Learning Picture Quality from Visual Distraction: Psychophysical Studies and Computational Models

Wei Zhang*, Hantao Liu

School of Computer Science and Informatics, Cardiff University, Cardiff, UK

Abstract

Visual saliency has been increasingly studied in relation to image quality assessment. Incorporating saliency potentially leads to improved ability of image quality metrics to predict perceived quality. However, challenges to optimising the combination of saliency and image quality metrics remain. Previous psychophysical studies have shown that distortion occurring in an image causes visual distractions, and alters gaze patterns relative to that of the image without distortion. From this, it can be inferred that the measurable changes of gaze patterns driven by distortion may be used as a proxy for the likely variation in perceived quality of natural images. In this paper, rather than using saliency as an add-on to image quality metrics, we investigate the plausibility of approximating picture quality based on measuring the deviation of saliency induced by distortion. First, we designed and conducted a large-scale eye-tracking experiment to clarify the knowledge on the relationship between the deviation of saliency and the variability of image quality. We then used the results to devise an algorithm which predicts perceived image quality based on visual distraction. Experimental results demonstrate this can provide good results of image quality prediction.

Keywords: Picture quality, visual distraction, saliency, eye-tracking, statistical learning

1. Introduction

Image quality metrics (IQMs) form the basis of algorithms that can automatically assess perceived quality of image material. They are nowadays widely available in many digital imaging

*Corresponding author
Email address: ZhangW71@cardiff.ac.uk (Wei Zhang)

systems in a broad range of applications, e.g., for fine tuning image and video processing pipelines, quality monitoring and control of displays. Substantial progress has been made on the development of IQMs over the last several decades. State of the art algorithms, however, demonstrate a lack of sophistication when it comes to dealing with real world complexity [1, 2], which makes image quality assessment a continued problem of interest. The fundamental challenge intrinsically lies in the fact that our knowledge about how the human visual system assesses image quality and how to express that in an efficient mathematical model remains rather limited. Being able to reliably predict image quality as perceived by humans requires a better understanding of functional aspects of the human visual system relevant to image quality perception, and optimal use of that to improve existing IQMs or devise more rigorous algorithms for IQMs.

In the literature, IQMs range from dedicated models that assess the annoyance of a specific type of visual distortion [3, 4, 5] to general models that measure the overall perceived quality [6, 7]. These IQMs can be categorised into full-reference, reduced-reference and no-reference metrics, depending on to what extent their algorithms utilise the undistorted reference. Full-reference metrics require the full access to the reference when assessing the quality of the distorted image [8, 9]. In the scenarios where the reference is partially available (e.g., in complex communication networks), reduced-reference metrics are meant to assess image quality by using information extracted from the reference [10, 11]. No-reference metrics attempt to predict the perceived quality purely based on the distorted image, and are useful for scenarios where there is no access to the reference at all [12, 13]. Generally, full-reference IQMs are more accurate and reliable than reduced-reference and no-reference IQMs. This is due to the quality of the distorted image can be directly measured against its original distortion-free reference [14]. However, full-reference IQMs are less practical in many real-world applications, where the reference is not fully available.

Successful IQMs in the literature mainly benefit from the advances in modelling early visual processing in the human visual system, e.g., contrast sensitivity [15], masking [16] and perception of structured information [17]. A growing trend in image quality research is to investigate the significance of visual attention on judging image quality. Visual attention exists as an important mechanism in the human visual system that allows effective selection of the most relevant information in a visual scene [18]. This attentional selection is known to be controlled by two

kinds of mechanisms: stimulus-driven, bottom-up mechanism and expectation-driven, top-down mechanism [19]. The former is often interchangeably referred to as visual saliency in the field of machine vision [20]. Eye movements of human observers serve as the empirical foundation of saliency modelling [21]. Saliency (i.e., bottom-up attention) models intend to explicitly predict fixations during (the first few seconds) freely viewing a visual stimulus [18]. A computational model of saliency aims to output a topographic map that represents conspicuousness of scene locations, where some parts of a scene appear to stand out relative to their neighbouring parts [22].

Incorporating saliency potentially leads to improved ability of IQMs to predict perceived quality [23, 24, 25]. The basic idea is to weight local distortions with local saliency, resulting in a more sophisticated means of image quality prediction. For example, in [26], a well-established saliency model (i.e., STB [27]) is integrated into a popular IQM (i.e., SSIM [17]) to improve its performance. However, finding ways to achieve such integration by means of a perceptually optimised strategy remains very challenging, and requires detailed mathematical models of distortion and saliency and of how both are perceived by the human visual system. In this paper, we hypothesize that distortion would cause visual distraction, i.e., variation in saliency, and that the extent of visual distraction should be related to the strength of distortion and can be used as a proxy for the likely variation in image quality. Therefore, rather than using saliency as a weighting function for IQMs, we investigate the plausibility of approximating image quality by means of measuring the saliency variation induced by distortion. The proposed investigation will provide a better understanding of how visual saliency plays a role in image quality assessment. Moreover, the proposed concept contains a new scope of developing saliency-based IQMs. It should, however, be noted that the proposed ideas fall into the full-reference (or reduced-reference) category of IQMs, where the saliency information of the reference image is needed.

2. Related work and contributions

2.1. Related work

Psychophysical studies have been undertaken to understand saliency in relation to image quality assessment [28, 29, 30, 31]. In [28], eye-tracking experiments were conducted to study how

human fixations might be affected by visual distortions in images. It is found that distortions can alter fixations relative to that obtained from the undistorted scene, and that the observed changes of fixations depend on the amount of distortion. A dedicated eye-tracking study was performed in [29] to investigate how JPEG compression affects fixations. The study shows that the impact of JPEG artifacts on fixations is more disruptive at low image quality than the high image quality. The eye-tracking data resulted from the study in [30] indicate that fixations change when visual distortions occur in an image, and that the extent of the change seems to be more related to the strength of artifacts rather than the type of artifacts. The eye-tracking study in [31] measured the gaze behaviour of viewers on original images and their distorted versions using three distortion types at two distortion levels. The results reveal that distortion added to an image leads to changes in gaze patterns and that this impact is predominately driven by distortion intensity rather than distortion type. In general, psychophysical studies as mentioned above show the tendency that distortion occurring in an image results in the deviation of fixations from their original patterns of deployment, and the extent of deviation largely depends on the strength of distortion.

It should, however, be noted that the generalisability of the empirical evidence found in above studies remains limited by the choices made in their experimentation, such as the use of a limited number of human subjects and a small degree of stimulus variability. For example, the experiments in [28, 29, 30, 31] all made use of a restrictive selection of visual stimuli (i.e., two distortion levels in [28] and [31], one distortion type in [29] and a total of six stimuli in [30]), which potentially affects the validity of these experiments in terms of making generalisations. The other drawback to existing studies is that they potentially suffer from an inherent bias in their data collection, where each observer had to view the same scene several times (with multiple variations of distortion) throughout a single session. In such a scenario, the viewers might have been forced to e.g., learn where to look for the impairments rather than observing the stimuli naturally. As a consequence, the recorded eye movements of fixations might have been more influenced by the visual distortions than the natural scenes . This implies that the eye-tracking data collected under such circumstances may potentially bias the actual relationship between saliency and distortion. In addition, some eye-tracking experiments (e.g., [30] and [31]) were performed in company with an image quality scoring task. Since a task given to the subjects is known to impact the deployment of fixations [32],

it is thus difficult to identify the extent to which the observed changes in gaze patterns are related to distortion rather than due to the perceptive and cognitive load of the assessor in need of completing a visual inspection task. Up till now, very little, if any, meaningful progress has been made on relating saliency variation (i.e., visual distraction) to image quality with detailed mathematical models, which is the topic to be investigated in this paper.

2.2. Contributions of the paper

(1) Obviously, existing psychophysical studies are either strongly biased in their data collection or limited by the generalisability of their results. To provide substantial and statistically sound empirical evidence, we start from designing and conducting a large-scale eye-tracking experiment by means of a reliable methodology. To this end, we propose a refined experimental design with an aim to eliminate bias induced by each subject having to view multiple variations of the same scene in a conventional experiment. This methodology allows reliably collecting eye-tracking data with a large degree of stimulus variability in terms of scene content, distortion type as well as degradation level.

(2) Unlike existing studies that have mainly focused on a limited dataset and rather qualitative analysis, we perform exhaustive statistical learning approaches using the "ground truth" data obtained from our eye-tracking experiment. This allows us to clarify the knowledge on the relationship between the distortion-driven saliency variation and perceived image quality, which provides a solid empirical foundation for computational modelling.

(3) We then use the results of our psychophysical studies to devise an algorithm which approximates perceived quality of natural images based on saliency deviation induced by distortion.

3. Psychophysical studies

To investigate the changes of saliency induced by distortion and their relation to image quality, we performed an eye-tracking experiment, where ground truth data of saliency were collected on natural scenes of varying quality. To be able to vary the perceived quality, each natural scene should be distorted with different types of distortion and at various levels of degradation. Asking

Table 1: Configuration of test stimuli extracted from the LIVE image quality database [33]. For each subset of the database, each interval (i.e., perceived quality level (PQL)) along the quality scale (i.e., DMOS) contains three images of different scenes.

	JPEG2000		JPEG		WN		GBLUR		FF	
PQL	image	DMOS	image	DMOS	image	DMOS	image	DMOS	image	DMOS
	img191	78.16	img91	78.98	img106	77.63	img121	74.67	img21	78.40
1	img79	79.17	img100	81.20	img61	78.47	img69	79.76	img18	81.02
	img107	79.85	img188	83.03	img50	79.44	img11	83.27	img92	81.44
	img220	70.84	img207	70.02	img96	69.26	img125	68.02	img112	70.08
2	img227	70.88	img175	70.50	img134	70.65	img118	70.26	img3	70.14
	img28	70.95	img134	72.36	img39	71.76	img120	71.84	img141	71.64
	img122	58.56	img156	59.65	img32	60.28	img53	59.57	img88	59.17
3	img160	58.75	img9	59.74	img124	61.46	img40	59.73	img66	59.22
	img137	60.61	img41	59.87	img26	61.63	img73	60.11	img98	60.48
	img91	48.72	img69	48.87	img25	49.83	img130	49.54	img81	48.85
4	img163	49.97	img21	50.12	img102	50.19	img38	49.57	img56	49.80
	img170	50.19	img128	52.32	img139	52.70	img30	50.78	img32	52.33
	img8	39.38	img15	41.37	img132	38.09	img76	40.32	img93	38.39
5	img133	40.05	img90	42.04	img90	39.71	img77	40.77	img135	38.98
	img120	40.34	img86	42.58	img70	41.32	img103	41.62	img123	41.42
	img187	29.92	img163	28.82	img22	29.05	img29	29.71	img38	30.78
6	img78	30.57	img63	30.00	img1	29.05	img35	29.95	img109	30.89
	img222	31.28	img56	32.01	img84	31.25	img13	30.51	img89	31.06
	img61	21.46	img38	18.03	img54	20.53	img32	20.42	img14	19.72
7	img21	22.46	img174	19.47	img138	22.15	img90	20.70	img44	20.70
	img80	20.57	img216	20.93	img103	22.88	img51	21.91	img98	20.90
	img198	9.47	img101	10.25	img59	11.04	img95	17.16	img119	8.32
8	img59	10.11	img58	10.55	img101	11.17	img137	17.75	img64	10.14
	img75	11.04	img130	11.02	img98	13.27	img63	18.64	img27	10.90

a human subject to view multiple variations of the same scene is likely to result in biased eye-tracking data [34]. To eliminate such potential bias, we proposed a refined experimental methodology for reliably collecting eye-tracking data, which is detailed below. Our experiment contains

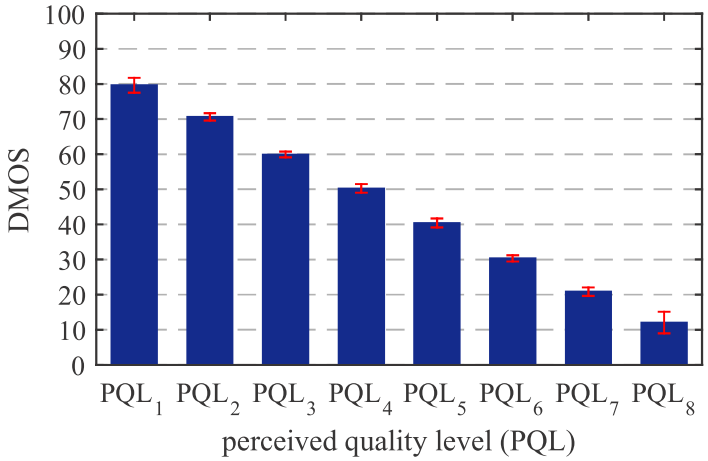

Figure 1: Illustration of the average DMOS of all stimuli for each perceived quality level (PQL) in our database. The error bars indicate the standard deviation.

149 test stimuli with a large degree of variability in terms of image content, distortion type as well as distortion level, and involves 100 human observers.

3.1. Proposed experimental methodology

3.1.1. Stimuli

To leave out expensive picture quality scoring experiments, we decided to construct our set of stimuli by systematically selecting images from the LIVE database [33], which already contains per image a ground truth quality score, i.e., difference mean opinion score (DMOS). A higher DMOS indicates a lower perceived quality. The LIVE database consists of five subsets: JPEG2000 compression (JPEG2000), JPEG compression (JPEG), white noise (WN), Gaussian blur (GBLUR) and simulated fast-fading Rayleigh occurring in wireless channel (FF). From each subset, we selected images to position at eight intervals along the DMOS scale; and each interval (i.e., perceived quality level (PQL)) accommodated three images of different scenes. This yielded a total of 149 test images (5 distortion type $\times$ 8 PQLs $\times$ 3 scenes + 29 originals). Table. 1 shows the configuration of the stimuli. Figure. 1 further illustrates the average DMOS of all stimuli for each PQL. It clearly visualises that our dataset contains eight distinct levels of perceived picture quality. Note that making sure the quality levels are perceptually distinct is an important prerequisite for the following study. Pairwise comparisons are performed with a Wilcoxon signed rank test

(i.e., an alternative for t-test for non-normal distributions) between two consecutive quality levels, selecting DMOS as the dependent variable, and the quality level as the independent variable. The results indicate that the difference between any pair of consecutive levels is statistically significant, with $P < 0.01$ at 95% confidence level.

3.1.2. Refined protocol

Note that stimuli used for image quality scoring purposely include per reference image its several versions of varying quality. Althought asking the same group of subjects to view all stimuli (i.e., so called "within-subjects") is customarily adopted in many eye-tracking studies [23, 28, 29, 30], this method potentially skews the results due to carryover effects [35]. This means every subject has to view multiple variations of the same scene, and may carry over undesirable effects (e.g., learning from experience) from viewing one stimulus to another. To overcome the bias, we propose to use an alternative method, in which multiple groups of subjects are to be randomly assigned to partitions of stimuli, each contains little or no stimulus repetition (i.e., so called "between-subjects"). In our experiment, the test dataset was divided into 5 partitions (i.e., one contains 29 images and the other four contain 30 images each); and only up to two repeated versions of the same scene were allowed in each partition. Stimuli assigned to each partition covered all distortion types and the full range of quality levels.

3.1.3. Experimental procedure

We set up a standard office environment as to the guidelines specified in [49] for the conduct of our experiment. The test stimuli were displayed on a 19-inch LCD monitor (native resolution: 1024×768 pixels). The viewing distance was approximately 60cm. The eye-tracking system (SensoMotoric Instrument RED-m) featured a sampling rate of 120Hz, a spatial resolution of 0.1 degree and a gaze position accuracy of 0.5 degree. Each subject was provided with instructions on the purpose and general procedure of the experiment (e.g., the task, the format of stimuli and timing) before the start of the actual experiment. A training session was carried out to familiarise the participants with the experiment, using 10 images that were different from those used in the real experiment. Each session per subject was preceded by a 9-point calibration of the eye-tracking equipment. The participants were instructed to experience the stimuli in a natural way ("view it

Figure 2: Illustration of the ground truth saliency maps for the original images used in our database (a) and for samples of their distorted versions (b). The stimuli in the first row of (b) are placed in order of perceived quality (with the corresponding DMOS values listed at the bottom of (b)), and the third row of (b) shows the image patches extracted from the stimuli (i.e., as indicated by the red boxes in the stimuli).

as you normally would"). Each participant viewed all stimuli (in his/her assigned partition) in a random order. Each stimulus was shown for 10 seconds followed by a mid-gray screen of 2 seconds.

We recruited a total of 100 participants in our experiment. The subject pool consisted of 50 male and 50 female university students and staff members. They were aged between 22 to 47 years, and all inexperienced with image quality assessment and eye-tracking recordings. The subjects were not examined for vision defects, and their verbal expression of the soundness of their vision was considered sufficient. The participants were first randomly divided into 5 groups of equal size, each with 10 males and 10 females; and the 5 groups of subjects were then randomly assigned to 5 partitions of stimuli. This gives a sample size of 20 subjects per test stimulus.

3.2. Results

3.2.1. Saliency map

A topographic saliency map that reflects the stimulus-driven, bottom-up aspects of visual attention is derived from free-viewing fixations. For a given stimulus, its saliency map is constructed by first accumulating fixations over all viewers (i.e., 20 subjects in our experiment) and then convolving the resulting fixation map with a Gaussian kernel. The width of the Gaussian kernel approximates the size of fovea (i.e., 2 degrees of visual angle, and 45 pixels in our experiment) [23, 30, 36]. The intensity of the resulting saliency map is linearly normalised to the range [0, 1]. Convolving a fixation map with a Gaussian kernel is conventionally used to construct a topographic saliency map. Both saliency map and the original fixation map are commonly used for various purposes, e.g., the former is often used for the visualization and comparison of the attentive regions [37] and the latter is needed for the comparison of fixation deployment [18]. Figure. 2 illustrates the saliency maps for the original images (i.e., referred to as scene saliency (SS)) used in our experiment, and for samples of their distorted versions (i.e., referred to as distorted scene saliency (DSS)). In Figure. 2 (b), we intentionally selected distorted stimuli from our dataset and showed them in order of perceived quality level (i.e., DMOS increases from 8.31 for the left end image to 81.44 for the right end image). As can be seen, the difference between SS and DSS increases as the perceived quality decreases (or DMOS increases), which will be further investigated below.

3.2.2. Reliability testing

Since the reliability of eye-tracking data strongly depends on, e.g., the number of viewers used and the strength of carryover effects, it is crucial to validate the reliability of the collected data before using them as a "ground truth". We propose and perform rigorous reliability testing to assess (1) whether the variances of eye-tracking data are homogeneous across different subject groups (in a between-subjects method); (2) whether the sample size (number of participants) used per stimulus is sufficient to achieve a stable saliency map.

Homogeneity of variance across groups. Since a between-subjects method is adopted, it is essential to make sure the variances of data across all groups are homogeneous. To be able to assess

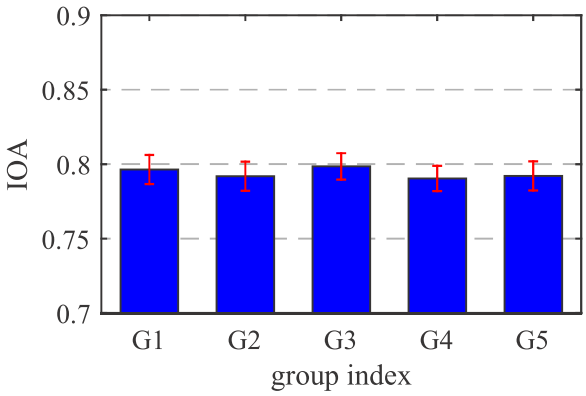

Figure 3: Illustration of the inter-observer agreement (IOA) averaged over all stimuli assigned to each group in our experiment. The error bars indicate a 95% confidence interval.

the homogeneity, we measure the inter-observer agreement (IOA), which refers to the degree of agreement among observers viewing the same stimulus [38, 39]. We implement IOA by comparing the saliency map derived from the fixations over all-except-one observers to the individual saliency map of the excluded observer; and by repeating this operation so that each subject serves as the excluded observer once. The similarity between two saliency maps is measured by the area under the receiver operating characteristic curve (AUC) [18]. Figure. 3 illustrates the IOA value averaged over all stimuli assigned to each subject group, showing that the IOA remains similar across five groups in our experiment. A statistical significance test (i.e., ANOVA) is performed and the results (i.e., $P > 0.05$ at 95% confidence level) show that there is no statistically significant difference between groups, suggesting a high degree of consistency in eye-tracking data across groups.

Data saturation. To determine the number of participants required for an eye-tracking experiment, researchers either follow the rule of thumb (i.e., use of 5-15 participants) or use "data saturation" as a guiding principle to ensure a chosen sample size suffices for leading to a "saturated" saliency map. The latter means a saliency map reaches the point at which no new information would be observed by adding more participants. We evaluate whether the sample size used in our experiment is adequate to reach saliency "saturation" (i.e., a proxy of sufficient degree of reliability). The validation is again based on the principle of IOA, which is extended to an inter-k-observer agree-

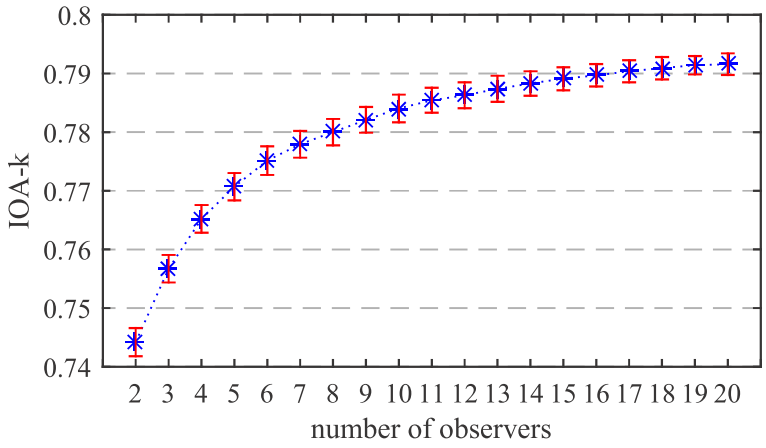

Figure 4: Illustration of the inter-*k*-observer agreement (IOA-k) averaged over all stimuli contained in our entire dataset. The error bars indicate a 95% confidence interval.

ment measure (i.e., referred to as IOA-*k*, and k=2, 3...20 in our experiment). More specifically, for a given stimulus, IOA-*k* is calculated by randomly selecting k observers among all. Figure. 4 illustrates the IOA-*k* value averaged over all stimuli contained in our entire dataset. It shows that "saturation" starts to occur with 18 participants, suggesting that our sample size (i.e., 20 observers per stimulus) gives a stable and saturated saliency map.

3.2.3. Significant findings: changes of saliency vs. changes of quality

As can be seen from Figure. 2, when comparing a DSS map to its corresponding SS map, there exist some consistent patterns, e.g., the highly salient regions tend to occur around the same places in both maps. The deviation of DSS from SS seems to be caused by distortion. Based on the observed trend, one may hypothesize that the degree of saliency deviation is associated with the strength of distortion (or level of perceived quality). To verify this hypothesis, we use the SS map derived from the original undistorted scene as the reference, and quantify the deviation of a DSS map from the reference (i.e., referred to as SS-DSS deviation), using three popular similarity measures: AUC, normalized scanpath saliency (NSS) [40, 41] and Kullback-Leibler divergence (KLD) [42]. These measures are already described in more detail in [43]; and we only briefly repeat their meaning in our context. AUC =1 means DSS is the same as its reference SS; whereas AUC = 0.5 corresponds to a chance level of finding similarity between DSS and SS. When NSS

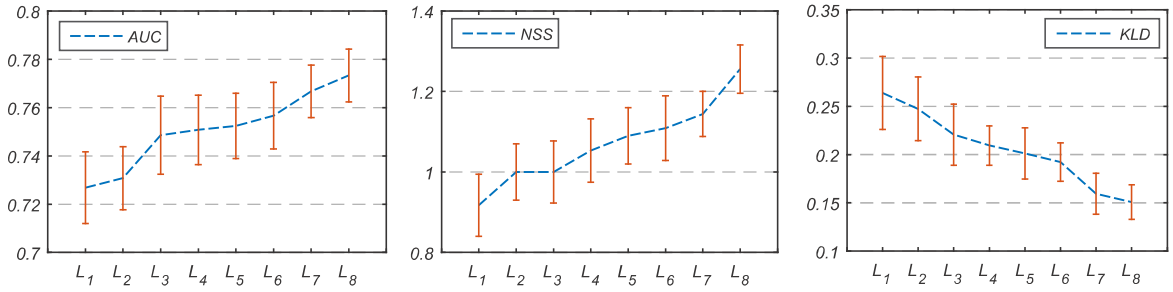

Figure 5: The measured SS-DSS deviation in terms of AUC, NSS and KLD for images of different perceived quality (or distortion strength). The error bars indicate a 95% confidence interval.

> 0, the higher the value, the more similar the DSS and SS. KLD = 0 means SS and DSS are absolutely equal. The higher the value, the more the DSS and SS are different.

Figure. 5 illustrates the SS-DSS deviation (in terms of AUC, NSS and KLD) averaged over all distorted stimuli within each distortion level. In general, the figure shows that distortion strength has a statistically significant effect on SS-DSS deviation, independent of the similarity measure used. The degree of SS-DSS deviation increases as the perceived quality decreases (or strength of distortion increases). An ANOVA test (preceded by a test for the assumption of normality) is further performed by selecting SS-DSS deviation as the dependent variable, and the perceived quality as the independent variable. The ANOVA results show that the image quality (or distortion strength) has a statistically significant effect ($P < 0.01$ at 95% confidence level) on the saliency deviation, independent of the similarity measure (i.e., AUC, NSS or KLD) used.

4. Computational modelling

Based on the findings of our psychophysical studies, it can be inferred that the changes of saliency induced by distortion are strongly associated with the changes of image quality. We further investigate the plausibility of modelling SS-DSS deviation as a proxy for the likely variation in perceived image quality. However, eye-tracking is cumbersome and impractical in many real-world applications. A more realistic and practical system will use a computational model of saliency rather than eye-tracking. Fundamental problems such as how to gauge the effectiveness of saliency models and to what extent the SS-DSS deviation measured by these models is useful for

Table 2: The structure of stimuli contained in three image quality databases

Image quality database	No. of source images	No. of distortion types	No. of distorted images
LIVE	29	5	779
CSIQ	30	6	866
TID2013	25	25	3000

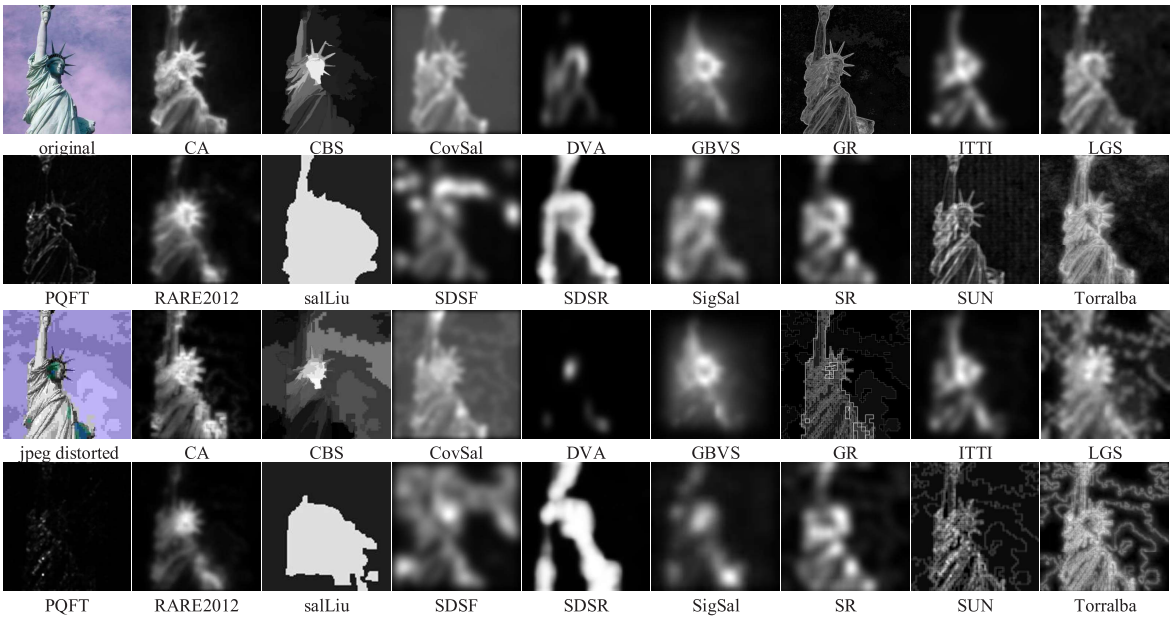

Figure 6: Illustration of saliency maps generated by 17 saliency models for one of the original images in CSIQ database and one of its jpeg distorted versions.

image quality prediction remain unsolved, and are the topics to be investigated below. Note that we focus on validating the plausibility of using computational saliency in place of eye-tracking data for SS-DSS deviation, rather than developing a new image quality metric. The latter is outside the scope of this paper, and will be treated in a separate contribution.

4.1. Investigation framework

To have sufficient statistical power, our evaluation is conducted using 17 state of the art saliency models and on three widely recognised image quality databases:LIVE [33], CSIQ [44] and TID2013 [45] as their structure of stimuli summarised in Table 2. The saliency models con-

sidered are CA, CBS, CovSal, DVA, GBVS, GR, ITTI, LGS, PQFT, RARE, salLiu, SDSF, SDSR, SigSal, SR, SUN and Torralba (as already detailed in [18, 46]). Figure. 6 shows the saliency maps generated by these models for one of the reference images and one of its distorted versions in our database.

To quantify the SS-DSS deviation using modelled saliency, KLD is used. Note since both AUC and NSS essentially require the access to the original fixation locations of SS which are, however, not available for a modelled SS, they become unrealistic for comparing two computed saliency maps and producing SS-DSS deviation. To compensate for the lack of appropriate similarity measures, we devise a new measure, saliency deviation measure (SDM), as follows:

$$SDM_{SS-DSS} = mean(\frac{2SS \cdot DSS + \varepsilon}{SS^2 + DSS^2 + \varepsilon}) \tag{1}$$

where the constant ε is to avoid instability when $SS^2 + DSS^2$ is very close to zero (i.e., $\varepsilon = 0.01$ is used in our experiment, note the value is somewhat arbitrary, but we find that the performance of the SDM algorithm is fairly insensitive to variations of this value).

Finally, the strength of the relationship between the SS-DSS deviation and the perceived image quality (i.e., DMOS scores) is quantified by the Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC) and Kendall rank order correlation coefficient (KROCC). Note the PLCC is customarily calculated after performing a nonlinear regression using the formula below (as also used in [47, 48, 49]):

$$f(x) = \alpha_1(\frac{1}{2} - \frac{1}{1 + e^{\alpha_2(x-\alpha_3)}}) + \alpha_4 x + \alpha_5 \tag{2}$$

where $\alpha_{1,2..5}$ are five fitting parameters.

4.2. Experimental results

4.2.1. Performance of saliency models on ground truth SS and DSS

Benchmarking saliency models against ground truth SS has been extensively attempted [18, 46], however, little is known about the performance of existing saliency models in terms of detecting DSS. Being able to detect both kinds of saliency would justify the applicability of a saliency model in the specific area of image quality, where stimuli are distorted. We use our eye-tracking

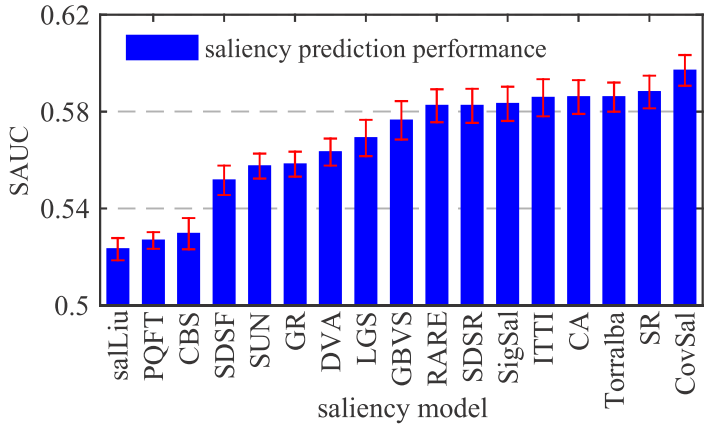

Figure 7: Illustration of the rankings of saliency models in terms of predictive power measured by SAUC. The error bars indicate a 95% confidence interval.

data that contain both SS and DSS to evaluate the performance of saliency models. The evaluation criterion is selected to be the shuffled AUC (SAUC) [18], which is a refined version of the traditional AUC, intent on precisely comparing the performance of different saliency models. Figure. 7 illustrates the rankings of saliency models in terms of the average SAUC. It shows that there is variation in performance among saliency models. The observed variation is further statistically analysed with an ANOVA, and the results (i.e., $P < 0.01$ at 95% confidence level) show that saliency model has a statistically significant effect on SAUC. This suggests the ability of predicting the ground truth saliency in terms of both SS and DSS is different for different saliency models.

4.2.2. Relationship between SS-DSS deviation and image quality

For each stimulus in a database, its SS-DSS deviation can be computed using a saliency model combined with a similarity measure. Table 3 and Table 4 illustrate the correlation between the SS-DSS deviations (i.e., measured by KLD or SDM) and image quality scores (i.e., DMOS) for three databases (i.e., LIVE, CSIQ and TID2013), using 17 different saliency models. Both tables show that the correlation varies significantly depending on the saliency model used. For example, in Table 3, the SROCC values range from 0.47 to 0.84 on LIVE, from 0.50 to 0.80 on CSIQ and from 0.28 to 0.73 on TID2013. However, the results are insensitive to the variations of similarity

Table 3: The correlation between SS-DSS deviation and image quality, using KLD as the similarity measure

	LIVE			CSIQ			TID2013		
model	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC
CA	0.7690	0.5734	0.7920	0.7468	0.5416	0.7639	0.5435	0.3747	0.5733
CBS	0.4664	0.3396	0.4795	0.5680	0.3940	0.5765	0.3793	0.2588	0.3910
CovSal	0.8052	0.6025	0.8099	0.7995	0.5947	0.7840	0.7343	0.5512	0.7512
DVA	0.6889	0.4862	0.7169	0.7068	0.500	0.7370	0.5242	0.3591	0.5189
GBVS	0.7895	0.5921	0.8177	0.7813	0.586	0.8003	0.5747	0.4027	0.5909
GR	0.5592	0.3848	0.5613	0.5251	0.3593	0.5438	0.3815	0.2609	0.4042
ITTI	0.8366	0.6394	0.8455	0.6602	0.4806	0.6965	0.6470	0.4639	0.6678
LGS	0.7986	0.5970	0.8078	0.6072	0.4334	0.6325	0.5298	0.3696	0.5505
PQFT	0.7900	0.5762	0.7924	0.6896	0.4922	0.7609	0.4659	0.3202	0.4949
RARE	0.6665	0.4722	0.6874	0.6124	0.4242	0.6746	0.4333	0.2965	0.4682
salLiu	0.4708	0.3258	0.4912	0.5008	0.3500	0.5329	0.2830	0.1934	0.3091
SDSF	0.7227	0.5313	0.7286	0.6280	0.4498	0.6327	0.4959	0.3445	0.4914
SDSR	0.7219	0.5256	0.7386	0.7117	0.5115	0.7417	0.6363	0.4526	0.6490
SigSal	0.6900	0.4948	0.7541	0.7065	0.5213	0.7758	0.6172	0.4414	0.6666
SR	0.7863	0.5952	0.7845	0.7803	0.5914	0.8033	0.6336	0.4552	0.6782
SUN	0.6728	0.4858	0.6818	0.5334	0.3656	0.5706	0.4060	0.2742	0.4064
Torralba	0.8193	0.635	0.8159	0.6759	0.4937	0.7367	0.6540	0.4750	0.6814

measure: both KLD and SDM yield consistent correlation coefficients. For example, Torralba and CovSal give the strongest correlation and salLiu and CBS give the weakest correlation, independent of the similarity measure used.

In order to provide the overall rankings of saliency models, we use the following formula:

$$Overall_{CC}(model) = \beta_1 \cdot CC_{LIVE} + \beta_2 \cdot CC_{TID2013} + \beta_3 \cdot CC_{CSIQ} \tag{3}$$

where β_1, β_2 and β_3 are weights assigned to each dataset depending on the number of distorted images in that dataset. For example, β_1 equals to 0.168 (i.e., 779/(779 + 866 + 3000)) according to Table 2. CC indicates the averaged correlation score over a certain correlation coefficient (i.e., either SROCC or KROCC or PLCC) and over two cases (i.e.,KLD and SDM).

Figure. 8 illustrates the saliency models in order of overall correlation for SROCC (the results

Table 4: The correlation between SS-DSS deviation and image quality, using SDM as the similarity measure

	LIVE			CSIQ			TID2013		
model	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC
CA	0.7699	0.5714	0.7813	0.7758	0.5617	0.7866	0.5732	0.3866	0.5797
CBS	0.6451	0.4584	0.6580	0.5633	0.3878	0.5647	0.3856	0.2616	0.3954
CovSal	0.7086	0.5116	0.7158	0.6605	0.4679	0.6508	0.6697	0.4855	0.6881
DVA	0.7317	0.5208	0.7438	0.6909	0.4804	0.7150	0.5751	0.3997	0.5762
GBVS	0.7876	0.5912	0.8031	0.7827	0.5833	0.7933	0.5500	0.3831	0.5527
GR	0.8300	0.6306	0.8337	0.7373	0.5335	0.7752	0.5078	0.3563	0.5598
ITTI	0.8305	0.6325	0.8340	0.6447	0.4665	0.6890	0.6449	0.4608	0.6582
LGS	0.7348	0.5357	0.7440	0.6130	0.4353	0.6299	0.4875	0.3377	0.5092
PQFT	0.6230	0.4333	0.6177	0.4933	0.3154	0.5169	0.2682	0.1775	0.3092
RARE	0.6718	0.4779	0.6949	0.6081	0.4214	0.6792	0.4270	0.2921	0.4570
salLiu	0.4678	0.3247	0.4867	0.4535	0.3160	0.4906	0.2593	0.1771	0.2844
SDSF	0.7251	0.5326	0.7304	0.6272	0.4430	0.6212	0.4875	0.3355	0.4836
SDSR	0.7507	0.5530	0.7596	0.7388	0.5345	0.7590	0.6537	0.4662	0.6639
SigSal	0.6903	0.4903	0.7399	0.6858	0.4933	0.7297	0.6077	0.4294	0.6311
SR	0.8846	0.7153	0.8766	0.8551	0.6706	0.8887	0.6649	0.4883	0.7250
SUN	0.5533	0.3808	0.5797	0.5747	0.3871	0.6003	0.3250	0.2174	0.3429
Torralba	0.8854	0.6937	0.8904	0.7204	0.5321	0.8007	0.6905	0.5029	0.7423

for KROCC and PLCC exhibit similar trends and thus not presented here). It shows that by use of saliency models, such as CovSal, SR and Torralba, the resulting SS-DSS deviation can reasonably predict the perceived image quality. To further justify of our idea of using SS-DSS deviation as a proxy for image quality prediction, we compare our best performing models (i.e., SS-DSS based on CovSal, SR and Torralba) to five state of the art image quality metrics in literature, including PSNR, SSIM [17], VIF [49], VSNR [50] and MAD [51]. Table 5 shows the comparison of overall performance (i.e., also calculated using Equation (3)) of these image quality metrics. It shows that the SS-DSS models are fairly comparable to some of the traditional image quality metrics, suggesting that measuring saliency deviation induced by distortion is a plausible method for assessing image quality.

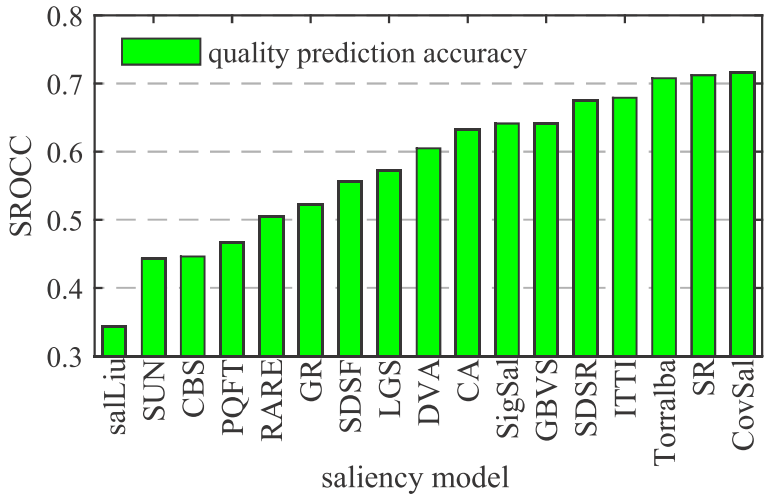

Figure 8: Illustration of the overall ability of saliency models in producing the correlation (in terms of SROCC) between SS-DSS deviation and image quality.

Table 5: The comparison of the overall performance (in terms of SROCC) between the best performing saliency models and some image quality metrics

	model			image quality metric				
	Torralba	SR	CovSal	PSNR	VSNR	VIF	SSIM	MAD
Overall SROCC	0.7079	0.7124	0.7166	0.7101	0.7466	0.7708	0.8012	0.8428

4.2.3. Impact of saliency models on the effectiveness of SS-DSS deviation measure

Having identified the benefits of SS-DSS deviation for image quality prediction, one could intuitively hypothesize that the better a saliency model can predict ground truth SS and DSS, the better the resulting SS-DSS model can predict image quality. To validate this hypothesis, we scatter plot in Figure. 9 the following two variables: the saliency predictive power of a saliency model (i.e., SAUC, based on the results of Figure. 7) and the quality predictive power of a corresponding SS-DSS model (i.e., SROCC, based on the results of Figure. 8). The Pearson correlation is equal to 0.87, suggesting a fairly strong positive relationship. In general, saliency models (e.g., CovSal, SR, Torralba and ITTI) that perform well in detecting ground truth SS and DSS lead to competitive performance of quality prediction when using these models for SS-DSS deviation measure, and vice versa (e.g., salLiu, Sun, CBS and PQFT). The findings above suggest that the accuracy of a

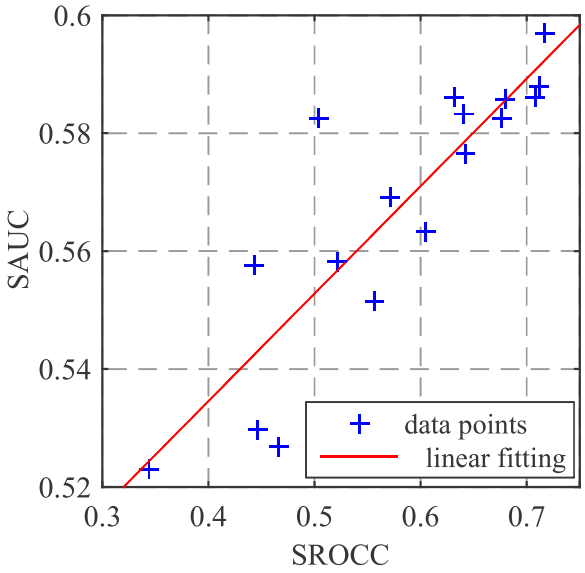

Figure 9: Scatter plot of two variables: the saliency predictive power of a saliency model (i.e., SAUC, based on the results of Figure. 7) and the quality predictive power of a corresponding SS-DSS model (i.e., SROCC, based on the results of Figure. 8

saliency model in detecting both SS and DSS should be used as a criterion to determine whether or not a specific saliency model is suitable for SS-DSS deviation measure and for image quality prediction.

5. Conclusion

This paper investigated the relationship between the changes of gaze patterns driven by distortion and the likely variability of perceived image quality. A large-scale eye-tracking experiment was conducted and an exhaustive statistical evaluation was performed on the resulting data. We found that the occurrence of distortion in an image causes the deviation of fixations from their original places in the image without distortion, and that the extent of distortion determines the amount of saliency deviation. We also considered how the findings can be used to devise an algorithm for image quality prediction. Experimental results show that it is highly plausible to approximate image quality by means of saliency deviation.

6. References

[1] D. M. Chandler, Seven challenges in image quality assessment: past, present, and future research, ISRN Signal Processing 2013. doi:10.1155/2013/905685.

[2] D. Ghadiyaram, A. C. Bovik, Massive online crowdsourced study of subjective and objective picture quality, IEEE Trans. on Image Process. 25 (1) (2016) 372–387. doi:10.1109/TIP.2015.2500021.

[3] S. Wang, C. Deng, B. Zhao, G.-B. Huang, B. Wang, Gradient-based no-reference image blur assessment using extreme learning machine, Neurocomputing 174, Part A (2016) 310 – 321. doi:http://dx.doi.org/10.1016/j.neucom.2014.12.117.

[4] L. Li, Y. Zhou, W. Lin, J. Wu, X. Zhang, B. Chen, No-reference quality assessment of deblocked images, Neurocomputing 177 (2016) 572 – 584. doi:http://dx.doi.org/10.1016/j.neucom.2015.11.063.

[5] O. A. Arqub, Z. Abo-Hammour, Numerical solution of systems of second-order boundary value problems using continuous genetic algorithm, Information Sciences 279 (2014) 396 – 415. doi:http://dx.doi.org/10.1016/j.ins.2014.03.128.
URL http://www.sciencedirect.com/science/article/pii/S0020025514004253

[6] P. Peng, Z.-N. Li, General-purpose image quality assessment based on distortion-aware decision fusion, Neurocomputing 134 (2014) 117 – 121. doi:http://dx.doi.org/10.1016/j.neucom.2013.08.046.

[7] H. wen Chang, Q. wen Zhang, Q. gang Wu, Y. Gan, Perceptual image quality assessment by independent feature detector, Neurocomputing 151, Part 3 (2015) 1142 – 1152. doi:http://dx.doi.org/10.1016/j.neucom.2014.04.081.

[8] W. Lu, T. Xu, Y. Ren, L. He, On combining visual perception and color structure based image quality assessment, Neurocomputing 212 (2016) 128 – 134. doi:http://dx.doi.org/10.1016/j.neucom.2016.01.117.

[9] Y. Yuan, Q. Guo, X. Lu, Image quality assessment: A sparse learning way, Neurocomputing 159 (2015) 227 – 241. doi:http://dx.doi.org/10.1016/j.neucom.2015.01.066.

[10] S. Decherchi, P. Gastaldo, R. Zunino, E. Cambria, J. Redi, Circular-elm for the reduced-reference assessment of perceived image quality, Neurocomputing 102 (2013) 78 – 89. doi:http://dx.doi.org/10.1016/j.neucom.2011.12.050.

[11] X. Wang, Q. Liu, R. Wang, Z. Chen, Natural image statistics based 3d reduced reference image quality assessment in contourlet domain, Neurocomputing 151, Part 2 (2015) 683 – 691. doi:http://dx.doi.org/10.1016/j.neucom.2014.05.090.

[12] L. He, D. Wang, Q. Liu, W. Lu, Fast image quality assessment via supervised iterative quantization method, Neurocomputing 212 (2016) 121 – 127. doi:http://dx.doi.org/10.1016/j.neucom.2016.01.116.

[13] Q. Li, W. Lin, Y. Fang, Bsd: Blind image quality assessment based on structural degradation, Neurocomputing 236 (2017) 93 – 103. doi:http://dx.doi.org/10.1016/j.neucom.2016.09.105.

[14] Y. Fang, W. Lin, S. Winkler, Review of existing objective qoe methodologies, Multimedia Quality of Experience

(QoE): Current Status and Future Requirements 29.

[15] P. G. J. Barten, Formula for the contrast sensitivity of the human eye, Vol. 5294, 2003, pp. 231–238. doi:10.1117/12.537476.

[16] S. J. Daly, Visible differences predictor: an algorithm for the assessment of image fidelity, Vol. 1666, 1992, pp. 2–15. doi:10.1117/12.135952.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Processing 13 (4) (2004) 600–612. doi:10.1109/TIP.2003.819861.

[18] A. Borji, D. N. Sihite, L. Itti, Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study, IEEE Trans. Image Process. 22 (2013) 55–69. doi:10.1109/TIP.2012.2210727.

[19] T. J. Buschman, E. K. Miller, Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices, Science 315 (5820) (2007) 1860–1862. doi:10.1126/science.1138071.

[20] J. H. Fecteau, D. P. Munoz, Salience, relevance, and firing: a priority map for target selection, Trends in Cognitive Sciences 10 (8) (2006) 382 – 390. doi:10.1016/j.tics.2006.06.011.

[21] A. Borji, L. Itti, State-of-the-art in visual attention modeling, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1) (2013) 185–207. doi:10.1109/TPAMI.2012.89.

[22] L. Itti, C. Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259. doi:10.1109/34.730558.

[23] H. Liu, I. Heynderickx, Visual attention in objective image quality assessment: based on eye-tracking data, IEEE Trans. Circuits and Syst. for Video Technol. 21 (2011) 971–982. doi:10.1109/TCSVT.2011.2133770.

[24] U. Engelke, H. Kaprykowsky, H. J. Zepernick, P. Ndjiki-Nya, Visual attention in quality assessment, IEEE Signal Processing Magazine 28 (6) (2011) 50–59. doi:10.1109/MSP.2011.942473.

[25] W. Zhang, A. Borji, Z. Wang, P. L. Callet, H. Liu, The application of visual saliency models in objective image quality assessment: A statistical evaluation, IEEE Trans. Neural Netw. and Learning Syst. 27 (6) (2016) 1266–1278. doi:10.1109/TNNLS.2015.2461603.

[26] X. Feng, T. Liu, D. Yang, Y. Wang, Saliency based objective quality assessment of decoded video affected by packet losses, in: Proc. of the 15th IEEE Intl. Conf. Image Process., 2008, pp. 2560–2563. doi:10.1109/ICIP.2008.4712316.

[27] D. Walther, C. Koch, Modeling attention to salient proto-objects, Neural Networks 19 (9) (2006) 1395 – 1407, brain and AttentionBrain and Attention. doi:http://dx.doi.org/10.1016/j.neunet.2006.10.001.

[28] E. Vu, D. M. Chandler, Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience, in: Proc.of IEEE Southwest Symposium Image Analysis and Interpretation, 2008, pp. 73–76. doi:10.1109/SSIAI.2008.4512288.

[29] X. Min, G. Zhai, Z. Gao, C. Hu, Influence of compression artifacts on visual attention, in: Proc. of the IEEE Int. Conf. Multimedia and Expo, 2014, pp. 1–6. doi:10.1109/ICME.2014.6890189.

[30] J. Redi, H. Liu, R. Zunino, I. Heynderickx, Interactions of visual attention and quality perception, in: Proc. SPIE, Human Vis. Electron. Imaging, San Francisco, USA, 2011, pp. 78650S–78650S–11. doi:10.1117/12.876712.

[31] F. Rhrbein, P. Goddard, M. Schneider, G. James, K. Guo, How does image noise affect actual and predicted human gaze allocation in assessing image quality?, Vision Research 112 (2015) 11 – 25. doi:10.1016/j.visres.2015.03.029.

[32] B. Tatler, M. Hayhoe, M. Land, D. Ballard, Eye guidance in natural vision: Reinterpreting salience, Journal of Vision 11 (5) (2011) 1–23. doi:10.1167/11.5.1.

[33] L. C. H.R. Sheikh, Z.Wang, A. Bovik, LIVE image quality assessment database release 2.
URL http://live.ece.utexas.edu/research/quality

[34] O. L. Meur, A. Ninassi, P. L. Callet, D. Barba, Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric, Signal Processing: Image Communication 25 (7) (2010) 547 – 558. doi:10.1016/j.image.2010.05.006.

[35] A. G. Greenwald, Within-subjects designs: To use or not to use?, Psychological Bulletin 83 (2) (1976) 314. doi:10.1037/0033-2909.83.2.314.

[36] A. Ninassi, O. L. Meur, P. L. Callet, D. Barba, Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric, in: Proc. of the 14th IEEE Intl. Con. Image Process., Vol. 2, 2007, pp. II – 169–II – 172. doi:10.1109/ICIP.2007.4379119.

[37] O. L. Meur, P. L. Callet, D. Barba, D. Thoreau, A coherent computational approach to model bottom-up visual attention, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (5) (2006) 802–817. doi:10.1109/TPAMI.2006.86.

[38] A. Torralba, A. Oliva, M. S. Castelhano, J. M. Henderson, Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search, Psychological review 113 (4) (2006) 766. doi:10.1037/0033-295X.113.4.766.

[39] T. Judd, F. Durand, A. Torralba, Fixations on low-resolution images, J. Vis. 11 (4) (2011) 14–14. doi:10.1167/11.4.14.

[40] M. Dorr, T. Martinetz, K. R. Gegenfurtner, E. Barth, Variability of eye movements when viewing dynamic natural scenes, J. Vis. 10 (10) (2010) 28. doi:10.1167/10.10.28.

[41] Q. Zhao, C. Koch, Learning visual saliency by combining feature maps in a nonlinear manner using adaboost, J. Vis. 12 (6) (2012) 22. doi:10.1167/12.6.22.

[42] B. W. Tatler, R. J. Baddeley, I. D. Gilchrist, Visual correlates of fixation selection: effects of scale and time, Vision Research 45 (5) (2005) 643 – 659. doi:10.1016/j.visres.2004.09.017.

[43] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, F. Durand, What do different evaluation metrics tell us about saliency models?, arXiv preprint arXiv:1604.03605.

[44] E. C. Larson, D. M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, Journal of Electronic Imaging 19 (1) (2010) 011006–011006–21. doi:10.1117/1.3267105.

[45] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, C.-C. J. Kuo, Image database tid2013: Peculiarities, results and perspectives, Signal Processing: Image Communication 30 (2015) 57 – 77. doi:10.1016/j.image.2014.10.009.

[46] A. Borji, M. M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Trans. on Image Processing 24 (12) (2015) 5706–5722. doi:10.1109/TIP.2015.2487833.

[47] H. R. Sheikh, M. F. Sabir, A. C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Trans. Image Process. 15 (11) (2006) 3440–3451. doi:10.1109/TIP.2006.881959.

[48] L. Zhang, L. Zhang, X. Mou, D. Zhang, FSIM: A feature similarity index for image quality assessment, IEEE Trans. on Image Process. 20 (8) (2011) 2378–2386. doi:10.1109/TIP.2011.2109730.

[49] H. R. Sheikh, A. C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (12) (2005) 2117–2128. doi:10.1109/TIP.2005.859389.

[50] D. M. Chandler, S. S. Hemami, VSNR: A wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Processing 16 (9) (2007) 2284–2298. doi:10.1109/TIP.2007.901820.

[51] E. C. Larson, D. M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, Journal of Electronic Imaging 19 (1) (2010) 011006–011006–21. doi:10.1117/1.3267105.