Creation of Individual Photo Selections: Read Preferences from the Users' Eyes

Tina Walber
Chantal Neuhaus
Steffen Staab
University of Koblenz-Landau, Germany
{walber, neuhaus, staab}@uni-koblenz.de

Ansgar Scherp
University of Mannheim, Germany
ansgar@informatik.uni-mannheim.de

Ramesh Jain
University of California Irvine, USA
jain@ics.uci.edu

ABSTRACT

The automated selection of satisfying subsets from large collections of photos is a central challenge in multimedia research. Objective criteria like the depiction of persons or the photo quality are met by existing approaches. But it is difficult to know the users' personal interest, which plays an important role in the selection process. The expected spread of devices with eye tracking support in the near future allows us to measure this interest in a new way. In an experiment with 12 participants, we derive the most interesting photos of a collection for every person from gaze information recorded during the free viewing of the photos. We can show that the eye tracking information delivers valuable information about the users' preferences by comparing the results to a manual selection. The selection based on gaze information significantly outperforms baseline approaches and improves the results by up to 17%. For photo sets of personal interest this improvement is even up to 23%.

Categories and Subject Descriptors

H.5.2 [**User Interfaces**]: Input devices and strategies

General Terms

Human Factors

Keywords

Eye tracking, personalized photo selection, personal interest

1. INTRODUCTION

Existing solutions for creating photo selections are based on content-information such as histograms [1], make use of clustering based on photo meta data like capture time [1] or GPS coordinates [2]. These tools find their limitation in considering the user's interest. Savakis et al. [3] investigated humans selecting photos from a collection with the result that the selections are subjective and differ between the users. They also determine that it is hard to identify the attributes on which the decision is based, as it is part of a high-level human cognitive process. In order to provide an extension of photo selection algorithms, a new approach is needed that can cover personal preferences. In this work, we use gaze as measure for capturing individual user's interest.

Systems that can detect the eyes and can calculate the viewing direction from cameras integrated in common devices like Tablet PCs are already on the market (e. g., Natural User Interface Technology, OKAO Vision[1]). One of the market leaders in eye tracking systems offers with the Tobii Rex System[2] the first device for less than 1000 USD. Given the recent development in low-cost eye tracking hardware, we investigate the possibilities to create selections based on gaze information collected from users while viewing photos. Central to this approach is that the users are not burdened with additional tasks such as annotating the images. The manual selection of photos is often perceived as cumbersome, e. g., during the creation of slide shows or photo books. At the same time, users usually enjoy viewing photos.

Our approach benefits from this viewing and creates valuable information about the interest in certain photos from the recorded gaze information. The fixations in gaze paths show the areas of the highest visual perception and they are an indicator for the users' attention. Eye movements are influenced by different factors from low-level information like contrasts and colors to high-level factors like a given task. We investigate if the influence of "interest" is high enough to derive valuable information from the gaze paths about the users' preferences when viewing photos *without* a specific task, besides getting an overview of the photo collection. The attention while viewing the photos is used for creating personalized photo selections. Photos with the highest attention, i. e., those that are fixated longest, are assumed being most interesting to the user and should be part of a selection. We compare the gaze-based selections to selections based on related work. The analysis results on our data sets show that we achieve a significantly better selection of photos compared to a random baseline and a baseline based on [1]. We also compare the quality of the selection for sets

MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright 2013 ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502165.

[1] http://www.omron.com
[2] http://www.tobii.com

of photos with a personal interest and sets that are less interesting. Personal interest in a photo collection significantly improves the results.

2. RELATED WORK

Approaches for the automatic creation of photo selections typically use low level image features and photo meta data. For example, Sinha et al. [2] present a framework for the generation of representative subsets of photos from large personal photo collections by using multidimensional content and context data like GPS coordinates or tags. Li et al. [1] create selective summaries based on time stamps and face features. Several approaches use eye tracking information to identify relevant images in search result lists and use this information as implicit user feedback to improve image search, e.g., [4, 5]. Kozma et al. [4] show that a comparison of the implicit gaze feedback with explicit user feedback by clicking on relevant images and a random baseline are promising. Pasupa et al. [6] apply a support vector machine algorithm using eye tracking information together with content-based features to rank images. Also Klami et al. [5] perform implicit user feedback on image search result lists by means of gaze information. They show that it is possible to use gaze information to detect image relevance in a controlled setup. Eye movements are strongly influenced by the task a human is performing as Yarbus [7] has already shown in 1967. In the existing work, concrete search tasks are given to the users. Images which are relevant concerning these tasks can be identified from the gaze information. In our approach, no concrete task for the viewing is given to the users. The subjects are asked to view the photos for getting an overview, but they were not told to concentrate, e. g., on the "best" photos. Thus, we do not measure how much a photo fits a given task but the personal interest of a user in a photo.

3. EXPERIMENT

We have conducted an experiment to compare our gaze-based approach for the creation of a selection S from a photo collection C to automatic approaches. After viewing the photos, a personal selection is created based on gaze information. These selections as well as baseline selections are compared to manually created selections, which serve as individual ground truth selection. 12 subjects (6 of them female) participated in our experiment. The age of the subjects ranged between 25 and 62 years (average: 34.4, SD: 13.4). Seven of them were graduate students, two post docs, the others were working in other professions.

Procedure.

In a first step, the subjects were asked to view all photos with the instruction to "get an overview". They were told that they would afterward, in a second step, create a selection of photos for their private photo collection. The instruction was to only view the photos, without time limit. While viewing the photos, the gaze paths were recorded by a Tobii X60 eye tracking device. Sets of 9 photos were presented in 3x3 photo grids. By clicking a button, the user advanced to the next set. The maximum height and width of the photos was set to 330 pixels. After having viewed all photos, the user created a manual selection S_m for every set by selecting exactly three photos from each set. The photos were selected in a ranked order (most important to third most important). No concrete instruction was given concerning the selection criteria. Thus, the subjects could apply their own (perhaps even not conscious) criteria. As motivation for the subjects, the manually created selection of photos was sent by email. Finally, the subjects completed a questionnaire.

Data set.

The experiment data set encloses two photo collections C_A and C_B taken during two different events A and B. Collection C_A consists of 164 photos, thus 18 sets of 9 photos. C_B has 126 photos (14 sets). The photos show events of two different research groups. The activities include team work situations, group meals, as well as leisure activities like bowling and hiking. The photos were taken by two resp. three different persons. Participants of the two events do not know each other, nobody participated in both events. The subjects in our experiment participated in event A or know participants of event A. Thus, the data set consists of photos that are of personal interest to the participants (C_A) and of photos of less interest, taken during event B. The sets were presented in chronological order but the photos in every set were displayed in random order. The first collection shown to the subjects is alternately C_A and C_B.

Gaze analysis.

The preprocessing of the raw eye tracking data for identifying fixations is performed with the fixation filter offered by Tobii Studio with its default thresholds. The obtained fixations are analyzed by the eye tracking measure *fixationDuration*. This measure calculates the sum of the durations of all fixations on a photo. The three photos with the highest results are selected as eye tracking selection S_e.

Baselines.

As the simplest possible baseline, we introduce a random baseline as used, e.g., by Kozma et al. [4]. It randomly selects three photos for every set. For the creation of the random selection S_r, the average over 10 random samples is built. Additionally, we implemented a baseline approach based on [1], which calculates a selection fully automatic based on three criteria:

- Many photos are taken within a short period of time during an important moment [1]. S_{bt} is built from photos taken from high concentrations in time.
- It is known that depicted persons play an important role in the selection of photos [1]. S_{bf} is built from photos best fulfilling three face conditions: a high number of depicted humans, a maximum area of the image is covered by human faces, and persons are depicted that frequently appear in the data set.
- Photo quality calculated by a sharpness score as presented by Xiao et al. [8]. The photos with the highest quality scores are selected for S_{bs}.

For each criterion a score is calculated for each photo. The scores are normalized for every set of 9 photos and lie between 0 and 1. The scores are summed up for each photo and the photos with the highest results build the baseline selection S_b. Clustering techniques as proposed by [1] delivered weak results on our data set and are

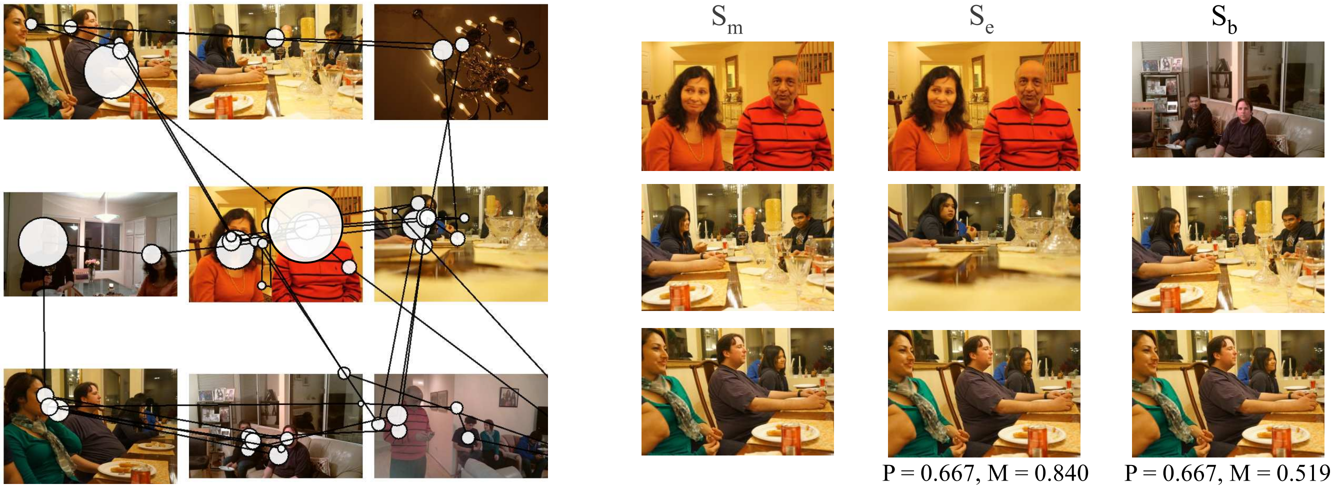

Figure 1: Sample photos including gaze path visualization (fixations displayed as circles) and three selections: manual S_m, eye tracking based S_e, and the baseline selection S_b.

therefore not included in the baseline. We also extend the baseline approach by information gained from gaze. For the selection S_{b+e}, *fixationDuration* values are added to S_b as an additional score.

Data analysis.

In the analysis of the data, we compare different selections in their capability to create selections similar to the manual selection S_m, which serves as "ground truth". For the evaluation, we calculate two measures for comparing selections S_x to the manual selection S_m. The first measure is precision P, which calculates the percentage of photos occurring in selection S_x as well as in S_m. The second one is the M measure, presented by Bar-Ilan et al. [9], which also considers the ranking of the photos in the selections. It is calculated as follows:

$$M' = \sum_{Z} \left| \frac{1}{S_m(i)} - \frac{1}{S_x(i)} \right| + \sum_{T_1} \left(\frac{1}{S_m(j)} - \frac{1}{k+1} \right) + \sum_{T_2} \left(\frac{1}{S_x(j)} - \frac{1}{k+1} \right)$$

where $Z = S_m \cap S_x$. T_1 and T_2 are the search result elements appearing only in one of the two selections: $T_1 = S_m \setminus S_x$, $T_2 = S_x \setminus S_m$. k is the number of investigated ranked result elements. Thus, $k = 3$ as the subjects selected the top three photos. Finally, the normalizing factor is the worst possible result, i. e., when the two lists do not share a single value. For $k = 3$ this factor is 3.12, thus $M = 1 - \frac{M'}{3.12}$.

4. RESULTS

Every set of nine photos was averagely viewed for 15.6 seconds (SD: 12.8). The shortest viewing time was 1.5 seconds and the longest 121.1 seconds. In average, 34 fixations were recorded per set (SD: 21.2). The highest number of fixations is 163. For one set, no fixation was recorded. Besides that the lowest number of fixations is 2.

Selections.

The results are evaluated by means of precision P and M measure as described in Section 3. The average results over all users and all sets can be found in Table 1. An example of a photo set with gaze path and three selections can be found in Figure 1.

Table 1: Precision P and M measure for eye tracking selection and different baseline selections

	S_r	S_e	S_b (S_{bt}, S_{bf}, S_{bs})	S_{b+e}
P	0.34	**0.42**	0.36 (0.36, 0.33, 0.33)	0.37
M	0.45	**0.53**	0.46 (0.45, 0.44, 0.44)	0.48

We can see that the best results are obtained by the gaze-based selection S_e. It improves the results compared to the baseline selection S_b by 17% for P and by 15% for M. The weak results for the baseline system show the difficulties of predicting the personal criteria for photo selections [3] by simple measures based on low-level features and meta information. The gaze-extended baseline approach S_{b+e} leads to a small improvement, compared to S_b. We have conducted a Friedman test with the average values per user to investigate the statistical significance of the results. We found that the differences between the selections S_r, S_e, and S_b are significant ($\alpha < .05$) for P ($\chi^2(2) = 11.167, p = .004$, $n = 12$) and M ($\chi^2(2) = 10.500, p = .005$, $n = 12$). Post-hoc analyses with pairwise Wilcoxon tests are conducted with a Bonferroni correction for the significance level (now: $\alpha < .017$). The eye tracking based selection S_e significantly outperforms S_r ($Z = -2,903, p = .004$) and S_b ($Z = -2,510, p = .012$) for the M measure as well as S_r ($Z = -2,667, p = .008$) for precision P. A comparison of precisions P for S_e and S_b does not show a significance ($Z = -2,275, p = .023$). Also no significance was determined comparing the results for S_r and S_b ($Z = -1,49, p = .136$). We did not find any correlations between the average viewing time on a set and the quality of the created selection.

The Influence of Personal Interest.

We distinguish between photo sets of personal interest and of less personal interest, as described in the Section 3. The results of the questionnaire support our approach. 10 of the subjects rated the collection C_A, which contains photos taken during their own event, as more interesting than the other collection. A Wilcoxon signed-rank test shows a statistically significant difference between the answers concerning the interestingness over all users ($Z = -2.292, p = .022$). The average viewing time for C_A and C_B differs only slightly; photos of C_B were viewed longer (15.3 vs. 16.1 seconds).

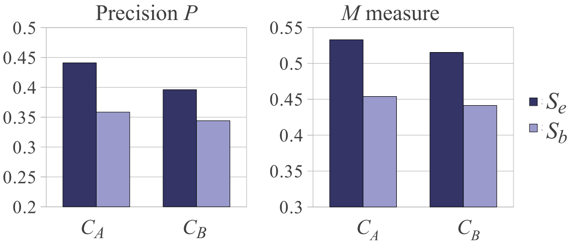

Figure 2: Results for interesting collection C_A and less interesting collection C_B.

P and M were calculated separately for C_A and C_B, the results can be found in Figure 2. The results are better for both measures when the subjects were viewing photos of personal interest. Compared to the baseline selection S_b an improvement of 23% for P and 17% for M can be stated. A Wilcoxon test showed a significant difference only for precision P ($Z = -2,353, p = .019$), but not for M ($Z = -1,098, p = .272$).

Analysis of Potential Bias.

Gaze paths could be subject to bias, e. g., from the starting position of a gaze path or from learned viewing patterns like a concentration on the center of the screen, where usually the content is displayed. Therefore, we investigated if such a bias occurs in our data. We can investigate a bias by analyzing the fixations depending on the positions of the images on the screen. In Figure 3 (a), the positions of the very first fixations on a photo set over all users are presented. One can see a clear concentration on the lower right corner and the center of the grid. The button for continuing with the next set is located in the right lower corner of the screen, which seems to have influence on the first fixation, as well a concentration on the center. In Figure 3 (b), the sum of all fixations over all users per image position in the grid is shown. We can see that the fixations are spread over the whole set quite evenly. We can also see a weak concentration in the center. However this concentration is small and does not play a role in our analysis as the photos were displayed in a random position.

5. CONCLUSION AND FUTURE WORK

In this work, we have shown that the usage of gaze information in the selection of photos delivers better results compared to random selection and baseline selections using low-level features and meta data. The results for a photo collection of personal interest are better compared to a collection of less personal interest.

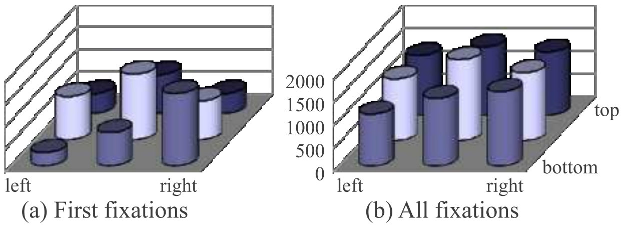

Figure 3: Number of fixations per image position for the first fixation of every user on every set (a) and all fixations (b).

In a next step, it has to be investigated if gaze information provided while viewing photos in less controlled environments, e. g., in a file system, delivers the same results. Additionally, the presented approach can be extended by the annotation of image regions describing the users' interest. This information can be used for authoring slide shows or photo books. The possibility to identify personal preferences from gaze information can also be adapted to other domains. One application could be the recommendation of products based on previous fixations on photos or objects in photos.

Acknowledgments

The research leading to this paper was partially supported by the EU project SocialSensor (FP7-287975). We thank all subjects for participating in our experiment.

6. REFERENCES

[1] J. Li, J.H. Lim, and Q. Tian. Automatic summarization for personal digital photos. In *Pacific Rim Conference on Multimedia*, volume 3, pages 1536–1540. IEEE, 2003.

[2] Pinaki Sinha, Sharad Mehrotra, and Ramesh Jain. Summarization of personal photologs using multidimensional content and context. *Multimedia Retrieval*, pages 1–8, 2011.

[3] A.E. Savakis, S.P. Etz, and A.C.P. Loui. Evaluation of image appeal in consumer photography. In *Electronic Imaging*, pages 111–120. International Society for Optics and Photonics, 2000.

[4] L. Kozma, A. Klami, and S. Kaski. GaZIR: gaze-based zooming interface for image retrieval. In *Multimodal interfaces*. ACM, 2009.

[5] A. Klami, C. Saunders, T.E. De Campos, and S. Kaski. Can relevance of images be inferred from eye movements? In *Multimedia information retrieval*, pages 134–140. ACM, 2008.

[6] K. Pasupa, C. Saunders, S. Szedmak, A. Klami, S. Kaski, and S. Gunn. Learning to rank images from eye movements. In *Workshops on Human-Computer Interaction*, 2009.

[7] A.L. Yarbus. *Eye movements and vision*. Plenum press, 1967.

[8] J. Xiao, X. Zhang, P. Cheatle, Y. Gao, and C.B. Atkins. Mixed-initiative photo collage authoring. In *ACM Multimedia*, pages 509–518. ACM, 2008.

[9] J. Bar-Ilan, M. Mat-Hassan, and M. Levene. Methods for comparing rankings of search engine results. *Computer Networks*, 50(10):1448–1463, 2006.