


Chapter 9

Study of Parameters Affecting Visual Saliency Assessment



Matei Mancas  and Nicolas Riche 

The computational modeling of visual attention has been developed and expanded considerably during the past 20 years. Many different saliency models are now available online (for still images and videos). At the same time, many popular image-video datasets with human gaze data or binary masks have been released to evaluate saliency models with commonly used evaluation metrics. The new challenges and future directions for this field are therefore to establish evaluation protocols and saliency benchmarks.

Although some evaluation studies (such as [1–3], and [4]) and online benchmarks (like [5] and [6]) have already been proposed and are major contributions, a key underlying issue is *how to fairly evaluate all these models*. In this chapter, we investigate this question with an evaluation, divided into four experiments, leading to the proposition of a new evaluation framework. Each experiment is based on an important aspect of visual saliency assessment in real-life images and is extended for videos in the validation framework. There are four questions that we will be carefully considered:

1. What are the differences between eye fixations and manually segmented salient regions?
2. What is the relation between model performances and the properties (e.g., the size) of the salient regions into images?
3. What is the effect of saliency map post-processing?
4. Is one metric enough to evaluate a saliency model?

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

First of all, there are mainly two ground-truth categories to assess a saliency map: human eye fixations obtained using an eye tracker device and manually segmented and labeled salient regions. In this chapter, we analyze the difference and the coherence between them. The second aspect of this chapter is about different categories of salient regions. Are saliency models equally efficient in predicting human gaze on three categories of salient regions: large, intermediate, and small? This is an important issue as real-life objects and scenes contain a very wide range of object sizes. The third experiment is about saliency map post-processing. Which ones increase the score of a saliency map? Finally, various evaluation measures exist to compare saliency and ground-truth maps. We study the redundancy of these metrics and propose, among them, three metrics which should be used to obtain a complete assessment of saliency model performance.

Statistical analysis is used here to answer each of these four questions.

9.1 Experiment 1: Effects of Ground-Truth

9.1.1 Goal

Nowadays, databases are coming with two ground-truths: eye fixations or labeled objects. Some databases have the interest of providing both approaches for the same set of images. Some saliency models will better model eye fixations while others focus on object detection and segmentation and are assessed with region-based labeled objects. The main idea of this first experiment is to assess the coherence between the region-based and eye fixation-based ground-truths.

9.1.2 Method

Database and Ground-Truth The database used here has been published by Li et al. [7] and provides both region ground-truth (human labeled) and eye fixation ground-truth (collected with an eye tracker). In this experiment, we use the whole database containing 235 color images.

Models Twelve state-of-the-art “classical” models (with handcrafted features and no deep learning) from a mix of eye tracking (ET)-based (80%) and object detection-based (20%) algorithms are used in this experiment. Some of the models are explained more in depth in the previous chapters, and a taxonomy is proposed in [8]. We use a wide range of recently published saliency models. FSM model [9] represents the cognitive approach. SUN [10] and SDLF [11] are Bayesian models. AIM [12], DVA [13], and RARE [1] are into the information theory category. SR [14], PFT [15], QDCT [16], SSAFD [17], and FTSD [18] use a spectral analysis

approach to compute their saliency map. Finally, AWS [19] that does not fit into Borji’s taxonomy represents the *other models* category.

Metrics In this study, the pAUC (Post-processing for Area Under the ROC Curve (2011) [7]) metric has been chosen. This metric can be applied to both eye tracking-based and region-based ground-truths and mainly measures the eye fixation or region locations.

Kendall’s W concordance measure is used for the statistical analysis. Kendall’s W concordance measure [20] is an effect size measure. It defines how big the discordance between two distributions is. Indeed, while common significant tests only assess if there is enough evidence to determine whether the null hypothesis is likely between two or more groups, they do not provide information about the size of this effect. The effect size measures by how much the detected effect is significant in practice; in other words, it defines, in our case, how big the discordance between the region-based and eye fixation-based ground-truths is.

It is defined as

$$W = \frac{12 * S}{m^2 * (n^2 - n)}, \quad (9.1)$$

where n is the number of models and m the number of metrics. So here $n = 12$ and $m = 2$ (pAUC on both ground-truths). S , the sum of squared deviations, is defined as

$$S = \sum_{i=1}^n (R_i - \bar{R})^2, \quad (9.2)$$

where R_i is the ranking given to model i . A ranking as used here replaces the mean score of each model based on one metric by the assignment of labels (first, second, third, etc.). \bar{R} is the mean value of those rankings.

Kendall’s W concordance is a coefficient measuring the degree of agreement between metrics. The value ranges from 0 (no agreement between model ranks) to 1 (full agreement, the same models ranking). Furthermore, some rules of thumb are provided to allow the researcher to interpret this measure as depicted in Table 9.1 [20].

However, in our study, the ranking range of 1–12 is small; therefore higher thresholds are required to keep on the interpretation. That is why we decided to

Table 9.1 Interpretation of Kendall’s W coefficient

Kendall’s W	Interpretation	Confidence in ranks
0.5	Moderate agreement	Fair
0.7	Strong agreement	High
0.9	Unusually strong agreement	Very high
1	Complete agreement	Very high

Table 9.2 Interpretation of Kendall’s W coefficient on mean scores

W	Interpretation	Rank Confidence
0.7	Moderate agreement	Fair
0.85	Strong agreement	High
0.93	Very Strong agreement	High
0.98	Unusually strong agreement with 2 or 3 switched models	Very high
0.99	Unusually strong agreement with one or two switched models	Very high
1	Complete agreement	Very high

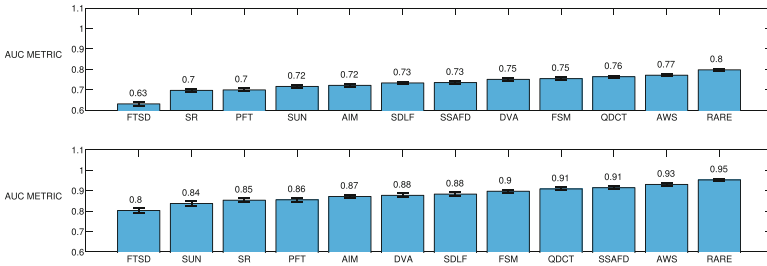


Fig. 9.1 Row 1: Eye fixation mean score for all the models on the whole database with their standard deviations. Row 2: Labeled regions mean score for all the models on the whole database with their standard deviations. The higher the pAUC is, the better the model

be much more selective than in Table 9.1: We interpret the Kendall coefficient as in Table 9.2. Indeed this interpretation shows that $W = 0.98$ means that only two or three models are switched between the rankings.

9.1.3 Results

The mean results of pAUC metric for each model are computed in Fig. 9.1 for the entire database and both ground-truths.

After this first score computation, a ranking-based statistical test is required. Considering our design, the 95% Confidence Interval (CI) Friedman test allows to respond to the H_0 hypothesis: Are the ranking of the individual results provided by the different models coherent between both ground-truth performance evaluations? As explained above, there is no specific effect size measure in case of the 95% CI Friedman test (only a binary response). Therefore, we use the presented Kendall’s W concordance measure, which basically fulfills our needs (response between 0 and 1).

As shown in Table 9.3, although differences between eye fixation and region results are significant (Friedman test), Kendall’s concordance between both ground-truths is very good. This means that there is a difference between both rankings, but

Table 9.3 Concordance based on Friedman test and Kendall Coefficient between eye fixation and region results

	Friedman test (p-value)	Kendall's concordance W
pAUC	≈ 0	0.82

the size of this difference based on Kendall's W coefficient is relatively small. In other words, if models have good results with one ground-truth, it is quite unlikely that these models completely fail with the other ground-truth except due to statistical fluctuation. A saliency model that is good in predicting human eye fixations will remain good in predicted human labeled regions and conversely.

These results depend on the experimental design. In our case, one database, 12 saliency models, and one metric have been chosen. However, the same experiment was conducted in our paper [21] based on another metric (NSS) and leads toward exactly the same conclusion. These results are not presented in this section to avoid redundancy information but validate the interpretations.

9.2 Experiment 2: Effects of the Size of Salient Objects

9.2.1 Goal

In this experiment, we want to compare the effectiveness of the models on three different images categories (large, medium, and small salient regions). In real-life images, all kinds of object sizes can be seen, and saliency models that are tuned for a given object size are not suitable. It should be noted that this study is divided into two parts: First, the experiment is computed on saliency models based on eye tracking, and second, the same experiment is calculated on saliency models based on salient object detection.

9.2.2 Method

Database and Ground-Truth The same database as in experiment 1 is used [7] with both region-based and eye tracking-based ground-truths. However, in this experiment, the whole database is not used. Only the first three categories are interesting for this study and therefore employed: 50 images with large salient regions, 80 with intermediate salient regions, and 60 with small salient regions.

Models In the first part of this experiment, nine state-of-the-art models from experiment 1 have been chosen. This is only eye tracking-based algorithms: FSM [9], SUN [10], SDLF [11], AIM [12], DVA [13], RARE [1], SR [14], QDCT [16], and AWS [19].

In the second part of this experiment, nine salient object detection-based state-of-the-art models have been chosen: FTSD [18], SSOI [22], SMSI [2], SLMC [23], SDHAS [24], SDAIR [25], SDBM [26], SIM [27], and SDWT [28].

Metrics As in the first study, the pAUC (Post-processing for Area Under the ROC Curve (2011) [7]) metric has been chosen for this experiment because it can be applied to both eye tracking-based and region-based ground-truths and mainly measures the eye fixation or region locations. Kendall’s W concordance measure is used for the statistical analysis.

9.2.3 Results

Models with Eye Tracking

Figure 9.2 shows the results for pAUC into the three categories for eye tracking-based algorithms. The mean trend can be computed by a linear regression (black line in Fig. 9.2). The general trend that can be highlighted is that the small regions have higher score than medium and large regions. This observation is correct for all models. We can also pay attention to SR model which significantly increases (in terms of scores rank) for small regions.

To assess the coherence between categories, the same ranking-based statistical test is required as in experiment 1; however, in this case it is applied to the means of each of the three classes (large, medium, and small). We use the averages because the number of images is different by categories. Kendall’s W coefficient as used in experiment 1 shows us a smaller concordance. As shown in Table 9.4, the p-value is significant. It means that the ranks between models are statistically different between the three categories, but the size of this difference in terms of ranking is relatively small. Indeed, Kendall’s concordance shows a moderate-strong agreement. In this experiment, the ranking is globally coherent (but less than

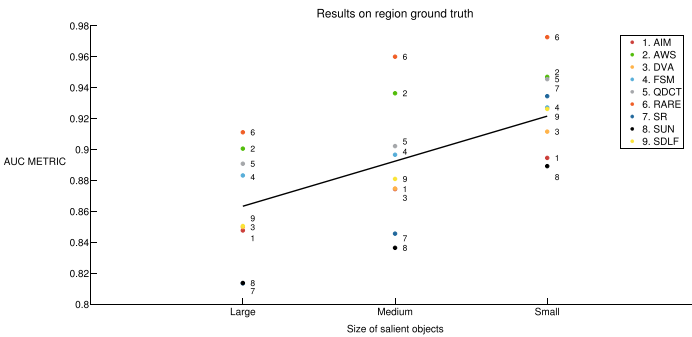


Fig. 9.2 Labeled region results on eye tracking-based algorithms on large, medium, and small regions for pAUC

Table 9.4 Concordance based on Friedman test and Kendall’s measure for large, medium, and small regions

	Friedman test (p-value)	Kendall’s concordance W
pAUC labeled regions	$6 * 10^{-4}$	0.74

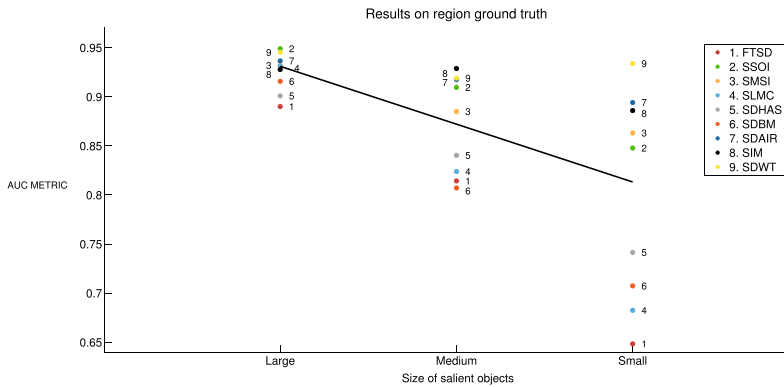


Fig. 9.3 Labeled region results on salient region detection algorithms on large, medium, and small regions for pAUC

between the two ground-truths). So, the size of the salient region can have a stronger impact on our assessment than the chosen ground-truth.

Models with Object Detection

Figure 9.3 shows the results on pAUC for the three categories for salient object detection-based algorithms. The mean tendency can be computed by a linear regression (black line in Fig. 9.3). The general trend that can be highlighted is the opposite of what we observed in Fig. 9.2. The large region has higher score than medium and small regions. This observation shows that most of the saliency models are tuned to their ground-truth (e.g., SOD-based models with the large binary masks and ET-based models with the small eye tracking distribution). It is correct for almost all models. However, SDWT, for example, is different: Its score is better with large salient regions than small ones, but its ranking is worse than both in medium regions. On the other hand, models with superpixels, like SDHAS, SDBM, and SLMC, significantly decrease (in terms of scores rank) for small regions.

To assess the coherence between categories, the same ranking-based statistical test is required as in the first part. We also use the means because the number of images is different by categories. Kendall’s W coefficient shows us a bigger concordance than in the first part. As shown in Table 9.5, the p-value is significant. It means that the ranks between models are statistically different between the three categories, but the size of this difference in terms of ranking is relatively small. Indeed, Kendall’s concordance shows a relatively strong agreement. In this

Table 9.5 Concordance based on Friedman test and Kendall coefficient for large, medium, and small regions

	Friedman test (p-value)	Kendall's concordance W
pAUC labeled regions	$8 * 10^{-4}$	0.81

experiment, the rankings are globally coherent (more so than in the first part and approximately equal to the one between the two ground-truths).

As mentioned for experiment 1, these results depend on the experimental design. In our case, one database, 18 saliency models divided into two groups, and one metric have been chosen. However, the first part of this experiment was conducted again in our paper [21] based on another metric (NSS) and leads toward exactly the same conclusion. These results are not computed for SOD models. Indeed, NSS is not a metric for object labeling.

It is important to notice that for deep learning models, where the result is mostly top-down and feature-dependent, the trend might not be the same for both ET and saliency object detection (SOD) as the logic is very different. For example, for ET models, the classical bottom-up models have the idea of rarity/surprise/contrast and, big objects have fewer chances to be very contrasted or rare as they take a lot of the space into the image. It is thus coherent to see that for classical ET models small rare objects are better highlighted than bigger ones. For deep learning models this is not obvious at all that the same trend would be true.

9.3 Experiment 3: Effects of Post-processing

9.3.1 Goal

In this experiment, only databases with eye fixations will be used. The purpose is to investigate which post-processing increases the score of a saliency map? Indeed, there are three aspects which should be considered: the blurring, the border cut, and the center effects.

First, we study the blurring that is used to better correlate the noisy human eye movement data. Indeed, the saliency maps obtained from a model usually score lower than smoother versions of these maps. However, based on [5], there is an optimal Gaussian blur level for each model.

Then, we investigated two other problems for fair comparison that are the center bias and border effect. Center bias means that a lot of fixations from natural images databases are located near the image center because when taking pictures, the amateur photographer often places salient objects in the image center. The computational saliency models that include a centered Gaussian use the prior knowledge of working on natural images and increase their score on some metrics compared with other models without this information. Moreover, Zhang et al. [10]

showed that metric scores are also corrupted by edge effects for the same reason. If we remove the edges of an image, metric scores usually increase as well. This is why a specific metric, called *sAUC*, has been designed to eliminate these undesirable effects. However, for other metrics (like NSS), these issues need to be taken into account.

These post-processing factors can dramatically influence some metric scores and affect the fairness of the validation. The main idea of this third experiment is to measure the impact of these factors on some saliency models.

9.3.2 Method

Database and Ground-Truth The database used here remains Jian Li's dataset [7], but only eye fixation ground-truth (collected with an eye tracker) will be employed. In this experiment, we use the whole database containing 235 images.

Models For this experiment, six state-of-the-art models have been chosen. This is only eye tracking-based algorithms: FSM [9], SUN [10], AIM [12], DVA [13], RARE [1], and AWS [19].

Metrics The NSS (Normalized Scanpath Saliency (2005) [29]) metric has been chosen for this experiment. Kendall's W concordance measure is used for the statistical analysis.

9.3.3 Results

Figure 9.4 shows an example of smoothing effect for the six saliency models used in this experiment. To find this optimal blur width, we use Y. Li's toolbox [30]. Some models such as FSM have already reached the optimal blur, while other models such as AIM, DVA, and SUN increase their score with smoother maps.

For the six saliency models with optimal blur (SM), we first cut the edges (8 pixels at each border) of each saliency map. Second, we multiply the output of every saliency model by a centered Gaussian to observe their improvement.

Figure 9.5 illustrates how the post-processing factors impact the score of each model based on NSS score. The general trend shows that all the scores increase. However, much depends on the saliency models.

Concerning the border cut, we observe that most of the saliency models such as AIM, FSM, DVA, or RARE have not improved their scores significantly. There are two reasons to this: Some methods already remove edges into their mechanism or some selective models often have low score on the border. At the opposite, SUN improves its scores. It means that this model often has high values on his edges and needs to be more selective.

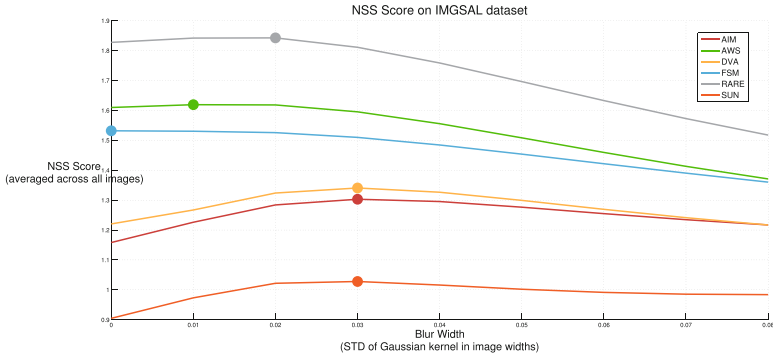


Fig. 9.4 Smoothing effect on six saliency models. The averaged NSS scores at all levels of blur widths are plotted and form a curve for each saliency model. The optimal blur is represented by a dot

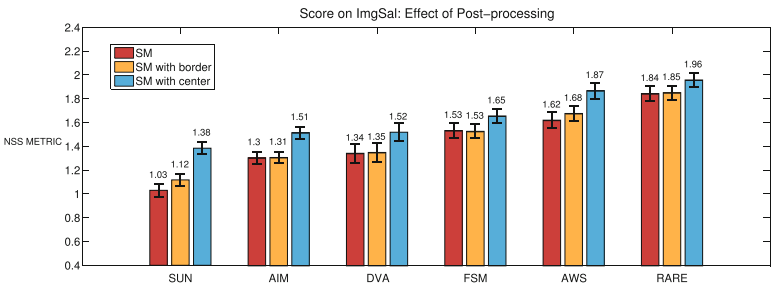


Fig. 9.5 Study of three post-processing factors: blurring, edge and center effects. The higher NSS is, the better the model

Concerning the 2D Gaussian center, we can see that all models improve their score. These results confirm that many fixations are located near the image center in Jian Li’s database [7]. These measures can help quantify the center bias of databases.

As mentioned above, these results depend on the experimental design. In our case, one database, six saliency models, and one metric have been chosen. However, other correlated results from literature can be found in [5, 6, 10, 17], etc. They lead toward the same interpretation.

9.4 Experiment 4: Effects of Metrics

9.4.1 Goal

Due to the diversity of available metrics for eye fixation prediction assessment, several benchmarks were proposed. In 2011, Toets proposed in [4] to compare

saliency models based on the Spearman’s rank correlation coefficient. In 2012, Borji built a benchmark [6] where three evaluation scores (PCC, NSS, and sAUC) are used. Finally, Judd et al. [5] proposed a platform using three different metrics: hAUC, S, and EMD. Although these benchmarks are major contributions, none of those studies deeply discussed the relevance of their similarity metrics mix.

The goal of this fourth experiment is twofold. First, it shows which metrics are close to each other. Second, it intends to reduce the dimensionality of the metrics we use and see which ones should be applied to do an efficient benchmark. Indeed, it is important to decide which metrics should be used together because they are complementary and which ones are useless to compute together because they will provide redundant information.

9.4.2 Method

Database and Ground-Truth The human eye fixation maps used are those in the database published by Li et al. [7] from experiment 1. This database provides eye fixation ground-truth (collected with an eye tracker) for 235 color images.

Models In this experiment, the same twelve state-of-the-art models from experiment 1 have been chosen: FSM [9], SUN [10], SDLF [11], AIM [12], DVA [13], RARE [1], SR [14], PFT [15], QDCT [16], FTSD [18], SSAFD [17], and AWS [19].

Metrics The 12 metrics presented in the previous chapter are used in this experiment. These metrics can be divided into three categories: value-based metrics that focus on saliency map values at eye gaze positions (NSS, P, and PF), distribution-based metrics that focus on saliency and gaze statistical distributions (PCC, KLD, SCC, EMD, and S), and location-based metrics that focus on location of salient regions at gaze positions (nAUC, pAUC, hAUC, and sAUC). A mean score by metric can thus be computed on the whole database for each model which leads to 12 different rankings of the 12 models, one for each comparison metric.

In the following, we will use the ranking between models and not their mean score values. This is due to the fact that the output of the metrics can be very different in terms of range of score value, and some of them should be maximized (correlation measures), while others should be minimized (divergence measures). Therefore, a direct score value comparison does not make a lot of sense. By contrast, the relative rank of the different models is a consistent measure common to all metrics, and its range is here between 1 and 12 (respectively, from the best model to the weakest).

To compare model rank according to the different metrics, Kendall’s W concordance measure [20] is used (as defined in Eq. 9.1 of experiment 1).

Kendall’s W concordance is a coefficient measuring the degree of agreement between metrics. The value ranges from 0 (no agreement between model ranks) to 1 (full agreement, the same model ranking). Furthermore, some rules of thumb are provided [20] to allow the researcher to interpret this measure as depicted in Table 9.2.

9.4.3 Results

9.4.3.1 Analysis of Consistency of Metrics

Intragroup Metrics The concordance is computed between all metrics into the three categories: value-based (amplitude), location-based, and distribution-based metrics (Table 9.6).

The concordance shows a moderate-strong agreement for location-based and distribution-based metrics. This means that these metrics provide some complementary information: They might provide different results for the same saliency map, and thus one of those metrics cannot just be ignored without a possible information loss about model ranking. However, one can see that the concordance between the amplitude metrics is high, which means that those measures are highly correlated and can therefore be summarized by a small subset of value-based metrics.

Intergroup Metrics Contrary to the intragroup study that does not achieve enough concordance, the intergroup suggests that some metrics are very close as it is shown in the Kendall matrix of Fig. 9.6a. NSS, P, PCC, and hAUC seem to be very close. On the opposite side, the KLD metric seems like an outlier in this matrix, and it is different from most of the other metrics in terms of model ranking.

To provide a better representation of the proximity in terms of model ranking among metrics, we apply, on Kendall’s coefficient, a classical Multidimensional Scaling (MDS) technique which visualizes and explores similarities or dissimilar-

Table 9.6 Kendall’s W coefficient of Intragroup Metrics

Group of metrics	W
Amplitude	0.9534
Distribution	0.7869
Location	0.8488

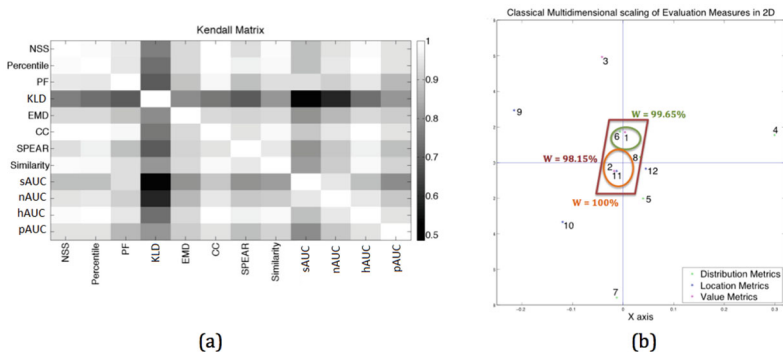


Fig. 9.6 Kendall’s analysis. (a) Kendall’s matrix on the 12 metrics. (b) Kendall’s measure on a group of metrics with Classical Multidimensional scaling of Evaluation Measures in 2D: 1. NSS, 2. P, 3. PF, 4. KLD, 5. EMD, 6. PCC, 7. SCC, 8. S, 9. sAUC, 10. nAUC, 11. hAUC, and 12. pAUC

ities in data. The results are displayed in Fig. 9.6b. In this representation, X-axis (equivalent to a first eigenvector) is more important than Y-axis (equivalent to a second eigenvector). From the figure, one can see, for example, that PF and NSS are closer than PF and sAUC.

9.4.3.2 Study of the Dimensionality

Based on the representation of Fig. 9.6 and in order to reduce the dimensionality of metric space, we decide to use a concordance of 98% as a threshold to fuse metrics (in terms of rank). By using this threshold, five metrics (NSS, P, PCC, S, and hAUC) can be fused into a single metric called *Cluster*. Indeed, as seen in Fig. 9.6b, the concordance between these metrics is 98.15%. It means that only the rank of two or three couples of models has been inverted on the 12 models between these metrics. The ranking of *Cluster* is defined as the mean ranking of all the metrics composing it.

For model validation, this *Cluster* means that one measure from those included in this set is enough, and the computation of the others inside this *Cluster* is useless in terms of new information about model ranking. In this case the five metrics can be summarized well enough by any of them.

To go further, a *Global* metric that acts like the barycenter of all metrics is also computed as the mean of the ranking of all metrics.

The same study as in the first part of section “Intergroup Metrics” is then applied but not on the same metrics. Indeed, we replace the five redundant metrics by the *Cluster* metric, and we add the *Global* one. Kendall’s matrix and the classical Multidimensional Scaling (MDS) technique are displayed in Fig. 9.7. We can observe that the *Cluster* and *Global* metrics are close. Moreover, along the X-axis (first eigenvector), the three metrics that cover most of the space are the Cluster, sAUC, and KLD.

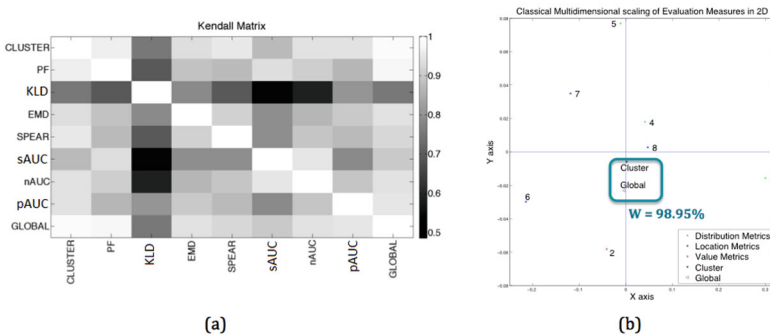


Fig. 9.7 Kendall’s analysis. (a) Kendall’s matrix on cluster, global, and seven metrics. (b) Kendall’s measure on group of metrics with Classical Multidimensional scaling of Evaluation Measures in 2D: 2. PF, 3. KLD, 4. EMD, 5. SCC, 6. sAUC, 7. nAUC, and 8. pAUC

These results depend on the experimental design. In our case, 1 database and 12 saliency models have been chosen. However, other online saliency benchmarks like [5, 6] lead towards the same interpretation.

9.5 Conclusion

In conclusion, there are many parameters affecting visual saliency assessment. Four experiments investigate basic questions to fairly evaluate saliency maps with human gazes or labeled regions.

To build a validation framework, first, a database with ground-truth needs to be chosen. Experiment 1 shows that there are significant differences between eye fixations and manually segmented salient region results, but the concordance between the rankings of models is strong. Moreover, the properties of the stimuli (e.g., in experiment 2: large, medium, and small salient regions) are addressed with different degrees of accuracy by the saliency models. For eye tracking-based models, small salient regions are better detected than medium and large salient regions. With object detection, the exact opposite behavior is observed. Therefore, the size of the salient region can have a stronger impact on our assessment than the chosen ground-truth. Having a large dataset with very different kind of objects of different sizes and exhibiting more bottom-up (surprising image regions) or top-down features (faces, people, etc.) are very important. It is also interesting to have different categories of viewers and different tasks.

Some metrics need to be chosen. For salient object detection, the gold standard F-measure is enough, but experiment 4 shows that one metric is not enough to evaluate the saliency model ranking on eye fixation data. The minimal set of similarity metrics which should be used is one of the metrics composing the cluster, sAUC and KLD. The use of those three metrics is enough to cover most of the space (along the first eigenvector) and provide a fair ranking result.

Finally, in terms of post-processing, experiment 3 shows that some factors centered bias, saliency map fuzziness, and border cut have an important influence on the final result and can dramatically improve the score, especially for the centered bias. The optimal parameters in terms of blur need to be assigned on each model to remain fair.

9.6 Summary

- Experiment 1 shows that the influence of the ground-truth is not crucial: If models have good results with one ground-truth, it is quite unlikely that these models completely fail with the other ground-truth except due to statistical fluctuation.
- Experiment 2 shows that the properties of the stimuli (e.g., large, medium, and small salient regions) are addressed with different degrees of accuracy by the

classical saliency models. For eye tracking-based models, small salient regions are better detected than medium and large salient regions. With object detection the exact opposite behavior is observed. So the size of the salient region can have a stronger impact on our assessment than the chosen ground-truth. However, deep learning models might be much less sensitive to the size of the important objects.

- Experiment 3 shows that several parameters centered bias, saliency map fuzziness, and border cut have important influence of the final result. It is thus possible to optimize a model by choosing the best parameters.
- Experiment 4 shows that the minimal set of similarity metrics which should be used is (a) one of the metrics composing the cluster in Fig. 9.6, (b) sAUC, and (c) KLD. The use of those three metrics is enough to cover most of space and provide a fair ranking result.

References

1. Riche, N., Mancas, M., Duvinage, M., Mibulumukini, M., Gosselin, B., & Dutoit, T. (2013). Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis. *Signal Processing: Image Communication*, 28(6), 642–658.
2. Vikram, T. N., Tscherepanow, M., & Wrede, B. (2012). A saliency map based on sampling an image into random rectangular regions of interest. *Pattern Recognition*, 45(9), 3114–3124.
3. Klein, D. A., & Frintrop, S. (2011). Center-surround divergence of feature statistics for salient object detection. In *2011 IEEE International Conference on Computer Vision (ICCV)* (pp. 2214–2219). IEEE.
4. 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.
5. Judd, T., Durand, F., Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. MIT tech report.
6. 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.
7. 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>
8. Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), 185–207.
9. Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), 1254–1259.
10. Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 32.
11. 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.
12. Bruce, N., & Tsotsos, J. (2006). Saliency based on information maximization. In *Advances in neural information processing systems* (Vol. 18, pp. 155–162).
13. Hou, X., & Zhang, L. (2008). Dynamic visual attention: searching for coding length increments. In *NIPS* (Vol. 5, p. 7).

14. Hou, X., & Zhang, L. (2007). Saliency detection: A spectral residual approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
15. Guo, C., Ma, Q., & Zhang, L. (2008). Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In *IEEE Conference on Computer vision and pattern recognition, 2008. CVPR 2008* (pp. 1–8). IEEE.
16. Schauerte, B., & Stiefelwagen, R. (2012). Predicting human gaze using quaternion DCT image signature saliency and face detection. In *Proceedings of the 12th IEEE Workshop on the Applications of Computer Vision (WACV)/IEEE Winter Vision Meetings, Breckenridge* (pp. 9–11).
17. Li, J., Levine, M. D., An, X., Xu, X., & He, H. (2012). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4), 996–1010.
18. 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.
19. Garcia-Diaz, A., Leborán, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision*, 12(6), 17.
20. Howell, D. (2012). *Statistical methods for psychology*. Cengage Learning.
21. Riche, N., Duvinage, M., Mancas, M., Gosselin, B., & Dutoit, T. (2013). A study of parameters affecting visual saliency assessment. arXiv preprint arXiv:1307.5691.
22. Rahtu, E., Kannala, J., Salo, M., & Heikkilä, J. (2010). Segmenting salient objects from images and videos. In *Computer Vision—ECCV 2010* (pp. 366–379). Springer.
23. Xie, Y., Lu, H., & Yang, M.-H. (2013). Bayesian saliency via low and mid-level cues. *IEEE Transactions on Image Processing*, 22(5), 1689–1698.
24. Fang, Y., Lin, W., Lee, B.-S., Lau, C.-T., Chen, Z., & Lin, C.-W. (2011) Bottom-up saliency detection model based on human visual sensitivity and amplitude spectrum. *IEEE Transactions on Multimedia*, 14(1), 187–198.
25. Fang, Y., Chen, Z., Lin, W., & Lin, C.-W. (2012). Saliency detection in the compressed domain for adaptive image retargeting. *IEEE Transactions on Image Processing*, 21(9), 3888–3901.
26. Xie, Y., & Lu, H. (2011). Visual saliency detection based on Bayesian model. In *2011 18th IEEE International Conference on Image Processing (ICIP)* (pp. 645–648). IEEE.
27. Margolin, R., Zelnik-Manor, L., & Tal, A. (2013). Saliency for image manipulation. *The Visual Computer*, 29(5), 381–392.
28. Imamoglu, N., Lin, W., & Fang, Y. (2013). A saliency detection model using low-level features based on wavelet transform. *IEEE Transactions on Multimedia*, 15(1), 96–105.
29. 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.
30. Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014). The secrets of salient object segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 280–287). IEEE.