Saliency, Visual Attention and Image Quality

Clément Fredembach, Jue Wang and Geoff J. Woolfe
Canon Information Systems Research Australia pty ltd.
{clement.fredembach,jue.wang,geoff.woolfe}@cisra.canon.com.au

Abstract

Measuring image quality is a complex process that often requires elements of subjective analysis in order to be reliable when the final judge is a human observer. Preference studies show that slight variations in colour have drastically different outcomes in quality perception depending on where *they occur. A few* ΔE *of difference in the colour of a wall may be unnoticed, while a shift of a single* ΔE *on a face will have a major impact. The perceived quality of images does not in general depend on the image as a whole, but on a few salient regions within. Measuring saliency in images is essential to identify which parts of an image are of particular importance to human observers for subsequent processing. Consequently, a number of algorithms have been developed, that purport to automatically predict an image's salient regions.*

Most saliency algorithms are based on low-level cues, be it the physiology of the human visual system or image statistics, and are designed with a broad scope in mind. However, studies of human attention and eye movements show that visual attention maps vary significantly depending on the task, due to the influence of high-level cognitive processes.

Identifying important regions for perceptual image quality measurement being a critical task, we devise an experimental framework to obtain visual attention maps and compare these to the saliency maps predicted by state-of-the-art algorithms. Measures of correlation and precision-recall curves indicate that automatic saliency measurement is not much better than random, and far from the performance of observers, perhaps suggesting that image quality assessment has more to do with high-level cognitive processes than with low-level vision.

Saliency and visual attention

The word saliency is frequently used in the literature, although it sometimes mistakenly defines a range of different phenomena to denote "regions or areas of interest". In fact, saliency has a much narrower definition, relating to bottom-up perception; the term visual attention is employed when perception is driven by a top-down process [14]. When a human views a scene, the bottom-up approach dictates that gaze fixation sites are the by-product of local scene statistics (e.g., contrast, colour) only. The saliency of an object is thus assumed to be directly linked to these local statistics, regardless of the scene content [23].

The top-down model, on the other hand, is a cognitive one and is task-driven instead of stimuli-driven, i.e., observers think about the task they have to address when viewing a scene and the points where they fixate their attention are determined by the task at hand, not only local scene statistics [24].

Being derived from scene statistics only, saliency is much easier to address than visual attention and it has been the goal of many computer vision algorithms to predict or measure saliency in real images. The most widely used method is the one of Itti and Koch [16, 18] where an image is analysed at multiple scales according to colour, luminance, and orientation features; a biologically plausible architecture proposed by Koch and Ullman [2]. Purely computational methods have also been proposed and at times exhibit a greater accuracy than biologically-inspired ones [1, 15, 17].

These algorithms have the advantage of being simple to implement and task-independent, but that independence is also their greatest limitation. In his seminal work on eye tracking and visual attention, Yarbus noted that observers viewing a painting would look at very different regions, depending on the task/question asked [27]. The visual attention map can greatly vary despite the saliency map being the same, a clear indication that a number of visual tasks do not follow a bottom-up approach. Predicting visual attention is significantly harder than saliency, because a new model has to be built for every task. In photography, one can however consider frameworks such as memory colour-based region detection [11] or face detection [25] to be top-down models addressing the question of identifying regions where a particular aspect of an image, e.g., colour accuracy, is assessed. Identifying whether a given task is more likely to be bottom-up or top-down driven can be challenging as one has to compare the predicted image saliency with measured observer data, usually by tracking the observers' gaze across a set of representative images. Additionally, the task set to the observers has to be precisely devised, because top-down visual attention results do not generalise well [23].

Parkhurst et al. [19] found that the Itti and Koch saliency model [16] could, to a certain degree, predict observers' gaze fixation locations. The experiment consisted in four observers viewing images freely, i.e., without being assigned a specific task, for five seconds. The images depicted fairly cluttered scenes (cityscapes, home interiors, fractals). Similar images employed in a task of memorisation led [9] to conclude that saliency maps were slightly better than random to predict gaze, but their relevance decreases when more complex tasks were implemented. The usefulness of salient maps has, however, been disputed for natural scene analysis. Einhauser [8] shows that natural scene analysis is a top-down process, and that local contrast on its own is unimportant: Contrast without context is no predictor. Similarly, Underwood et al. [24] concludes that (bottom-up) saliency can be relevant but is easily overridden by cognitive influences once observers do anything else than free viewing, a conclusion shared in [14] for the task of counting people in images and in [6] for the detection of print artefacts.

This work focuses on subjective image quality assessment. The image capturing, rendering, and printing processes, introduce a number of defects and artefacts whose influence on per-

ceived image quality is not constant. For instance, noise perception depends on the frequency content of both the noise and the image, and colour shifts are more easily seen and more disturbing when occurring in the memory colours [4]. A comprehensive image quality measure has to take these elements into account. Additionally, people rarely scan an image unless instructed to do so [27] but only look at specific regions, an attitude reinforced by the photographer's rendition of the scene, where local contrast, depth of field, colour balance, and composition are used to guide the onlooker towards a particular portion of the image [5]. It follows that an image defect occurring in a region of "lesser importance" will be less noticed by observers, and its influence on the perceived quality of the image will be less significant, as illustrated in Fig. 1.

Figure 1. The same image defect applied to different parts of the same image. The difference in image quality is significant, decreasing from "satisfactory" (4/5) to "unacceptable" (2/5).

While physiological or evolutionary aspects are generally put forward to explain the varying level of sensitivity of humans to certain defects, e.g., the contrast sensitivity function threshold or the once important ability to distinguish sickness from the colour of a person's face, large unknowns remain regarding where people look in an image.

This paper evaluates the performance of state-of-the-art saliency prediction algorithms when observers are asked to assess an image's technical quality. A high degree of correlation between algorithms and observers indicates that people judge an image based on regions that are not necessarily the most conspicuous from an image statistics perspective, while a low correlation would indicate that technically image quality is, to an extent, independent from the scene content.

Eye tracking and visual attention map

Measuring visual attention without the use of external devices is challenging as observers do not provide reliable *a posteriori* information [3, 26], and asking observers where they look during the viewing experiment would alter the results. A better alternative is to employ an eye-tracking device to record observers' eye movements during the experiment without interference. An eye tracker will, however, only output snapshots of an observer's gaze sampled at regular intervals, and such a collection of points is not equivalent to a visual attention map.

Eye movements can be thought of as a succession of fixation points separated by saccades. To be considered a fixation point, the eye has to look at a given location for a time exceeding 100ms. The actual threshold used in the literature actually varies between 100 and 400ms [13]. Fixation points are of particular importance because it is only during a fixation that information is properly assimilated. Saccades, on the other hand, are rapid unconscious eye movements and are categorised either as stationary or travelling saccades, whether they occur around a fixation point to create surround awareness, or in between two distant fixations. Saccadic movement is discriminated by its high velocity ($\geq$ 30 degrees/s) and initial acceleration (3000-8000 degrees/s^2). While visual information resulting from saccades is used at a certain level of the brain to "fill in the blanks" during rapid eye movement, observers have neither control nor reminiscence of what has been seen by the eye during a saccade. Thus, saccade points are often removed from the raw eye tracking data prior to computing a visual attention map. We refer the reader to [7] for additional information.

Two other elements of interest when mapping eye tracking data are fixation duration and scanpath, however, there is no general consensus regarding their use in mapping visual attention, as the usefulness of the information they yield depends on the experiment. The rationale for taking fixation duration into account is that the longer an observer spends looking at an area of the image, the more important that area is for the task. A scanpath refers to the order in which the different parts of the image are viewed. That sequence contains valuable information when the task is highly specific -target acquisition, reading, finding differences between images- [14, 16], but in general image viewing, the scanpath is strongly observer-dependent and is of little value [22].

The sheer number of possible experimental variations implies that there is very little agreement in experimental protocols in the literature. Our method, detailed in the experimental section and illustrated in Fig. 2, maps visual attention with conservative parameters, implicitly takes into account fixation duration but ignores the order in which the different regions are viewed.

Figure 2. Left: The raw eye tracking data; Right: fixation points after pruning saccades.

We note that many more variables pertaining to eye tracking exist, although they are rarely encountered because of their task-dependency or difficulty to be accurately computed. For a comprehensive list and description, we point the reader to [20] and [7].

Experimental Methodology

Twelve observers, nine men and three women aged 25-50, took part in the experiment. All had normal or corrected to normal visual acuity and all were experts in imaging, albeit with different specialties. The observers' gaze was tracked using SeeingMachines near-infrared video cameras and software, the illumination is provided by an array of high-power, invisible, near-infrared LEDs. Reliable eye tracking performance was a prerequisite for participation in the image quality assessment test. Two observers, one male one female, had to be discounted because their glasses (both varyfocal) were only partly transparent to near-infrared light, decreasing their pupil/iris contrast too much to allow accurate tracking. A nine-point calibration target was used to verify the tracker's accuracy, as recommended by the manufac-

turer.

The stimuli comprises 20 high-quality colour images displayed at a resolution of 1400x932 pixels on a 27 inch IBM 9503-DG3 screen colour-calibrated to a D50 white point. The screen native resolution is 1900x1200 pixels and the image is surrounded with a neutral $L^* = 50$ grey boundary for display. The viewing distance was fixed at 62cm, the screen spanning 30^o of viewing angle horizontally and 22.5^o vertically, and a chin rest was placed to ensure that the distance between the observer and the screen remains constant during the experiment.

Each image was shown a total of three times to each observer. The 60 total images were displayed in a random order in six sequences of 10 images. In between each sequence, the nine-point calibration was performed. If the calibration error is greater than 3^o of subtended visual angle, the entire sequence is discarded from the results. Each image was displayed for eight seconds, and a neutral grey background with a centred cross was displayed for two seconds in between each image. The entire experiment duration was about 25 minutes.

Observers were given written instruction regarding the experiment and were asked to judge the technical quality of each image (i.e., taking into account elements such as contrast, sharpness, and colour balance, but discounting elements of composition or aesthetics) on a scale ranging from 1 (unacceptable, "unfixable" artefacts) to 5 (technically perfect image). The observers were asked to pass their judgment orally after viewing each image and were not made aware that the images being repeated were identical to the original ones.

A note on the stimuli

Figure 3 shows 18 of the 20 different images used in the experiment. To some extent, exhibiting strong differences between automatic saliency algorithms and observers-based visual attention map can be done by selecting images with known bias such as "captcha" text or people wearing bright colours (the observers focus on the letters and the faces, while the algorithms mostly detect the background and the clothing). This typical limitation of saliency algorithms has been partially addressed in [17] where face and people detectors are used in addition to low-level cues. We are, however, interested in the process involved in evaluating image quality rather than simply criticising saliency algorithms. As such, we have selected images that are similar to the ones of [19] and [1], using the categories of "city", "landscape", "animals" and "flowers" so that the results obtained herein can objectively be compared to prior art, instead of simply be attributed to a difference in stimuli.

Results: visual attention maps and image quality assessment

Visual attention maps are calculated for each observer-image pair where the calibration data error was lower than 3^o, which yielded 530 valid observations out of 600.

First, we remove saccadic points from the eye tracking data. A point is considered a saccade if its velocity is greater than 30^o/s, which for our 60Hz system is equivalent to having no neighbour inside a 2^o radius. Due to the relative low temporal sampling frequency of the tracker, onset acceleration was not taken into account as it cannot be reliably computed.

Non-saccadic points are then assessed to determine whether or not they are fixation points, by filtering the non-saccadic data with a time window of 200ms (a conservative minimum fixation time) and a spatial window equivalent to the calibration error. Non-fixation points are discarded.

The visual map proper is computed by creating an empty grid of identical size to the image (1400x932 pixels). For every fixation point, we add a count of one at the corresponding location in the grid. Once all the fixation points have been addressed, we convolve the grid with a 1^o gaussian (the manufacturer's stated precision of the eye tracker) and a function of foveal eccentricity, as measured in [21], to take into account the physiology of the human visual system. Sample images and their corresponding visual attention map (averaged over all observations and subsequently normalised) are shown in Fig. 4

All observers are experts in the field of photography or image processing. As such, we expect their judgment on image quality to be coherent with each other. This assumption is corroborated by looking the image quality assessment results; there is little variability across observers. We note that selected test images were in general of good technical quality, none of them showing major artefacts such as compression artefacts or noise; their average score ranges form 2.95 to 4.23 with a standard deviation between 0.2 and 0.4, see Fig. 5 for details.

Composition, experimental, and observer bias

A common aspect of all our visual attention maps is the prominence of the centre as a region of significant attention, see Fig. 4. This "centre bias" effect has been previously reported [9, 19, 22] and has to be taken into account prior to comparing our visual attention maps to the saliency maps automatically generated that are designed to be free from bias.

To correctly address this behaviour, we first have to distinguish the three plausible causes of such bias: composition, experimental, and observer-based. Composition bias is the tendency for many of photographers to centre the main subject in the image. Accurately measuring this bias necessitates to analyse a large number of images, but a couple of studies [10, 12] on a database of 10,000 images showed that while its influence was not negligible, centre-surround type of decompositions were not especially common. We further argue that composition bias, whatever its influence, should not be removed; if the most salient object happens to be in the centre, it should be detected by algorithms and humans alike.

Experimental bias, on the other hand, is unrelated to image content but directly created by the experimental setup. In our case, the "cross image" showed in between each stimuli fixates the gaze of the observers at the centre of the image, potentially biasing the resulting visual maps. Finally, observer bias has been proposed in [22] as a natural tendency for people to pay attention at the centre of the image before selecting an area to look at.

To measure the influence of experimental and observer bias, we devise an experiment with two identical sets of 10 stimuli images. For the first set we replicate the original experimental protocol, while for the second the "cross image" was altered to appear in the top-right corner instead of the centre. These two "control " sets were shown to all observers directly after the six sets of stimuli. The images were chosen so as not to have any saliency or point of interest in either the centre or the top-right corner. They

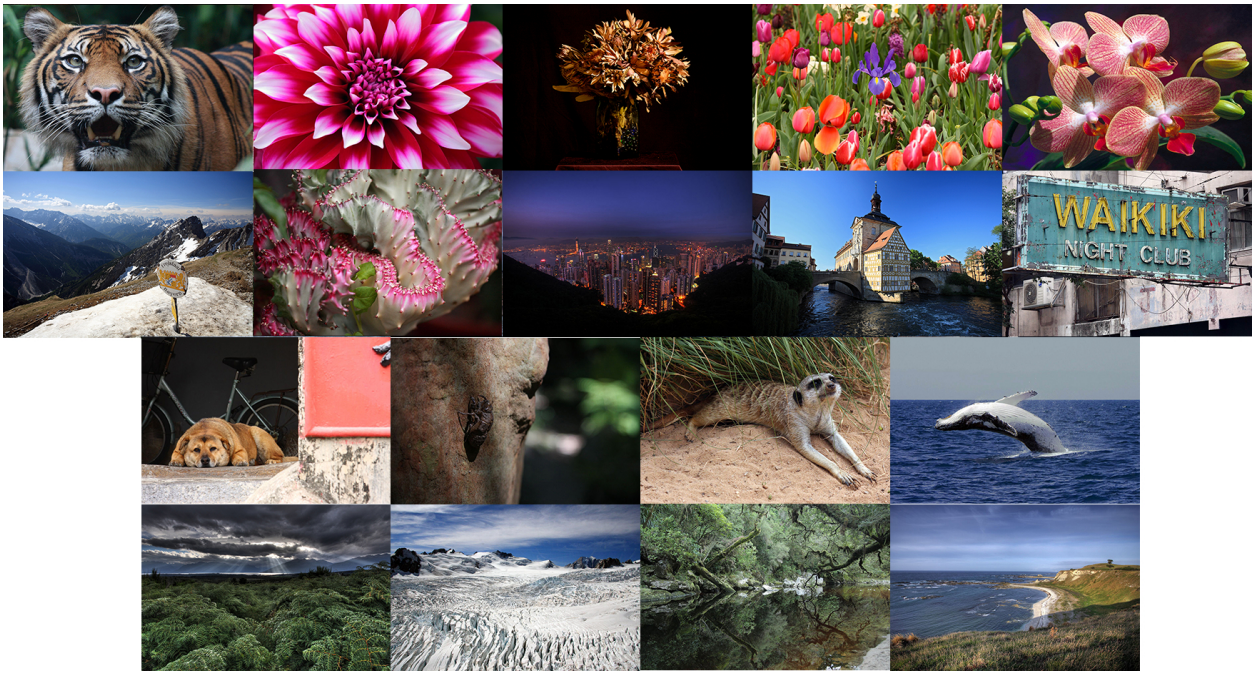

Figure 3. 18 of the 20 different stimuli sorted according to their ascending scores (best rated image: 4.23/5, worst rated image: 2.96/5). Images courtesy of Barry Drake, David Morgan-Mar and Scott Rudkin.

Figure 4. Two sample images and their average visual attention maps.

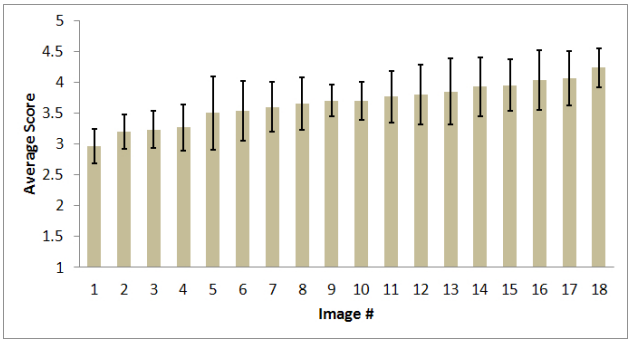

Figure 5. Average scores and standard deviations of the 18 test images. Images are all of similar quality, a feature reflected in the compactness of the scores and the low variations across observers. Error bars are two σ wide.

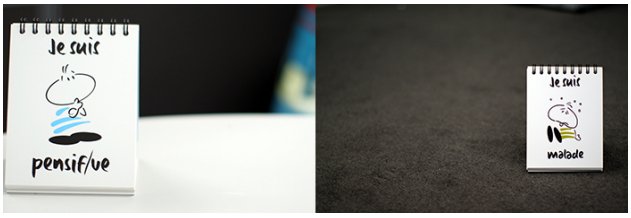

Figure 6. Sample images used in the bias-measure experiment.

consist in a drawing and some text cast against a uniform and background blurred (by photographing the scene with an aperture of f/1.2), such as the ones shown in Fig. 5.

In the visual maps from the control set with a centre cross, despite no object or saliency being present, significant central visual attention was observed. The top-right cross control set, however, exhibit a significant visual attention in the top-right corner, but none in the centre, see Fig. 6 for an illustration. There is therefore no observer-centric bias, and as the images were chosen so as not to have any central composition bias, we can conclude that the entire observed bias of the test comes from the experimental design.

We propose to remove the bias from our test images by calculating the average length of time observers gaze at the location of the cross after the image was shown on the screen, also called "dwell time". Crucially, while the average dwell time was 800ms, the variance across observers was of 300ms, a significant difference. To be faithful to the data, we opted to remove the experimental bias on an observer-basis, i.e., calculating the average dwell time for each observer and discarding it from the test image data. As the average bias time across different images for the

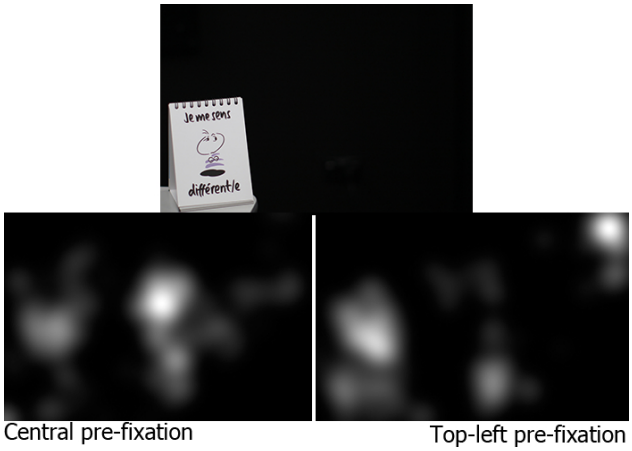

Figure 7. Average eye tracking data and visual attention map for a central control image and a top-right control image. While both maps show significant visual attention where the cross was, the top-right control set has no central visual attention.

same observer was constant, we surmise that it is a latency issue, linked to the observer's reaction time to the presentation of a new stimulus.

Saliency vs. Visual Attention

In this section, we compare the performance of the visual attention maps generated by human observers to state-of-the-art saliency prediction algorithms. Itti and Koch (IK) [16] is the *de facto* benchmark of saliency prediction. IK employs low-level vision cues (intensity, orientation and colour contrast) at various scales to identify potential regions of interest. Achanta et al. (AHES) [1] detects salient regions by calculating the ΔE of images filtered by two different gaussian kernels. Finally, Judd et al. (JEDT) [17] constructs a feature vector of both low- and high-level (face and person detector, composition bias) features and train an SVM detector on the eye tracking data of 1000 images freely viewed (i.e., no task was given) by 15 observers (assumed to be naïve).

For every algorithm, saliency maps are generated using the code provided by the authors, with default parameters. All maps are normalised between 0 (no saliency) to 1 (maximum saliency) to allow a realistic comparison.

The first measure we employ is cross-correlation. For each one of the 20 images, a ground truth attention map is created by averaging the attention maps of nine observers (out of 10). Cross-correlation is then computed between this ground truth and the tenth observer, IK, AHES, and JEDT. The procedure is repeated 10 times, selecting a different observer in a "leave one out" fashion. The results, Fig. 7, show that human observers exhibit a much greater correlation to the ground truth than either of the algorithms. Importantly, the correlation between observers is quasi-constant over the entire image set, while AHES is not significantly better than random.

In a second test, we compute ROC curves for each of the observers and algorithms. The ground truth images are the same, but we binarise it with a threshold corresponding to one standard deviation of the Gaussian function used to build the attention maps. True and false positive rates are then calculated for AHES, IK, and the observers by varying their threshold value from 1 (no salient regions are detected) to 0 (the entire image is considered salient). The curves are plotted in Fig. 8 where we observe that the performance of both algorithms is significantly lower than the observers'.

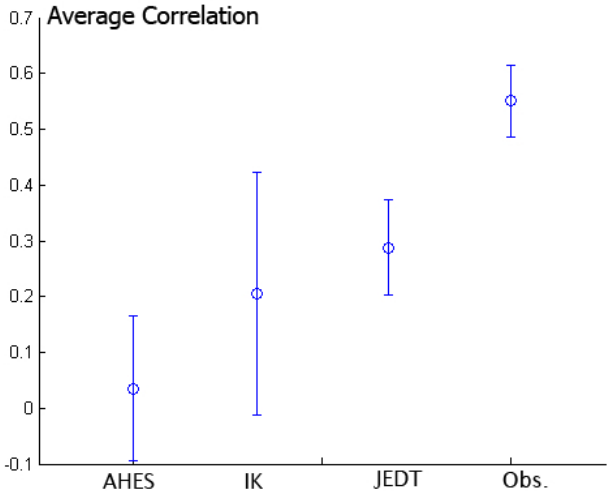

Figure 8. Average cross correlation between the ground truth visual attention map and the maps from the observers, IK, AHES, and JEDT. The correlation between observers is almost independent form the image content and much higher than either IK, AHES, or JEDT. Error bars represent two standard deviations.

The relative low performance of prediction algorithms in this test can partly be attributed to having expert observers judging technical image quality, while most algorithms are designed with free viewing by naïve observers in mind. The present experiment, due to being more specific, is harder to predict using low-level features. In addition, the similar performance of IK and JEDT indicate that the effectiveness of high-level features significantly depends on the task (and data) at hand.

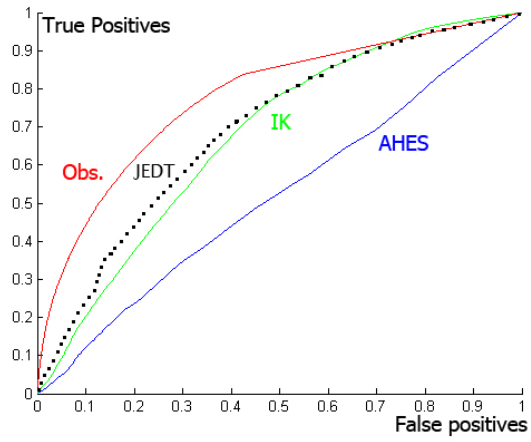

Figure 9. ROC curves for observer data, IK, and AHES, compared to the ground truth. The observer performance, while not perfect, is significantly higher than either of the two algorithms.

Conclusion

Saliency and visual attention are essential to understand and measure for applications that have a human observer in mind. Prior art demonstrates that different processes, from bottom-up saliency to top-down cognitive control, are employed by the human brain for scene viewing depending on the type of images being shown and the task an observer has to address.

In this work, we considered the important topic of perceived image quality, to find out whether judging technical image quality is stimuli- or cognition-driven. A panel of 10 expert observers was assembled and visual attention maps were calculated using standard techniques. Comparing these visual maps to the saliency maps produced by three state-of-the-art algorithms lead us to conclude that bottom-up saliency is not a good predictor of where observers look when assessing image quality and that free viewing gaze fixations are not necessarily similar to the ones obtained when assessing image quality.

Importantly, the degree of agreement or correlation between observers does not appear to be image-dependent. However, other factors can greatly influence the results of such a study, such as the task given to observers and their level of expertise. Factors that have to be accurately identified and addressed at the experiment design stage.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk. Frequency-tuned salient region detection. In *IEEE conf. on Computer Vision and Pattern Recognition*, 2009.

[2] C. Koch andS. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, pages 219–227, 1985.

[3] J. S. Babcock, J. B. Pelz, and M. D. Fairchild. Eye tracking observers during rank order, paired comparison, and graphical rating tasks. In *Proc. IS&T PICS conference*, 2003.

[4] P. Bodrogi and T. Tarczali. Colour memory for various sky, skin, and plants colours: effect of the image context. *Color Research and Application*, pages 278–289, 2000.

[5] Guy T. Buswell. *How people look at pictures: A study of the psychology of perception in art*. University of Chicago Press, 1935.

[6] G. Cao, M. Pedersen, and Z. Baranczuk. On the use of saliency maps for the detection of print artifacts. In *IS& T european conference on Colour in Graphics, Imaging, and Vision*, 2010.

[7] Andrew T. Duchowski. *Eye tracking methodology: theory and practice*. Springer Verlag, 2007.

[8] W. Einhauser and P. König. Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, pages 1089–1097, 2003.

[9] T. Foulsham and G. Underwood. What can saliency models predict about eye movements? spatial and sequential aspecs of fixations during encoding and recognition. *Journal of Vision*, pages 1–17, 2008.

[10] C. Fredembach, F. Estrada, and S. Süsstrunk. Memory colour segmentation using class-specific eigenregions. *Journal of the Society for Information Display*, pages 921–931, 2009.

[11] C. Fredembach, M. Schroeder, and S. Süsstrunk. Region-based image classification for automatic color correction. In *IS&T/SID 11th Color Imaging Conference*, 2003.

[12] C. Fredembach, M. Schroeder, and S. Süsstrunk. Eigenregions for image classification. *IEEE Trans. On Pattern Analysis and machine Intelligence*, pages 1645–1649, 2004.

[13] J.M. Henderson. Human gaze control during real-world scene perception. *TRENDS in Cognitive Sciences*, pages 498–504, 2003.

[14] J.M. Henderson, J. R. Brockmole, M. S. Castelhano, and M. Mack. Visual saliency does not account for eye movements during visual search in real-world scenes. In *Eye Movements: A window on the brain, Gompel, Fischer, Murray, and Hill Eds*, 2007.

[15] X. Hou and L. Zhang. Saliency detection: a spectral residual approach. In *IEEE conf. on Computer Vision and Pattern Recognition*, 2007.

[16] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, pages 1489–1506, 2000.

[17] T. judd, K. Ethinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision*, 2009.

[18] C. Koch L. Itti and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE trans. on Pattern analysis and Machine Intelligence*, pages 1254–1259, 1998.

[19] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocution of overt visual attention. *Vision Research*, pages 107–123, 2002.

[20] A. Poole and L. J. Ball. Eye tracking in human-computer interaction and usability research: current status and future prospects. In *Encyclopedia of human computer Interaction, C. Ghaoui Editor*, 2005.

[21] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Photographic granularity and graininess iii. *Journal of the Optical Society of America*, pages 217–263, 1947.

[22] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision research*, pages 643–659, 2005.

[23] G. Underwood and T. Foulsham. Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology (Colchester)*, pages 1931–1949, 2006.

[24] G. Underwood, T. Foulsham, E. van Loon, L. Humphreys, and J. Bloyce. Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of cognitive Psychology*, pages 321–342, 2006.

[25] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, pages 137–154, 2004.

[26] J. Wang, D. Chandler, and P. Le Callet. Quantifying the relationship between visual salience and visual importance. In *Proc SPIE/IS&T Electronic Imaging vol. 7525*, 2010.

[27] Alfred L. Yarbus. *Eye movements and Vision*. Plenum Press, 1967.

Author Biography

Clément received his M.Sc. from EPFL, Switzerland in 2003 with internships in Fujifilm Japan and Gretag Imaging, and his Ph.D. from the University of East Anglia, Norwich, U.K. in 2007 on illuminant estimation, shadow detection and removal. From 2007 to 2009 he was a post-doc of the IVRG/LCAV at EPFL working on novel aspects of near-infrared imaging. He is now a senior research engineer with Canon Research (CiSRA) in Sydney, Australia. He is a member of the IS & T.