sLab: Smart Labeling of Family Photos Through an Interactive Interface

Ehsan Fazl-Ersi, I. Scott MacKenzie, and John K. Tsotsos
Dept. of Computer Science and Engineering
York University
Toronto, ON, Canada
{efazl,mack,tsotsos}@cse.yorku.ca

ABSTRACT
A novel technique for semi-automatic photo annotation is proposed and evaluated. The technique, *sLab*, uses face processing algorithms and a simplified user interface for labeling family photos. A user study compared our system with two others. One was Adobe *Photoshop Element.* The other was an in-house implementation of a face clustering interface recently proposed in the research community. Nine participants performed an annotation task with each system on faces extracted from a set of 150 images from their own family photo albums. As the faces were all well known to participants, accuracy was near perfect with all three systems. On annotation time, *sLab* was 25% faster than *Photoshop Element* and 16% faster than the face clustering interface.

Categories and Subject Descriptors
D.0 General

General Terms
Algorithms, Performance, Design, Experimentation

Keywords
Photo annotation, face detection and recognition, face ranking and clustering

1. INTRODUCTION
With the rapid development of digital cameras and mobile camera-phones, personal digital photo collections are growing explosively. Therefore, effective photo management interfaces are required to facilitate browsing and searching. In such interfaces, the most challenging step is creating semantically meaningful labels and associating them with photos. The method for doing this directly depends on how photos are searched. Rodden et al. [4] showed that when people search for older photos in a family collection, there are three types of queries:

Type 1: photos from a particular event (e.g., Halloween 2007)

Type 2: a specifically remembered photo

Type 3: photos sharing a common property (e.g., containing a certain person)

A straightforward method for photo annotation, capable of supporting the above, is to use semantic keywords describing *who is in the photo*, *where and when the photo was taken*, and *what is happening*.

Almost all digital photos encode the date and time they were captured. Today, we have *GPS-ready* digital cameras that provide the location of a photo at the time of capture. Therefore, the main label left for automatic extraction is the names of people in the photos. Unfortunately, it is not yet possible to do this automatically since state-of-the-art face recognition methods are not accurate and reliable enough. On the other hand, manual annotation, although accurate, is labour intensive (depending on the number of the photos and their content). Because of this heavy workload, people are less motivated to annotate large collections of photos [4].

The approach used in many commercial photo annotation systems is to allow users to simultaneously annotate multiple photos, rather than labeling photos one by one. An interface provides thumbnails and the user selects photos deserving the same label (e.g., containing a certain person). The selected photos are annotated all at once by a one-click assignment of a label to the selected photos. These systems are often combined with a drag and drop interface to further facilitate annotation. Well-known applications using this approach are ACDSee's *Photo Manager*[1], iView's *MediaPro*[2], Microsoft's *Digital Image Suit*[3], and Google's *Picasa*[4]. Although group annotation mitigates the workload of manual annotation, users must still manually locate similar photos and make decisions for each.

Recent progress in computer vision, combined with users' desire to retrieve photos containing particular people, has motivated industry to provide solutions for efficient face annotation. Adobe's *Photoshop Element 6.0*[5] is the pioneer commercial software for face annotation. It automatically detects faces in all photos and allows users to manually select and annotate detected faces, rather than photos. Upon annotating a face with a label (person name), the software automatically associates the entire photo with that label as well.

Although interesting and innovative, this contributes little in reducing the labour of annotation, since an interface similar to the traditional photo annotation software is used for face annotation, and this requires the user to select and annotate manually.

JCDL'08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.

[1] http://www.acdsee.com/

[2] http://www.iview-multimedia.com/

[3] http://www.microsoft.com/products/imaging

[4] http://picasa.google.com/

[5] http://www.adobe.com/products/photoshopelwin/

This paper presents the design and evaluation of *sLab*, a novel user interface for *semi-automatic* face annotation. Next, we briefly review available approaches to semi-automatic face annotation and discuss their strengths and weaknesses. We then describe *sLab* and show how it uses face recognition technology to facilitate annotation. Following this, an evaluation is presented.

2. RELATED WORK

To automatically annotate faces in family photos, face detection and recognition are two essential steps. Over the past ten years, face detection was extensively studied in computer vision research, and, as a result, efficient and accurate *face detection* techniques (e.g., [7]) are available. However, for the more difficult problem of *face recognition*, current techniques are not robust enough to automatically annotate faces in family albums. Therefore, researchers focus on semi-automatic frameworks to assist in face annotation instead of attempting to label faces automatically.

In practical semi-automatic photo annotation systems, face recognition occurs in two ways: *face ranking* and *face clustering*.

Influential work on *face ranking* is that of Chen et al. [8]. Given an un-annotated detected face, their system generates a list of candidate labels based on the similarity between an unlabeled face and previously annotated faces. The user then selects the correct label from the list. A drawback is that the user must confirm or select the label for all faces one-by-one. In other work, Girgensohn et al. [9] proposed a system where users manually select and annotate several faces in an album, so the system can build a model for each person. With these models, the system displays unlabeled faces ranked by similarity, so the user can add more faces to the selected model. One problem is the need to train the system with several face images for each person. Furthermore, at each step, the user can only annotate faces from the selected model. This is an added cost in comparison to the manual *Photoshop Element* software where the user annotates faces from any category at a time.

In *face clustering*, recently proposed by Cui et al. [5] for semi-automatic photo annotation, similar faces are automatically grouped as clusters, enabling the user to select them in one operation. However, due to limitations of face recognition and data clustering algorithms, user interaction is required to confirm or correct each cluster. The faces of the same person may be scattered in several clusters, and different persons may exist in the same cluster. Therefore, a number of operations (in addition to selection and annotation) were introduced, such as merging and splitting. In comparison to *Photoshop Element*, this helps the user in selecting a group of similar un-annotated faces by one selection operation (e.g., a mouse click) at the expense of a *supervision cost* (validation of clustering results and correction of mis-clustered items). Although this cost was not measured or analyzed by the authors, if the number of mis-clustered faces is higher than some threshold, the manual *Photoshop Element* software will likely perform faster than Cui et al.'s [5] semi-automatic face clustering method.

3. *sLab*

In this paper we present *sLab*, a novel semi-automatic face annotation approach for family photo management. *sLab* takes advantage of face ranking and face clustering, while avoiding their shortcomings. As with Girgensohn's work (face ranking), our system puts together similar faces and lets the user group them instead of automatically clustering them. However, unlike the Girgensohn's approach and similar to the work of Cui's group (face clustering), our system does not require a model for each person beforehand. As well, *sLab* gives the user the freedom of annotating faces from any category (person) at a time.

The first step is to automatically detect faces from photos (or new photos) in the album. For each detected face, a description vector for facial features is computed. Then, the similarities between each face and other extracted faces are pre-calculated (by comparing their description vectors) and used for annotation and retrieval. Because selection reflects the user's intention, our interface automatically arranges similar unlabeled faces close to selected face(s). Therefore, the user may quickly select additional similar faces and annotate them all at once. This occurs without building or updating any model for people in the album. Therefore, all system interactions occur in real-time.

3.1 User Interface (UI)

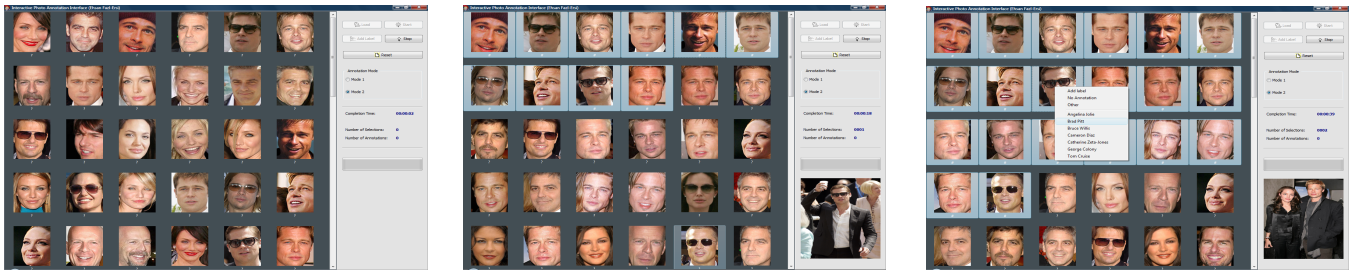
Figure 1 shows several snapshots of the *sLab* user interface (UI) for experimental studies. In the right panel are buttons for loading photos, adding annotation labels (person names), and starting, stopping and restarting an experiment. When photos are imported to the application, faces are extracted from the photos and their similarities to each other computed. Thumbnails of the detected faces are displayed in a List View component, occupying a large portion of the screen. The List View is enabled when the *start* button is clicked; then the user performs annotations until all faces are annotated and clicks *stop*.

The user interface is simple. Faces are selected using the mouse or keyboard, just as files and folders are selected in *Windows Explorer*. Multiple faces are selected by pressing and holding the CTRL key while clicking on target faces. Similarly, clicking on a selected face while holding CTRL deselects that item. To select a large sequential list of faces (i.e., similar faces located side by side), the user selects the first face, presses and holds SHIFT, then clicks on the last item (the user can also use SHIFT with arrow keys). Upon selecting each new face, the system displays the photo that the face was extracted from. This helps the user recognize the selected face in context and is particularly useful if the displayed face is poorly cropped or partly covered.

Upon releasing SHIFT or CTRL (in multiple selection) or the left mouse button or keyboard arrow keys (in single selection), faces are rearranged by displaying the selected faces at the top of the List View followed by the remaining un-annotated faces sorted by their similarity to the selected ones (Figure 1b). After re-arranging un-annotated faces, the user either continues selecting similar faces or annotates all selected faces by right clicking the mouse (Figure 1c). Right clicking produces a popup menu with a list of annotation labels (sorted alphabetically). The user selects a label or inputs a new label to annotate all selected faces at once.

The *Add annotation label* button in the right panel creates labels for one or more desired people before starting the annotation. Besides the annotation labels added to the system, there are two more labels: "no annotation", to be assigned to false positive responses (if any), and "other", to be assigned to faces the user is not interested in. All annotated faces move to the bottom of the List View, so more un-annotated faces can percolate to the top.

We also added hot keys, including *BACKSPACE* for undoing the last annotation operation and *DELETE* to remove the annotation of selected face(s).



(a) (b) (c)

Figure 1: Face annotation using *sLab*. (a) The initial state of the interface. (b) The state of the interface after the user selected all visible faces of "Brad Pitt" using the CTRL key and clicking the left mouse button. As seen, the interface brought 10 more faces of that person to the visible area of the List View. (c) The state of the interface once the other 10 faces of the target person were selected. Since no more faces of the person were available in the visible area of List View, the user chose to annotate the selected face. Therefore, 20 faces were annotated by two multi selections and one annotation.

4. METHOD

4.1 Participants

We recruited nine volunteer participants (5 male, 4 female) for the study. Ages ranged from 23 to 46 (*mean* = 29.7, *SD* = 7.7). All were daily users of computers, with 1 to 8 hours usage per day (*mean* = 6.3, *SD* = 2.7). None of the participants indicated prior experience with face annotation software. According to our pre-test questionnaire, however, all were interested in easy-to-use and efficient software to annotate family photos by the names of people in the photos.

4.2 Apparatus

We used a standard *Windows Vista* PC with a 2.4 GHz CPU and 2 GB memory. The UI was designed using Borland *C++ Builder*. The method of Viola-Jones [7] is used in our system for face detection. Pair-wise similarities between extracted faces are pre-computed using a face recognition technique similar to the work of Fazl-Ersi et al. [10]. The processing operations of face detection, face description, and face similarity computation are performed offline (before the annotation starts) and the computed face similarities are provided to the UI for managing the interaction.

To validate the efficiency of *sLab*, we compared it with *Photoshop Element*, as a representative of existing commercial software for photo annotation, and a face clustering interface [5], as a recent solution in the research community. For the face clustering interface, since no prototype, demo, or implementation was publicly available, we implemented face clustering[6] in our UI as a different operational mode ("*Mode 1*"). Upon selection of any face in the List View in this mode, the system automatically selects all the faces belonging to the similar cluster as the selected face. Similar to [5], selected clusters in the clustering mode (*Mode 1*) can be split or merged by the user, and annotated in one operation. Since the face description and recognition methods in both modes (*sLab* and the clustering approach) are the same, differences in participant behavior are attributed to inherent differences in the interfaces (which is the point of our study).

[6] Face clustering was achieved by applying agglomerative clustering on the computed similarity values between all faces.

4.3 Procedure

A pre-test questionnaire was presented to solicit computer usage information and experience with face annotation software. Before the test began, participants were tutored on the use of *Photoshop Element* and both modes of our experimental software to annotate faces in a sample album of celebrities. All participants were asked to practice on both modes of our software and also on *Photoshop Element* until they were familiar with the interfaces.

Participants were then asked to complete an annotation task using face clustering (*Mode 1*), *sLab* (*Mode 2*), and *Photoshop Element*. Task completion time and accuracy were measured. Participants were asked to annotate all faces "as quickly and accurately as possible". The experiment was a within-subjects design with one factor (interface) having three conditions: *Mode 1*, *Mode 2,* and *Photoshop Element*. The order of conditions was counterbalanced to offset potential learning effects.

We asked participants to bring their own family photo albums. The rationale behind this, as noted by Rodden et al [9], is that the main usage of photo annotation software is labeling photos of family members and friends. The user has likely seen these people thousands of times in different locations and with different facial expressions, clothing, etc. Therefore, the recognition of photos of familiar people is instant. Note that most user studies on face and photo annotation (including [5]) use a single dataset of images unfamiliar to participants. Clearly, this compromises external validity and exacerbates efforts to measure user performance and compare different interfaces.

A set of photos from each participant's album was randomly selected producing about 150 (*SD* = 2.3) faces, from eight individuals in the album. This was done to bring the annotation of all test albums to about the same level of difficulty. For each task, the software (in both modes) recorded the completion time, and the annotation workload as the number of selection and annotation operations. For *Photoshop Element*, only the completion time was measured, as we did not have control over that software.

After testing with each method, annotation accuracy was measured (by the experimenter). Given the true annotation for the test album of each participant, the accuracy was measured in our experiments using the harmonic average of the proportion of

annotated faces to all faces (p_1) and the proportion of correctly annotated faces to all annotated faces (p_2):

$$p = \frac{2}{\frac{1}{p_1} + \frac{1}{p_2}} \quad (1)$$

Participants were encouraged to take a short break between conditions to avoid fatigue. The total time was about 50 minutes per participant. At the end of the experiment, a post-test questionnaire was presented to solicit subjective impressions on the three interfaces.

5. RESULTS AND DISCUSSION

Analyzing the annotation accuracy of participants, we observed that seven participants achieved 100% accuracy with all three systems. This is a consequence of participants using their own family albums. In a similar experiment using an unfamiliar dataset [5], no perfect annotation by participants was reported.

Figure 2 shows the task completion time for the participants who achieved perfect annotation accuracy in the three systems. The mean task completion times were 295 s using *sLab* (*Mode 2*), 331 s using face clustering (*Mode 1*) and 376 s using *Photoshop Element*. The differences were statistically significant ($F_{2,6} = 7.6$, $p < .005$). For all participants, the overall completion times were reduced using *sLab* – by about 25% compared to *Photoshop Element*, and by about 16% compared to face clustering (*Mode 1*).

We compared the annotation workload of *sLab* with the face clustering technique, by comparing the number of selection/de-selection and annotation operations during each annotation task. The number of selection/de-selection operations using *sLab* (*Mode2*) was about 11% less on average for all participants. However, there was substantial variation across participants, and, so, the reduction was not statistically significant ($F_{1,8} = 1.21$, $p > .05$). The number of annotation operations using *sLab* was about 17% less on average for all participants. The reduction in this case was statistically significant ($F_{1,8} = 11.17$, $p < .005$).

Based on the post-test questionnaire responses, eight of nine participants believed our method for face annotation was better than the other methods.

6. Conclusion

We have presented *sLab*, a semi-automatic system to assist in annotating family photo albums. By using face processing algorithms and a simplified user interface, *sLab* simplifies arranging similar unlabeled items close to selected items, and therefore allows users to quickly annotate similar items together. This is done without building or updating a model for persons in the album, and without requiring the user to learn additional operations beyond the simple selection and annotation operations in manual annotation systems. A user study, compared our system with two others, confirmed the efficiency of our method.

The direction for future work is mainly on improving the user interface to add more annotation techniques, including drag-and-drop and additional specialized hot keys.

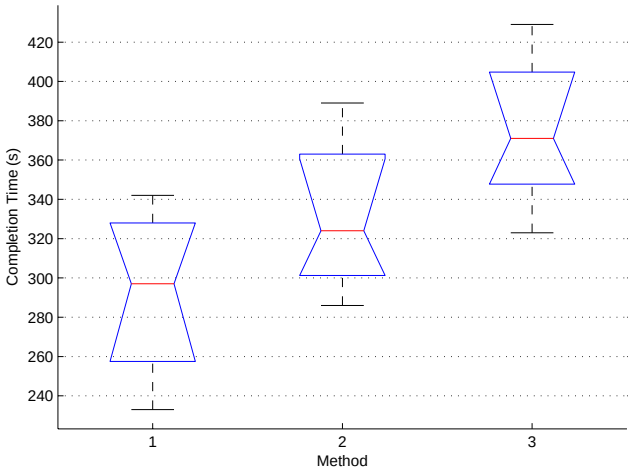


Figure 2: Box plot of the task completion time for the participants who achieved perfect annotation accuracy in the three systems. Methods 1, 2, and 3, refer to our method, face clustering labeling technique, and Photoshop Element software, respectively. The lower and upper lines of the "box" are the 25th and 75th percentiles of the sample, and the line in the middle is the sample median.

7. REFERENCES

1. Kuchinsky, A. et al. 1999. A consumer multimedia organization and retrieval system. Proceedings of the ACM CHI '99. ACM Press, New York, NY, 496-503.
2. Lim, J.-H., Tian, et al. 2003. Home photo content modeling for personalized event-based retrieval. IEEE Multimedia, 10(4), pp. 28-37.
3. Mills, T. J., et al. 2000. ShoeBox: A Digital Photo Management System. Technical report, 2000.10, AT&T Laboratories, Cambridge.
4. Rodden, K., et al. 2003. How do people manage their digital photographs? Proceedings of the ACM CHI '03. ACM Press, New York, NY, 409-416.
5. Cui, J., et al. 2007. EasyAlbum: An interactive photo annotation system based on face clustering and re-ranking. Proceedings of the ACM CHI '07. ACM Press, New York, NY, 367-376.
6. Suh, B., et al. 2004. Semi-automatic Image Annotation Using Event and Torso Identification. Technical report, Computer Science Department, University of Maryland, MD.
7. Viola, P., et al. 2001. Rapid object detection using a boosted cascade of simple features. Proceedings of the IEEE CVPR, 511-518.
8. Chen, L., et al. 2003. Face annotation for family photo album management. International Journal of Image and Graphics, 3(1), 1-14.
9. Girgensohn, A., et al. 2004. Leveraging face recognition technology to find and organize photos. Proceedings of the ACM SIGMM MIR. ACM Press, New York, NY, 99-106.
10. Fazl-Ersi, E., et al. 2007. Robust face recognition through local graph matching, Journal of Multimedia, 2(5), 31-37.