SPATIO-TEMPORAL COMBINATION OF SALIENCY MAPS AND EYE-TRACKING ASSESSMENT OF DIFFERENT STRATEGIES

C. Chamaret, J.C. Chevet

TECHNICOLOR Corporate Research
1 Avenue de Belle-fontaine
35518 Cesson-Sevigne, France

*O. Le Meur**

University of Rennes 1
IRISA/TEMICS
35042 Rennes, France

ABSTRACT

The modeling of the human visual attention into a computational attention model leads to the split of visual features into several independent channels. Then, a difficult problem arises to combine these maps, having different dynamic ranges or distribution. When several maps are considered, such process is mandatory in order to compute a single measure of interest for each location, regardless of which features contributed to the salience. Several strategies of cue combination are proposed in this paper for the spatial cues as well as the temporal saliency. Finally, some user tests on still image and video databases leads to highlight one operator.

Index Terms— Visual attention, computational model, map fusion, user experiments, eye-tracker.

1. INTRODUCTION

A visual attention model yields an attention (or saliency) map, indicating the visually important areas of a picture. In this paper, a fundamental problem is addressed regarding the creation of final saliency map. Indeed, one of the most difficult problem concerns the fusion of the different saliency maps, which are created when modeling several independent visual features. A biological model comprises a number of parallel channels for processing different visual features such as the luminance, the orientation and the color. Each visual feature is processed to provide a feature-dependent attention-map. These maps are then combined to produce a single feature-independent map. But, how can we mix together these maps to take into account relevant information of each feature map? Different visual dimensions with several dynamic ranges are competing with each other.

R. Milanese [1] gave early the rationale of a fusion process. His approach intended to formalize the data fusion problems. He clearly pointed out the most important issue of a fusion algorithm: competition that must exist between visual feature maps and within each map. H. Nothdurft [2] shown that a stimulus that presents some saliency in more than one visual dimension is generally more conspicuous than a stimulus having only some salience in one dimension. For example, the conjunction of several visual features (luminance contrast and orientation contrast for example) leads to a more salient target. The different visual dimensions generally all interact and contribute simultaneously to the representation of saliency. However, H. Nothdurft [2] noticed the sum of the salience carried by different dimensions is greater than the final observed salience. L. Itti proposed in [3] several algorithms to mix together several saliency maps. Nevertheless, all the proposed algorithms only deal with the intra map competitions. T. Jost [4] compares more specifically a one-cue gray-level model to a two-cues color model; the competitions is inter-map, but only focusing on spatial saliency maps. [5, 6] have restricted their human vision tests to a still image database. They validated only spatial cues of their saliency computation.

This paper proposes to assess the relevance of several fusion techniques with two ground truth databases. This study rests on the biologically plausible model of bottom-up attention, proposed by Le Meur et al. [7]. The first part of this paper describes the important features of this bottom-up model. From this framework, different techniques to mix together different maps are proposed and tested by competition to two databases of still color pictures and videos. Results are expressed to know how the fusion process may influence the attention-map relevance in case of spatial as well as spatio-temporal saliency.

2. COMPUTATIONAL MODELISATION OF VISUAL ATTENTION (VA)

Originally, Le Meur et al. have defined a computational salience model bottom-up visual attention model based on a hierarchical architecture, such as described in [7]. An input image is decomposed in a number of feature channels (color, luminance, orientation) in a parallel manner. A single saliency map is then formed by the combination of the outputs of these channels.

The novelty introduced in [7] mainly concerns the extraction of the different visual features. In the previous visual attention models, the visual features (color, luminance, orientation) were extracted by dedicated mechanisms. For example, Itti's model [3] computed 6 visual features for the luminance channel, 12 for the color channel and 24 for the orientation channel. In term of dynamic range, these three sets of maps are completely different. The consequence is that it is required to normalize all the feature maps in the same dynamic range before performing the fusion. To cope with this issue, Le Meur et al. [7] propose to express all the data in term of visibility. Early visual features, luminance and color, are then first normalized to their own visibility threshold, by using both Contrast Sensitivity Functions (CSF) and visual masking.

Due to new considerations on the model, the original architecture has been simplified, such as depicted in figure 1. A wavelet transform instead of a Fourier transform is applied on luminance component due to computational constraints. This hierarchical decomposition simulates the different populations of cortical cells. A center surround filter for the Y component as well as the sum of absolute differences for the U and V component removes redundant information between the subbands. The fusion between the subbands is performed within

*This work was done while the author was at Thomson R&D

the center surround filter, such as depicted in equation 1, 2.

$$R_{OnOff} = \beta_1 \times R_{OnOff}(l+1) + \beta_2 \times max(0, S_C - \beta_{OnOff} \times S_S) \quad (1)$$

$$R_{OffOn} = \beta_1 \times R_{OffOn}(l+1) + \beta_2 \times max(0, S_S - \beta_{OffOn} \times S_C) \quad (2)$$

where l represents the current level of wavelet decomposition. β_x are some gains determined by using CSF curves. S_C and S_S are the summation of coefficients in the area of filters, that means for the $Center$ and for the $Surround$ surface, respectively. Only the low frequency subband is used per level and the total number of level L is related to the original picture resolution.
Finally, the saliency map SM_Y is obtained by competing the two previous responses: $SM_Y = \max(1.5 \times R_{OffOn}, R_{OnOff})$. As a consequence to this new implementation choice, the normalization problem of saliency maps has to be solved again in addition to the fusion issue.

3. FUSION STRATEGIES

The mentioned visual attention model creates three, purely spatial, saliency maps related to the luminance and color components Y, U and V. A first fusion step is performed for those three maps as depicted in figure 1. Afterwards, the fusion operators are applied between the spatial and temporal saliency maps. Before summarizing the different operators of fusion, the notations used in the paper are introduced.

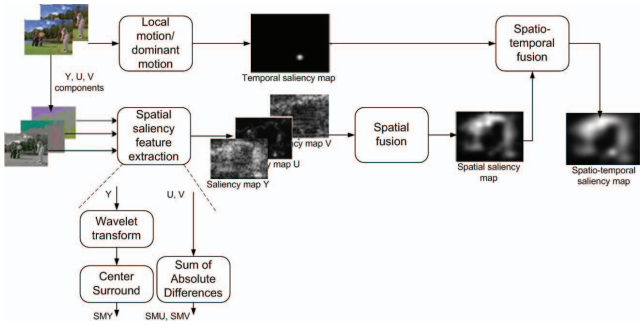

Fig. 1. Flow chart of the fusion implementation regarding the involved maps.

3.1. Notations

Different operators, more or less sophisticated, are proposed to merge the feature-dependent attention-map into a final attention map Out. The following notations are used to describe them:

- SM_i: attention-map of the i^{th} early visual feature,
- $\mathcal{N}$: normalization operator. This operator uses the global maximum of each map to obtain a dynamic range between 0 and 1,
- $\mathcal{N}_c$: normalization operator driven by a priori knowledge. Instead of using the global maximum of each map, this operator uses an empirical value. These empirical values are deduced from adequate experiments. After, a quantization is done to rescale the data on a specific range,
- WTA, short for Winner-Take-All. The basic WTA paradigm selects the winning location based on its salience. The WTA is used with a localized inhibitory spread. The WTA is an iterative approach repeating the following steps: first, find the global maximum. Second, the spatial neighborhood of the previous maximum is inhibited. Then, these two steps are repeated until to reach the stop condition. In the proposed design, both number of maximum and ratio between two consecutive maximum are used to break the loop.
- $NEAREST\,(SM)$: this function returns the maximum value that is the nearest regarding the current position in the map SM.

3.2. Operators and methods

3.2.1. NS: Normalized and Sum

The most simple method to solve the fusion problem is to normalize all attention-maps to the same dynamic range (between 0 and 1) and then to sum all maps into the final saliency map.

$$Out = \mathcal{N}\left(\sum_i \mathcal{N}\,(SM_i)\right) \quad (3)$$

The relative importance between the different maps is lost, as they are normalized to the same dynamic range. Moreover, irrelevant information, due to a noisy map, can be promoted leading to a wrong result. Moreover, there is no spatial competition: not in the map itself nor between the maps. The sole advantage refers to its simplicity.

3.2.2. NM: Normalized and Maximum

Compared to NS method, the summation is replaced by the maximum operator.

$$Out = \max_i \mathcal{N}\,(SM_i) \quad (4)$$

The drawbacks and advantages are the same than the previous method.

3.2.3. CNS: Coherent Normalization and Sum

A normalization seems unavoidable in the process of saliency maps fusion. For this approach, the maximum saliency value for each visual dimension is empirically determined. Particular pictures (pictures with a high contrasted color, black picture with a white patch...) are used.

$$Out = \sum_i \mathcal{N}_c\,(SM_i) \quad (5)$$

The most important advantage concerns the conservation of the relative importance between maps. Obviously, if the empirical maximum values are erroneous, the final result will be erroneous too. This approach is not very biologically plausible because this method tends to favor a unique location per map. For example, if two saliency peaks are present in the map, the highest peak will be promoted to the detriment of the second.

3.2.4. CNM: Coherent Normalization and Maximum

Comparatively to CNS method, the summation operator is replaced by the maximum operator.

$$Out = \max_i \mathcal{N}_c(SM_i) \quad (6)$$

The drawbacks and advantages are roughly the same as the previous one.

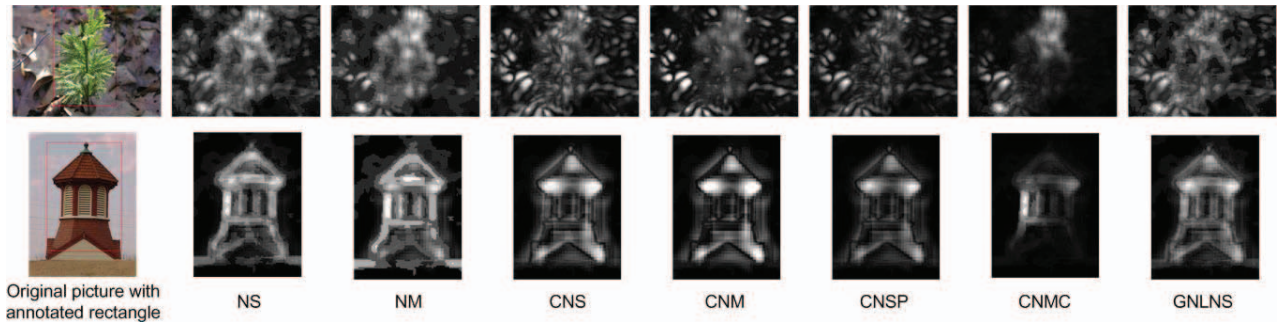

Fig. 2. A set of fusion results for all the proposed methods. Red rectangle is the manual annotated rectangle from gound truth.

3.2.5. CNSP: Coherent Normalization, Sum plus Product

The local inter-map redundancies and the local inter-map incoherence are the heart of this new approach. A particular areas of a picture can generate some saliency in several visual dimensions (luminance, color), even if the feature maps are independently obtained. To deal with the inter-map redundancies, the product operator is used, while the summation product is used to consider the local inter-map incoherence.

$$Out = \sum_i \mathcal{N}_c(SM_i) + \prod_i (1 + \mathcal{N}_c(SM_i)) \tag{7}$$

The advantage is the competition between the different maps. All the visual dimensions are used to promote a local part of the picture. An item that generates saliency in several dimensions will be favor compared to another one. Although that this approach brings something new, an intra-map competition is always lacking.

3.2.6. CNMC: Coherent Normalization, intra and inter Map Competition

To solve the aforementioned drawbacks, the $CNSP$ approach is upgraded by using a WTA algorithm with localized inhibitory spread. The local maxima are then detected and used to locally favor some parts of the picture. The number of maximum peaks, their values and the difference value between two consecutive maximum peaks are required to keep only the most interesting areas.

$$Out = \sum_i \tilde{SM}_i + \prod_i (1 + \tilde{SM}_i) \tag{8}$$

where $\tilde{S}_i = \frac{\mathcal{N}_c(S_i)}{NEAREST(WTA(\mathcal{N}_c(S_i))}$

This approach is interesting because a sparse distribution of maximum is computed in order to favor and to promote certain locations in the scene. This property is close to the biological behavior. Unfortunately, several thresholds are required to drive the algorithm. The two most important difficulties are: first, a threshold α is used to decide whether a local maximum is relevant or not. If this threshold is too high, we can miss some regions of interest, while if this threshold is too small, regions of non interest can be considered as of interest. The value α is finally used to prevent the promoting of an area having a weak saliency. Second, how many local maximum β should be taken into account? The ratio between two consecutive maximum is an important criterion. To test the relevance of this ratio, a third threshold γ is used. In the preliminary design, α, β and γ are respectively set to $\frac{L-1}{2}$, 3 and 1.6. L is the maximum dynamic obtained after the coherent normalization (typically, L = 8).

3.2.7. GNLNS: Global Non-Linear Normalization followed by Normalization

This method implemented by L. Itti [3] consists in promoting the maps having few saliency peaks and in removing the maps having an uniform distribution and a lot of saliency peaks.

$$Out = \sum_i \left[\left(\mathcal{N}\left(SM_i\right) \right) \left(M_i - m_i \right)^2 \right] \tag{9}$$

with M_i: the maximum value of map i and m_i: the average of all the other maximum values of map i.
Due to the global normalization, this method is highly sensitive to noise in the maps.

4. RESULTS

All the fusion operators have been evaluated in a first time for the spatial maps fusion and in a second time for the spatio-temporal fusion. The evaluation of performance is different depending on the available database.

4.1. Spatial evaluation: MSRA data

Regarding the spatial fusion, the Microsoft MSRA[1] database has been used. 500 pictures have been manually annotated with a region-of-interest area. The visual attention model provides a certain number of points of interest within a picture through the saliency map. This percentage is computed by summation of all saliency values per pixel within the rectangle-of-interest. This quantity is divided by the global saliency within the entire picture. These percentages are proposed in table 1.

	mean	st. dev.	min.	max.
NS	52.38	19.42	7.29	98.95
NM	50.28	18.78	8.34	98.36
CNS	52.52	19.07	8.64	98.26
CNM	51.18	19	9.06	97.61
CNSP	53.36	19.24	8.50	98.39
CNMC	**71.05**	25.07	0	100
GNLNS	52.79	19.29	7.49	98.62

Table 1. Percentage of visual significance (or computed saliency points inside the rectangle-of-interest).

[1] http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient-object.htm

Methods	NS			NM			CNS			CNM			CNSP			CNMC			GNLNS		
Metrics	CC	KL	ROC	CC	KL	ROC	CC	KL	ROC	CC	KL	ROC	CC	KL	ROC	CC	KL	ROC	CC	KL	ROC
canoa	0.44	11.5	0.71	0.44	12.1	0.70	0.44	11.5	0.71	0.44	12	0.70	0.55	10.3	0.84	**0.61**	**8.5**	**0.87**	0.59	8.5	0.84
kayak	0.48	12.3	0.76	0.44	12.9	0.73	0.49	12.2	0.76	0.44	12.9	0.73	0.59	10.8	0.86	**0.63**	**8.6**	**0.87**	0.57	9.6	0.82
pat	0.59	8.8	0.83	0.55	9.7	0.82	0.59	8.9	0.83	0.54	9.9	0.81	0.61	8.2	0.87	**0.63**	**6.6**	**0.88**	0.55	8.5	0.83
patVit	**0.65**	**7.6**	**0.85**	0.61	9.3	0.83	0.65	7.8	0.85	0.57	10	0.81	0.61	8.6	0.83	0.6	8.5	0.83	0.52	10.4	0.77
stefan	0.53	9.9	0.8	0.48	11.1	0.79	0.52	9.8	0.8	0.5	10.6	0.8	0.53	9.7	0.81	**0.54**	**9**	**0.82**	0.49	10.1	0.8
table	0.54	9.3	0.8	0.52	9.9	0.79	0.54	9.3	0.8	0.52	9.9	0.79	0.55	9.0	0.8	**0.57**	**8.4**	**0.8**	0.53	9.9	0.8
titleist	**0.62**	8.5	0.85	0.6	9.4	0.84	0.62	8.8	**0.85**	0.57	10.1	0.83	0.6	9.0	0.84	0.61	**8.1**	0.85	0.55	10.2	0.81
mean	0.55	9.7	0.8	0.52	10.6	0.78	0.55	9.7	0.8	0.51	10.8	0.78	0.58	9.4	0.83	**0.6**	**8.2**	**0.85**	0.54	9.6	0.81
st. dev	0.07	1.5	0.04	0.06	1.3	0.04	0.06	1.4	0.04	0.04	1.1	0.04	0.03	0.8	0.02	0.03	0.7	0.02	0.02	0.7	0.02

Table 2. Results of spatio-temporal fusion for a set of sequences: three statistical metrics CC, KL and ROC.

One strategy outperforms the other one in terms of percentage. The $CNMC$ method reachs more than 70 % of predicted points while the other methods stay around 50 %. One can notice that the MSRA database induces a top-down behavior of users, although the considered VA model is based on bottom-up concepts of attention.
Qualitative evaluation is also proposed in figure 2. As expected, the CNMC minimizes the global quantity of saliency around the picture due to the intra and inter-map competitions. The other methods tend towards the enhancing of too many saliency peaks coming from several components without inter-map correlation.

4.2. Spatio-temporal evaluation: eye-tracking data

The spatio-temporal fusion (proposed in figure 1) has been evaluated by means of eye-tracking database. The fixations of unpaid observers for seven sequences have been recorded by an Cambridge Eye tracking apparatus. More details on the protocol are described in [7].
Several statistical metrics have been used to quantitatively assess or classify the different fusion methods. Results are summarized in the table 2. CC stands for the linear correlation coefficient which has been computed between the saliency map of the VA model and a ground truth saliency map averaged for all observers. The Kullback-Liebler divergence, noted KL, estimates the overall dissimilarity between two probability density functions. These two metrics are sensitive to saliency dynamic range and to the salience distribution. ROC analysis is complementary to the two previous ones. The ROC value of table 2 is the area under the ROC curve, also noted AUC. An area of 1 means the VA model is fully conformed to the ground truth. More details about these metrics may be found in [7].

One can notice the $GNLNS$ performances. The mean results are pretty similar to the simplest method NS, although this method is more elaborated.

The three metrics point to the same conclusion. The $CNMC$ method is the most efficient one, because it is the closest one to the users experiments. Indeed, the mean values over all sequences per metric is in favor of this method. This conclusion is in line with the one for spatial fusion and expected due to the considerations mentioned in section 3.2.6.
Moreover, one can notice the low values of standard deviation for the three metrics of this method as well. The simplest methods (NS, NM, CNS, CNM) are less stable when focusing on their standard deviation values.

5. CONCLUSION

This paper proposed a number of fusion schemes for saliency maps. Previous works have compared their fusion scheme in the framework of spatial competition, but never considered the validation of spatio-temporal maps.
The most effective operator of fusion has been identified coherently as being the same for still images and video sequences. A database with 500 pictures annotated by users with areas-of-interest has been confronted to spatial operators of fusion. In a second time, the spatio-temporal fusion has been tested with a ground truth of saliency maps recorded from several users by means of an eye-tracking. Both experiments lead to the same classification of fusion methods.
The novelty of this paper rests on the assessment of temporal fusion of saliency maps. Previous papers did not evaluate if the temporal saliency may be merged in the same way as the spatial saliency components. To go one step further, the computation of the similarity degree between a ground truth and a prediction for video sequences shoud be based on a given temporal horizon rather than on a unique instant.

6. REFERENCES

[1] R. Milanese, "Detecting salient regions in an image: from biological evidence to computer implementation," in *PhD thesis*. Geneve University, 1993.

[2] H. Nothdurft, "Salience from feature contrast: additivity accross dimensions," in *Vision Research*, 2000, vol. 40, pp. 1183–1201.

[3] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," in *Journal of Electronic Imaging*, Jan 2001, vol. 10, pp. 161–169.

[4] T. Jost, N. Ouerhani, R. van Wartburg, R. Muri, and H. Hugli, "Assessing the contribution of color in visual attention," in *Computer Vision and Image Understanding, Special Issue on Attention and Performance in Computer Vision, Elsevier*, Oct-nov 2005, vol. 100, pp. 1–248.

[5] N. Ouerhani, A. Bur, and H. Hugli, "Linear vs nonlinear feature combination for saliency computation: a comparison with human vision," in *Pattern recognition, Springer Verlag*, 2006, vol. 4174., pp. 314–323.

[6] A. Bur and H. Hugli, "Optimal cue combination for saliency computation: a comparison with human vision," in *Lecture Notes in computer science, Springer-Verlag GmbH*, 2007, vol. 4528, pp. 109–118.

[7] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," in *Vision Research*, September 2007, vol. 47, pp. 2483–2498.