

Chapter 8

Metrics for Saliency Models Validation



Matei Mancas  and Nicolas Riche 

Different scores have been used in the literature to validate saliency models. Those scores have not evolved much and the same metrics are used for saliency validation on both images or videos. In this chapter, we will explain the standard measures used to evaluate the salient (1) object detection and (2) eye-tracking models.

There are the so-called static metrics which focus on a static saliency map represented as an image (2D map) and “dynamic” metrics where saliency maps are given in terms of dynamic scan paths of the eye.

In this chapter we focus mainly on static metrics, but we also provide a view on eye scan-path dynamic metrics [1, 2]. In the first section we will describe the static metrics for object detection ground truth. In the second section the static metrics for the eye-tracking ground truth. Finally, we will rapidly cover the dynamic metrics for eye-tracking ground truth and conclude.

8.1 Literature Review of Metrics for Object Detection

In this section, all metrics that have been used to assess salient object detection models are presented. Indeed, there are several ways to measure the agreement between salient object detection models and binary masks (bounding boxes or pixel-wise masks). Sometimes, metrics do not agree with each other.

However, contrary to the eye-tracking-based category, all the salient object detection benchmarks use very close gold standard location-based metrics. Moreover, in

M. Mancas (✉)
Numediart Institute, University of Mons, Mons, Belgium
e-mail: matei.mancas@umons.ac.be

N. Riche
Carneuse S.A., Louvain-La-Neuve, Belgium

85 % of the publications on salient object detection models, the authors use one gold standard metric (F-score from Precision-Recall curve) to compare their models with other state-of-the-art models.

8.1.1 Location-Based Metrics: Focus on Location of Salient Regions and Binary Masks

For all location-based metrics, we retrieve the concept and terminology from a confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) that compare the predicted results (saliency map) with the reference results (binary mask). Therefore, the saliency maps need to be converted into a binary map. To do that, several thresholds are defined. Table 8.1 illustrates their definition.

From this concept, two metrics can be computed. A new one, the F-score, calculates a score from Precision-Recall and the Area Under the receiver operating Characteristic (AUC), like with eye-tracking-based metrics, computes a score from the False and True Positive Rate. All these notions will be described in detail below.

A third metric called MAE exists in the literature as described in [3]. The purpose is to consider the true negative (TN) when a pixel is correctly marked as non-salient.

Finally, recently, we find some variations of F-score which propose a weighted calculation of Precision and Recall. The objective is to provide a more reliable evaluation. In [4], the authors start by identifying three causes of inaccurate evaluation: interpolation flaw, dependency flaw, and equal-importance flaw. By amending these three assumptions, they propose a new reliable measure available for images.

8.1.1.1 F: F-score from Precision-Recall (2009)

Authors R. Achanta, S. Hemami, F. Estrada and S. Susstrunk [5].

Description Many authors like [6–9] also used the F-score (Fig. 8.1) metric to compare saliency maps and binary masks. Precision is the number of relevant points compared with the total number of points found [Eq. 8.1 (left)]. Recall is the number of relevant points compared with the total number of important points in the reference [Eq. 8.1 (right)].

Table 8.1 Definitions of four concepts to compute Precision/Recall and FPR/TPR

	Reference results	
Predicted results	TP: Correct result	FP: Unexpected result
	FN: Missing result	TN: Correct absence of result

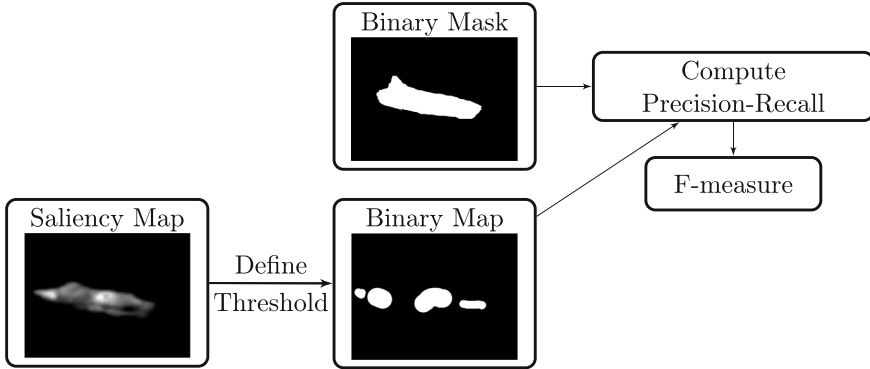


Fig. 8.1 F-score between saliency map and binary mask

$$Precision = \frac{tp}{tp + fp} \quad Recall = \frac{tp}{tp + fn} \quad (8.1)$$

A usual way to combine Precision and Recall is to use the F-score defined as in Eq. 8.2 where as suggested by several salient object detection benchmarks [5], β^2 is set to 0.3 to give more importance to the Precision value.

$$F\text{-score} = \frac{(1 + \beta^2) * Precision * Recall}{\beta^2 * Precision + Recall} \quad (8.2)$$

8.1.2 AUC: Area Under the ROC Curve (2011)

Authors J. Li, L. Martin, A. Xiangjing and H. Hangen [10].

Description Many authors like [11, 12] also used AUC metric (Fig. 8.2) to compare saliency maps and binary masks. The true positive rate, also called sensitivity, measures, as the Recall, the proportion of true positive under all the positive reference results [Eq. 8.3 (left)]. The false positive rate measures the proportion of false positive under all the negative reference results [Eq. 8.3 (right)].

$$TPR = \frac{tp}{tp + fn} \quad FPR = \frac{fp}{fp + tn} \quad (8.3)$$

A usual way to combine them is to plot the true positive rate (TPR) vs. false positive rate (FPR) to form an ROC curve. Then, the area under the ROC can be computed.

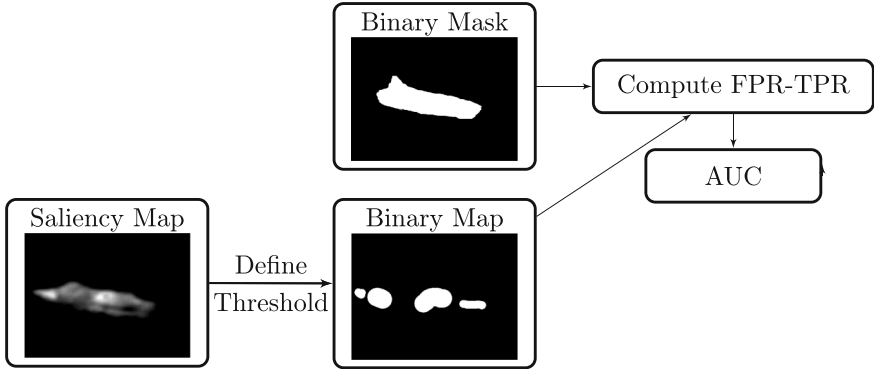


Fig. 8.2 AUC between saliency map and binary mask

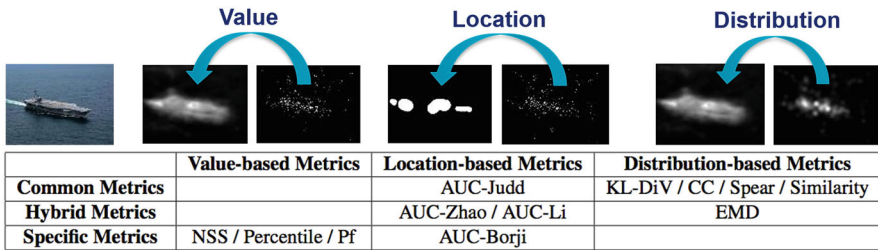


Fig. 8.3 Eye-tracking-based static metrics families

8.2 Literature Review of Metrics for Eye Tracking

Contrary to salient object detection validation, a gold standard metric does not exist for eye-tracking maps and a lot of metrics have been proposed to validate eye-tracking saliency models. Therefore, we propose here a taxonomy to classify them. The classification is related to the nature of the similarity metric and can be divided into three categories: value-based metrics which focus on saliency map values at eye gaze positions, distribution-based metrics which focus on saliency and gaze statistical distributions, and location-based metrics which focus on location of salient regions at gaze positions (Fig. 8.3).

All these metrics will be described in detail in this section and will be used in the next chapter to study their similarity. They take two distributions as input: the prediction (noted SM for Saliency Map) and the ground truth (noted FM for Fixation Map).

It is important to note that a discrete fixation map is used for location-based and value-based metrics, while a continuous one is used for distribution-based metrics. The continuous fixation map is obtained by convolving the fixation map with a 2D Gaussian function. The parameters of this function depend on the database.

8.2.1 Value-Based Metrics: Saliency Map Values at Eye Positions

This first category of metrics compares values or amplitudes of the saliency maps with the corresponding eye fixations maps.

8.2.2 NSS: Normalized Scan Path Saliency (2005)

Authors R. Peters, A. Iyer, L. Itti and C. Koch [13].

Description The idea is to quantify the saliency map values at the eye fixation locations and to normalize it with the saliency map variance (Fig. 8.4):

$$NSS(p) = \frac{SM(p) - \mu_{SM}}{\sigma_{SM}} \quad (8.4)$$

where p is the location of one fixation and SM is the saliency map which is normalized to have a zero mean and unit standard deviation. Indeed, the NSS score should be decreased if the saliency map variance is large or if all values are globally similar (small difference between fixation values and mean) because it shows that the saliency model will not be very predictive, while it will precisely point a direction of interest if the variance is small or if the difference between fixation values and means is high.

The NSS score is the average of $NSS(p)$ for all fixations:

$$NSS = \frac{1}{N} * \sum_{p=1}^N NSS(p) \quad (8.5)$$

where N is the total number of eye fixations.

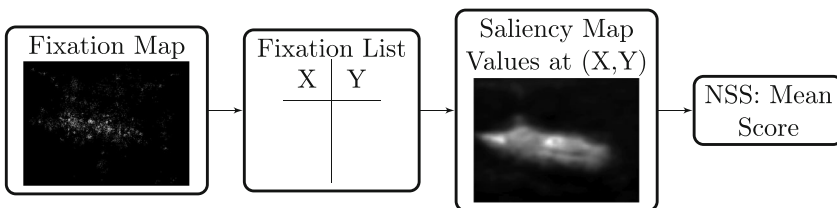


Fig. 8.4 NSS between saliency and fixation map

8.2.3 PF: Percentage of Fixations into the Salient Region (2006)

Authors A. Torralba, A. Oliva, M. Castelhana and J. Henderson [14].

Description Its purpose is to measure the percentage of fixations into the salient region (Fig. 8.5). In a first step, saliency maps are thresholded at $T = 0.8$ where the saliency is normalized between 0 and 1. The threshold is set so that the selected image region occupies a fixed proportion of the image size. In a second step, the percentage of fixations in this area is computed and called PF.

8.2.4 P: Percentile (2008)

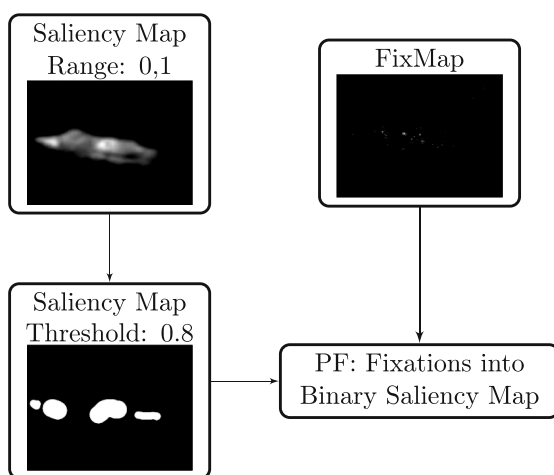
Authors R. Peters and L. Itti [15].

Description The percentile metric (Fig. 8.6) is, for each pixel p on the eye fixation map, a ratio between the number of pixels in the saliency map with values smaller than the one corresponding to pixel p from the eye fixation map and the total number of pixels (Eq. 8.6).

$$P(p) = \frac{|x \in X : SM(x) < SM(p)|}{|SM|} \quad (8.6)$$

where X is the set of all pixels of the saliency map SM , p is the location of one eye fixation, and $|SM|$ indicates the total number of pixels. Like in the case of NSS, the global percentile score is the average of $P(p)$ for all the eye fixations.

Fig. 8.5 PF between saliency and fixation map



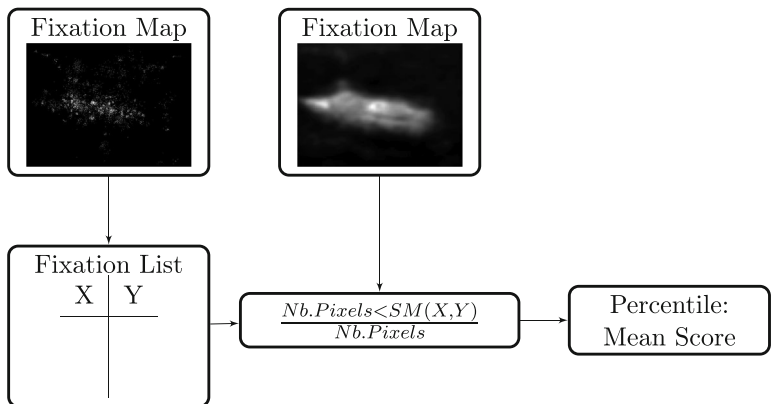


Fig. 8.6 P between saliency and fixation maps

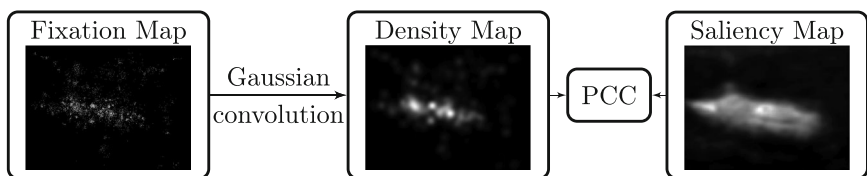


Fig. 8.7 Pearson's correlation coefficient between saliency and density maps

8.2.5 Distribution-Based Metrics: Focus on Saliency and Gaze Statistical Distributions

In the literature, there are two kinds of distribution-based metrics. Those which compute a similarity between two distributions and those which compute a dissimilarity. Moreover, some metrics which are not a distance are nonsymmetric. It means that by first considering the saliency map (SM) as the first input and secondly the fixation map (FM) as the first input, the results are not the same. This should be taken into account for the comparison. Two dissimilarity and three similarity metrics are proposed and described in the following subsections.

8.2.6 PCC: Pearson's Correlation Coefficient (2004)

Authors N. Ouerhani, R. Von Wartburg, H. Hugli and R. Muri [16].

Description The Pearson's correlation coefficient (Fig. 8.7) also named linear correlation coefficient was first used in [16] as a metric. Other authors also used it such as in [17]. The linear CC output range is between -1 and 1 . When the

correlation value is close to -1 or 1 , there is almost a perfect linear relationship between the two variables:

$$CC = \frac{cov(SM, FM)}{\sigma_{SM} * \sigma_{FM}} \quad (8.7)$$

8.2.7 KLD: Kullback-Leibler Divergence (2004)

Authors U. Rajashekar, L. Cormack and A. Bovik [18].

Description The Kullback-Leibler divergence (Fig. 8.8) is a commonly used metric to estimate an overall dissimilarity between two distributions. Many authors like [19] and [17] also used this metric to compare saliency maps with human eyes fixations. The KLD is a measure of the information lost when the saliency maps probability distribution (called SM) is used to approximate the human eye fixation map probability distribution (called FM).

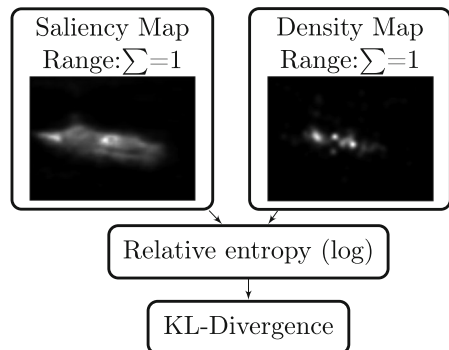
$$KLD = \sum_{x=1}^X FM(x) * \log \left(\frac{FM(x)}{SM(x) + \epsilon} + \epsilon \right) \quad (8.8)$$

where X is the number of pixels and ϵ is a small constant to avoid $\log(0)$ and division by zero. SM and FM distributions are both normalized as in Eq. 8.9.

$$SM(x) = \frac{SM(x)}{\sum_{x=1}^X SM(x) + \epsilon} \quad FM(x) = \frac{FM(x)}{\sum_{x=1}^X FM(x) + \epsilon} \quad (8.9)$$

When the two maps are strictly equal, the KL-divergence value is zero.

Fig. 8.8 KLD between saliency and density maps



8.2.8 SCC: Spearman’s Correlation Coefficient (2011)

Authors A. Toet [20].

Description The Spearman’s rank correlation coefficient metric [20] is defined as the CC metric (Eq. 8.7) but on ranked variables (Fig. 8.9). This can be understood as a nonlinear correlation. Toets uses this metrics in [20] to evaluate 13 models.

8.2.9 EMD: Earth Mover’s Distance (2012)

Authors T. Judd, F. Durand and A. Torralba [21].

Description Earth Mover’s distance (Fig. 8.10) is a measure of the distance between two probability distributions over a region. Judd [21] used this metric in her benchmark which is now available online. She uses a fast implementation of EMD provided by Pele and Werman [22, 23], but without a threshold. It computes the minimal cost to transform the probability distribution of the saliency maps SM into the one of the human eye fixations FM .

Fig. 8.9 Spearman’s correlation coefficient between saliency and density maps

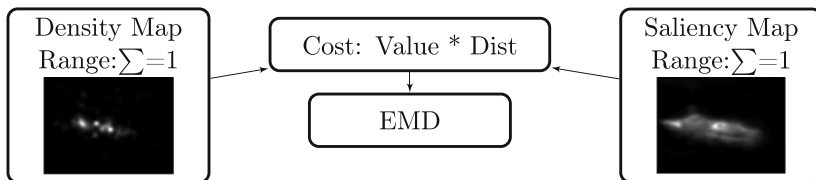
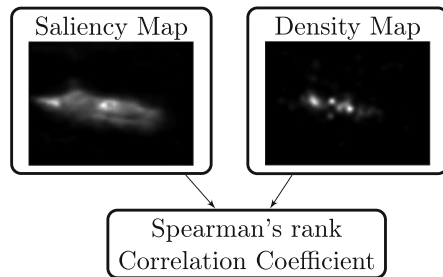


Fig. 8.10 EMD between saliency and density maps

$$EMD = \left(\min_{f_{ij}} \sum_{i,j} f_{ij} d_{ij} \right) + \left| \sum_i FM_i - \sum_j SM_j \right| \max_{i,j} d_{ij}$$

$$s.t. f_{ij} \geq 0, \sum_j f_{ij} \leq FM_i, \sum_i f_{ij} \leq SM_j, \tag{8.10}$$

and

$$\sum_{i,j} f_{ij} = \min \left(\sum_i FM_i - \sum_j SM_j \right)$$

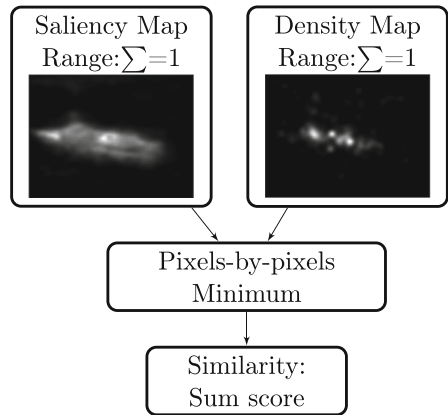
where each f_{ij} represents the amount transported from the i_{th} supply to the j_{th} demand. d_{ij} is the ground distance between bin i and bin j in the distribution. A larger EMD indicates a larger overall difference between the two distributions. An EMD of zero indicates that two distributions are the same.

8.2.10 S: Similarity (2012)

Authors T. Judd, F. Durand and A. Torralba [21].

Description The similarity metric [21] also uses the normalized probability distributions of the saliency map SM and human eye fixation map FM . The similarity is the sum of the minimum values at each point in the distributions (Fig. 8.11). Mathematically, the similarity between two maps SM and FM is

Fig. 8.11 S between saliency and density maps



$$S = \sum_{x=1}^X \min(SM(x), FM(x)) \tag{8.11}$$

where $\sum_{x=1}^X SM(x) = \sum_{x=1}^X FM(x) = 1$.

A similarity score of one indicates that the distributions are the same. A similarity score of zero indicates that they do not overlap at all and are completely different.

8.2.11 IG: Information Gain (2015)

Authors M. Kümmerer, T. S. Wallis, and M. Bethge [24].

Description The information gain metric [24] computes how much better a saliency model predicts human fixations on a given image than an image-independent baseline. The difference between the information gains provides a saliency metric.

As shown in Fig. 8.12, both the saliency map and the density map are divided by the center effect (which is here the image-independent baseline). The Log images are then multiplied by the Density Map giving the IG. The IG will quantify how much information bring the Density Map and the Saliency Map compared to the center effect. The difference shows then a comparison between the Saliency Map and Density Map. This difference can also be shown as an image (Diff to IG on the right) showing in red the regions where the saliency model overestimates the density map and in blue the regions where it underestimates it.

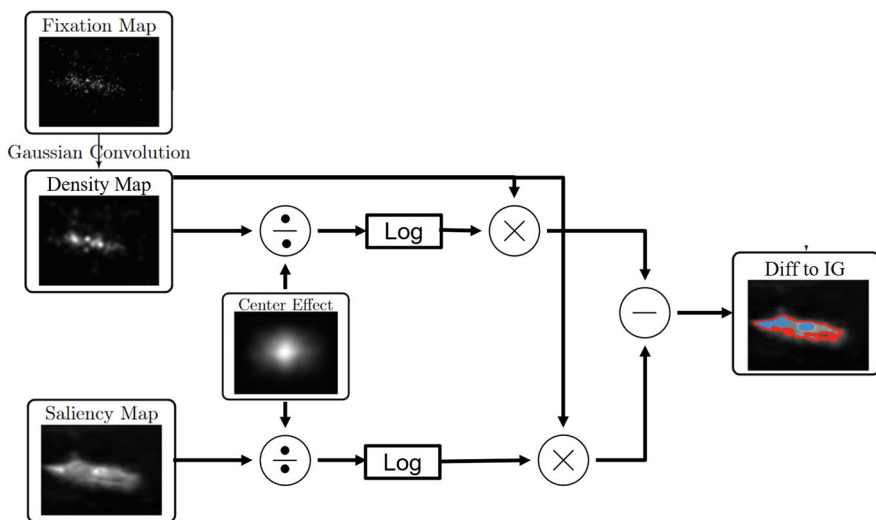


Fig. 8.12 IG between saliency and density maps

8.2.12 *Location-Based Metrics: Focus on Location of Salient Regions at Gaze Positions*

Location-based metrics are very popular to evaluate saliency maps. They are based on the notion of Area under the Receiver Operating Characteristic curve coming from signal detection theory. Four main different implementations are available dealing with some limitations of the classical approach.

8.2.13 *nAUC: Normalized Area Under the ROC Curve (2011)*

Authors Q. Zhao and C. Koch [25].

Description Zhao used a normalized AUC (Fig. 8.13). The idea is that saliency algorithms perform less well (on average) than the area under the ROC curve coming from intersubject variability for each image. Zhao computes an ideal AUC by measuring how well the human fixations of one subject can be predicted by those of the other $n - 1$ subjects, iterating over all n subjects and averaging the result with an upper limit of one. Finally, the AUC of the saliency map is normalized by this ideal AUC.

8.2.14 *pAUC: Post-processing for Area Under the ROC Curve (2011)*

Authors J. Li, L. Martin, A. Xiangjing and H. Hangen [10].

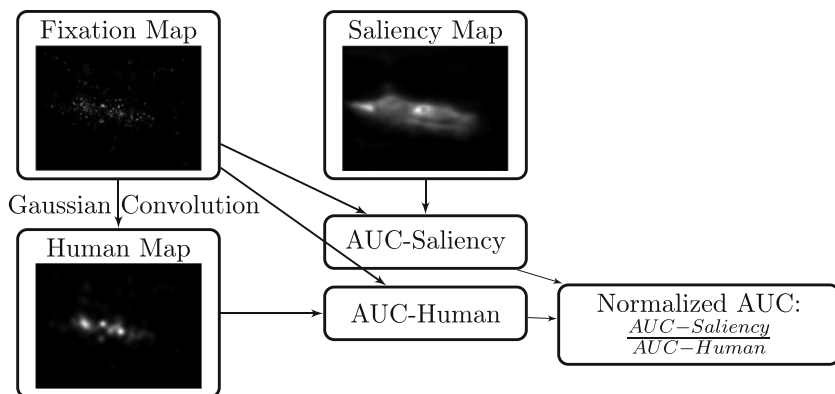


Fig. 8.13 nAUC between saliency and density maps

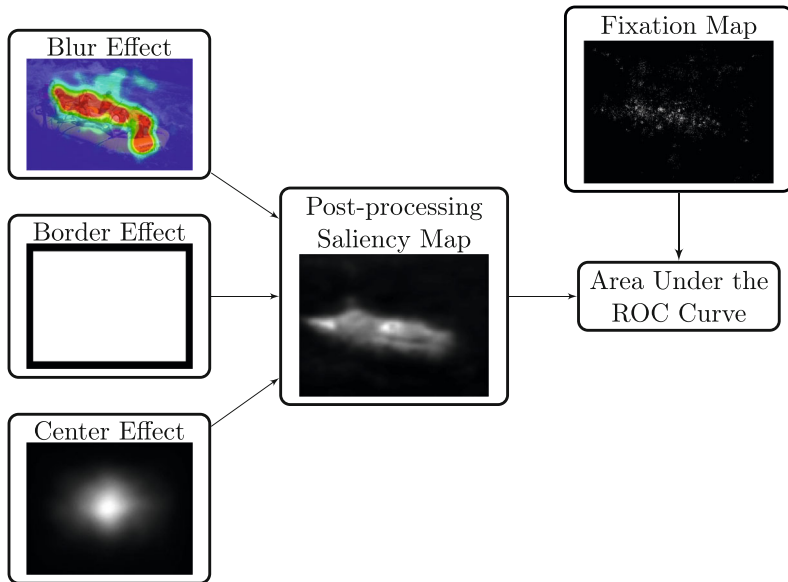


Fig. 8.14 pAUC Between saliency and density maps

Description Li set the border cuts for all models to be of equal size and avoids in that way to artificially increase the AUC scores for the models which already do this pre-processing in comparison with those which do not (Fig. 8.14). The border cut post-processing affecting the fairness during the assessment is thus eliminated.

8.2.15 *hAUC: Hit Rate for Area Under the ROC Curve (2012)*

Authors T. Judd, F. Durand and A. Torralba [21].

Description Judd proposed another version of AUC to validate saliency models (Fig. 8.15). First, fixation pixels were counted once and the same number of random pixels is extracted from the saliency map. For one given threshold, saliency pixels can be seen as a classifier, with all points above threshold indicated as “fixation” and all points below threshold as “background.”

For any particular value of the threshold, there is some fraction of the actual fixation points which are labeled as True Positives (TP), and some fraction of points which were not fixation but labeled as False Positive (FP). This operation is repeated one hundred times. Then the ROC curve can be drawn and the Area Under the Curve (AUC) computed. An ideal score is 1.0, while random classification provides 0.5.

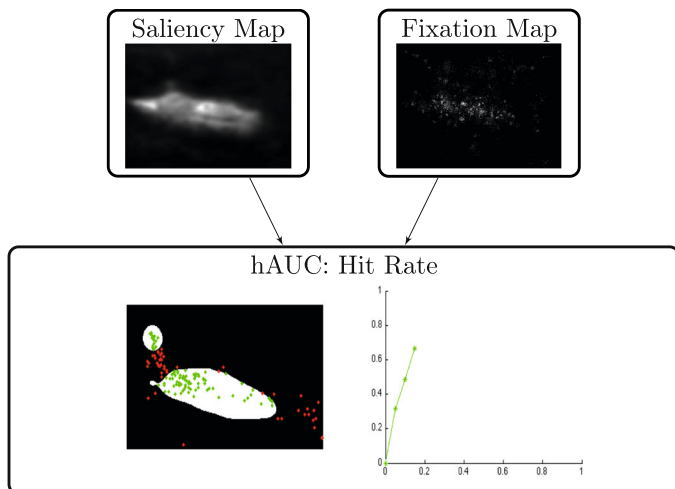


Fig. 8.15 hAUC between saliency and density maps

8.2.16 *sAUC: Shuffled Area Under the ROC Curve (2012)*

Authors A. Borji, D. Sihite and L. Itti [26].

Description Borji applied to saliency maps validation a suitable AUC metric called *shuffled AUC*. In his classical AUC, saliency map values from random points from the image are addressed to create a binary mask. In the *shuffled AUC* metric, saliency values and fixations from another image (instead of random) of the same dataset are taken into account (Fig. 8.16). In that way, the more or less centered distribution of the human fixations of the database is taken into account in the AUC computation. This point is important because the AUROC scores can dramatically increase if a saliency map is weighted by a centered Gaussian. Indeed, human eye fixations are rarely near the edges of general test images, and the amateur photographer often places salient objects in the image center.

8.3 Literature Review of Dynamic Metrics

Here we propose a short synthetic overview of the dynamic metrics which try to measure a similarity between a scan path generated by an attention model and an eye scan path based on [2].

Indeed, some visual attention models can provide a scan path in addition to the static saliency map. An example can be seen in Fig. 8.17 where there is an image with a saliency map and a provided scan path with fixations (circles) and saccades

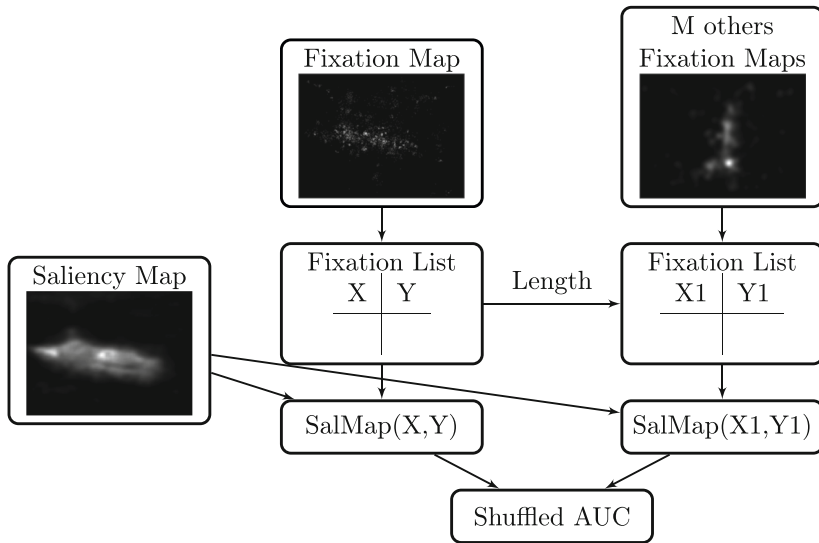


Fig. 8.16 sAUC between saliency and fixation maps

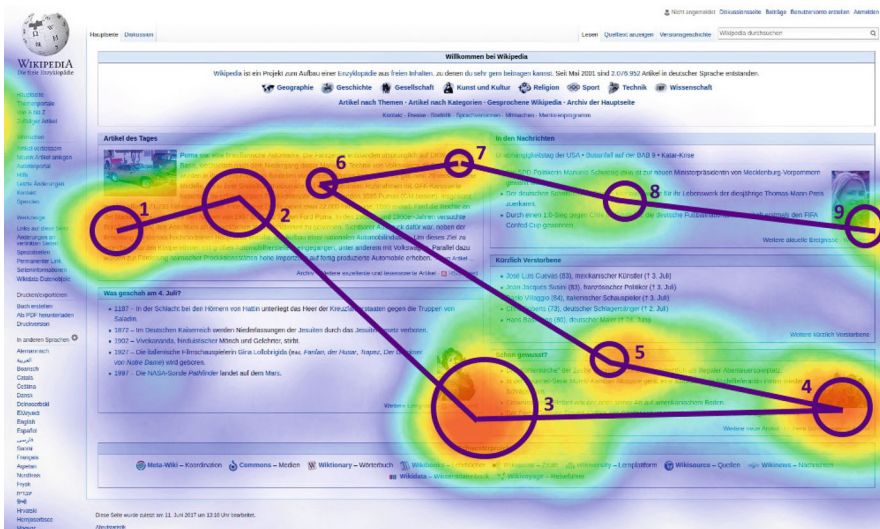


Fig. 8.17 Eye scan path superimposed to saliency map. The scan path is made from sequential saccades and fixations with a time spent on each fixation proportional to the circle size

(lines) having a given dynamics (numbers show the fixation order). The evaluation of scan paths needs specific metrics or methods. We can organize those metrics in four big classes: distance-based, time series/vectors, recurrence and density maps metrics.

8.3.1 *Distance-Based Metrics*

We can find here classical Euclidean distances as the sum of the distance between the pairs of fixations between each scan path (if of equal number of fixations).

The Hausdorff distance [27] will measure the bigger difference between two subsets (here the spatial scan-paths curves). This measure will also lose the sequential information between the two scan paths.

The Mannan distance [28] will check each fixation of one scan path compared to all fixations of the other giving an interesting measure but losing the sequential nature of the scan path. Eye analysis [29] is an improvement of Mannan distance tending to provide at some extent some sequential information.

The Frechet distance [30] will compare in a sequential way two curves by always keeping the shortest distance between them (acceleration of the speed on the longer curve). This gives both spatial and sequential information about the scan paths. However there is no straightforward way to include fixation time in the algorithm.

The Levenshtein distance [31] was initially made to compare string sequences based on the minimum number of deletions, insertions or substitutions needed to turn one sequence into the other. For scan paths, this metric needs first to turn fixation coordinates into strings. A discrete grid is superimposed on the scan path, and strings are assigned to different fixations. While this metric is good for the sequential information, it might be incorrect for the spatial information as depending on the bin in the grid sometimes close locations can have different strings. Also the fixation durations are not taken into account. Some issues of the Levenshtein distance are fixed in an improved version called ScanMatch [32]. Here the interesting point is that fixations and fixation time are taken into account for the final result. However, the main issue as Levenshtein distance is still there as the exact position of the fixation depends on the grid bins.

8.3.2 *Time Series and Vector Metrics*

Dynamic time warping (DTW) [33] is known to measure in a dynamic way the distance between two time series (with possibly different lengths). It considers both scan path (1) fully, (2) without jumping in time and the possibility to miss specific patterns and (3) sequentially.

Time Delay Embedding (TDE) [34] takes into account the idea of using subparts of the scan path, compute distance metrics on those subparts and then average those metrics. The size of those subparts is of crucial importance to decrease possible noise/errors in the metric computation.

The MultiMatch approach [35] is based on five saccades features: shape, length, direction, position, and duration. Depending on the application, one of the five measures will be more interesting than the others and MultiMatch let people choose the most interesting one. MultiMatch implies a scan-path simplification (fixations

grouping, etc.) which might have an impact on the final metrics. Scan paths are then aligned and vector similarities are applied and normalized:

- Length similarity: absolute difference in the amplitude of aligned saccades pairs
- Position similarity: Euclidean distances between aligned fixations pairs
- MultiMatch similarity: vector difference between aligned saccade pairs
- Direction similarity: angle difference between aligned saccades pairs
- Duration similarity: absolute difference in fixation durations of aligned fixations pairs

8.3.3 *Recurrence Analysis Metrics*

In recurrence quantification analysis (RQA) [36] let us analyze a scan path but can be extended to scan-paths comparison. The idea is to build a recurrence matrix where two fixations are recurrent if they are close enough (closer than a given threshold). To measure this recurrence, several previous metrics can be used like Euclidean, ScanMatch, or Levenshtein distances. From this recurrence matrix, four measures can be computed:

- Cross-recurrence (REC): percentage of fixation matching between the two scan paths
- Determinism (DET): percentage among REC which is on the matrix diagonal to provide information on the local fixation sequential matching between scan paths
- Center of recurrence mass (CORM): distance of REC from the diagonal barycenter to provide information on the global sequential matching between scan paths
- Laminarity (LAM): the number of REC where the fixation time is not matching (long fixation on one scan path versus short on the other one)

Those four measures provide interesting both local and global and spatial and sequential on the two scan paths. However, the scan paths need to have equal length.

8.3.4 *Density Maps Metrics*

One straightforward approach of measuring the accuracy of a set of scan paths is to convert the predicted eye fixations into a saliency map and use already existing metrics for saliency assessment. This will exclude the issues due to specific scan-path metrics and also their specific problems. However, the dynamic nature of scan path is lost in the case of using this approach.

Kümmerer et al. [37] proposed another approach which is based on a saliency (or priority) map but still provides dynamic information. Indeed, all models which provide eye scan-path data will at some point compute a priority map which will be used to find the possible position of the next fixation. The author proposes to

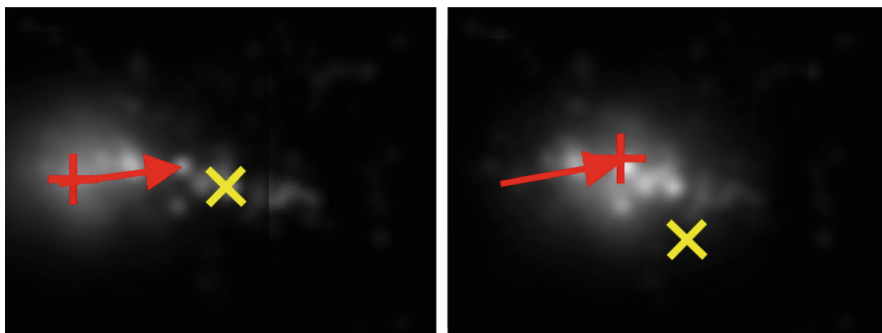


Fig. 8.18 Dynamic priority maps metric

use this map and the next real fixation (from the eye-tracking data) to compute a metric at each new fixation. The final metric will be the average metric for all the eye fixations recorded for this image in the dataset.

In Fig. 8.18 we can see at the left the cross which is the initial position of the fixation. The “x” is the next eye fixation. The arrow shows the next fixation proposed by the model. On the right, the second fixation of the model as a cross and an “x” for the next eye fixation. At each step different metrics adapted from those for saliency maps can be measured:

- AUC metric: Here, the priority map has several thresholds, and at some rank, the real next eye fixation will match.
- NSS metric: Here, the priority map value at the eye fixation location is used.
- IG metric: Information gain between the priority map and the eye fixation can also be computed.

8.4 Discussions and Conclusions

There are a large variety of metrics in the literature which provide a score between saliency map and ground-truth data which have been processed into a two-dimensional map. These metrics depend on the nature of the ground truth and what authors want to measure: amplitude, location, distribution, or all three.

It shows the importance of choosing appropriate metrics for a validation framework. The authors need to clarify why they choose these metrics. Moreover, the framework validation needs a preliminary study to investigate the relevance of the chosen metrics mix.

We also made a short list of scan-path validation metrics from distance metrics to density-based metrics and going through vector/time series and recurrence metrics. You can find more in [1, 2] or [37] on scan-paths assessment. This one requires

taking into account a number of factors, such as the temporal dimension or the alignment procedure.

8.5 Summary

- For object detection validation, all the metrics are based on the notion of true positives (TP)/false positives (FP), true negatives(TN)/false negatives (FN), to compute F-score and weighted F-score.
- For eye-tracking ground truth, there are dozens of metrics (amplitude-based, location-based, distribution-based). However it is not mandatory to compute them all for fair validation. In Chap. 9 we will see that around 3 different metrics are enough.
- For scan-path ground truth, there are also dozens of metrics (distances, density, vector/time series, recurrence). It is not easy to find one which is better on both spatial and sequential planes. However, the arrival of metrics adapted from eye-tracking metrics such as IG opens new possibilities in a fair comparison of models providing dynamical saccades sequences. Another track is about a fair fixation time duration validation as some new models can provide fixation locations but also fixations' durations.

References

1. Le Meur, O., & Baccino, T. (2013). Methods for comparing scanpaths and saliency maps: Strengths and weaknesses. *Behavior research methods*, 45(1), 251–266.
2. Fahimi, R., & Bruce, N. D. (2021). On metrics for measuring scanpath similarity. *Behavior Research Methods*, 53, 609–628.
3. Borji, A., Cheng, M.-M., Hou, Q., Jiang, H., & Li, J. (2019). Salient object detection: A survey. arXiv preprint arXiv:1411.5878.
4. Margolin, R., Zelnik-Manor, L., & Tal, A. (2014). How to evaluate foreground maps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 248–255). IEEE.
5. Achanta, R., Hemami, S., Estrada, F., & Susstrunk, S. (2009). Frequency-tuned salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009* (pp. 1597–1604). IEEE.
6. Cheng, M.-M., Zhang, G.-X., Mitra, N. J., Huang, X., & Hu, S.-M. (2014). Global contrast based salient region detection. In *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 409–416). IEEE.
7. Perazzi, F., Krahenbuhl, P., Pritch, Y., & Hornung, A. (2012). Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 733–740). IEEE.
8. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., & Shum, H.-Y. (2010). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2), 353–367.
9. Cheng, M.M., Warrell, J., Lin, W.-Y., Zheng, S., Vineet, V., & Crook, N. (2013). Efficient salient region detection with soft image abstraction. In *2013 IEEE International Conference on Computer Vision (ICCV)* (pp. 1529–1536). IEEE.

10. Li, J., Levine, M., An, X., He, H. (2011). Saliency detection based on frequency and spatial domain analyses. In *Proceedings of the British Machine Vision Conference* (pp. 86.1–86.11). BMVA Press. <https://doi.org/10.5244/C.25.86>
11. Borji, A., Sihite, D. N., & Itti, L. (2012). Salient object detection: A benchmark. In *Computer Vision—ECCV 2012* (pp. 414–429). Springer
12. Borji, A. (2014). What is a salient object? A dataset and a baseline model for salient object detection.
13. Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision research*, *45*(18), 2397–2416.
14. Antonio Torralba, Aude Oliva, M. C. & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search. *Psychological Review*, *113*(4), 766–786.
15. Peters, R. J., & Itti, L. (2008). Applying computational tools to predict gaze direction in interactive visual environments. *ACM Transactions on Applied Perception (TAP)*, *5*(2), 9 (2008)
16. Ouerhani, N., Von Wartburg, R., Hugli, H., & Muri, R. (2004). Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, *3*(1), 13–24.
17. Le Meur, O., Le Callet, P., Barba, D., et al. (2007). Predicting visual fixations on video based on low-level visual features. *Vision research*, *47*(19), 2483–2498.
18. Rajashekar, U., Cormack, L. K., & Bovik, A. C. (2004). Point-of-gaze analysis reveals visual search strategies. *Proceedings of SPIE*, *5292*, 296–306.
19. Tatler, B. W., Baddeley, R. J., Gilchrist, I. D., et al. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659.
20. Toet, A. (2011). Computational versus psychophysical bottom-up image saliency: A comparative evaluation study. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(11), 2131–2146.
21. Judd, T., Durand, F., & Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. MIT tech report.
22. Pele, O., & Werman, M. (2008). A linear time histogram metric for improved sift matching. In *Computer Vision—ECCV 2008* (pp. 495–508). Springer
23. Pele, O., & Werman, M. (2009). Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision* (pp. 460–467). IEEE
24. Kümmerer, M., Wallis, T. S., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, *112*(52), 16054–16059.
25. Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of Vision*, *11*(3), 9.
26. Borji, A., Sihite, D. N., & Itti, L. (2013). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, *22*(1), 55–69.
27. Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(9), 850–863.
28. Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, *10*(3), 165–188.
29. Mathôt, S., Cristino, F., Gilchrist, I. D., & Theeuwes, J. (2012). A simple way to estimate similarity between pairs of eye movement sequences. *Journal of Eye Movement Research*, *5*(1), 1–15.
30. Aronov, B., Har-Peled, S., Knauer, C., Wang, Y., & Wenk, C. (2006). Fréchet distance for curves, revisited. In *Algorithms—ESA 2006: 14th Annual European Symposium, Zurich, Switzerland, September 11–13, 2006. Proceedings 14* (pp. 52–63). Springer

31. Privitera, C. M., & Stark, L. W. (2000). Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9), 970–982.
32. Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, 42, 692–700.
33. Berndt, D. J., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (pp. 359–370).
34. Wang, W., Chen, C., Wang, Y., Jiang, T., Fang, F., & Yao, Y. (2011). Simulating human saccadic scanpaths on natural images. In *CVPR 2011* (pp. 441–448). IEEE.
35. Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior Research Methods*, 44, 1079–1100.
36. Anderson, N. C., Bischof, W. F., Laidlaw, K. E., Risko, E. F., & Kingstone, A. (2013). Recurrence quantification analysis of eye movements. *Behavior Research Methods*, 45, 842–856.
37. Kümmerer, M., & Bethge, M. (2021). State-of-the-art in human scanpath prediction. arXiv preprint arXiv:2102.12239.