IMPACT OF IMAGE APPEAL ON VISUAL ATTENTION DURING PHOTO TRIAGING

Syed Omer Gilani,[1] *Ramanathan Subramanian,*[2] *Huang Hua,*[1] *Stefan Winkler,*[2] *Shih-Cheng Yen*[1]

[1] Department of Electrical and Computer Engineering, National University of Singapore
[2] Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore

ABSTRACT

Image appeal is determined by factors such as exposure, white balance, motion blur, scene perspective, and semantics. All these factors influence the selection of the best image(s) in a typical photo triaging task. This paper presents the results of an exploratory study on how image appeal affected selection behavior and visual attention patterns of 11 users, who were assigned the task of selecting the *best* photo from each of 40 groups. Images with low appeal were rejected, while highly appealing images were selected by a majority. Images with higher appeal attracted more visual attention, and users spent more time exploring them. A comparison of user eye fixation maps with three state-of-the-art saliency models revealed that these differences are not captured by the models.

Index Terms— Image appeal, Photowork, Triaging, Eye fixations, Visual attention, Saliency

1. INTRODUCTION

Since the advent of inexpensive digital cameras with extensive storage capacity, users tend to take multiple pictures of the same scene, and select the best picture(s) afterwards. It is now fairly standard practice to take 2-3 shots of a scene, and up to 8-10 shots for certain situations [1], *e.g.*, family photos with kids or wedding ceremonies. Consequently, photo selection or "triaging" is an important task when working with digital photo collections.

Prior research [2] has noted that aspects such as lighting conditions (exposure and white balance), scene framing and perspective, face and body pose, actions of people in the scene as well as basic image quality collectively contribute to ***image appeal*** (IA) and play an important role in the selection of a representative set of photos from an image collection. Building an automated photo triaging system would entail predicting interesting regions in each photo, while considering IA-related aspects, and then determining the most appealing image(s). Therefore, we set out to better understand how people engage their visual attention during a typical photo triaging task involving multiple images with differing IA. This paper attempts to answer the following questions:

1. Are visual attention characteristics influenced by IA-related differences?
2. If certain aspects of visual attention are indeed sensitive to the degree of image appeal, how well do existing saliency models reflect these phenomena?

We performed a photo triaging study where 11 users were asked to select the *best* image from each of 40 groups. Each group comprised images having similar content; apart from some object/camera motion across images, they also differed with respect to one or more IA factors. As users performed the task, their eye movements were recorded using an eye tracker. Our experiments suggest that, while the general regions fixated upon by participants are consistent across images, there are significant differences between the visual attention patterns observed for good and bad quality photos.

We then employed three static saliency models incorporating low and high level features to generate saliency maps for the image groups, and compared the results against the observed eye fixation maps. The saliency models do not effectively reflect the changes in attention patterns when quality differences exist, motivating the need for further research in this regard.

2. RELATED WORK

Pioneering work on what interests people as they compare images is described in [3]. Apart from image quality issues such as motion blur and exposure, this study observes that local structure changes (*e.g.*, pose and facial expression changes, occlusions and appearance changes) are identified as key regions of interest during comparisons. These changes are modeled using parameters such as optical flow field divergence along with factors influencing single image saliency to develop a co-saliency model, which is then used to create collection-aware crops. While this work explores factors affecting photo triaging, an actual triaging task is not posed to users, and the comparison is restricted to pairs of images.

Another study investigating how people compare retargeted (intelligently resized) images against originals is [4]. The main findings of this work are that (a) when semantically

This work was supported by research grants from ADSC's Human Sixth Sense Programme and the Human Factors Engineering Programme, both from Singapore's Agency for Science, Technology and Research (A*STAR).

relevant image content is removed during retargeting, there are significant differences between eye fixation maps for the original and retargeted images, and (b) even if the retargeting procedure induces large artifacts in semantically less salient regions, such changes go unnoticed. However, this study also does not involve a typical triaging task.

The influence of a free-viewing task vs. a picture quality assessment task on visual attention patterns is examined in [5]. This study observes that a quality comparison task has a significant effect on eye movements– higher fixation durations are observed on unimpaired pictures compared to quality-impaired images, suggesting that observers tend to memorize some image parts. While [5] relates to image quality and impairments, we are concerned with image appeal, which involves a host of other factors as mentioned above.

A user study that investigates what constitutes image appeal is [6]. This study reveals that an image may be appealing because certain image regions are appealing. From the user study, a number of low-level factors are identified and combined to measure a set of image appeal metrics. Two image appeal metrics are discussed, one that ranks images based on image appeal and correlates very well with human ranking, and a second which does well at retrieving highly appealing images from a collection. However, this user-study only involves interviews with photographers rather than end-users and does not consider an image triaging task as such.

3. PHOTO TRIAGING EXPERIMENT

3.1. Test Material

40 image groups typical of photo collections were used in our study, all of which depicted family or social scenes (people performing various activities), representative of personal photo albums or online photos. Each group represented a scene captured by 3-5 shots taken in sequence, which varied mainly due to (a) *lighting* variations owing to changes in camera exposure and white balance (7 groups), (b) *motion blur*, where some scene objects appear blurred due to camera or object motion (8 groups), (c) *perspective changes* (8 groups), where the semantically relevant scene objects appeared either closer/farther from the camera (zoom variations) or near the center/periphery of the scene, and (d) *scene semantics* relating to the pose and/or facial expressions of persons in the scene (8 groups). The remaining 9 groups involved combinations of these factors. Fig. 1 presents one exemplar group for each of these factors.

3.2. Experimental Set-up and Protocol

11 university students (5 male, 6 female) aged 20-33 years with normal or corrected eyesight participated in the photo triaging study. They were paid a token fee for participation. All participants were naive to the purpose of the experiment and were simply asked to 'select the *best* or most appealing image' from each of the groups. To this end, they viewed the images of size 800×600 pixels on a 21.5" LED monitor from a distance of $\approx$60 cm with their heads firmly placed on a chin-rest. The order of presentation of the 40 groups was randomized across subjects. As participants viewed the images, their eye movements were recorded using the *EyeLink 1000* eye-tracker. The eye-tracker had a sampling frequency of 2 kHz and was accurate to within $0.25°$ of visual angle.

Participants could interactively access the next/previous image in the group using the right/left arrow keys for comparing images, and select the *best* image by pressing the return key. While subjects could view an image for as long and for as many times as they desired, a maximum selection time limit of 30 seconds per group was set, failing which a null selection was assumed. To ensure eye-tracker accuracy, re-calibration was performed after every 10 groups.

4. DATA ANALYSIS

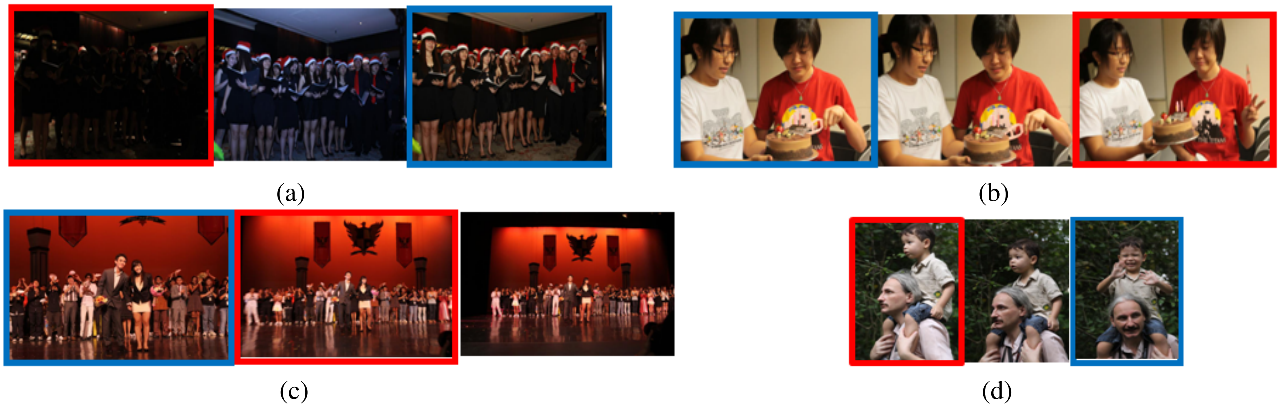
4.1. Image Selections

Denoting the presentation of each image group as a trial, there were totally 11 (subjects)$\times$40 (groups) = 440 trials. Of these, a valid *best* image selection was made in all-but-two trials. Subjects carefully compared images in each group before selecting the *best* image: in 365 trials, at least one image per group was revisited by users, and the mean proportion of images reviewed prior to selection was found to be 0.84. For 32 out of the 40 groups, over 50% of the participants concurred on their *best* image selections.

For 19 of the 32 groups with majority agreement, the last image in the group was chosen as the *best*. This was perhaps because most groups depicted snapshots of an event, and photographers usually attempt to capture progressively better representations of the event of interest. However, when the final image was affected by IA-related quality issues, it was rejected in favor of a preceding and more appealing image (Fig.1b,c). Overall, users selected those images that (a) were devoid of lighting and blur defects, (b) contained the objects of interest captured at maximum resolution and near the image center (center bias), (c) captured most people in the scene facing the camera and exhibiting prominent facial expressions. We also determined the least appealing or *worst* image from each group based on the ratings of three independent experts. The images with *best* and *worst* appeal for the groups in Fig.1 are highlighted using blue and red borders respectively.

4.2. Visual Attention Characteristics

Human eye movements comprise *saccades* (rapid eye movements enabling selective attention) and *fixations* (rest state during which visual information is assimilated). Given the nature of images used in the study, most eye fixations were observed on people (and predominantly on their faces) and



(a)

(b)

(c)

(d)

Fig. 1. Images varied mainly due to (a) lighting changes, (b) motion blur, (c) perspective, and (d) scene semantics. The blue and red borders respectively highlight the *best* (based on majority agreement) and *worst* images in each group.

the objects they were interacting with (*e.g.*, ball, knife, *etc.*).

While analyzing eye movements, we considered only those 32 groups which had majority consensus on the *best* image. Per subject, we computed the proportion of fixations, saccades, and total fixation duration for each image in the group. On average, 44.1% fixations and 45.9% saccades were made on the photo the user selected; that image also accounted for 41.2% of the duration fixated over the entire group. Therefore, participants' visual attention resources were biased towards the image they found most appealing.

Intuitively, one would expect users to also give considerable attention to the first image, in order to understand the visual content, and subsequently focus on differences during picture comparisons. We aggregated the data for all subjects and computed how much visual attention was devoted to the first image, *best* image (as per majority vote), and remaining images in the group. For 23 groups where the *best* photo was different from the first, these accounted for 35.8%, 40.1%, and 23.8% of the group fixation duration respectively, whereas for the remaining 9 groups, the first (*best*) photo accounted for 50.8%.

We then focused on the *best* (most appealing) and *worst* (least appealing) images. If certain aspects of visual attention are influenced by differences in image appeal, one can expect them to differ at least for the *best* and *worst* images. A two-sample Kolmogorov-Smirnov (KS) test confirmed a significant difference in fixation duration for the *best* and *worst* images ($p < 10^{-7}$). Another visual attention aspect we were concerned with was the fixation *entropy*, which measures the breadth of scene details covered by user fixations. Upon pooling user fixation maps, we measured entropy of the cumulative map for the *best* and *worst* photos. Across all pairs, mean entropy for the *best* image was higher than the *worst* image (2.4 vs. 2) and a two-sample t-test showed a significant entropy difference between the *best*-*worst* pairs ($p < 10^{-4}$).

Since IA differences can be construed as a change in the level-of-detail between the *best* and *worst* images, this result mirrors the finding of [7], which explores how image resolution affects eye fixation patterns. The authors observe that when the image size is gradually reduced (thereby simulating a blur at normal viewing resolution), the fixation entropy reduces until that point where people can still perceive the image gist. The fact that more scene details are explored for the higher-quality (or *best*) images is also evident from Fig. 2, where the average fixation map is overlaid on the *worst* and *best* images (rendered as alpha channel for see-through effect) for three of the groups shown in Fig. 1.

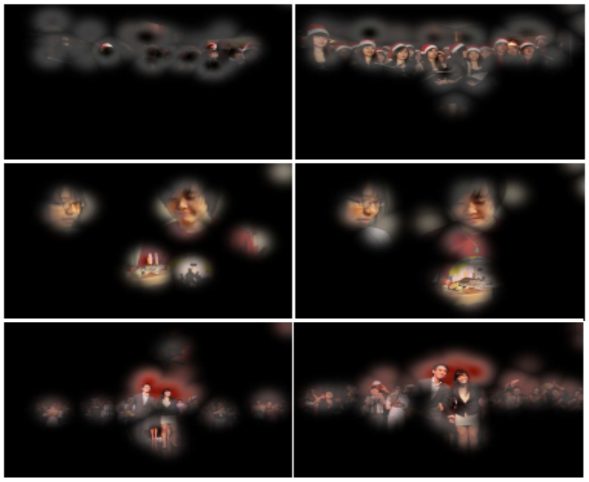


Fig. 2. Average fixation maps overlaid on the *worst* (left) and *best* (right) images from the groups shown in Fig.1a–c.

In summary, our analysis showed that (a) participants largely concurred on their choice of the *best* image for most (32/40) photo groups, (b) the *best* or most appealing image in each photo group attracted maximum visual attention, accounting for at least 40% of the time spent examining the entire group, and (c) the breadth of scene details explored was significantly higher for the *best* image as compared to the *worst*. All of these suggest that visual attention patterns were reflective of differences in image appeal.

5. EYE FIXATION MAPS VS. SALIENCY MODELS

In this section, we examine how well three state-of-the-art saliency algorithms model IA-related visual attention characteristics. We consider only spatial saliency algorithms, even though the image groups involve object/camera motion, as image appeal is intrinsic to a particular photo.

We generated saliency maps corresponding to the models described in [8, 9, 10] for the 32 image groups. The Attention based on Information Maximization (AIM) model [8] defines bottom-up saliency based on maximizing the information sampled from the image, upon learning an independent component analysis basis from a set of natural images. The Saliency Using Natural Statistics (SUN) model [9] combines bottom-up and top-down information to compute the saliency map using the difference-of-Gaussians (DoG) and ICA filters. The Judd *et al.* saliency model [10] combines bottom-up features [11], mid-level gist [12], and top-down face [13], person, and car features [14] to compute the final saliency map.

5.1. Evaluation Metrics

To compare the user eye fixation and saliency maps, we used similarity score [15] and area under the Receiver Operating Characteristic (ROC) curve or AUC. The similarity score measures how similar two maps are. In each map, the distribution of saliency values at pixel locations (i, j) is normalized so as to sum to one, and the similarity score is computed as the sum of the minimum values at each point:

$$S = \sum_{\forall i,j} \min(P_{i,j}, Q_{i,j}) \text{ with } \sum_{\forall i,j} P_{i,j} = \sum_{\forall i,j} Q_{i,j} = 1.$$

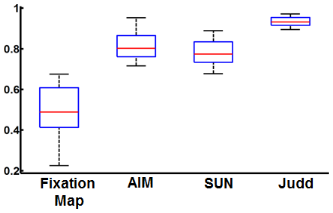
The AUC indicates how well a saliency map predicts human fixations. In the AUC scoring method, one set of values are sampled from human fixated locations (S_{fix}). A second set of values are obtained by randomly drawing samples from a uniform distribution over the saliency map (S_{rand}). These two sets of values are then thresholded using different cutoffs τ to determine true positives ($S_{fix} > \tau$) and false positives ($S_{rand} > \tau$), and generate the ROC curve. Area under the ROC curve gives the AUC score. An AUC score of 1.0 means that saliency values at *all* fixated locations are higher than those at randomly selected locations, whereas AUC≤0.5 implies that saliency values at fixated locations are similar to, or lower than, those at random locations.

5.2. Observations

We compared the eye fixation maps and saliency maps generated by the three models for the *best* and *worst* image in each group based on the similarity and AUC scores. Eye fixations were predicted at better than chance level by all saliency models in both cases. Nevertheless, we found that the similarity and AUC score distributions for the *best* and *worst* images did not vary significantly for the three saliency models as per the KS test (Fig. 3a). This is surprising as one would expect a saliency model to work better for the *best* images as compared to the *worst* images, which were impaired by factors such as lighting and blur. Given that significant differences were observed between the user fixation patterns for the *best* and *worst* images, one would also expect some dissimilarity between the saliency maps for the *best*-*worst* image pairs, if the saliency algorithms perfectly modeled human visual attention. Therefore, we compared the similarity between (i) fixation maps for the *best*-*worst* image pairs, and (ii) saliency maps for the same (Fig. 3b). The median similarity score between user fixation maps for the *best* and *worst* images was found to be slightly less than 0.5. In contrast, the saliency maps for the *best*-*worst* image pairs were found to be quite similar (median similarity score of 0.78 or higher). Exemplar saliency maps generated using [10] for the *best* and *worst* images from two photo groups (Fig. 4) support this observation. Cumulatively, the above results imply that the saliency models do not effectively capture differences in user fixation patterns owing to IA-related differences, motivating the need for further research in this direction.

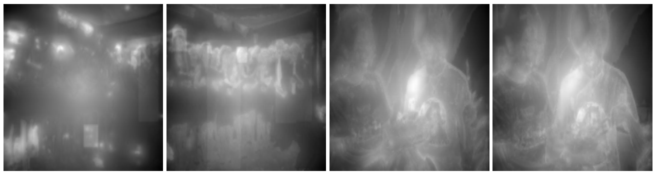
Model	p_{sim}	p_{AUC}
AIM	0.06	0.58
SUN	0.3	0.67
Judd et al.	0.13	0.62

(a)



(b)

Fig. 3. (a) Significance levels obtained on comparing similarity/AUC score distributions for the *best* and *worst* images. (b) Boxplot for the distribution of similarity between fixation/saliency maps for the *best*-*worst* image pairs.



(a) (b)

Fig. 4. Saliency maps for the *worst* (left) and *best* (right) images from the photo groups shown in Fig. 1a (a) and Fig. 1b (b) as computed from the Judd *et al.* saliency model [10].

6. CONCLUSION

The present study clearly demonstrates that image appeal (IA) plays a crucial role in photo triaging. Visual attention is sensitive to IA-related differences, such that attention patterns significantly differ for the most and least appealing photos in a group. However, these differences are not captured by current saliency models. Future work involves extending the dataset to include a wider range of scene types, and developing saliency models sensitive to IA-related aspects.

7. REFERENCES

[1] Seon J. Kim, Hongwei Ng, Stefan Winkler, Peng Song, and Chi-Wing Fu, "Brush-and-drag: A multi-touch interface for photo triaging," in *Proc. International Conference on Human-computer Interaction with Mobile Devices and services (MobileHCI)*, San Francisco, CA, 2012, pp. 59–68.

[2] Andreas E. Savakis, Stephen P. Etz, and Alexander C. P. Loui, "Evaluation of image appeal in consumer photography," in *Proc. SPIE Human Vision and Electronic Imaging*, San Jose, CA, 2000, vol. 3959, pp. 111–120.

[3] David E. Jacobs, Dan B. Goldman, and Eli Shechtman, "Cosaliency: Where people look when comparing images," in *Proc. ACM Symposium on User Interface Software and Technology (UIST)*, New York, NY, 2010.

[4] Susana Castillo, Tilke Judd, and Diego Gutierrez, "Using eye-tracking to assess different image retargeting methods," in *Proc. Symposium on Applied Perception in Graphics and Visualization (APGV)*, Toulouse, France, 2011, pp. 7–14.

[5] Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, Dominique Barba, and Arnaud Tirel, "Task impact on the visual attention in subjective image quality assessment," in *Proc. European Signal Processing Conference (EUSIPCO)*, Florence, Italy, 2006.

[6] Pere Obrador, "Region based image appeal metric for consumer photos," in *Proc. IEEE Workshop on Multimedia Signal Processing (MMSP)*, Cairns, Australia, 2008, pp. 696–701.

[7] Tilke Judd, Frédo Durand, and Antonio Torralba, "Fixations on low-resolution images.," *Journal of Vision*, vol. 11, no. 4, pp. 1–20, 2011.

[8] Neil D. B. Bruce and John K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.

[9] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell, "SUN: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, 2008.

[10] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, "Learning to predict where humans look," in *Proc. International Conference on Computer Vision (ICCV)*, Kyoto, Japan, 2009.

[11] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[12] Aude Oliva and Antonio Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[13] Paul Viola and Michael J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[14] Pedro F. Felzenszwalb, David A. McAllester, and Deva Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage,Alaska, 2008, pp. 1–8.

[15] Tilke Judd, Frédo Durand, and Antonio Torralba, "A benchmark of computational models of saliency to predict human fixations," Tech. Rep., Massachusetts Institute of Technology, 2012.