Saliency aggregation: Does unity make strength?

Olivier Le Meur[a] and Zhi Liu[a,b]

[a] IRISA, University of Rennes 1, FRANCE
[b] School of Communication and Information Engineering, Shanghai University, CHINA

Abstract. In this study, we investigate whether the aggregation of saliency maps allows to outperform the best saliency models. This paper discusses various aggregation methods; six unsupervised and four supervised learning methods are tested on two existing eye fixation datasets. Results show that a simple average of the TOP 2 saliency maps significantly outperforms the best saliency models. Considering more saliency models tends to decrease the performance, even when robust aggregation methods are used. Concerning the supervised learning methods, we provide evidence that it is possible to further increase the performance, under the condition that an image similar to the input image can be found in the training dataset. Our results might have an impact for critical applications which require robust and relevant saliency maps.

1 Introduction

In 1985, Koch and Ullman proposed the first plausible architecture for modelling the visual attention [1]. This seminal paper has motivated much of the following work of computational models of attention. Today there exist a number of saliency models for predicting the most visually salient locations within a scene. A taxonomy composed of 8 categories has been recently proposed by Borji and Itti [2]. The two main categories, encompassing most existing models, are termed as *cognitive models* and *information theoretic models*. The former strives to simulate the properties of our visual system whereas the latter is grounded in the information theory. Although all existing models follow the same objective, they provide results which could be, to some extent, different. The discrepancies are related to the quality of the prediction but also to the saliency map representation. Indeed some models output very focused saliency maps [3–5] whereas the distribution of saliency values is much more uniform in other models [6, 7]. Others tend to emphasize more on the image edges [8], the color or luminance contrast. This saliency map manifold contains a rich resource that should be used and from which new saliency maps could be inferred. Combining saliency maps generated using different models might enhance the prediction quality and the robustness of the prediction. Our goal is then to take saliency maps from this manifold and to produce the final saliency map.

To the best of our knowledge, there is no study dealing with the fusion of saliency maps. In the context-of-object of interest detection, we can mention two related studies. Borji et al. [11] combined the results of several models and found out that the simple average method performs well. Mai et al. [10] combined results of models detecting object-of-interest on simple images (mainly composed of one object-of-interest withsimple background). They use simple methods as well as the trained methods. The main drawback of the aforementioned studies concerns the choice of the tested models, which are not The best-in-class. Consequently, the room for improvement is still important and can be obtained, to some degree, by aggregating different results. However, we draw attention to a crucial difference between our work and the two aforementioned studies [11, 10]. The saliency maps that are aggregated in this study are computed using computational models of visual attention for eye fixation prediction. In [11, 10], the saliency maps are the outputs of saliency models which aim to completely highlight salient objects such as [12].

Keeping all these points in mind, we investigate whether we could improve on the prediction quality by aggregating a set of saliency maps or not. Eye fixation datasets will be used as the ground truth.

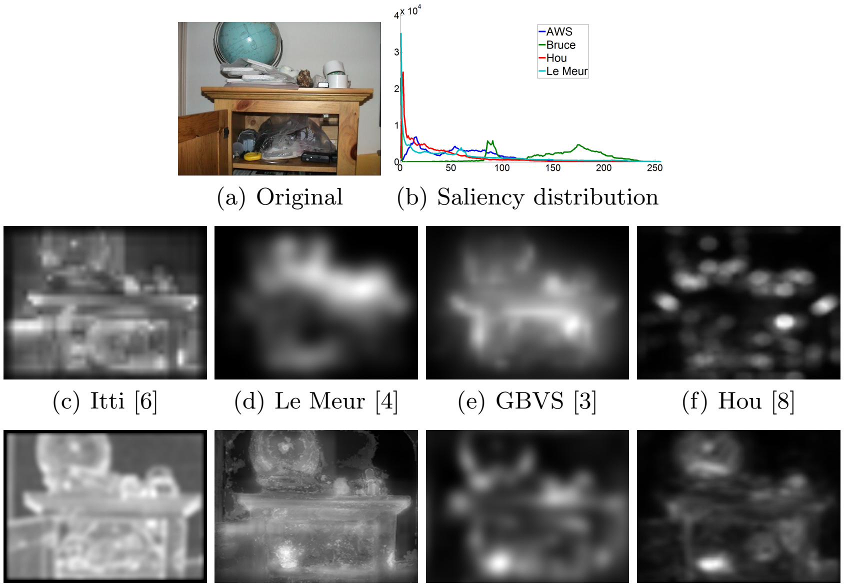
The paper is organized as follows. Section 2 presents the methods we use for aggregating saliency maps. Section 3 shows the performance of the saliency models, taken alone, and the performance of the aggregation functions. Finally, we draw some conclusions in Section 4.

2 Saliency aggregation

2.1 Context and problem

As illustrated by Figure 1, the predicted saliency maps do not exhibit similar characteristics. Figure 1 (b), which plots the distribution of saliency values for four models, clearly shows the discrepancy that exists between saliency maps. Some are very focused whereas others are much more uniform. We can also notice that the contrast between salient and non-salient areas can be either high or very low. This discrepancy between maps can be considered as noise but also as an important cue that needs to be exploited. Combining saliency models may enhance the similarity between human and the predicted saliency maps. Human saliency maps, as we will see in Section 3, will be inferred from publicly-available eye fixation datasets.

To investigate this point, we select 8 state-of-the-art models (GBVS [3], Judd [14], RARE2012 [15], AWS [5], Le Meur [4], Bruce [7], Hou [8] and Itti [6]) and aggregate their saliency maps into a unique one. The following subsections present the tested aggregation methods. Two categories of methods have been tested. The methods in the first category are unsupervised, meaning that there is neither optimization nor prior knowledge on saliency maps. The methods in the second category are supervised. Different algorithms are used to train the best way to combine together saliency maps.



(a) Original (b) Saliency distribution

(c) Itti [6] (d) Le Meur [4] (e) GBVS [3] (f) Hou [8]

(g) Bruce [7] (h) Judd [14] (i) AWS [5] (j) RARE2012 [15]

Fig. 1. (a) Original picture; (b) saliency distribution obtained using 4 models (AWS, Bruce, Hou and Le Meur); (c) to (j) the predicted saliency maps from the 8 state-of-the-art saliency models considered in this study.

2.2 Unsupervised methods

Six aggregation methods are tested here. The first 4 functions are based on a simple weighted linear summation:

$$p(x|M_1, \cdots, M_K) = \sum_{k=1}^{K} w_k \times p(x|M_k) \tag{1}$$

where $p(x|M_1, \cdots, M_K)$ is the probability of an image pixel x ($x \in \Omega$, with $\Omega \subset \mathcal{R}^2$) to be salient after the combination; $p(x|M_1, \cdots, M_K)$ is positive or null. M_k is the saliency map produced by model k. $p(x|M_k)$ is the probability of an image pixel x from the saliency map M_k to be salient. w_k is the weighting coefficient, given that $\sum_{k=1}^{K} w_k = 1$ and $w_k \geq 0, \forall k$. K is the number of saliency maps ($K = 8$ in our case).

The main goal is to compute the weighting coefficients in order to improve the degree of similarity between the ground truth and the aggregated saliency map. These weights are computed thanks to the following methods:

- Uniform: weights w are uniform and spatially invariant, $w_k = \frac{1}{K}$;
- Median: weights w are locally deduced from the saliency values. All weights are null, except for the one which corresponds to the median value of the saliency values for a given location. In this case, weights are spatially variant;

- M-estimator: weights w are computed by a weight function commonly used for robust regression. The weight function aims to limit the influence of outlier data by decreasing their contributions. We consider three weight functions. They are defined by the second derivatives of the Welsh, the L_1L_2 and the Geman-McClure functions. The first one requires a parameter whereas the other two functions are non-parametric:

$$g_{Welsh}(e(x|M_k)) = exp\left(-\frac{e(x|M_k)^2}{\sigma^2}\right) \tag{2}$$

$$g_{L_1L_2}(e(x|M_k)) = \frac{1}{\sqrt{1+e(x|M_k)^2/2}} \tag{3}$$

$$g_{Geman}(e(x|M_k)) = \frac{1}{(1+e(x|M_k)^2)^2} \tag{4}$$

where, the error $e(x|M_k)$ for a location x and model k represents the deviation between the current location and the average saliency value computed locally over all saliency maps:

$$e(x|M_k) = p(x|M_k) - \frac{1}{Z}\sum_{y\in\nu}\sum_{k=1}^{K} p(y|M_k) \tag{5}$$

where ν is a 3×3 local neighbourhood centred on x. Z is a normalization factor. Functions g_{Welsh} and g_{Geman} further reduce the effect of large errors compared to function $g_{L_1L_2}$. Weights of equation (1) are finally given by $w_k = g_{Welsh}\left(e\left(x|M_k\right)\right)$, $w_k = g_{L_1L_2}\left(e\left(x|M_k\right)\right)$ and $w_k = g_{Geman}\left(e\left(x|M_k\right)\right)$ for Welsh, L_1L_2 and the Geman-McClure function, respectively. For the function g_{Welsh}, the standard deviation is locally estimated using the K saliency maps.
- The last tested method is based on a global minimization of an energy function. Let $\mathcal{I}$ be the set of pixels in the final saliency map and $\mathcal{L}$ be the finite set of labels. The labels correspond to the final saliency values (coming from one given model) that we want to estimate at each pixel. A labeling f assigns a label $f_x \in \mathcal{L}$ to each pixel x of image $\mathcal{I}$. The best labeling minimizes the energy given below

$$E(f) = \sum_{p\in\mathcal{L}} D(p) + \lambda \sum_{(x,y)\in\mathcal{N}} V(f_x, f_y) \tag{6}$$

where $\mathcal{N}$ is a 3×3 square neighbourhood, λ is a positive constant that controls the trade-off between $D(p)$ and $V(f_x, f_y)$. $D(p)$ is the data cost and $V(f_x, f_y)$ is the smoothness term. They are defined as

$$\begin{aligned} D(p) &= \sum_{n\in\mathcal{L}}\sum_{u\in\nu} \left(p(x+u|M_p) - p(x+u|M_n)\right)^2 \\ V(n,m) &= \parallel p(x|M_n) - p(y|M_n) \parallel^2 + \parallel p(x|M_m) - p(y|M_m) \parallel^2 \end{aligned} \tag{7}$$

The minimization of the energy E is achieved using loopy belief propagation [16], and the number of iterations is set to 10.

2.3 Supervised learning methods

In this section, the weights w_k are computed by minimizing the residual r between the actual and the predicted saliency values:

$$r(x) = \|p(x) - \sum_{k=1}^{K} w_k p(x|M_k)\|^2 \tag{8}$$

where $p(x)$ is the actual saliency deduced from the eye fixation dataset and $p(x|M_k)$ is the saliency at location x for model M_k. The actual saliency map which represents the ground truth is classically obtained by convolving the fixation map (considering all visual fixations of all observers) by a 2D Gaussian function, having a standard deviation of one degree of visual angle. More details can be found in [17].

For a given image $\mathcal{I}$ defined over $\Omega \subset \mathcal{R}^2$, and given that the number of unknowns, i.e. K, which is much smaller than the number of locations in $\mathcal{I}$, the optimal vector of weights W^* can be computed by the least-squares method as follows:

$$W^* = \arg\min \sum_{x \in \Omega} r(x) \tag{9}$$

In this study, W^* is computed by the following methods: The first one is the classical least-squares method, noted as LS, which minimizes the residual error between the actual and the aggregated saliency maps. One drawback is that the weights do not sum to 1 and can be positive or negative. This makes the interpretation difficult. This is the reason why three other methods have been tested. Two methods are constraint least-squares problems. Adding constraints aims to ease the interpretation of the computed weights. However, it is important to keep in mind that introducing constraint will reduce the solution space. The first constraint is that the weights have to sum to one. The sum-to-one constraint of the weights moves the LS problem onto the Locally Linear Embedding (LLE) [18]. Another constraint is that the weights are positive; this problem is similar to the problem of Non-negative Matrix Factorization (NMF) [19]. Finally, we also test a robust least-squares problem, noted as LSR. Instead of minimizing a sum of squares of the residuals, we use a Huber-type M-estimator [20] to reduce the influence of outliers. The algorithm simply consists in re-weighting iteratively the residuals according to the Cauchy weighting function given that a higher residual leads to a lower weight.

3 Performance

The performance of the aggregation functions have been evaluated on Bruce's [7] and Judd's [14] eye fixation datasets. Four metrics were used: linear correlation coefficient, Kullback-Leibler divergence, normalized scanpath saliency and hit rate. The linear correlation coefficient, noted as CC, computes the linear relationship between the ground truth saliency map and the predicted saliency

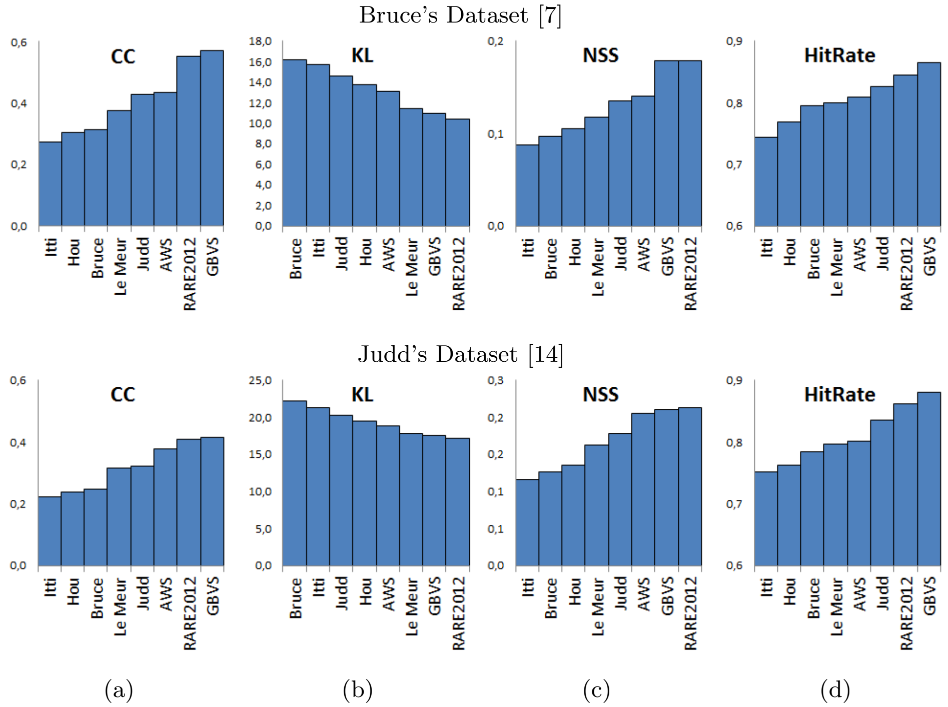


Fig. 2. Ranking visual saliency models over two datasets. Top row: Bruce's dataset; Bottom row: Judd's dataset. Four metrics are used. From left to right: correlation coefficient (CC), Kullback-Leibler divergence (KL), NSS (normalized scanpath saliency) and HitRate. Models are ranked in the increasing order according to their performance.

map. There is a perfect linear relationship when $CC = 1$. The Kullback-Leibler divergence, noted as KL, computes an overall dissimilarity between two distributions. The first step is to transform the ground truth saliency map and the predicted saliency maps into 2D distributions. The KL-divergence is positive or null. The perfect similarity ($KL = 0$) is obtained when the two saliency maps are strictly equal. The normalized scanpath saliency (NSS) proposed by [21] involves a saliency map and a set of fixations. It aims at evaluating the saliency values at fixation locations. The higher the NSS value, the better the predicted saliency maps. The hit rate measure used in this study is similar to the measure used in [14]. It involves a binarized saliency map and a set of fixations. It aims at counting the number of fixations falling within the binarized salient areas. By varying the binarization threshold, a hit rate curve is plotted. The hit rate measure is simply the area under the curve. The chance level is given by 0.5, whereas the highest similarity is given by 1. More details can be found in [17].

3.1 Performance of state-of-the-art models

Figure 2 illustrates the performance of the 8 selected models (GBVS [3], Judd [14], RARE2012 [15], AWS [5], Le Meur [4], Bruce [7], Hou [8] and Itti [6]) over Bruce and Judd datasets. According to our results, we find that the top 2 models are GBVS and RARE2012, the top 4 models are GBVS, RARE2012, Judd and AWS. This result is consistent with the recent benchmark of Borji et al. [9].

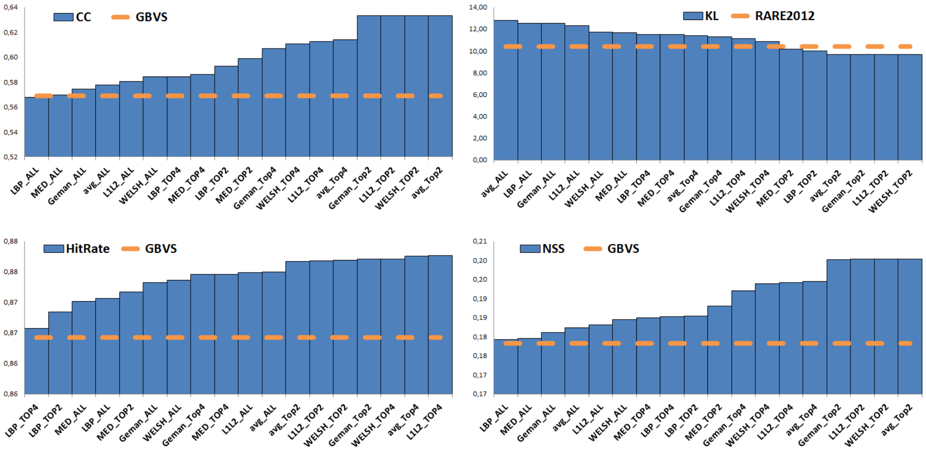


Fig. 3. Performance on Bruce's dataset of the six aggregation functions: Avg (for average), Geman (for g_{Geman}), L1L2 (for $g_{L_1 L_2}$), Welsh (for g_{Welsh}), LBP (for Loopy Belief Propagation) and MED (for Median operator). These functions are tested when considering the top 2, top 4 and all models. For each metric, namely CC, KL, HitRate and NSS, the aggregation functions are ranked from the lowest to the highest performance. The orange bar indicates the performance of the best saliency model. For instance, GBVS model achieves the best results on Bruce's dataset for CC, KL and NSS metrics.

3.2 Performance of saliency aggregation

The aggregation functions described in Section 2 are applied on the top 2 (GBVS and RARE2012), top 4 (GBVS, RARE2012, Judd, AWS) and the 8 saliency models. Figure 3 gives the performance of the saliency aggregation on Bruce's dataset. The performance of the best saliency model (out of the 8 models) for each metric is also indicated by the orange bar. From these results, we can draw several conclusions.

1. Except for the KL-divergence, the aggregation of saliency map outperforms the best saliency models in all cases. For instance, in terms of HitRate, the best aggregation function, i.e. L1L2 TOP4 (meaning that the function

$g_{L_1L_2}$ is applied on the top 4 saliency models) performs at 0.878 whereas the best saliency model performs at 0.864 (note that this gain is statistically significant (paired t-test, $p < 0.01$)). For the KL-divergence, only the LBP $TOP2$, avg $TOP2$, $L1L2$ $TOP2$ and $WELSH$ $TOP2$ aggregation function perform better than the best saliency model, i.e. RARE2012 model;

2. The second observation is related to the number of saliency models required to get the good performance. It is indeed interesting to notice that the aggregation functions using all saliency maps get the lowest scores. At the opposite, the best performances are obtained when the top 2 models are used for the CC, KL and NSS metrics. For the hit rate metric, the aggregation of the top 4 models is ranked first. However, the performances between the aggregation of the top 2 and top 4 models are not statistically significant. Considering more models tend to decrease the performances, the worst case occurring when all models are considered;
3. The third observation is related to the aggregation functions. The average, L1L2 and Welsh functions perform similarly and better than the median and LBP functions (considering the top 4 and top 2 models). The low performance of the LBP method can be explained by the obvious difference and the lack of spatial coherency between saliency maps as illustrated in Figure 1.

To conclude, a simple aggregation function, such as the average function, operating on the top 4 or top 2 models is a good candidate to improve significantly the performance of saliency models. For the sake of simplicity, we could only consider GBVS and RARE2012 models and average their saliency maps. Note that on Judd's dataset, we get similar trends (results are not presented here due to the page limit). The best performance is given by the average of the top 4 and the top 2 models.

Figure 4 (a) presents some results of the aggregation methods for a given image. For this example, it is difficult to see a significant difference between the average, Welsh and L1L2 methods. This is consistent with our previous findings (see Figure 3). However, concerning the LBP method, we notice a lack of spatial consistency, especially when all saliency maps are taken into account.

3.3 Performance of supervised methods

The optimal weights for aggregating the saliency maps are learned on Bruce's dataset. The different methods, namely LS, LSR, LLE and NMF, are evaluated for the top 2, top 4 and all models. Figure 5 illustrates the results on Bruce's dataset. The orange bar indicates the performance of the best saliency model.

As expected, the performance increases when the weights are learned. This is perfectly normal since we seek for the weights minimizing the prediction error (error between the ground truth and the aggregated saliency maps). Whatever the regression methods, the learning process outperforms the best saliency model, taken alone, in most of the tested configurations. There are only 3 cases out of 48 for which the weight optimization does not bring any improvement. Compared to the average of the top 2 models (performances were presented in

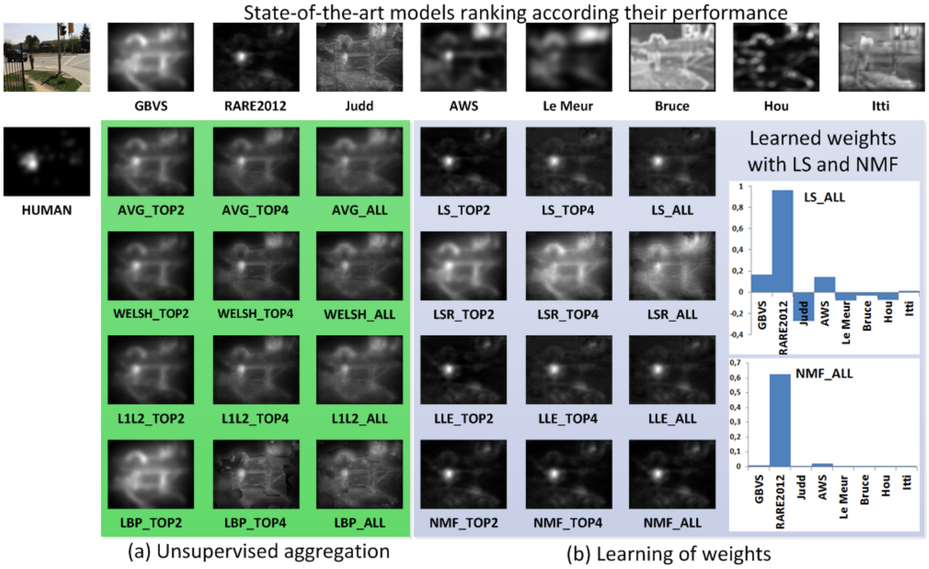


Fig. 4. Results of the aggregation obtained by (a) unsupervised and (b) supervised approaches. The original image and the human saliency map (i.e. the ground truth) are given on the top-left corner. On the top, the predicted saliency maps, obtained with the 8 tested saliency models, are illustrated. Results of the average, Welsh, L1L2 and LBP functions are shown in the green box (a) when the top 2, top 4 and all models are considered. Results of the LS, LSR, LLE and NMF learning methods are shown in the light blue box (b). On the right hand-side of the light blue box, the weights computed by the LS and NMF methods (considering all the maps of saliency models) are given. As we can see, RARE2012 model gets, for this particular example, the highest weights. Notice that for the NMF method, weights are positive.

the previous section), results are more contrasted (see the green horizontal line in Figure 5): only the simple least-squares method involving all models (noted as LS ALL) performs significantly better than the average of the top 2 models (except for the KL metric).

As soon as a constraint is added, such as the sum-to-one constraint of the weights (LLE) or the positivity of the weights (NMF), performance tends to decrease. This observation is valid when we consider the top 2, top 4 and all models. Figure 4 (b) presents some results of the supervised aggregation methods for a given image.

Similar results have been observed on Judd's dataset. The best learning function is the simple least-squares method involving all saliency maps; for instance, in terms of HitRate, it achieves 0.91, whereas the average of the top 2 models and the best model (Judd in this case) achieves 0.88. Figure 6 presents the results on Judd's dataset.

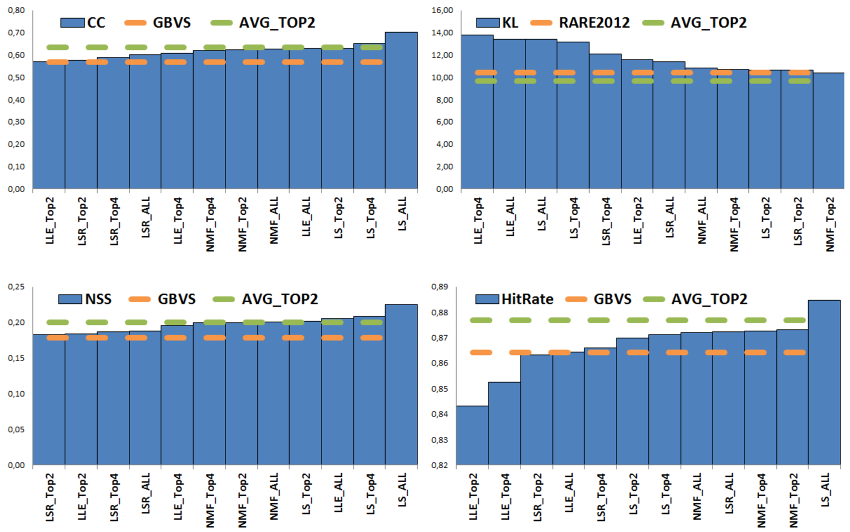


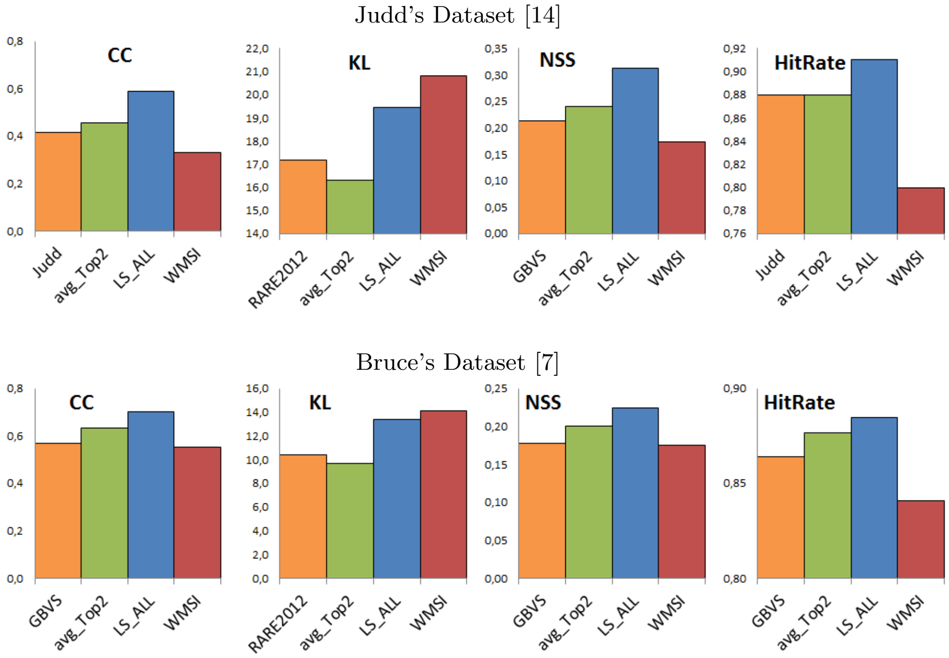
Fig. 5. The performance for the four methods (LS, LSR, LLE and NMF) in terms of (a) CC, (b) KL, (c) NSS and (d) HitRate on Bruce's dataset. We combine all saliency maps coming from the top 4 models (GBVS, RARE2012, Judd, AWS) and maps coming from the top 2 models (GBVS and RARE2012). Methods are ranked from the lowest to the highest performance.

The learning results presented so far in Figures 5 and 6 have to be considered as the upper-bound on the performance we can achieve by using a learning method. Obviously, in practice, we do not know the ground truth represented by the human saliency map, which is exactly what we want to predict.

To overcome this problem, we learn the weights for all pictures of Bruce's and Judd's datasets. The method chosen is the simple least-squares method which is applied on the 8 saliency maps. This strategy is called LS ALL, in previous paragraphs. As illustrated by Figures 5 and 6, this method provides the best results.

Once all the weights have been computed (8 weights per image), we compute the aggregated saliency map of an input image by using the pre-computed weights corresponding to the nearest neighbor image of the input image. In other words, we assume that the discrepancy between weights of two similar images is not significant. Figure 7 (a) presents the synoptic of the proposed method.

Given an input image, the first thing to do is to retrieve its nearest image from the dataset. This problem can be efficiently handled by using the VLAD (Vector of Locally Aggregated Descriptors) method introduced by Jégou et al. [22]. VLAD is an image descriptor which has been designed to be very low dimensional: only 16 bytes are required per image. The computation of VLAD descriptor is based on the vector quantizing a locally invariant descriptor such as


Judd's Dataset [14]

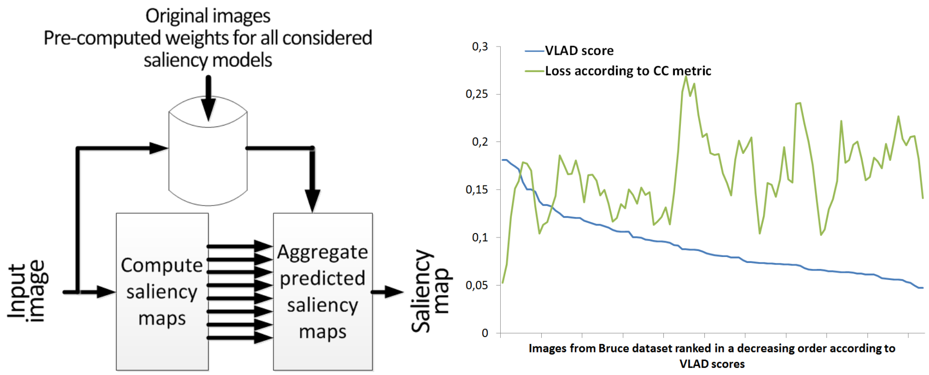
Bruce's Dataset [7]

Fig. 6. Performance of LS and WMSI methods on Bruce's and Judd's dataset. To ease the comparison, the performance of the best saliency model and the performance of the best aggregated method (average of the top 2 models) are displayed. The color code is the same as Figure 5: orange for the best saliency model, green for the best aggregation method (*avg Top*2) and blue for the learning method (*LS ALL*). The red bar, called WMSI (Weight of Most Similar Image), indicates the performance of the aggregation method when the weights of the most similar image in the dataset are used.

SIFT. From the weights of the most similar image, the aggregated saliency map is computed. We call this method WMSI, for Weight of Most Similar Image.

Figure 6 presents the results of the WMSI method (see the rightmost red column). Whatever the metrics, the WMSI method gets the lowest performance compared to the best saliency model, taken alone, the average of the two best saliency maps and, as expected, the learning method *LS ALL*. These results suggest that the initial assumption does not hold, i.e. similar images do not have the same distribution of weights. However, it is necessary to tone down this conclusion. Figure 8 illustrates this point. Two pairs of images ((a)-(b) and (c)-(d)) are given: the image (b) is the most similar image to the image (a). The VLAD score is equal to 0.22. The second pair of images (c)-(d), for which image (d) is the image which is the most similar to image (c) presents a low VLAD score, i.e. 0.06.

In the first case, when the VLAD score is high, the two sets (optimal versus weights of the similar image) of weights are strongly correlated $r = 0.94$. The



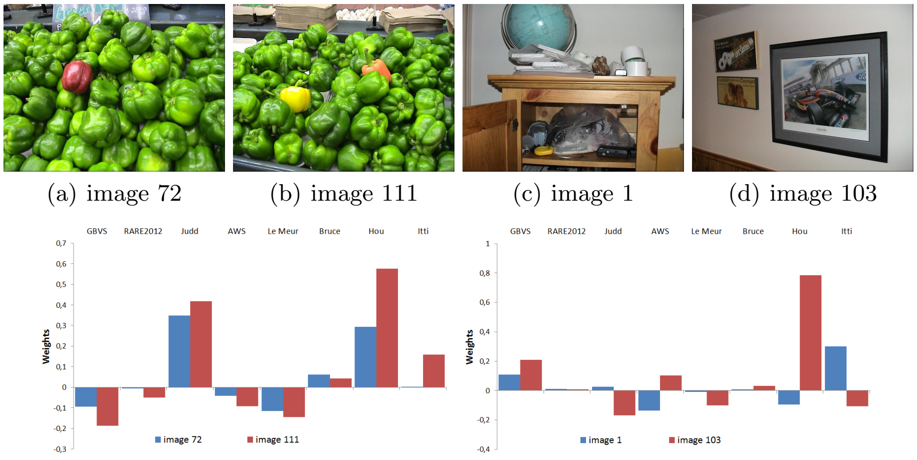
(a) (b)

Fig. 7. (a) From a dataset composed of a number of still color images for which the vector of weights W^* is known, we compute the aggregated saliency map for any input image. The first step is to compute the 8 saliency maps according to the 8 saliency models. We search into the dataset the image which is the most similar to the input image. This search is performed by using the VLAD method. The result of this search query is a set of optimal weights. They are used to combine the 8 saliency maps. (b) Loss of performance for the metric CC in function of VLAD score on Bruce's dataset. We ranked these images in the decreasing order according to the similarity score VLAD.

difference between the method $LS\ ALL$ (which represents the upper-bound of performance) and the WMSI method is limited: -0.012 and -0.015 for the metrics CC and HitRate, respectively. For this case, the WMSI method provides better results than GBVS and $avg\ TOP2$ methods.

The loss of performance is much more significant when the similarity score is low. For image (c) and (d) in Figure 8, the gap between $LS\ ALL$ and WMSI becomes much more significant: -0.237 and -0.144 for the metrics CC and HitRate, respectively. The two sets of weights are here not well correlated, $r = -0.42$, and are negatively correlated. To go one step further on this point, Figure 7 (b) plots the relationship between performance loss and VLAD score on Bruce's dataset. The Y-axis displays the loss of performance when considering the CC metric. To display the trend line, we smooth the raw data with a sliding average window using the two past and two next values. We observe that the loss of performance in terms of CC metric is correlated to the similarity score VLAD (the correlation coefficient is $r = -0.41$). The more similar the two images, the less important is the loss.

These results suggest that a supervised method might improve the quality of saliency map, provided that we succeed in finding an image similar to the input one. However, regarding the trade-off quality of prediction versus complexity, our study suggests that the simple average of the two best saliency maps is already a good candidate.



(a) image 72 (b) image 111 (c) image 1 (d) image 103

(e) Weights of (a) and (b) (f) Weights of (c) and (d)

Fig. 8. Weights differences: (a) to (d) represent four images extracted from Bruce's dataset. (b) is the nearest neighbors of (a) and (d) is is the nearest neighbors of (c) according to the VLAD score. (e) and (f) are the weights for the pair of images (a) & (b) and the pair of images (c) & (d), respectively.

4 Conclusion

In this paper, we investigate whether the aggregation of saliency maps can improve the quality of eye fixation prediction or not. Simple aggregation methods are tested as well as the supervised learning methods. Our experiments, requiring the computation of more than 100,000 saliency maps, show that saliency aggregation can consistently improve the performance in predicting where observers look within a scene. Among the 6 tested unsupervised methods, the best method is the simple average of the saliency maps from the top 2 best models. Considering more saliency maps do not allow to further improve the performance. Concerning the supervised learning approaches, they do not succeed in improving the performance on average. This is mainly due to the image matching: if the similarity score between the input image and its most similar image is low, the trained weights for combining the predicted saliency maps are not appropriate. However, when the similarity score is high, we provide evidence that the loss is limited, compared to the upper bound for which the weights are estimated by minimizing the prediction error.

For critical applications for which the relevance and robustness of the saliency map are fundamental such as video surveillance [23], object detection [13], clinical diagnostic [24], implementation of traffic sign [25], the conclusion of this study is interesting; the robustness of the prediction can be indeed enhanced by either averaging the saliency maps of the top 2 models or by considering a dedicated training dataset.

Future works will deal with the improvement of the learning methods as well as other retrieval methods, given a query image. In this context, it will be also required to define and build a very large database of eye tracking data.

Acknowledgement. This work is supported in part by a Marie Curie International Incoming Fellowship within the 7th European Community Framework Programme under Grant No. 299202 and No. 911202, and in part by the National Natural Science Foundation of China under Grant No. 61171144.
We thank Dr. Wanlei Zhao for his technical assistance for computing the VLAD scores.

References

1. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology **4** (1985) 219–227
2. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Trans. on Pattern Analysis and Machine Intelligence **35** (2013) 185–207
3. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Proceedings of Neural Information Processing Systems (NIPS). (2006)
4. Le Meur, O., Le Callet, P., Barba, D., Thoreau, D.: A coherent computational approach to model the bottom-up visual attention. IEEE Trans. On PAMI **28** (2006) 802–817
5. Garcia-Diaz, A., Fdez-Vidal, X.R., Pardo, X.M., Dosil, R.: Saliency from hierarchical adaptation through decorrelation and variance normalization. Image and Vision Computing **30** (2012) 51 – 64
6. Itti, L., Koch, C., Niebur, E.: A model for saliency-based visual attention for rapid scene analysis. IEEE Trans. on PAMI **20** (1998) 1254–1259
7. Bruce, N., Tsotsos, J.: Saliency, attention and visual search: an information theoretic approach. Journal of Vision **9** (2009) 1–24
8. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: CVPR. (2007)
9. Borji, A., Sihite, D.N., Itti, L.: Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. IEEE Transactions on Image Processing **22** (2012) 55–69
10. Mai, L., Niu, Y., Feng, L.: Saliency aggregation: a data-driven approach. In: CVPR. (2013)
11. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: a benchmark. In: ECCV. (2012) 414–429
12. Liu, Z., Zou, W., Le Meur, O.: Saliency tree: A novel saliency detection framework. IEEE Transactions on Image Processing **23** (2014) 1937–1952
13. Liu, Z., Zhang, X., Luo, S., Le Meur, O.: Superpixel-based spatiotemporal saliency detection. IEEE Transactions on Circuits and Systems for Video Technology **24** (2014) 1522–1540
14. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where people look. In: ICCV. (2009)
15. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., Dutoit, T.: Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. Signal Processing: Image Communication **28** (2013) 642 – 658

16. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. IEEE Trans. On PAMI **26** (2004) 1124–1137
17. Le Meur, O., Baccino, T.: Methods for comparing scanpaths and saliency maps: strengths and weaknesses. Behavior Research Method **1** (2012) 1–16
18. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **5500** (2000) 2323–2326
19. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Process. Syst. (NIPS). (2000)
20. Huber, P.: Robust regression: Asymptotics, conjectures and monte carlo. Ann. Stat. **1** (1973) 799–821
21. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision Research **45** (2005) 2397–2416
22. Jégou, H., Perronnin, F., Douze, M., Sánchez, J., Pérez, P., Schmid, C.: Aggregating local image descriptors into compact codes. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 1704–1716
23. Yubing, T., Cheikh, F., Guraya, F., Konik, H., Trémeau, A.: A spatiotemporal saliency model for video surveillance. Cognitive Computation **3** (2011) 241–263
24. Mamede, S., Splinter, T., van Gog, T., Rikers, R.M.J.P., Schmidt, H.: Exploring the role of salient distracting clinical features in the emergence of diagnostic errors and the mechanisms through which reflection counteracts mistakes. BMJ quality and safety (2011)
25. Won, W.J., Lee, M., Son, J.W.: Implementation of road traffic signs detection based on saliency map model. In: Intelligent Vehicles Symposium. (2008) 542 – 547