *Image Processing, IEEE Transactions on*, vol. 18, no. 11, pp. 2572–2583, 2009.

[5] Christof Koch and Shimon Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of intelligence*, pp. 115–141. Springer, 1987.

[6] John K Tsotsos, Scan M Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo, "Modeling visual attention via selective tuning," *Artificial intelligence*, vol. 78, no. 1, pp. 507–545, 1995.

[7] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 11, pp. 1254–1259, 1998.

[8] Neil Bruce and John Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2005, pp. 155–162.

[9] Jonathan Harel, Christof Koch, and Pietro Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[10] Nicolas Riche, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, and Thierry Dutoit, “Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis,” *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 642–658, 2013.

[11] Qi Zhao and Christof Koch, “Learning saliency-based visual attention: A review,” *Signal Processing*, vol. 93, no. 6, pp. 1401–1407, 2013.

[12] Junting Pan and Xavier Giró-i Nieto, “End-to-end convolutional network for saliency prediction,” *arXiv preprint arXiv:1507.01422*, 2015.

[13] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu, “Deepfix: A fully convolutional neural network for predicting human eye fixations,” *arXiv preprint arXiv:1510.02927*, 2015.

[14] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao, “Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 262–270.

[15] Guanbin Li and Yizhou Yu, “Visual saliency based on multiscale deep features,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[16] Ali Borji and Laurent Itti, “Cat2000: A large scale fixation dataset for boosting saliency research,” *arXiv preprint arXiv:1505.03581*, 2015.

[17] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao, “Salicon: Saliency in context,” in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 1072–1080.

[18] Matei Mancas, Julien Leroy, Nicolas Riche, and Pierre Marighetto, “Attention website,” http://tcts.fpms.ac.be/attention/.

[19] Laurent Itti and Christof Koch, “Comparison of feature combination strategies for saliency-based visual attention systems,” in *Electronic Imaging’99*. International Society for Optics and Photonics, 1999, pp. 473–482.

[20] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit, “Saliency and human fixations: state-of-the-art and study of comparison metrics,” in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1153–1160.

[21] Nicolas Riche, Matei Mancas, Bernard Gosselin, and Thierry Dutoit, “Rare: A new bottom-up saliency model,” in *Image Processing (ICIP), 2012 19th IEEE International Conference on*. IEEE, 2012, pp. 641–644.

[22] Jianming Zhang and Stan Sclaroff, “Saliency detection: A boolean map approach,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 153–160.

[23] Anton Garcia-Diaz, Victor Leboran, Xose R Fdez-Vidal, and Xose M Pardo, “On the relationship between optical variability, visual saliency, and eye fixations: A computational approach,” *Journal of vision*, vol. 12, no. 6, pp. 17–17, 2012.

[24] Antón Garcia-Diaz, Xosé R Fdez-Vidal, Xosé M Pardo, and Raquel Dosil, “Saliency from hierarchical adaptation through decorrelation and variance normalization,” *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012.

[25] Fernando López-García, Raquel Dosil, Xosé Manuel Pardo, and Xosé Ramón Fdez-Vidal, *Scene Recognition through Visual Attention and Image Features: A Comparison between SIFT and SURF Approaches*, INTECH Open Access Publisher, 2011.

[26] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal, “Context-aware saliency detection,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 10, pp. 1915–1926, 2012.