Visual Attention Modeling: Region-of-Interest Versus Fixation Patterns

Ulrich Engelke, Hans-Jürgen Zepernick, Anthony Maeder

27th Picture Coding Symposium

2009 Chicago

VISUAL ATTENTION MODELING: REGION-OF-INTEREST VERSUS FIXATION PATTERNS

Ulrich Engelke[†]*, Hans-Jürgen Zepernick*[†]*, and Anthony Maeder*[*]

[†]Blekinge Institute of Technology, PO Box 520, 372 25 Ronneby, Sweden, E-mail: uen@bth.se
[*]University of Western Sydney, Locked Bag 1797, Penrith South DC, NSW 1797, Australia

ABSTRACT

Visual attention (VA) is an integral property of the human visual system. The deployment of a VA model can be beneficial for many image and video applications, such as, compression, transmission, and quality assessment. However, the design of a VA model is highly dependent on the subjective VA data used as ground truth and the application that the model is intended for. In this paper, we discuss two ways of obtaining subjective VA data that can subsequently be used to develop VA models; selective regions-of-interest and visual fixation patterns. The feasibility of both methods will be discussed, in particular, with respect to visual quality assessment.

Index Terms— Visual attention, region-of-interest, visual fixation pattern, image quality.

1. INTRODUCTION

In visual content there are typically objects or regions that particularly draw the viewer's attention, usually referred to as regions-of-interest (ROI). The underlying phenomenon is known as visual attention (VA) and is an integral property of the human visual system. The level of attention can vary strongly and is influenced by many factors, such as size, shape, colour, and location of the object [1]. Furthermore, humans and their faces were found to strongly draw attention. However, an image can usually not be apprehended with a single observation since the human retina is highly space variant in processing and sampling of visual information with the highest accuracy in the central point of focus, the fovea. Therefore, rapid eye movements (called saccades) carry the focus to the salient parts of an image. The factors guiding the eye movements are generally considered to be either bottom-up (i.e. task-independent, saliency-driven, and fast) or top-down (i.e. task-dependent, volition-controlled, and slower) [2].

There are a number of image and video applications that would highly benefit from the incorporation of a VA model including, for instance, compression where a higher bit-rate can be assigned to the ROI, thus maintaining a better visual quality of the salient information. In image or video communication, unequal error protection (UEP) can be deployed by means of stronger channel codes for the ROI and as such, better protection against error-prone transmission channels. Finally, in objective quality evaluation distortions that occurred in the ROI may receive more impact on the overall quality metric. The complexity of the deployed VA model plays a vital role depending on the application it is intended for. In some objective quality evaluation tasks, for instance, a complex saliency map [2] or importance map [3] may improve quality prediction performance. On the other hand, in UEP one may prefer a more simple representation of VA in order to enable an efficient assignment of channel codes.

Given the above, the selection of an appropriate VA model is a crucial but nevertheless difficult task. One important factor is the validation of a particular model. The question here is; what can actually serve as the ground truth for the VA model to be validated on? Gaze patterns obtained in eye tracking experiments [4] are widely utilized for model design and validation. In [5] we discussed an alternative subjective VA information based on the selection of ROI. Here, human observers actively choose particular ROI in a set of images that are presented to them. In this context, it would be interesting to gain some insight into the relationship between such an active selection of ROI and the gaze patterns that are passively recorded in eye tracking experiments. In this paper, we will present some initial results of an eye tracking experiment that we conducted. We will analyze the relationship between the gaze patterns and selective ROI that we obtained in an earlier experiment. However, given the very recent eye tracking results and also the space restrictions in this paper, this analysis is by no means attempted to be complete, but rather intended to raise some open issues and to stimulate discussion.

The paper is organized as follows. Section 2 and Section 3, respectively, summarize an ROI experiment and an eye tracking experiment that we conducted. In Section 4 the relationship of ROI and visual fixation patterns is discussed. Conclusions are drawn in Section 5.

2. SELECTIVE REGIONS-OF-INTEREST

We conducted a subjective ROI experiment, here referred to as EXP_{ROI}, at the Blekinge Institute of Technology, Sweden. The experiment is explained and analyzed in detail in [5] and will be summarized in the following.

Thirty viewers participated in EXP_{ROI}. Their task was to

Fig. 1. Visual attention through region-of-interest selections of 30 human observers obtained in EXP_{ROI}.

select a region in a number of images that drew most of their attention. For this purpose, the viewers were presented one training image, two stabilization images, and seven widely used test images. All images were of dimension 512×512 pixels and presented in grey scale. The training image was used to explain the simple selection process. The stabilization images then further allowed for the viewer to adapt to the selection process. The actual test images were used for the evaluation. We did not put any restrictions on the size of the ROI and on the time given for the selection process. However, for simplicity we only considered a single rectangular ROI.

The outcomes of EXP_{ROI} are visualized in Fig. 1 where the ROI selections of all 30 viewers are plotted as intensity shifts (with a blue hue) within the corresponding images. The top row shows humans and, in particular, human faces and the bottom row shows more complex images and natural scenes. The brighter the area, the more ROI are overlapping and as a result, the stronger the VA for a particular region. One can see that the VA is very high for human faces and especially their eyes. In the 'Tiffany' image (top right) the mouth seems to also attract a lot of attention. In the 'Barbara' image (top left) one can see that the VA is more focused on the whole face rather than its details, which may be explained by the smaller area that it covers compared to the other faces. For more complex scenes such as 'Goldhill' (bottom left) and 'Peppers' (bottom right) the VA seems to be less concentrated.

3. VISUAL FIXATION PATTERNS

In order to determine visual fixation patterns (VFP) for the images from EXP_{ROI}, we recently conducted an eye tracking experiment, here referred to as EXP_{VFP}, at the University of Western Sydney, Australia. Fifteen viewers participated in EXP_{VFP}. In addition to the images from EXP_{ROI}, we also presented distorted versions of these images and asked the viewers to rate the quality using a single stimulus assessment method. However, given the context of this paper, in which we want to compare the VFP to the ROI, we will for now focus on the gaze patterns of the seven images from EXP_{ROI}.

We used an EyeTech TM3 eye tracker [6] to record the viewers' gaze patterns while observing the images. The TM3 is equipped with two infrared light sources and an infrared camera to track the gaze of both eyes. It was installed under the screen and the viewers were seated at a distance of approximately 60 cm. The eye tracker was calibrated independently for each viewer. Gaze points (GP) were recorded at a frequency of about 40-45 GP/sec. Each image was presented for 8 seconds, resulting in about 320-360 GP per person and image. A mid-grey screen with a central fixation point was shown between images, for the viewers to focus on.

Due to the high recording frequency of the eye tracker, the gaze patterns do not only contain GP that belong to fixations but also GP recorded during saccades. Vision, however, is suppressed during saccades and as such, the corresponding GP do not contribute to VA. It is thus common practice to cluster the GP of close temporal and spatial proximity into fixations to create VFP. The VFP for all 7 images are presented in Fig. 2. As in Fig. 1, the brightness indicates the amount of VA, which can either be due to a long fixation of a single viewer or to aggregated fixations of several viewers. It can be seen for all images that the GP are widely spread, indicating, that the viewers' eye movements covered large parts

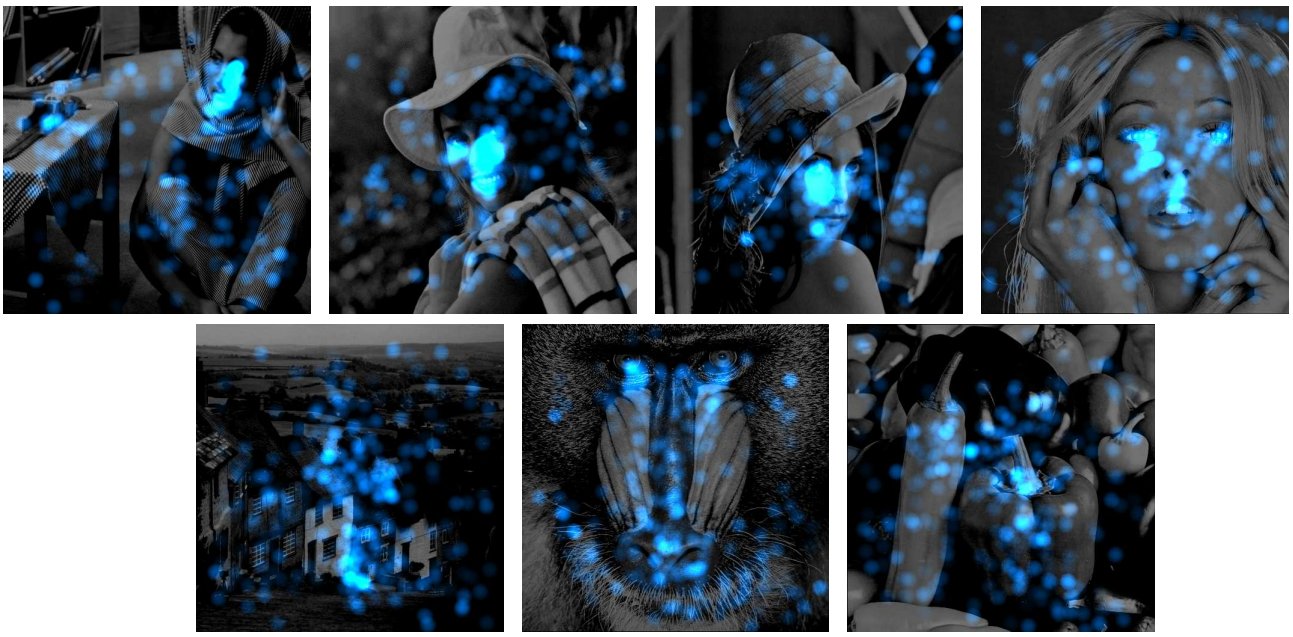

Fig. 2. Visual attention through visual fixation patterns of 15 human observers obtained in EXP_{VFP}.

of each image. However, it is also apparent that the fixations are clustered around the human faces and their eyes, whereas such clusters are less prevalent in the complex scenes.

4. DISCUSSION

4.1. ROI or VFP for visual attention modeling

The comparison of the VA information in Fig. 1 and Fig. 2 shows that the selective ROI and the VFP agree very well. In both cases, the attention is focused in the face images and more spread in the complex scenes. The following more detailed observations are also interesting to point out. In the 'Barbara' image the object on the table on the left receives a noticeable amount of attention in both EXP_{ROI} and EXP_{VFP}. In the 'Tiffany' image, the mouth receives substantial attention with the VFP, as was already observed with the ROI. Finally, the man walking the streets in the 'Goldhill' image receives much VA in both EXP_{ROI} and EXP_{VFP}.

Despite these strong relationships, there are some substantial differences in the way that the VA information is obtained. Firstly, the ROI are actively selected by the viewers whereas the VFP were passively recorded. Thus, an informed decision is made in EXP_{ROI}, whereas one may be uncertain about the reliability of the VFP which depend highly on the alertness of the viewer. As such, there may be fixations that may not be due to high attention but, for instance, due to the viewer being unfocused or 'dreaming'. On the other hand, one may argue that the bottom-up attention gets lost in case of the selective ROI, since the selection is volition-controlled and the saliency-driven, early attention is not captured. This issue is discussed in more detail in the following section.

4.2. Early versus late visual attention

Unlike with the selective ROI, as we deployed it, there is a temporal factor captured with the VFP, meaning that information is available as to in which order the different regions drew the viewer's attention. In this respect it is interesting to evaluate which fixations actually relate more to the ROI; the saliency-driven, early fixations or the volition-controlled, late fixations. In fact, we found that there is indeed a distinct difference between the early VFP of a viewer and the late VFP. An illustrative example is given in Fig. 3 for the 'Goldhill' image. Here, we show the first three and the last three fixations of each viewer in the left image and the right image, respectively. It is apparent that the VFP are very different. The saliency-driven, early attention is in high agreement between many viewers and is focused on the man walking down the hill. On the other hand, the late VA is much more spread. In addition, some long fixations can be observed for the late VA (visualized by bright blue) indicating that the late attention may be used to analyze the image in greater detail. Similar observations have also been made for the other images.

Comparing the images in Fig. 3 to the 'Goldhill' image in Fig. 1 reveals that the maximum VA from the ROI coincides better with the early attention VFP, indicating that most viewers selected the ROI according to what they looked at first. As such, it would be interesting to not just be able to predict the spatial locations of visual fixations but also the sequential order, which in turn would provide valuable information as to where the ROI is located. However, in [7] it was found that the prediction of spatial gaze patterns can be done with reasonable accuracy, but it was concluded that the sequential order of the pattern could not be predicted.

Fig. 3. Early (left) and late (right) visual attention.

Fig. 4. Visual attention for a distorted image.

4.3. VA modeling for visual quality assessment

One application in which VA is often neglected is visual quality assessment. Most contemporary metrics usually assess the quality uniformly over the whole image. Given the discussion in the previous sections one may, however, expect that distortions in salient regions would impact stronger on quality degradations as compared to the rest of the image. In [8], VA was considered but no general improvement in quality prediction accuracy could be achieved. It was concluded that the saliency information and the degradation intensity have to be jointly considered in a pooling function. We have recently proposed a framework [5] in which we used a mean ROI computed from the 30 selected ROI (see Section 2) to include a simple VA model into an objective quality metric that we previously designed. We achieved a significant improvement in quality prediction performance, not just of our metric, but also of two other contemporary quality metrics. From these examples one can see that the VA modeling has to be conducted with great care.

In the context of quality assessment it is also interesting to evaluate how distortions in an image actually influence the VA. In [9] it was found that uniformly distributed distortions, such as blur, do not cause the VFP to change very much as compared to the VFP of the undistorted image. On the other hand, distortions caused by JPEG and JPEG2000 coding were found to disturb the VFP. It was concluded that especially large amounts of spatially localized distortions drew the viewer's attention. A preliminary result from EXP_{VFP} provides some more evidence for this conclusion in terms of the distorted 'Lena' image shown in Fig. 4. It can be observed that much attention is drawn away from Lena's face to the distorted row in the lower half of the image. In future work, we will investigate further how much the VA shift is impacted by the distortion being located inside or outside the ROI.

5. CONCLUSIONS

We revealed strong relationships between two different kinds of subjective VA information; selective ROI and VFP. Indications were found that early VFP more precisely reflect VA as represented by the ROI. The prediction of the sequential order of visual fixations, however, seems to be an open research topic. It can be further concluded that the integration of the subjective VA information into an objective VA model is a task that has to be handled with great care. This does not necessarily mean that the model has to be complex, as suggested by the example given in relation to visual quality prediction.

6. REFERENCES

[1] W. Osberger and A. M. Rohaly, "Automatic detection of regions of interest in complex video sequences," in *IS&T/SPIE HVEI VI*, Jan. 2001, vol. 4299, pp. 361–372.

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. PAMI*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[3] A. J. Maeder, "The image importance approach to human vision based image quality characterization," *Pattern Recog. Lett.*, vol. 26, no. 3, pp. 347–354, Feb. 2005.

[4] A. L. Yarbus, *Eye Movements and Vision*, Plenum, 1967.

[5] U. Engelke and H.-J. Zepernick, "Optimal region-of-interest based visual quality assessment," in *IS&T/SPIE HVEI XIV*, Jan. 2009, vol. 7240.

[6] EyeTech Digital Systems, "TM3 eye tracker," http://www.eyetechds.com/, 2009.

[7] C. M. Privitera and L. W. Stark, "Algorithms for defining visual regions-of-interest: Comparison with eye fixations," *IEEE Trans. PAMI*, 2000.

[8] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," in *IEEE ICIP*, Oct. 2007, vol. 2, pp. 169–172.

[9] C. T. Vu, E. C. Larson, and D. M. Chandler, "Visual fixation patterns when judging image quality: Effects of distortion type, amount, and subject experience," in *IEEE SSIAI*, Mar. 2008, pp. 73–76.