Smart Photo Selection: Interpret Gaze as Personal Interest

Tina Walber
WeST Institute
University of Koblenz, Germany
walber@uni-koblenz.de

Ansgar Scherp
Kiel University, Germany
Leibniz Information Center for Economics, Kiel, Germany
mail@ansgarscherp.net

Steffen Staab
WeST Institute
University of Koblenz, Germany
staab@uni-koblenz.de

ABSTRACT

Manually selecting subsets of photos from large collections in order to present them to friends or colleagues or to print them as photo books can be a tedious task. Today, fully automatic approaches are at hand for supporting users. They make use of pixel information extracted from the images, analyze contextual information such as capture time and focal aperture, or use both to determine a proper subset of photos. However, these approaches miss the most important factor in the photo selection process: the user. The goal of our approach is to consider individual interests. By recording and analyzing gaze information from the user's viewing photo collections, we obtain information on user's interests and use this information in the creation of personal photo selections. In a controlled experiment with 33 participants, we show that the selections can be significantly improved over a baseline approach by up to 22% when taking individual viewing behavior into account. We also obtained significantly better results for photos taken at an event participants were involved in compared with photos from another event.

Author Keywords

Photo selection; eye tracking; usage-based image selection.

ACM Classification Keywords

H.5.2 User Interfaces: Input devices and strategies

INTRODUCTION

The large number of personal digital photos makes the management of one's photo collection an increasingly challenging task. Users easily take hundreds of photos during vacation or personal events such as weddings or birthday parties. Often, selections of "good" photos are created to reduce the amount of photos stored or shared with others [5, 7, 13, 16]. While users enjoy certain photo activities like the creation of collages for special occasions such as anniversaries or weddings, these tasks are seen as "complex and time consuming" for normal collections [5]. In order to alleviate this situation, different approaches that allow for the automatic selection of a subset of photos from a large collection have been developed in the past. *Content-based approaches* initially used pixel information of the photos to compute exposure, sharpness, and other photo properties in order to determine a photo subset [24, 23, 3, 25, 26]. These approaches were followed by *context-based approaches*, which exclusively or additionally analyze the contextual information of the photos. This information could be technical parameters of the digital still camera, such as capture times or GPS coordinates, or information gained from social networks like blogs [10, 22, 12, 15]. While acknowledging the achievements made by content- and context-based approaches, we claim that they miss the most important factor in the photo selection process: the user's interests. Capturing the user's interests is important as the human photo selection process is assumed to be guided by very individual factors and is highly subjective [20].

Rodden and Wood showed that photo collections are browsed frequently [16]. However, the frequency of browsing decreases over time. Thus, the viewing of photos usually happens shortly after the capturing or downloading to a computer. We present a *usage-based approach* [18], where the very individual viewing behavior of the user is captured and used in the selection process. We record the user's eye movements while viewing a photo collection. Fixations in the gaze paths show the moments of the highest visual perception and indicate the user's attention. Different eye tracking measures have been used to identify important photos. They consider for example the duration of fixations, how frequently a photo is fixated, and the pupil reaction. Photos with high measure values are assumed to be the most interesting to the user and thus should be part of a selection. Our approach is reasonable as we expect a higher availability of eye tracking hardware in the future, as indicated by recent developments in the directions of less expensive professional hardware and eye tracking with low-cost hardware like webcams.

In our experiment participants first viewed a collection of photos and then manually created personal selections. The manual selections served as ground truth and allowed for the evaluation of different selection approaches. As we assume that the eye movements are strongly influenced by interest, we also investigated if the personal relevance of viewed photo sets influences the quality of the gaze-based photo selection results. To this end, we showed photos of an event the user took part in or in which the user knew the participants ("home collection") and photos of an event the user was not personally involved in ("foreign collection"). Also, the different selection tasks for manual selection are compared (selections

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.

ACM 978-1-4503-2473-1/14/04..$15.00.
http://dx.doi.org/10.1145/2556288.2557025

for private collection, for friends and family, and for publication on social networks). The results of our experiment show that photo selection criteria are often subjective and manually created selections are diverse. The usage-based approach, taking the individual user interest into account, significantly improves the automatically created photo selections based only on content and context information [10, 25, 26] by up to 17%. Considering only photo sets of the home collection, we even achieve an improvement of up to 22%. The three different tasks for manual selection had only little influence on the performance of the usage-based approach. The related work is discussed below. Subsequently, we describe the experiment design and the applied methods for creating photo selections. We compare the participants' behavior when viewing and selecting photos and show the distribution of the selected photos. Finally, the gaze selection results are presented and discussed before we conclude the paper.

BACKGROUND LITERATURE

Approaches for the automatic creation of photo selections typically use low-level image features (content-based information) and/or photo meta data (context-based information). For example, Chu and Lin [3] selected representative photos by identifying near-duplicate photos. Xiao et al. [25] presented a semiautomatic collage creating tool that makes a selection of photos purely based on content-based information, such as color histograms, sharpness, and near duplicates. Li et al. [10] created summaries of photo collections based on time stamps and facial features. Sinha et al. [22] introduced a framework for generating representative subsets of photos from large personal photo collections by using multidimensional content and context measures like GPS coordinates or tags. Rabbath et al. [15] used content and context information from blogs to automatically create photo books.

Eye tracking technology can be used as an explicit input device. The users explicitly control software by moving their eyes as presented, for example, in the evaluations of gaze interaction by Mollenbach et al. [11] and Sibert and Jacob [21]. We differentiate our approach from gaze-controlled approaches as the user is not asked to control his or her gaze in interacting with the system, e.g., by fixating on a photo to select it or by concentrating only on the "best" photos. In contrast, we obtain information from users freely viewing the photos without concrete instructions. Several approaches use eye tracking to identify attractive or important images in a search results list and use this information as implicit user feedback in improving the image search, e.g., [6, 8, 9]. From these works, we know that it is possible to use gaze information to detect images relevant to a given search task. Support vector machines have been applied on eye tracking data together with content-based features to rank images [14]. Santella et al. [19] presented a method for image cropping based on gaze information. Their goal was to find the most important image regions in order to exclude these regions from cropping. This approach does not have the goal of creating selections of photos. However, it shows that relevant information on the user's interest can be obtained from gaze.

It can be assumed that eye tracking will be available to the average user in the near future. One reason for this assumption is the rapid development of sensors in IT hardware: While the cost for eye tracking devices was about USD $35,000$ just two years ago, embedded and low-cost solutions are available now for less than USD 100[1]. Nowadays, eye trackers can also be developed using low-cost hardware. San Agustin et al. [17] compared a commercial eye tracker and a webcam system. The results for the webcam system are satisfactory and comparable to the commercial system, although still with limitations concerning the comfort of use. This rapid development in eye tracking hardware will allow using gaze information in everyday tasks like photo viewing.

EXPERIMENT

We developed an experiment application that allowed the participants to view and select photos from a collection $C = \{p_1, p_2, ...p_n\}$. In the first part of the experiment, eye tracking data was collected from the participants while viewing photos. Subsequently, ground truth data was collected by asking the participants to manually create three personal selections of these photos.

Participants

A total of 33 participants (12 of them female) completed the first part of the experiment. Twelve were associated with a research lab A in North America and 21 with institute B in Europe. Members of institute A and institute B did not know one another. Their age ranged between 25 and 62 years (M: 33.5, SD: 9.57). Twenty of them were graduate students and 4 postdocs. The remaining 9 participants worked in other professions, such as secretaries or veterinary assistants. Eighteen of the 33 participants (7 of them female) completed the second part of experiment. Six of them were associated with institute A and 12 of them with institute B. Their average age was 31.7 (SD: 8.74).

Materials

The experiment photo collection C consisted of two collections of photos taken during two social events, one organized by each of the two research institutes the participants were associated with. The activities during the events included teamwork situations, group meals, as well as leisure activities like bowling and hiking. Event A lasted half a day and event B three days. The photos were taken by different people: three people for collection C_A and two for collection C_B. The photos were not preselected but taken directly from the camera. Only two extremely blurry photos were removed. The photo collection of the participants' own institute is called "home collection" and the other one "foreign collection" (cf. Figure 1). Collection C_A (photos were taken during the event of institute A) consisted of 162 photos and C_B 126 photos. The photo collection $C = C_A \cup C_B$ was split chronologically into sets of nine photos $c_i = \{p_{i \cdot 9+1}, ..., p_{i \cdot 9+9}\}$. Each set c_i contained only photos of one of the collections. The complete collection of 288 photos was thus split into 32 sets (18 sets of C_A and 14 sets of C_B).

[1] http://www.tobii.com/eye-experience/ (last visited Jan. 5, 2014)

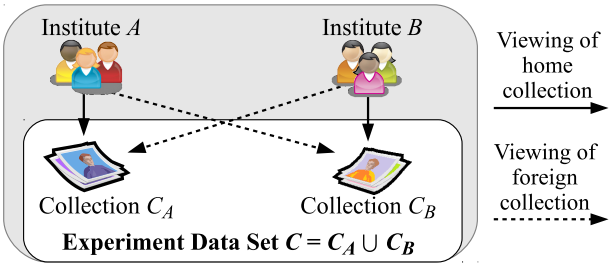

Figure 1. Structure of the experiment data set.

We assumed that the photos of the home collection are of higher personal interest for the participants than the photos of the foreign collection. This assumption is supported by results from the questionnaire. The participants were asked to indicate how interesting the two photo collections were using a Likert scale from 1 ("Not interesting") to 5 ("Very interesting"). For the home collections, the question was answered with an average of 4.36 (SD: 0.6) and for the foreign collection with an average of 2.72 (SD: 1.14). A chi-square test was applied for testing the significance of the differences as the data was not normally distributed (shown by a Shapiro-Wilk test of normality with $p < .001$ for the home set ratings and $p < .018$ for the foreign set ratings). The chi-square test showed a statistically significant difference between the answers, $X^2(5, N = 66) = 34.594, p < .001$.

Apparatus

The experiment was performed either on a 22-inch or a 24-inch monitor for the two research groups (cf. section Participants). The participants' gazes were recorded with a Tobii X60 eye tracker at a data rate of 60 Hz and an accuracy of 0.5 degree. The distance between the participants and the computer screen was about 60 cm. The setup (including a laptop, the eye tracking device, and a standard computer mouse) was the same for both groups.

Procedure

The experiment consisted of four steps, one viewing and three selection steps. In the first step ("Photo Viewing" in Figure 2), the participants were asked to view all photos of collection C with the goal "to get an overview." Eye tracking data was recorded only during this photo viewing step. Thus, unlike other item selection experiments [2], we clearly separated the viewing step from the selection steps. This was crucial to avoid an impact of the selection process on the viewing behavior. The order in which the two collections C_A and C_B were presented to the participants in the experiment was alternated. No time limit was given for viewing the photos. The participants were told that they would afterward, in the second step, create selections of the photos. No details about the selection process were given at this stage of the experiment.

Each photo set c_i was presented on a distinct page; the photos were arranged in 3×3 grids in the center of the screen. The photos' maximum height and width were set to 330 pixels, corresponding to about 9° at the visual angle. The minimum distance between the photo was 22 pixels (0.6°). By clicking on a button, the next set of nine photos was presented. The

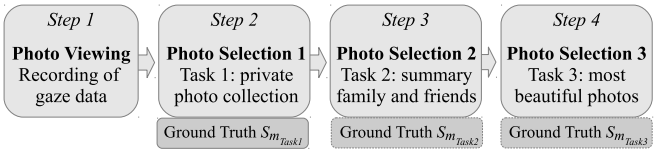

Figure 2. Experiment setup with the photo viewing step and the three selection steps.

photos in each set c_i were arranged in random order, whereas the sets themselves and the order of the sets remained the same to preserve the chronological order of the events.

After having viewed all photos, the participants were asked to select exactly three photos of each set c_i in the second step ("Photo Selection 1" in Figure 2). The photos were selected by means of a drag-and-drop interface as depicted in Figure 3. The same sets as in the viewing step were again presented to the participants, but the photos were rearranged in a new random order. The participants were asked in this second step to select the photos as they would do for their private photo collection. We gave no specific instructions regarding the selection criteria for choosing the photos. Thus, the participants could apply their own (perhaps even unconscious) criteria. Also, in the third and fourth steps ("Photo Selection 2" and "Photo Selection 3" in Figure 2), the participants performed manual selections. In the third step (Task 2), the participants were asked to "select photos for their friends or family that provide a detailed summary of the event." The fourth step (Task 3) was to "select the most beautiful photos for presenting them on the web, e.g., on Flickr." In the experiment steps 3 and 4, the users performed the manual selections only for the photo sets belonging to their home collections, not the complete collection C. Eighteen of the participants completed these tasks. The manual selections served as ground truth in the later analysis. Finally, the participants filled in a questionnaire. It comprised questions about demographical user data (age, profession), the experiment data set, and the experiment task as well as a rating on different selection criteria.

Figure 3. Photo selection interface with one selected photo.

METHODS FOR CREATING PHOTO SELECTIONS

The aim of the photo selection methods is to create a subset $S \subset C$ that best suits the user's preferences. The capabilities of each method are evaluated by comparing the calculated selection with the manual selection for each set c_i created

during the experiment. A "perfect" selection would be a selection identical to the manual selection. The photos C were displayed in sets of nine photos c_i. Selections of $j = 3$ photos are created for each set. An overview of the different photo selection approaches is shown in Figure 4. They are presented in detail in the following sections. We start by describing the content-based and context-based measures for photo analysis used in our baseline system. Subsequently, we present the eye tracking based measures and then the combination of different measures by means of logistic regression. Finally, we describe the calculation of precision P for comparing the selections with the ground truth selections $S_{m_{Task1}}$, $S_{m_{Task2}}$, and $S_{m_{Task3}}$.

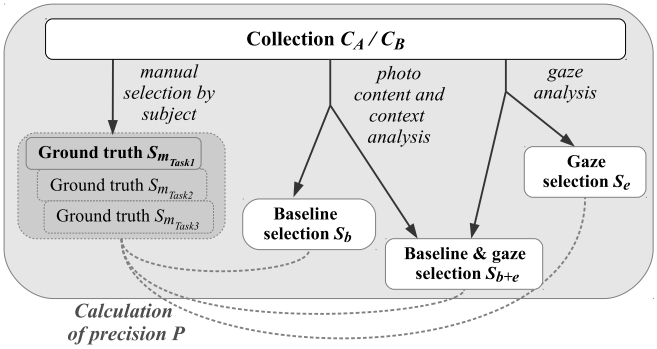

Figure 4. Overview of the investigated photo selection approaches and calculation of precision P.

Content and Context Analysis Baselines

We make use of six measures as baselines that analyze the context or the content of photos. An overview is shown in Table 1. The measures are motivated from related work, and details on their implementations can be found in the cited papers.

No	Measure	Description
1	concentrationTime	Photo p was taken with other photos in a short period of time [10]
2	sharpness	Sharpness score [25]
3	numberOfFaces	Number of faces
4	faceGaussian	Size and position of faces [10]
5	personsPopularity	Popularity of the depicted persons [26]
6	faceArea	Areas in pixels covered by faces

Table 1. Baseline measures based on content and context analysis.

The first measure, (1) concentrationTime, relies on the assumption that many photos are taken within a short period of time when something interesting happens during an event [10]. This measure is context-based as the information when a photo was taken is obtained from the photos' meta-information. Li et al. [10] created a function f_k indicating the number of photos taken for a point in time. By means of the first derivation of this function, a temporal representative value for each photo is calculated. The next four measures are content-based as they analyze the photos' content at pixel level. The photos' quality is considered in measure (2) sharpness by calculating a sharpness score as presented by Xiao et al. [25]. The score is calculated as $Q = \frac{strength(e)}{entropy(h)}$ with $strength(e)$ as the average gradient edge strength of the top 10% strongest edges and $entropy(h)$ as the entropy of the normalized gradient edge strength histogram. The edge strength is calculated by the well-known Sobel operator from computer vision.[2]

Related work, presented by Boll et al. [18], showed that depicted persons play an important role in the selection of photos. The four measures (3) to (6) are based on the analysis of depicted persons. Measure (3) numberOfFaces simply counts the number of faces on a photo. The detection of faces is done using OpenCV's Haar Cascades[2]. Also, measure (6) faceArea is based on this calculation. It considers the size in pixels of the photo areas covered by human faces. A Gaussian distribution of the face areas as proposed by Li et al. [10] is considered by measure (4) faceGaussian, identifying photos with large depicted faces in the photo's center. Measure (5) personsPopularity considers a persons' popularity in the data set as presented by Zhao et al. [26]. It assumes that faces appearing frequently are more important than the ones appearing less often. The calculation is performed by the OpenCV's face recognition algorithm and considers persons appearing in each set c_i of nine photos. This measure is context-based as well as content-based.

Gaze Analysis

Human gaze paths consist of fixations and saccades. Fixations are short time periods when the gaze is focused on a specific point on the screen. Saccades are the fast movements between the fixations. Most of the visual perception takes place during the fixations. Thus, we mainly analyze these parts of the gaze paths. Fixations are extracted from the raw eye tracking data by applying a fixation filter. This preprocessing is performed with the fixation filter offered by Tobii Studio[3] with the default velocity threshold of 35 pixels and a distance threshold of 35 pixels. A visualization of a sample gaze path can be found in Figure 5. Fixations are visualized as circles, and the diameter indicates the duration of a fixation.

The obtained fixations are analyzed by means of eye tracking measures. Each measure assigns a value to each photo p. An overview of all measures can be found in Table 2. Measure (7) fixated determines if a photo was fixated or not. Measure (8) fixationCount indicates how often a photo was fixated. Measures (9) to (12) consider the durations of fixations on a photo. Measures (13) to (15) are based on "visits." A visit is a series of fixations on a photo. Measure (16) saccLength considers the saccade lengths before fixating a photo. The three measures (17) to (19) rely on the size of the user pupil while fixating a photo. Related work shows that the pupil size can vary with emotional reactions [1]. This reaction could appear more often for interesting photos. To compensate the inaccuracy of the eye tracking data, fixations in the surrounding of 11 pixels ($0.3°$ at the visual angle) of a photo are also

[2]http://opencv.org/ (last visited Sept. 17, 2013)
[3]http://www.tobii.com (last visited Sept. 18, 2013)

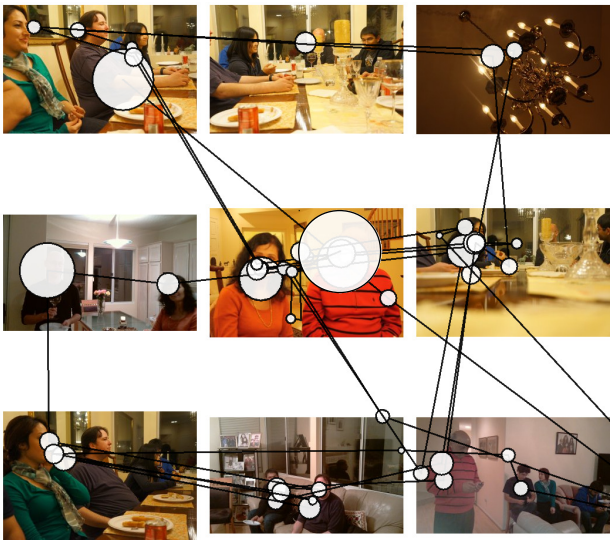

Figure 5. Visualization of a gaze path on a photo set.

considered as being on a photo (the smallest distance between two photos is 22 pixels, or $0.6°$).

No	Measure	Description
7	fixated	Indicates if p was fixated or not
8	fixationCount	Counts the number of fixations on p
9	fixationDuration	Sum of the duration of all fixations on p
10	firstFixationDuration	Duration of the first fixation on p
11	lastFixationDuration	Duration of the last fixation on p
12	avgFixationDuration	Average of the durations of all fixations on p
13	maxVisitDuration	Maximum visit length on p
14	meanVisitDuration	Mean visit length on p
15	visitCount	Number of visits within p
16	saccLength	Mean length of the saccades before fixating on p
17	pupilMax	Maximum pupil diameter while fixating on p
18	pupilMaxChange	Maximum pupil diameter change while fixating on p
19	pupilAvg	Average pupil diameter while fixating on p

Table 2. Eye tracking measures for photo p.

Combining Measures Using Logistic Regression

Different combinations of the content-based and context-based measures and eye tracking measures are investigated. To this end, all measure values are normalized per set c_i by subtracting the mean of the nine values per set and dividing it by the standard derivation σ. The measures are combined by means of a model learned from logistic regression as presented by Fan et al. [4]. The data of all users is split into a training set and a test set. About 15% of the data are selected as test data, which correspond to five sets of nine photos for every user as test data and 27 sets of nine photos as training data. The test sets are randomly chosen. Only complete sets c_i are selected for training and testing, respectively. When analyzing subsets of the data (e.g., when analyzing only the photos that are part of the home collection for each user) less data is available. The test data size is reduced to three sets of nine photos. The model is trained with the training data of all 33 users. That corresponds to $33 * 27 * 9 = 8,019$ training samples, when using the whole data set C. This number reduces to 3,699 samples when training the model only with those photos of the home sets. 1,998 samples were used when performing the training for the data from the experiment steps 3 and 4, which were completed by less participants. The default parameter settings of the LIBLINEAR library [4] are used for training. For every analysis, 30 iterations with different random splits are performed and the average results of all iterations are presented in this paper.

Three different measure combinations are investigated. Selection S_b takes only the baseline measures (1) to (6) into account. For the selection S_{b+e}, all 19 measures are considered in the logistic regression. For S_e exclusively the gaze measures (7) to (19) are used in the learning algorithm. The logistic regression predicts a probability of being selected for each photo in set c_i of nine photos. The three photos with the highest probability are chosen for the selection and compared with the ground truth selections $S_{m_{Task1}}$ to $S_{m_{Task3}}$.

Computing Precision P

For comparing a computed selection to the ground truth, the percentage of correctly selected photos of all selected photos is calculated (precision P). This calculation is conducted for each set c_i. Precision P for a selection approach is the average precision over all sets c_i. As three of nine photos are selected, a random baseline selecting three photos by chance would have an average precision of $P_{rand} = 0.\overline{3}$.

USERS' PHOTO VIEWING AND SELECTION BEHAVIOR

In this section, we first investigate the users' photo viewing times and photo selection times in our experiment. Subsequently, the distribution of the manual photo selections of our participants is presented. Finally, we show the users' rating regarding the importance of different photo selection criteria.

Viewing and Selection Times

The sets c_i of nine photos were viewed on average for 12.6 s (SD: 11.9 s). The shortest viewing time was below a second and the longest 121.1 s. The viewing times were on average higher for the sets belonging to the home collection with 13.3 s (SD: 12.2 s) compared with 11.8 s (SD: 11.5 s) for the foreign collection. These values are calculated from the time the participants looked at the photo viewing pages in the experiment application. The distribution of the viewing times significantly deviated from a normal distribution (shown by a Shapiro–Wilk test of normality with $p < .001$ for the home set and foreign set, respectively). Thus, we applied a Mann–Whitney U test in comparing the viewing durations for the sets belonging to the home collection and the foreign collection. The result is that the viewing durations are significantly longer for the home sets compared with the foreign sets ($U = 138462, Z = -3.194, p = .001$).

The average selection time per set was 20.9 s (SD: 11.6 s) for Task 1. The selection times were slightly shorter for the foreign sets with an average of 20.1 s (SD: 10.5 s) compared with those of the home collection with an average of 21.7 s (SD: 12.6 s). Like the viewing times, the distribution of the selection times also significantly deviated from a normal distribution (shown by a Shapiro–Wilk test with $p < .001$ for the home set and foreign set, respectively). Applying a Mann–Whitney U test on the selection durations showed that the differences are not statistically significant ($U = 125877, Z = -1.013, p = .311$). The selection process clearly took longer than the viewing step (+66%). Although the selection process was different from selections usually performed in daily life, it shows that the selection of photos is more time-consuming than the viewing.

The participants rated how difficult the creation of the selection was on a Likert scale from 1 ("It was hard to select the photos") to 5 ("It was easy to select the photos"). The ratings were performed separately for the home collection and the foreign collection. The results show that the ratings were on average higher for the home set with 3.85 (SD: 0.94) versus 3.06 (SD: 0.94). Shapiro–Wilk tests revealed that the data was not normally distributed ($p < .001$ for the home set ratings and $p < .015$ for the foreign set ratings). A chi-square test was applied, which showed that the difference is significant ($\alpha < 0.05$) with $X^2(4, N = 66) = 9.714, p < .046$.

Distribution of the Users' Manual Photo Selections

In Figure 6, the numbers of selections for all photos are displayed. On average, every photo was selected 3.7 times. The highest number of selections was 24. Approximately 75% of the photos were selected five times or less. Thus, most of the photos were selected only by a minority of the participants. We conclude that photo selections were very individual in our experiment and confirm results from previous work [20].

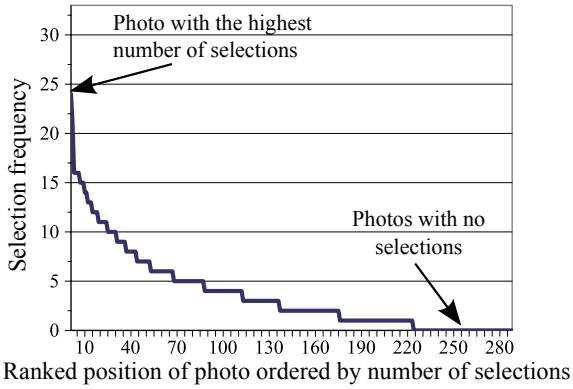

Figure 6. The number of selections for all photos in data set C, ordered by the number of selection.

Cohen's kappa k was calculated for all possible user pairs with $k = \frac{q_x - q_r}{1 - q_r}$. In this formula, q_x is the observed agreement between two users. This corresponds to the percentage of photos that were selected by both users. The value $q_r = 0.556$ is the probability of a by-chance agreement of two users on their photo selections. As the number of selected photos is high compared with the total number of photos per page (we select three out of nine), the value for q_r is already quite high. The obtained results for Cohen's kappa comparing all user selections have a minimum of $k = 0.5$ and a maximum of $k = 0.757$. The average Cohen's kappa over all users is $k = 0.625$. The average result lies only about 12% above the by-chance probability of $q_r = 0.556$. This further confirms that the photo selections are very diverse.

Ratings of Photo Selection Criteria

In the second experiment step, where a manual selection was created for Task 1, no specific criteria regarding the selection of photos were given to the participants. They were just asked to create selections for their private photo collection and could apply their own criteria. In the questionnaire, we asked the participants to indicate how important different criteria were for their selections. Nine criteria were rated on a five-point Likert scale. Additionally, the users were given the option to add criteria as free text. The selection criteria were taken from related work [22, 18, 25, 7, 16]. An overview of the criteria rated by the participants can be found in the following list:

1. Attractiveness — the photo is appealing
2. Quality — the photo is of high quality (e.g., it is clear, not blurry, good exposure)
3. Interestingness — the photo is interesting to you
4. Diversity — there are no redundant pictures
5. Coverage — all locations/activities of the event are represented
6. Depiction of the persons most important to me
7. Depiction of all participants of the event
8. Refreshment of the memory of the event
9. Representation of the atmosphere of the event

Figure 7 shows the results of the ratings on a Likert scale between 1 ("Not important") and 5 ("Very important"). The criteria are ordered by their mean results. One can see that some of the criteria have a wide range of ratings, from 1 to 5. Every criterion has at least one rating with five points.

The criteria were classified as "rather objective" (striped bars) and "rather subjective" (solid bars), expressing if a criterion is an objective measure and could (theoretically) be calculated by computer algorithms. Although this classification could be a subject of discussion, it serves the goal to better understand the nature of selection criteria. In Figure 7, we can see that three of the five criteria with the largest range in the answers (8, 4, 6, 7, 5) belong to the objective criteria. Also, the two criteria with the lowest mean results are rather objective criteria. It is remarkable that the two criteria with the highest average rating and the smallest deviation, *3. Interestingness* and *1. Attractiveness*, are rather subjective criteria. Also, four of the five highest-rated criteria are subjective. Eight participants provided additional criteria as free comments like "the picture makes me laugh" or "the photo is telling a story." All criteria added by the participants were very subjective.

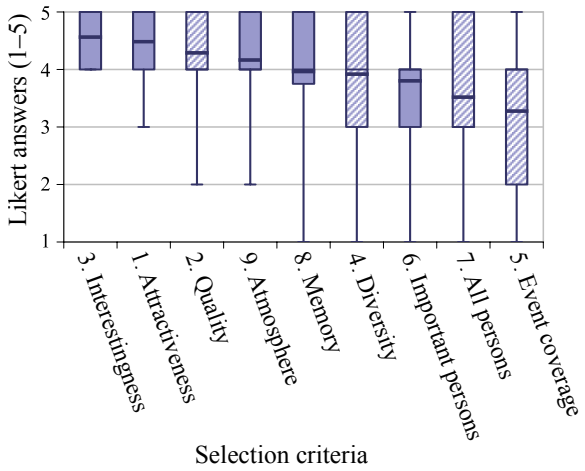

Figure 7. Selection criteria sorted by mean value.

GAZE SELECTION RESULTS

We present the results for selections based on single measures, followed by the results from combining the measures with machine learning. Subsequently, we show the influence of personal interest in the photos in the selection results. Finally, the weak influence of different selection tasks is shown.

Selection Results for Single Measures

Figure 8 shows some sample photos with the highest and lowest measure results for three baseline measures. The samples show that the measures basically succeeded in analyzing the content of the photos. For example, the first row shows the most blurred photo (left) and the photo with the highest sharpness (right). But it also shows the limitations of today's computer vision approaches as, e. g., the photo with the highest number of faces is determined with 7 faces, although almost 20 people are captured in this shot. Please note that for measure (4) faceGaussian the examples with the lowest result of 0 (no faces) are not considered in this overview.

As described in the previous section, we randomly split the data set into a training set and a test set in 30 iterations. For the analysis of the performance of single measures in this section, no training was needed. Thus, the training data set was not considered, but for ensuring compatibility to the following sections, we applied the measures only on the test data sets. Figure 9 shows the average results for each user over the 30 iterations. Precision P was calculated by using only a single measure for creating the selections of the test data sets. The photos in the selections were the three photos with the highest measure values. The results strongly vary between $P = 0.202$ and $P = 0.56$ for different users and measures. Of all baseline measures, (6) faceArea performed best with a mean precision of $P = 0.365$. One can see that the face-based baseline measures (3) to (6) show high variances in precision P. (2) sharpness delivered a mean result that lies with $P = 0.344$, which is close to random selection with $P_{rand} = 0.\overline{3}$. It is interesting that photo quality as a selection criterion was ranked very high by the users (third important measure, see previous section), but the sharpness score, considering the photo quality, did not deliver good results. On average, 29.3 fixations were recorded per set (SD: 19.97). The average fixation number per photo is 3.25 (SD: 3.15). The highest median precision results are obtained by the three eye tracking measures (9) fixationDuration ($P = 0.419$), (13) maxVisitDuration ($P = 0.42$), and (14) meanVisitDuration ($P = 0.421$). The pupil-based eye tracking measures (17) to (19) did not deliver good results. They are close to the precision results for a random selection $P_{rand} = 0.\overline{3}$ or even slightly below for (19) pupilAvg with $P = 0.32$.

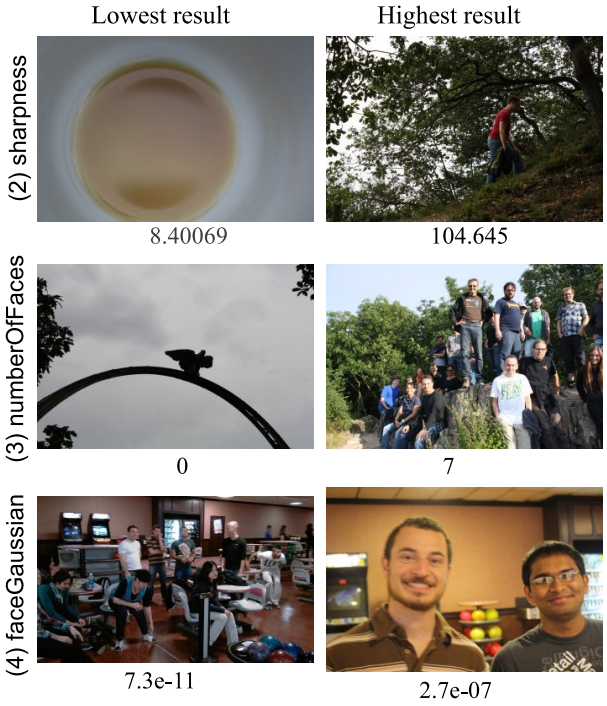

Figure 8. Sample photos with the highest and lowest results for three of the baseline measures.

Selection Results for Combined Measures

We combined the measures by means of logistic regression. Pairwise Pearson correlation tests showed that all correlation coefficients were below 0.8. Thus, the correlations between the single measures were not too high, and we, therefore, decided not to exclude measures from the logistic regression. We obtained the best average precision result of $P = 0.428$ for S_{b+e}, the selections created based on baseline measures and eye tracking measures. The result for S_e (only eye tracking measures) is $P = 0.426$ and $P = 0.365$ for S_b (only baseline measures). Using gaze information improves the baseline selection by 17%. The results of all users averaged over 30 iterations are shown in Figure 10. Statistical tests were applied on the average precision values obtained from the 30 random splits for each user for investigating the significance of the results. A Mauchly's test showed that sphericity had been violated ($X^2(2) = 27.141, p < .001$). Consequently, the nonparametric Friedman was used for the analysis. We found that the differences between the three selections are significant ($\alpha < 0.05$) for P with $\chi^2(2) = 49.939, p < .001$, $n = 33$. For post hoc analysis, pairwise Wilcoxon tests

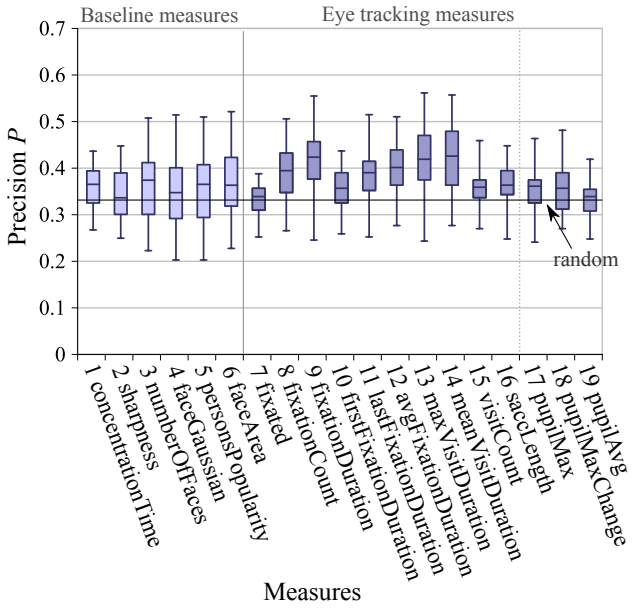

Figure 9. Precision results for all users averaged over 30 random test sets when selecting the photos based on single measures.

were conducted, with a Bonferroni correction for the significance level ($\alpha < 0.017$). The tests showed that baseline selection S_b was significantly outperformed by the gaze including selections S_{b+e}, $Z = -4.297, p < .001$, and S_e, $Z = -3.600, p < .001$. No significant difference was detected between S_{b+e} and S_e, $Z = -0.019, p < .496$.

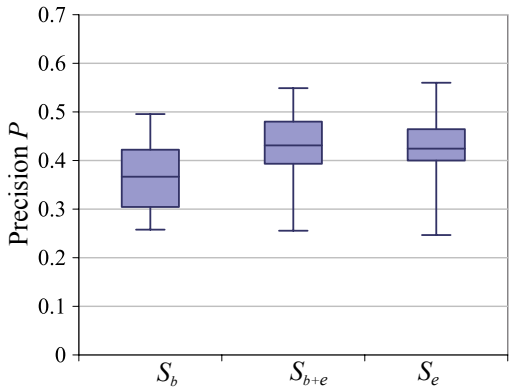

Figure 10. Precision results for all users averaged over 30 random splits obtained from combining measures by logistic regression. The results are based on baseline measures S_b, eye tracking measures S_e, and all measures S_{b+e}.

Figure 11 shows the results for the 30 random splits for one single user. Precision results are between $P = 0.267$ and $P = 0.6$ and show the strong influence of the training data and test data splits. The user selected for this example is the one with the precision result closest to the average precision over all users.

Influence of Personal Involvement

For each user, we distinguished between photo sets c_i that were part of the home collection and those that were part of

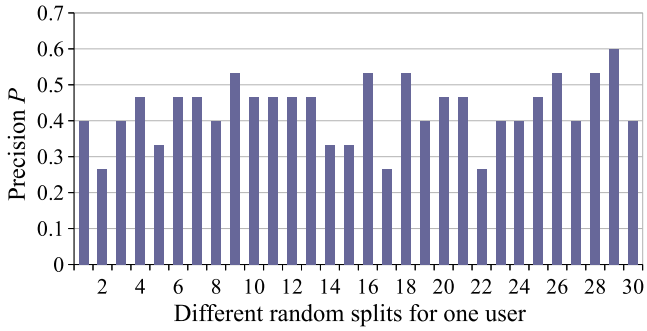

Figure 11. Precision results for S_{b+e} over 30 different random splits for one user.

the foreign collection as described in the section Experiment. Precision of selection S_{b+e} was calculated separately for both collections. The results can be found in Figure 12. They show that P results for the foreign photo set have a larger range, and the average precision is lower with $P = 0.404$ compared with $P = 0.446$ for the home set. Comparing the precision result for the home sets with the results for S_b leads to an improvement of 22%. A Wilcoxon test showed a significant difference between the precision values of all users for the home and foreign photo sets, $Z = -2.842, p < .004$.

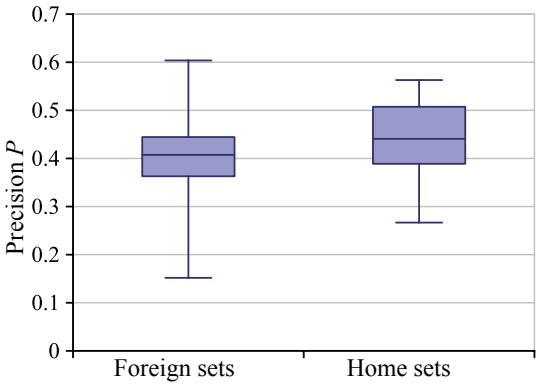

Figure 12. Results for S_{b+e} for foreign and home sets.

Influence of the Selection Task

In the experiment, the participants were first asked to create a "selection for their private photo collection" (Task 1). Subsequently, we asked them to perform further selections for the task: "Select photos for giving your friends or family a detailed summary of the event" (Task 2) as well as "Select the most beautiful photos for presenting them on the web, e.g., on Flickr" (Task 3). The participants created the selections in Task 2 and Task 3 only for the photo sets of personal interest (the "home sets"), which were taken during the event they participated in.

As we were interested in the differences between the performances of the automatic selection compared with these three manual selections, we computed the precision results of the selections under each task (Tasks 1 to 3). The results are shown in Figure 13. The average precision results for

the 18 participants that took part in this part of the experiment are $P = 0.456$ for Task 1, $P = 0.432$ for Task 2, and $P = 0.415$ for Task 3. A Friedman test revealed no statistical significance between the three tasks with $\alpha < 0.05$ for P, $\chi^2(2) = 0.778, p < .678, n = 18$.

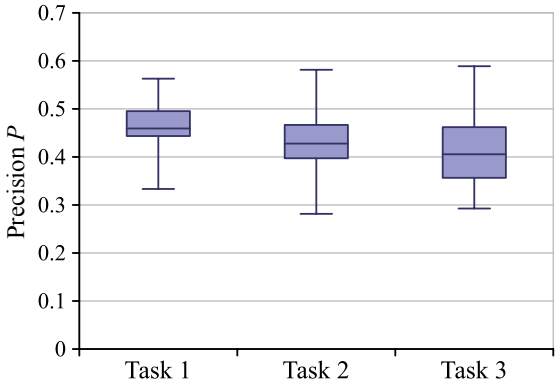

Figure 13. Results for S_{b+e} for different selection tasks.

DISCUSSION

This work has two main contributions. (1) We found that users created highly individual photo selections based on very individual selection criteria. From the analysis of the selection criteria, we conclude that the criteria judged by the users as most important are rather subjective. At the same time, the more objective criteria which could at least theoretically be calculated by algorithms, such as the number of faces depicted or the sharpness of a photo, are less important to most users. In addition, the manually created selections are very diverse; only few photos were selected by most of the users. Thus, there is no "universal" selection that fits the preferences of all users. (2) We found that previous attempts to automatically select photos solely based on content information and context information are not sufficient. Rather, a system supporting users in automatic photo selections by applying eye tracking data significantly outperformed these approach by 17%. Considering only photo sets that were of personal interest, the improvement increased to 22% over the baseline approach. Thus, our approach performed better for photos that are personally related to the user viewing them. The overall best selection result with a mean precision of 0.428 were obtained when combining all measures (content, context, and gaze) by machine learning. It is noteworthy that a single eye tracking measure already delivered competitive results with a mean precision of 0.421 without any machine learning.

In our experiment application, users viewed sets of nine photos and navigated through the sets by clicking on a "Next" button to avoid scrolling. This viewing behavior is different from real life photo viewing, where it is more likely that photos are viewed in a file viewer environment or in full screen mode. It could be that the analysis of viewing behavior in these settings has to be adapted. Bias effects like the concentration on the first photo of a page would be necessary to be considered.

The results strongly vary between users and between different partitions of the data into training set and test set for the machine learning. It is possible that this effect depends on the users and their individual viewing behavior or on the characteristics of the viewed photo sets. For example, for sets including many interesting and good photos the viewing behavior is less obvious, because it is likely that several photo are intensively fixated, and it is more difficult to create a selection.

Automatic approaches, even when including gaze data, may probably be not sufficient for a "perfect photo selection," because of the complexity of human decision processes. We think that the decision on how much support a gaze-based system should offer has to be made by the user. Assistance in the creation of selections by suggesting photos is an option as well as applications that fully automatically create photo selections for the user without additional interaction. One participant in our study concluded: "Dealing with only half of the photos of a collection would already be an improvement."

The viewing and the selection times were longer for photo sets of personal interest. At the same time, the ratings from the questionnaire showed that the selection was rated as being less difficult for the photos of personal interest. This indicates that on the one hand, users like viewing photos of personal interest, but on the other hand, the selection process seems to be even more time-consuming for these sets. Our approach delivers significantly better results for photo collections of personal interest than for photo sets of less personal interest. With other words, the prediction of the photo selections performs better when the photos' content is personally related to the users. This suggests that our approach could work even better in real life with users viewing photos of strong personal interest, e. g., one's wedding, summer vacation, or a family gathering, compared with the data set in this experiment, which is taken from a working group situation. Finally, we compared the results for different manual selections created under different selection tasks. We found that the results are about the same. This result indicates that the information gained from eye movements can be useful in diverse scenarios where photo selections are needed.

Based on our results, others features such as photo cropping based on gaze data [19] may be integrated into future research. The results of our findings may be implemented in authoring tools such as miCollage [25] to enhance an automatic photo selection for creating multimedia collections. We hope that our approach enhances research in the direction of helping users in their photo selection tasks and allowing them to spent more time on the pleasurable aspects of creating photo products like slide shows or collages.

Acknowledgments. The research leading to this paper was partially supported by the EU project SocialSensor (FP7-287975). We thank Chantal Neuhaus for implementing the baseline system and all volunteers for participating in our study.

REFERENCES

1. Bradley, M. M., Miccoli, L., Escrig, M. A., and Lang, P. J. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology 45*, 4 (2008), 602–607.
2. Brumby, D. P., and Howes, A. Strategies for guiding interactive search: An empirical investigation into the consequences of label relevance for assessment and selection. *Human-Computer Interaction 23*, 1 (2008), 1–46.
3. Chu, W.-T., and Lin, C.-H. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection. In *MULTIMEDIA*, ACM (2008), 829.
4. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. LIBLINEAR: A library for large linear classification. *JMLR 9* (2008), 1871–1874.
5. Frohlich, D., Kuchinsky, A., Pering, C., Don, A., and Ariss, S. Requirements for photoware. In *Computer supported cooperative work*, ACM (2002), 166–175.
6. Hajimirza, S., and Izquierdo, E. Gaze movement inference for implicit image annotation. In *Image Analysis for Multimedia Interactive Services*, IEEE (2010).
7. Kirk, D., Sellen, A., Rother, C., and Wood, K. Understanding photowork. In *CHI*, ACM (2006), 761–770.
8. Klami, A., Saunders, C., De Campos, T., and Kaski, S. Can relevance of images be inferred from eye movements? In *MIR*, ACM (2008), 134–140.
9. Kozma, L., Klami, A., and Kaski, S. GaZIR: gaze-based zooming interface for image retrieval. In *Multimodal interfaces*, ACM (2009).
10. Li, J., Lim, J. H., and Tian, Q. Automatic summarization for personal digital photos. In *Pacific Rim Conf. on Multimedia*, IEEE (2003), 1536–1540.
11. Mollenbach, E., Stefansson, T., and Hansen, J. P. All eyes on the monitor: gaze based interaction in zoomable, multi-scaled information-spaces. In *Intelligent user interfaces*, ACM (2008), 373–376.
12. Naaman, M., Yeh, R. B., Garcia-Molina, H., and Paepcke, A. Leveraging context to resolve identity in photo albums. In *JCDL*, M. Marlino, T. Sumner, and F. M. S. III, Eds., ACM (2005), 178–187.
13. Neustaedter, C., and Fedorovskaya, E. Understanding and improving flow in digital photo ecosystems. In *Proceedings of Graphics Interface 2009*, Canadian Information Processing Society (2009), 191–198.
14. Pasupa, K., Saunders, C., Szedmak, S., Klami, A., Kaski, S., and Gunn, S. Learning to rank images from eye movements. In *Workshops on Human-Computer Interaction* (2009).
15. Rabbath, M., Sandhaus, P., and Boll, S. Automatic creation of photo books from stories in social media. *TOMCCAP 7*, 1 (2011), 27.
16. Rodden, K., and Wood, K. How do people manage their digital photographs? In *CHI*, ACM (2003), 409–416.
17. San Agustin, J., Skovsgaard, H., Hansen, J., and Hansen, D. Low-cost gaze interaction: ready to deliver the promises. In *Extended abstracts on Human factors in computing systems*, ACM (2009), 4453–4458.
18. Sandhaus, P., and Boll, S. Social aspects of photobooks: Improving photobook authoring from large-scale multimedia analysis. In *Social Media Modeling and Computing*. Springer, 2011, 257–277.
19. Santella, A., Agrawala, M., DeCarlo, D., Salesin, D., and Cohen, M. Gaze-based interaction for semi-automatic photo cropping. In *CHI*, ACM (2006), 771–780.
20. Savakis, A. E., Etz, S. P., and Loui, A. C. Evaluation of image appeal in consumer photography. In *Electronic Imaging*, SPIE (2000), 111–120.
21. Sibert, L. E., and Jacob, R. J. Evaluation of eye gaze interaction. In *CHI*, ACM (2000), 281–288.
22. Sinha, P., Mehrotra, S., and Jain, R. Summarization of personal photologs using multidimensional content and context. *Multimedia Retrieval* (2011), 1–8.
23. Tan, T., Chen, J., and Mulhem, P. Smartalbum - towards unification of approaches for image retrieval. In *ICPR (3)* (2002), 983–986.
24. Wenyin, L., Sun, Y., and Zhang, H. Mialbum - a system for home photo managemet using the semi-automatic image annotation approach. In *MULTIMEDIA*, ACM (2000), 479–480.
25. Xiao, J., Zhang, X., Cheatle, P., Gao, Y., and Atkins, C. B. Mixed-initiative photo collage authoring. In *MULTIMEDIA*, ACM (2008), 509–518.
26. Zhao, M., Teo, Y. W., Liu, S., Chua, T.-S., and Jain, R. Automatic person annotation of family photo album. In *Image and Video Retrieval*. Springer, 2006, 163–172.