Erschienen in: PQS 2016 : 5th ISCA/DEGA Workshop on Perceptual Quality of Systems / Möller, Sebastian; Egger, Sebastian (Hrsg.). - Baixas, France : ISCA, 2016. - S. 117-121
https://dx.doi.org/10.21437/PQS.2016-25

Reported Attention as a Promising Alternative to Gaze in IQA Tasks

Vlad Hosu, Franz Hahn, Igor Zingman, Dietmar Saupe

Department of Computer and Information Science
University of Konstanz, Germany

Abstract

We study the use of crowdsourcing for self-reported attention in image quality assessment (IQA) tasks. We present the results from two crowdsourcing campaigns: one where participants indicated via mouse clicks the image locations that influenced their rating of quality, and another where participants chose locations they looked at in a free-viewing setting. The results are compared to in-lab eye tracking experiments. Our analysis shows a strong connection between the in-lab and self-reported IQA locations. This suggests that crowdsourced studies are an affordable and valid alternative to eye tracking for IQA tasks.
Index Terms: crowdsourcing, visual attention, image quality assessment

1. Introduction

Which parts of an image do we attend to during a viewing task? How well aware are we of the locations that we inspect? To answer these questions we compared two methods for probing visual attention: eye tracking and self-reporting. We are interested in the similarities and differences between them with respect to a task-specific setting, considering the trade-offs in the experimental procedure.

A strong connection has been shown between eye-gaze and what people pay attention to [1]. An alternative to recording eye-gaze data, observers can report which parts of an image they remember to have attended. Collecting self-reported attention has been done on several occasions [2–7]. A strong connection was shown between self-reporting and eye tracking during free-viewing tasks. The effects of free-viewing relative to task-specific conditions have only been briefly studied [7,8]. In some cases, the two sources of information show a significant difference [9]. We consider whether the observed differences also apply for image quality assessment (IQA) tasks.

Gaze is well known to be directed by both subconscious and conscious processes, while self-reported attention involves mostly conscious aspects. On one hand eye tracking helps probe the way we direct our gaze, however this doesn't always relate to covert attention. An observer can look at a spot, but pay attention to something else. On the other hand, self-reporting can capture covert attention. Nonetheless, observers have to be able to correctly report it. We are attempting for the first time to study the use of self-reported attention during IQA tasks.

We acquired our data using crowdsourcing experiments. Crowdsourcing offers a low-cost solution for large scale data collection. This simplifies the procedure allowing for fast turn-around experiments. Crowdsourcing is an unreliable data source, requiring us to devise effective strategies of ensuring the quality of the results. We devised ways to compare different data sources and tasks without assuming that either eye tracking or self-reporting constitute a ground-truth for attention.

Our main contribution is the design and evaluation of a crowdsourced IQA experiment collecting self-reported attention. A secondary contribution is the evaluation and interpretation of the data. This amounts to comparing gaze data from an eye tracker [10] with self-reported attention from our crowdsourcing experiments. We do this both for data captured during an IQA and a free-viewing task using the same set of 40 images. The results suggest that self-reported attention is a good alternative to eye tracking during IQA tasks.

2. Methodology

We started with an appropriate image dataset, for which there is pre-existing eye tracking information. One such dataset (TUD) was captured during an IQA task by Alers et al. [10–12]. The set consists of 40 original images to which 4 levels of distortion have been applied. This amounts to a total of 160 distorted images with varying degrees of JPEG compression. We extended the data set by collecting self-reported attention for all 160 images.

Several image location annotation methods have been considered in the literature. Character charts were used by [4, 5] allowing the user to enter labels corresponding to image locations. Engelke et al. [1] asked users to select rectangular regions with their mouse, whereas Salvador et al. asked users to point-and-click to specify locations on images [13]. More complex foveated exploration strategies have also been proposed [2, 7]. These involve an interactive exploration using the mouse. Due to its simplicity we chose to work with a simple point-and-click interface. The aggregated results collected from many users look similar to eye tracking fixations: the familiar 2D point cloud.

We performed two crowdsourced experiments on crowdflower.com. Both used the same input method: workers were asked to click locations of interest. They did this while performing an IQA task or during a free-viewing context. In addition to image locations, for the IQA task users were also asked to rate the quality of each image. We computed a mean opinion score (MOS) for each image based on the image ratings.

A crowdflower experiment is set up such that contributors can be tested. The questions are set by the experimenter for each task. Users are screened based on their performance on both the current task, and on their overall accuracy on the crowdflower platform. In both of our experiments we only allowed users with a cumulative accuracy rating of more than 80% to participate.

2.1. Free-viewing experiment

When performing the free-viewing experiment participants were given very brief instructions. They were asked to select between 3 to 10 locations on each image by clicking on them

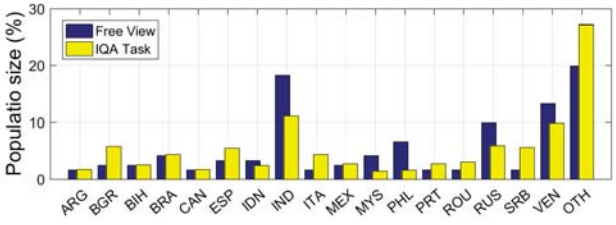

Figure 1: Distribution of participants based on their country of origin in our crowdsourcing experiments: free-viewing (dark/blue), IQA task (light/ yellow). Countries with less than 1% participants are grouped into "OTH".

using the mouse. The task is not time-constrained. Each contributor needed to select a minimum of 3 locations per image, and was allowed to annotate at most 50 images. This ensures sufficient user diversity. Participants were also not required to take a qualification test. If users were screened based on a predefined set of valid annotations, the natural distribution of reported attention might have been skewed.

A total of 121 workers from 36 countries participated in this experiment with an average of 39.7 answers, generating 30 answers for each stimulus. Workers were paid $0.01 for 10 annotations, tallying up to a total price of $5.83. The experiment completed in approximately 90 minutes. This shows the task to be simple to perform, allowing a fast turn-around time, and involving very low costs on the part of the experimenter. Compared to setting up an eye tracking experiment, crowdsourced point-and-click experiments are extremely simple, but as we will show, effective at recording attended regions.

2.2. IQA experiment

In the IQA task contributors were asked to rate the quality of an image on a 5-point scale and indicate at least 3 image locations via mouse clicks that influenced their rating. More precisely, the phrasing that we used was: "What is the overall quality of the following image?" and "Click on the points you looked at when deciding the quality of the image." The experiment also included a detailed description of the task and showed a simple example of poor and good quality images. Workers were allowed as much time as they needed. The definition of "quality" that we presented to contributors was the following: "Quality relates to the level of degradation a processed image shows compared to a perfect capture. A high quality image will not show any processing artifacts or defects, it will be clear and without any distortions. Quality is not the same as the aesthetic appeal of an image. For instance a beautiful image can be displayed at a low quality, and an excellent quality photo can be unappealing."

We asked contributors to follow these steps: first "pay close attention to each image trying to find visible defects", then "decide on the quality of each image and rate it", and lastly "point out the areas you've looked at when making your decision, click 3 to 5 points on each image". Contributors were reminded they would see different versions obtained from the same original images. They needed to click regions of interest even if the picture was of excellent quality.

In order to ensure the quality of the results, we set 10 test questions and we allowed a maximum of 50 answers per worker. For each image, contributors were only tested on the quality rating. All click locations were considered valid. Each question had multiple allowed answers to capture the variability of the crowd. Contributors were presented a short qualification quiz,

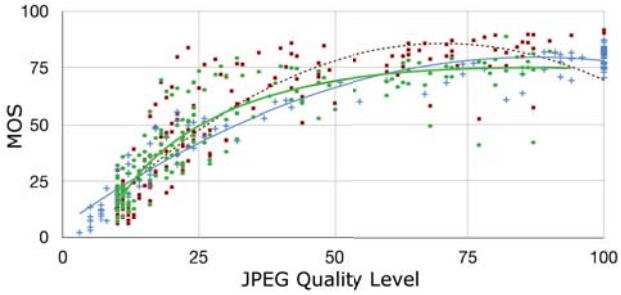

Figure 2: The relation between the JPEG quality level and the MOS given by observers is shown: Our crowdsourcing results (green dots) and lab results from Alers et al. [12] (red squares) on same the dataset in Sec. 2, as well as lab results taken from the LIVE dataset [14, 15] (blue crosses).

the completion of which allowed them to perform the rest of the experiment. Throughout the experiment random test questions were presented. These were not explicitly marked as tests questions such that contributors would pay close attention to each one. Whenever a contributor fell below 70% accuracy in answering test questions he was not allowed to continue. In this case all his previous ratings were discarded.

A total of 620 workers participated in this second experiment, with 562 (90.64%) passing the qualification test, and 556 (89.68%) staying above 70% test question accuracy throughout the course of the 44 hour experiment. On average each participant performed 14.6 judgments with a test question accuracy of 95.5%, yielding approximately 51 trusted answers per stimulus. Workers were paid $0.02 for 5 annotations, tallying up to a total price of $56.80. The contributors spanned a wide range of countries, their distribution is shown in Fig. 1.

2.3. Data processing

We generated the saliency maps following an approach similar to Liu et al. [8]. We pooled the point coordinates from click or eye tracking fixation data of all participants and applied a Gaussian filter with a fixed variance that corresponds to the size of the fovea, which is about two degrees of visual angle. This pooled self-reported saliency map is computed as

$$S_i(x,y) = \sum_{j=1}^{T} \exp\left(-\frac{(x-x_j)^2+(y-y_j)^2}{2\sigma^2}\right),$$

where S_i denotes the saliency map for stimulus i, (x, y) spans the dimensions of the stimulus, (x_j, y_j) is the location of the jth fixation/click ($j = 1 \ldots T$), T is the total number of fixations/clicks over all subjects, and σ denotes the standard deviation of the Gaussian.

3. Results and discussion

The main point of the discussion, is establishing the connection between eye tracking data and self-reported attention in two contexts: free-viewing and a specific IQA task. Before we get to this, we need to ensure the quality of our results.

3.1. Control variable for the IQA task

Our IQA experiment is similar in nature to that proposed by Alers et al. [12]. They performed an eye tracking study during an IQA task. The difference in our case is that we use crowdsourcing with self-reported locations. For both experiments quality ratings were collected. Thus, we can compare the

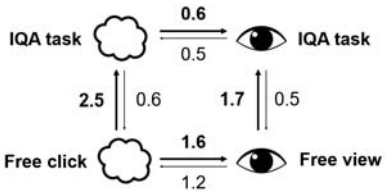

Figure 3: Kullback-Leibler divergences for comparing saliency maps. The nodes symbolize the four experiments compared: two eye tracking (taken from Liu et al. [8]) and two crowdsourcing experiments (clouds). We compare each pair of saliency maps from the two experiments, for each data-source: eye tracker and crowd. The D_{KL} value is shown on the arrows pointing from the prior to the posterior (aka true) distribution.

results of our experiments on the common variable: the mean opinion scores (MOS) for the quality of each image. The crowdsourcing data is similar in nature to that of a corresponding controlled lab experiment. The results of our IQA study closely resemble those obtained by Alers et al. [12], MOS values being similarly distributed, see Fig. 2.

Alers et al. [12] compared their results for the TUD data set to those obtained by Sheikh et al. [14] for the LIVE image data set (red squares and blue crosses in Fig. 2). With our crowdsourcing results we obtain a similar distribution (green dots). This argues for the quality of our results, suggesting that contributors performed the entire task fairly.

3.2. Self-reported and eye tracked attention

We investigate the agreement between saliency maps from different sources and task-domains by computing several measures: the Pearson correlation coefficient (PCC), Bhattacharyya distance (Bhat) and Kullback-Leibler divergence (KL). The averages over corresponding pairs of saliency maps associated with the same image version are presented in Table 1. We are considering multiple versions for an image, each one at a different quality level. The subscripts of the column names indicate the saliency maps that are compared, e.g. 'Free' refers to the comparison of eye tracking and crowdsourced free-viewing saliency maps, while 'Eye' compares free-viewing and IQA task saliency maps gathered using an eye tracker.

There is a strong agreement between the results of the similarity measures (PCC and Bhat). Pearson correlation coefficients are high across all pairings with a mean value of 0.82, indicating that the saliency maps are well correlated regardless of data source or task-domain.

PCC (high) and Bhat (low) provide a hint of how well the saliency maps overlap, whereas Kullback-Leibler divergence tells us also how they differ. We can study the dissimilarity between the saliency maps of the same stimuli by using the asymmetry of the Kullback-Leibler divergence (D_{KL}) [16]. We apply it in both directions to assess the relationship between two distributions. In this case we treat saliency maps as distributions of gaze fixations and clicks respectively.

The mean divergences between the saliency maps obtained from free viewing and IQA task experiments are presented in Fig. 3. As D_{KL} is asymmetric, $D_{KL}(P||Q) \neq D_{KL}(Q||P)$ if $P \neq Q$. In this notation, for $D_{KL}(P||Q)$, P is the posterior distribution, whereas Q is the prior.

- Whenever the divergences $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$ have comparable, small values, the two distributions P and Q are closely related.

	$\mathrm{P_{Free}}$	$\mathrm{P_{IQA}}$	$\mathrm{P_{Eye}}$	$\mathrm{P_{Crowd}}$
$\mathrm{D_{KL}}$	1.4	**0.6**	1.1	1.6
$\mathrm{D_{Bhat}}$	0.13	**0.08**	0.11	0.14
PCC	0.83	0.79	**0.85**	0.79

Table 1: Average value for distance and similarity measures between saliency maps associated with the same image version (there are multiple quality levels). The entries for the Kullback-Leibler divergence are the means $(D_{KL}(P||Q) + D_{KL}(Q||P))/2$. The highest average similarity and lowest distance between saliency maps for each measure is indicated in boldface. For P_{Free} and P_{IQA} the results are between saliency maps based on eye-tracking and self reporting experiments, for P_{Eye} and P_{Crowd} the results are between maps from free-viewing and IQA task experiments.

- When the divergences $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$ have a high asymmetry ($D_{KL}(P||Q) \gg D_{KL}(Q||P)$ and $D_{KL}(Q||P)$ is small) the distribution P is broader than and overlaps with Q well in terms of information content.

There is generally a high asymmetry between free-viewing and IQA tasks, irrespective of the data source. As we expect, viewers' responses are more comprehensive (higher information) when they are given a task compared to when they are not. In other words, a contributor's response to a task is a superposition of free-viewing and task-specific information.

From Fig. 3 it follows that the resulting saliency maps from the crowdsourced free viewing task are similar to the those from the eye tracker. A similar result has been reported before by Jiang et al. [2]. The relation appears to be stronger in the IQA task. This argues for the use of crowdsourcing for the IQA task even more so than for free viewing.

Most of the images in our experiment have a clear subject, which is a limitation of the dataset. It makes it more likely to self-report points of attention on the main subject which usually agrees with eye tracking data. To compensate for this bias, we adopted the point-and-click interface. Thus, contributors can choose more specific locations. More so, the blur kernel used on the saliency maps has a small sigma, to allow for narrower distributions. We notice that users indeed choose very particular locations in both the free-viewing and task-specific contexts. In the IQA task, they pick many off-subject locations, which are mostly related to image quality e.g. compression artifacts.

Even though reported attention correlates well with eye tracking data, they are not perfect substitutes. They capture different aspects of attention. We notice that in both free-viewing and task-specific situations, users preferred to click on meaningful locations i.e. parts of objects. In the same situations the corresponding gaze data is pointing at nearby, non-overlapping locations. This is partly due to the expectations of crowd contributors. As their performance is evaluated, users try to avoid unjustified clicks. It becomes clearer on a closer inspection of Fig. 4. On many of the images, crowd contributors consistently clicked on the same location of the sky, whereas the eye tracking data suggests a more spread out distribution of fixations.

4. Conclusions

Our experiments confirmed that crowdsourcing is a good alternative to eye tracking for studying attention during an IQA task. We showed that a simple point-and-click annotation strategy is sufficient and accurate when applied in the crowd. The distribution of attention points correlated well with eye tracking data

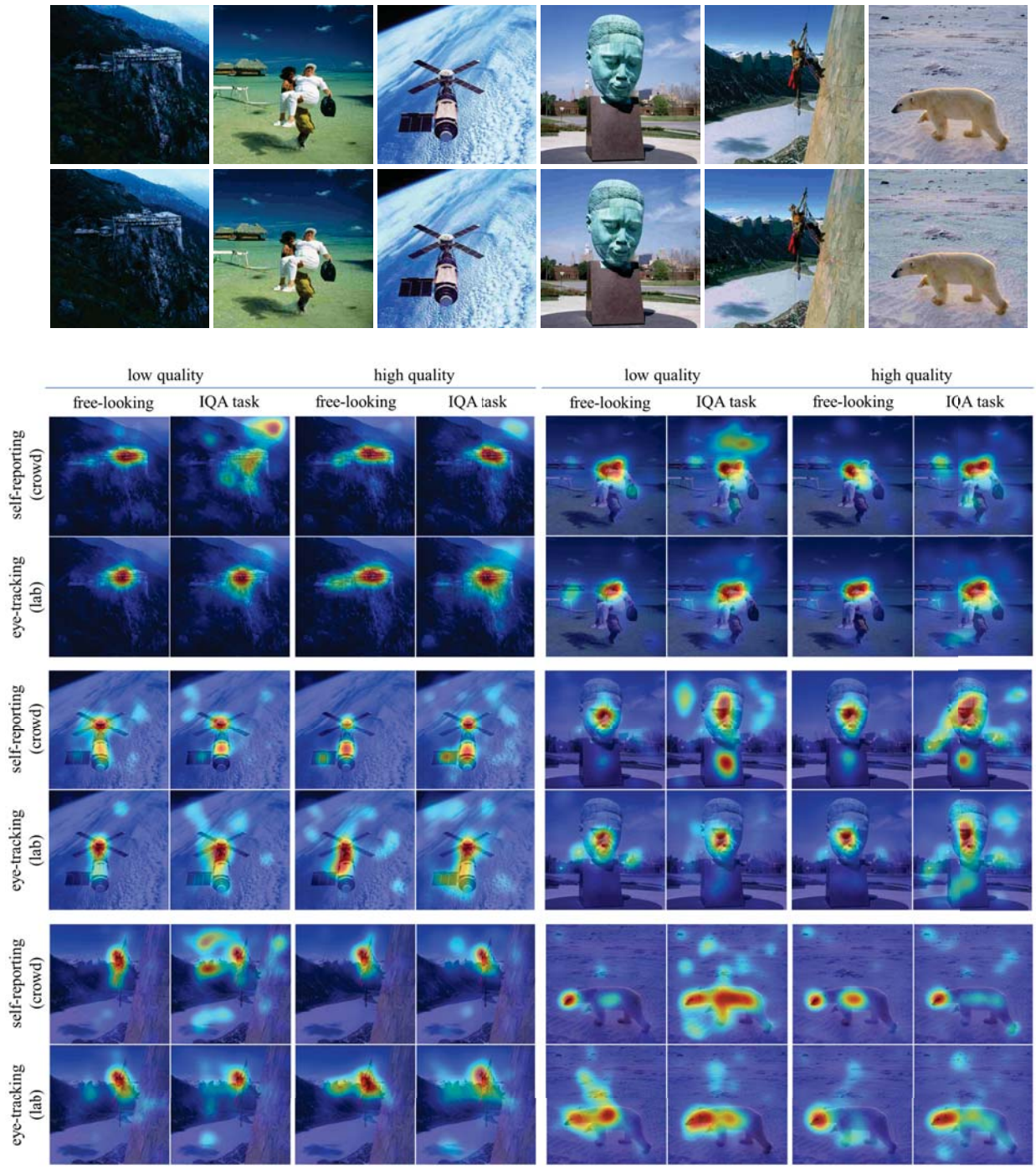

Figure 4: Example images from the TUD data-set [10] (top 2 rows). Two JPEG quality levels are displayed: highest quality on the 1st row and lowest on the 2nd. Below, there are 6 blocks of saliency maps, each consisting of two sub-blocks of 4 images. Each of the smallest blocks of 4 belongs to one source image: the low quality JPEG on the left and high on the right. The saliency maps were captured during our crowdsourced self-reporting experiments and eye-tracking [11], while users were performing a task or free-looking.

in both free-viewing and task-specific conditions. In the latter case, the connection between the two data sources was stronger. This suggests our strategy could be used for other tasks such as the detection or identification of compression artifacts or of other degradations.

Simple experiments for crowdsourced data collection have significant advantages over eye tracking (specialized equipment or web camera based [9,17]). Our approach can easily and cost-effectively scale to large numbers of images. This can be a good data source for machine learning methods that require generous annotation for task-specific attention.

Acknowledgements. This work was supported by the German Research Foundation (DFG) within project A05 of SFB/Transregio 161. We thank the anonymous reviewers for their insightful comments.

5. References

[1] U. Engelke and P. Le Callet, "Perceived interest and overt visual attention in natural images," *Signal Processing: Image Communication*, vol. 39, pp. 386–404, 2015.

[2] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Conf. on Computer Vision and Pattern Recognition (CVPR, Boston)*, 2015, pp. 1072–1080.

[3] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 50–59, 2011.

[4] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Crowdsourcing gaze data collection," *Conf. Collective Intelligence*, 2012.

[5] S. Cheng, Z. Sun, X. Ma, and J. Forlizzi, "Social eye tracking: gaze tecall with online crowds," *ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 454–463, 2015.

[6] J. Huang and R. White, "User see, user point: gaze and cursor alignment in web search," *Proc. ACM SIGCHI Conf. Human Factors in Computing Systems (Austin, Texas)*, pp. 1341–1350, 2012.

[7] N. W. Kim, Z. Bylinskii, M. A. Borkin, A. Oliva, K. Z. Gajos, and H. Pfister, "A crowdsourced alternative to eye-tracking for visualization understanding," *Proc. ACM Conf. Extended Abstracts on Human Factors in Computing Systems (CHI EA, Seol)*, pp. 1349–1354, 2015.

[8] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: Based on eye-tracking data," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, 2011.

[9] P. Lebreton, I. Hupont, T. Mäki, E. Skodras, and M. Hirth, "Eye tracker in the wild: studying the delta between what is said and measured in a crowdsourcing experiment," *Proc. ACM International Workshop on Crowdsourcing for Multimedia (CrowdMM, Brisbane)*, pp. 3–8, 2015.

[10] H. Alers, H. Liu, J. Redi, and I. Heynderickx, "TUD Image quality database: eye-tracking (release 2)." [Online]. Available: http://mmi.tudelft.nl/iqlab/eye_tracking_2.html

[11] H. Alers, L. Bos, and I. Heynderickx, "How the task of evaluating image quality influences viewing behavior," in *Int. Workshop Quality of Multimedia Experience (QoMEX, Mechelen)*. IEEE, 2011, pp. 167–172.

[12] H. Alers, H. Liu, J. A. Redi, and I. Heynderickx, "Studying the effect of optimizing the image quality in saliency regions at the expense of background content," in *Proc. SPIE 7529 Image Quality and System Performance*, S. P. Farnand and F. Gaykema, Eds. International Society for Optics and Photonics, 2010.

[13] A. Salvador, A. Carlier, X. Giro-i Nieto, O. Marques, and V. Charvillat, "Crowdsourced object segmentation with a game," *Proc. ACM Int. Workshop on Crowdsourcing for Multimedia (CrowdMM, Barcelona)*, pp. 15–20, 2013.

[14] H. Sheikh, Z.Wang, L. Cormack, and A. Bovik, "LIVE Image quality assessment database (release 2)." [Online]. Available: http://live.ece.utexas.edu/research/quality

[15] Z. Wang, A. C. Bovik, and H. R. Sheikh, "Image quality assessment: From error measurement to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600 – 612, 2004.

[16] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: state-of-the-art and study of comparison metrics," *IEEE International Conference on Computer Vision (ICCV, Sydney)*, pp. 1153–1160, 2013.

[17] P. Xu, K. Ehinger, and Y. Zhang, "TurkerGaze: crowdsourcing saliency with webcam based eye tracking," *arXiv:1504.06755*, 2015.