Automatic Person Annotation of Family Photo Album

Conference Paper *in* Lecture Notes in Computer Science · July 2006

DOI: 10.1007/11788034_17 · Source: DBLP

Automatic Person Annotation of Family Photo Album

Ming Zhao[1], Yong Wei Teo[1], Siliang Liu[1], Tat-Seng Chua[1], and Ramesh Jain[2]

[1] Department of Computer Science, National University of Singapore,
21 Lower Kent Ridge Road, Singapore 119077
{zhaom, chuats}@comp.nus.edu.sg
[2] Donald Bren Professor in Information & Computer Sciences
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine, CA 92697-3425
jain@ics.uci.edu

Abstract. Digital photographs are replacing tradition films in our daily life and the quantity is exploding. This stimulates the strong need for efficient management tools, in which the annotation of "who" in each photo is essential. In this paper, we propose an automated method to annotate family photos using evidence from face, body and context information. Face recognition is the first consideration. However, its performance is limited by the uncontrolled condition of family photos. In family album, the same groups of people tend to appear in similar events, in which they tend to wear the same clothes within a short time duration and in nearby places. We could make use of social context information and body information to estimate the probability of the persons' presence and identify other examples of the same recognized persons. In our approach, we first use social context information to cluster photos into events. Within each event, the body information is clustered, and then combined with face recognition results using a graphical model. Finally, the clusters with high face recognition confidence and context probabilities are identified as belonging to specific person. Experiments on a photo album containing over 1500 photos demonstrate that our approach is effective.

1 Introduction

Digital cameras are widely used by families to produce a huge amount of photos everyday. With vastly growing number of family photos, efficient management tools are becoming highly desirable. To achieve efficient management, photos should be indexed according to when, where, who, what and etc. Although time and location can be available in cameras, the annotation of "who" is left to users, which is a tedious task. In this paper, we assume that the information of time and location in terms of GPS is available and we attempt to automatically annotate family photos with "who".

Intuitively, persons can be annotated with their faces. While current face recognition systems perform well under relatively controlled environments [1], they tend to suffer when variations in pose, illumination or facial expressions

H. Sundaram et al. (Eds.): CIVR 2006, LNCS 4071, pp. 163–172, 2006.

are present. As real life family photographs tend to exhibit large variance in illuminations, poses and expressions of face images, it is difficult to detect, align and hence recognize faces in such photographs. Thus, automatic annotation of family photos cannot be solved by face recognition alone.

In fact, the human perception does not make use of facial structure alone to recognize faces. It also uses cues such as color, facial motion, and visual contextual information. Color and motion information have been studied to show their effectiveness in face recognition [2,3]. Visual contextual information has also been successfully used for object and face detection [4], but has not been carefully studied yet for face recognition. In addition, social context is another clues for inferring the presence of specific persons. Social context information takes advantage of the fact that in a family setting, the same group of people tend to appear in the same social events, and they tend to wear the same clothes in the same events. Such information can be used to induce the people's presence when other examples of the same person are recognized or other group members are recognized.

Several semi-automatic annotation systems [5,6,7] have been proposed to help users to annotate faces in each photo by suggesting a list of possible names to choose. Zhang *et al.* [5] formulated face annotation in a Bayesian framework, in which face similarity measure is defined as the maximum a posteriori (MAP) estimation with face appearance and visual contextual features. With this similarity measure, they generated the name list for a new face based on its similarity to the previously annotated faces. Instead of using the visual content information, Naaman *et al.* [7] used only the (social) context information including the time and location of persons' occurrence. Based on time and location, clustering is applied to form events. They then proposed several estimators to estimate the probabilities of each person's presence. The name list is generated based on the combined probability. In the mobile phone environment, Davis *et al.* [8] used time, location, social environment and face recognizer to help automatic face recognition. In particular, they used the identities of mobile phones to detect the presence of specific people in the environment, and used this information for effective person identification. This information, however, is not available in most family photo album environment.

In this paper, we propose a fully automatic framework for person annotation in family photo album. We employ face detection and recognition, in conjunction with visual context and social context to induce the presence of persons in photos. The main contribution in the research is in developing a framework that utilizes all available information for person annotation. The unique features of our system are: (1) Our system is fully automated. This is different from the semi-automatic systems reported in [5,6,7] that suggest a list of probable names for users to select. (2) We improve on face recognition techniques by using eye alignment, delighting and a systematic approach to increasing the number of training samples. This technique helps to maintain the recognition rate even when user is not able to provide sufficient number of training samples. (3) For body detection and recognition, our system uses image segmentation and body clustering, which is more accurate as compared to that reported in [5].

The rest of the paper is organized as follows. The whole framework is described in Section 2. Social context information is discussed in Section 3. Visual information, including face and visual context, is discussed in Section 4. Experiments are performed in Section 5 before conclusions are drawn in Section 6.

2 Automatic Family Photo Album Annotation

This section discusses the overall framework of combining face, visual context (body) and social context (time and location) information for automatic family album annotation.

To obtain face information, we first utilize face recognition to recognize the faces. Even though we use only frontal faces for face recognition, the results for even trained faces are still not very accurate due to the large variation of illumination and expression. However, we know that within a short duration and in nearby places, the same group of people tend to appear together in most pictures and they usually wear the same clothes. This social context information is used to cluster photos into events, so that the visual context (body) of the recognized faces can be used to find other presence of the same person. In fact, both face recognition and body information should be used to complement each other to achieve more reliable results with minimum false detection. In this paper, we propose a graphical model to combine the face and body information. The choice of graphical model is because it provides a natural framework for handling uncertainty and complexity through a general formalism for compact representation of joint probability distribution [9]. The overall framework works as follows:

(1) Cluster family photographs into events according to social context information based on time and location.

(2) Perform face and eye detection, followed by rectification and delighting to provide good alignment and illumination for face recognition.

(3) Perform face recognition on all detected faces.

(4) Extract the visual context information (body) for all detected faces of all persons. For each event, visual context information (body) is first clustered. The resulting clusters are then combined with the face recognition results using a graphical model to provide better person clustering.

(5) Based on face recognition results, build social context estimators, which estimate the probability of people's presence in each photo.

(6) Select the clusters according to the cluster recognition score by combining face recognition and context estimation. For each cluster r, we denote the average face recognition score for person i as $\bar{S}_{FR}(r,i)$ and average context estimation score as $\bar{S}_{CON}(r,i)$. The final recognition score of person i for cluster r is

$$S_C(r,i) = \alpha \bar{S}_{FR}(r,i) + (1-\alpha)\bar{S}_{CON}(r,i) \tag{1}$$

where α is heuristically chosen. This score is used to annotate persons in the photos.

3 Social Context Information

The social context information is used in two ways: first, it is used to cluster photos into events; second, it is used to estimate the probability of people's presence based on the results of face and body recognition.

3.1 Event Clustering

Event is the basic and important organizational unit for family photo album. Although there is not strict definition of event, it usually represents a meaningful happening within a short time duration and in nearby places, such as a birthday party and a visit to the park etc. Event is important as it provides the basis for using the visual context information (body) for person annotation. This is because the visual context information is likely to be consistent within an event, but not so across events. Event is also important for constructing contextual estimators for estimating the probability of the presence of a person in a photo. Figure 1 shows examples of photos in an album event.

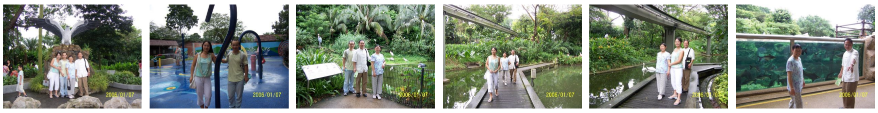


Fig. 1. Photo Examples in an Album Event

The automatic organization and categorization of personal photo albums into meaningful events is an important problem intensively explored in recent years [10,11]. In this paper, we adopt an adaptive event clustering method based on time and location. It consists of an initial time-based clustering, and a location-based post-processor that analyzes the location names of photos. Our time-based clustering is heuristic-based, and is based on observations not previously utilized: (a) the probability of an event ending increases as more photos are taken; and (b) the probability of an event ending increases as the time span increases. Photos are processed sequentially in temporal order. A new photo p belongs to cluster C_k if

$$ATD(C_k, p) \leq F(C_k) \tag{2}$$

where $ATD(C_k, p)$ is the average time difference between all photos in C_k and photo p; and $F(C_k)$ is an adaptive function that dynamically predicts the time gap which would possibly indicate the start of a new cluster, here

$$F(C_k) = I - T_w * T_{C_k} - S_w * S_{C_k} \tag{3}$$

where I is the initial value, T_w is the time weight, T_{C_k} is the time span of cluster C_k, S_w is the size weight and S_{C_k} is the size of cluster C_k. Based on observations (a) and (b), with more photos and larger time span of cluster C_k, the chances of adding new photos to this event will be lower as $F(C_k)$ is smaller. Currently, T_w and S_w are heuristically chosen.

3.2 Person Context Estimators

Context information can be used to estimate the probability of person's presence in a photo. We adopt 4 context estimators as proposed in [7]: global, event, time-neighboring and people-rank estimators related to person i. To build the estimators, face recognition results are used in this paper, which is different from [7] where manual annotation results are used. The estimation is based on the following observations:

- Popularity. Some people appear more often than others.
- Co-occurrence. People that appear in the same photos or events may be associated with each other, and have a higher likelihood of appearing together in other photos or events.
- Temporal re-occurrence. Within a specific event, there tend to be multiple photos of the same person.

The first three estimators are modeled in similar ways. The probability of photo p containing person i is modeled as $P_Q(p,i)$:

$$P_Q(p,i) = \frac{\sum_{q \in Q(p)} K_q(i)}{|Q(p)|} \tag{4}$$

where $Q(p)$ represents the set of photos containing photo p, and $K_q(i)$ is 1 if person i is contained in photo q. The form of $Q(p)$ determines the type of estimator. If $Q(p)$ contains all the photos, $P_Q(p,i)$ is the global estimator. If $Q(p)$ only contains photos of the event of photo p, $P_Q(p,i)$ is the event estimator. If $Q(p)$ contains photos of the neighboring time span of photo p, then $P_Q(p,i)$ is the time-neighboring estimator.

Next, we derive the people-rank estimator by making use of the cooccurrence of persons as

$$PeopleRank(j_2|j_1) = \frac{W(j_1,j_2)}{\sum_{i \in I} W(j_1,i)} \tag{5}$$

where $W(j_1,j_2)$ is the number of events or photos where person j_1 and j_2 appear together in the training set.

These four estimators are linearly combined as follows:

$$S_{CON}(p,i) = \sum_{j=1}^{3} \alpha_j P_{Q_j}(p,i) + \sum_{j \in I(p,i)} \beta_j PeopleRank(i|j) \tag{6}$$

where P_{Q_j} represents the global, event, time-neighboring estimators, and $I(p,i)$ is set of persons that appear together with person i in the same photo p or in the same event containing photo p. Currently, we assign the weights heuristically, where higher weights are given to event, time-neighboring and people-rank estimators as the person's presence in a photo is more likely to be inferred from his presence in other photos of the same event or neighboring events, or from related persons' presence in the same event. Further details of the estimators can be found in [7].

4 Visual Information

4.1 Face Recognition with Photograph

Face recognition is a long-studied research topic and many methods have been proposed [1]. However, face recognition is not effective with photos having no restriction on the pose, illumination and expression. So, measures must be taken to circumvent these problems. Face detection [12] is first applied to detect the near frontal faces. To alleviate the pose problem, eye detection is used to rotate the faces so that the two eye are horizontal. The eye detector is trained with AdaBoost, which has been used successfully in face detection [12]. To overcome the illumination problem, we employ the generalized quotient image [13] for delighting. Finally to tackle the pose problem, we use translation and rotation to generate more training faces for three views, *i.e.* left-view, front-view and right-view, for each person.

We then employ the pseudo 2DHMM [14] to perform face recognition. We build three 2DHMMs to model the left-view, front-view and right-view of the face respectively. For each testing face f, the Viterbi Algorithm is used to calculate the recognition probabilities for person i. We consider the top three recognition probabilities P_{M_1}, P_{M_2} and P_{M_3} from models M_1, M_2 and M_3. The face recognition score is

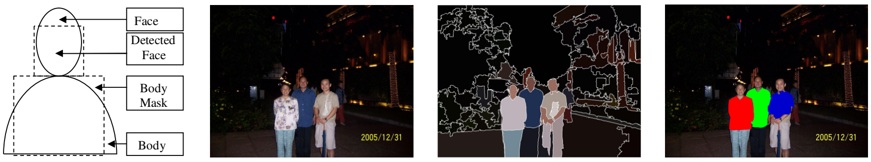
$$S_{FR}(f,i) = 10^3\delta(M_1) + 10^2\delta(M_2) + 10^1\delta(M_3) + (P_{M_1} - P_{M_2})(2\delta(M_1) - 1) \quad (7)$$

where $\delta(M_X)$ is 1 if M_X is one of the three models of the person i, otherwise 0. Further details of the face recognition algorithm can be found in [15].

4.2 Body Detection and Clustering

The body detection uses body mask along with the results of image segmentation, which is performed with mean shift-based feature space analysis [16]. An example of the resulting segmented image is shown in Figure 2(c). With the help of the detected face region in a training data set, a body mask, shown in Figure 2(a), is created to approximate the region of body. For each image segmentation region, we first combine the overlap region between the segmentation region and mask region. We then compute two overlap ratios: the ratios of the overlap region with the segmentation region and the mask region. The eventual body region is extracted based on these two ratios. One example of body detection is shown in Figure 2.

The detected bodies in an event are then grouped into clusters using the constrained clustering method [17]. We employ the affine image matching and feature points matching [18] to identify body regions that are highly similar and should be clustered together. They are the set of “Must-Link” body regions. The set of “Cannot-Link” regions come from the fact that the bodies within the same photo cannot be clustered together as one person cannot appear more than once in a photo. We use LUV color histogram and edge directional histogram



(a) Body Mask (b) Photograph (c) Segmentation (d) Body Detection

Fig. 2. Body Detection

for similarity computation. We employ average-link hierarchical clustering to cluster the body regions. The merging process stops when the average similarity falls below a threshold, which will introduce over-clustering. However, this is better than under-clustering as different persons may have similar contextual information and we want to differentiate the persons. The over-clustered persons will be merged with the help of face recognition information using the graphical model to be discussed in Section 4.3.

4.3 Graphical Model for Combining Face and Body

As discussed in Section 2, body information can help to detect unrecognized faces and reject false recognized faces, and face recognition can help to differentiate persons with similar body information. Obviously, we must combine them to achieve more reliable results. The relationship between face recognition and body information can be properly modeled by graphical model, which is suitable to model the complex relationship between variables. The proposed graphical model for the combined clustering is shown in Figure 3. For a given event k, b_k is the set of body information; c_k is the set of body clusters, i.e. clusters according to body information; f_k is the set of face recognition results while r_k is the resulting clusters combining the body clusters c_k and face recognition results f_k. The reasons for employing this graphical model are as follows:

1) $c_k \to b_k$: the body clustering provides a set of clusters with relatively small variations for body information.
2) $c_k \to r_k$: in order to identify the person's cluster, the small clusters in c_k are encouraged to group into larger cluster with face recognition results from f_k.
3) $r_k \to f_k$: a cluster is encouraged to be split into several clusters if there are several face recognition clusters in it.

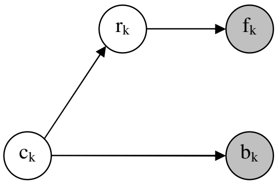


Fig. 3. Graphical Model for Combined Clustering

Given the observation of body information and face recognition results, our goal is to estimate the joint conditional probability of body clustering and combined clustering.

$$p(c_k, r_k|f_k, b_k) = \frac{p(c_k, r_k, f_k, b_k)}{p(f_k, b_k)} \tag{8}$$

To get the optimal clustering results, we maximize the posterior probability:

$$\begin{aligned}(\hat{c_k}, \hat{r_k}) &= \arg\max_{(c_k,r_k)} p(c_k, r_k|b_k, f_k) = \arg\max_{(c_k,r_k)} p(c_k, r_k, f_k, b_k) \\ &= \arg\max_{(c_k,r_k)} p(c_k)p(b_k|c_k)p(r_k|c_k)p(f_k|r_k)\end{aligned} \tag{9}$$

For each cluster $r \in r_k$ in the combined clusters r_k, the average face recognition score for person i is

$$\bar{S}_{FR}(r, i) = \sum_{f\in F(r)} S_{FR}(f, i)/|F(r)| \tag{10}$$

where $F(r)$ is the set of faces contained in cluster r, and $|F(r)|$ is the number of faces; the average context estimation score for person i is

$$\bar{S}_{CON}(r, i) = \sum_{p\in P(r)} S_{CON}(p, i)/|P(r)| \tag{11}$$

where $P(r)$ is the set of photos contained in cluster r, and $|P(r)|$ is the number of photos.

5 Experiments

We evaluate our approach using a personal photo album, containing about 1500 photos with 8 family members. It is taken within 15 months. Each person appears about one hundred times. The album is clustered into 46 events according to the time and location information. The experimental precision/recall results are shown in Figure 4. Four experiments are carried. The curves with “Face” (Equ.(7)), “Face+Body(A)” (Equ.(10)) and “Face+Body(A)+Context” (Equ.(1)) respectively represent the results of employing face recognition only; face recognition with body information; and face recognition with body information plus social context information. The curve with “Face+Body(M)” represents the result of face recognition with manual body clustering. The “Face+Body(M)” is used to provide an indication of upper bound performance if the body information is detected and clustered to 100% accuracy.

It is clear from Figure 4 that the adding of body information to face is very effective. Both precision and recall are improved. The precision is improved because the body information can help to reject the false face recognition through the graphical model clustering. The recall is improved because the body information can get more faces that cannot be recognized by face recognition alone.

The adding of social context information “Face+Body(A)+Context” contributes also to the overall performance. But the improvement is not so big. We

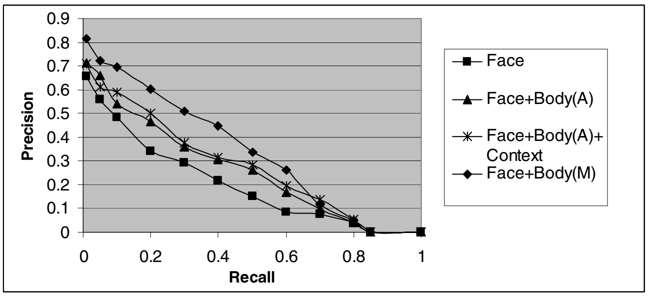


Fig. 4. Recall vs. Precision Performance

can see that its precision is slightly lower than that of "Face+Body(A)" in the low recall range. This is because the contextual information cannot accurately estimate the identity of each detected person. It can only estimate the probability with which each photo contains a person. However, the use of context helps to improve the recall. For example, if no face is recognized by face recognition in an event, the context information can estimate the person's presence if the person appears in nearby events or other related persons appear in this event. Although context estimation will make mistake if "Face+Body(A)" is not accurate, we found that the "Face+Body(A)" detector provides fairly good results, and hence the use of context information improves the overall performance.

We notice that there is a big gap between the manual body clustering and automatic body clustering. This indicates the challenge for body clustering. It also implies that if the body clustering can be improved, the overall performance of person annotation can be improve significantly. Another observation is that the precision becomes zero when the recall reaches about 84%. This is because the face detector can only find about 84% of the presence of persons on average. Again, the improvement in face and body detection results to cover profile faces, faces with sunglasses and backs of bodies etc, would improve the overall performance.

6 Conclusions

In this paper, we proposed an automated method to annotate the names of persons in family photos based on face, content and social context information. The body content information can be used to identify other instances of the recognized persons. Body content information can also improve the recognition accuracy as it can reject falsely recognized faces by performing combined clustering using a graphical model. Also, social context information can be used effectively to estimate the person's presence, though our results indicate that the improvement is minor as it is unable to pinpoint the identity of persons in photo, hence leading to low precision. However, social context information can help to improve the recall. Overall, our results show that the body information is the key to improving the performance of person annotation. However, our current body clustering technique is still preliminary and is far from ideal performance. Future work includes improving the performance of body clustering, the ability to detect bodies without detected faces, and better use of social context information.

References

1. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.: Face recognition: A literature survey. ACM Computing Surveys **35**(4) (2003) 399–458
2. Yip, A.W., Sinha, P.: Contribution of color to face recognition. Perception **31**(5) (2002) 995–1003
3. O'Toole, A.J., Roark, D.A., Abdi, H.: Recognizing moving faces: A psychological and neural synthesis. Trends in Cognitive Science **6** (2002) 261–266
4. Murphy, K., Torralba, A., Freeman, W.T.: Using the forest to see the trees: a graphical model relating features, objects and scenes. In Thrun, S., Saul, L., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 16, Cambridge, MA, MIT Press (2004)
5. Zhang, L., Chen, L., Li, M., Zhang, H.: Automated annotation of human faces in family albums. In: Proceedings of the 11th ACM International Conference on Multimedia. (2003) 355–358
6. Zhang, L., Hu, Y., Li, M., Ma, W.Y., Zhang, H.: Efficient propagation for face annotation in family albums. In: Proceedings of the 11th ACM International Conference on Multimedia. (2004) 716–723
7. Naaman, M., Yeh, R.B., Garcia-Molina, H., Paepcke, A.: Leveraging context to resolve identity in photo albums. In: JCDL. (2005) 178–187
8. Davis, M., Smith, M., Canny, J.F., Good, N., King, S., Janakiraman, R.: Towards context-aware face recognition. In: ACM Multimedia. (2005) 483–486
9. Jensen, F.B.: Bayesian Networks and Decision Graphs. Springer (2001)
10. Cooper, M., Foote, J., Girgensohn, A., Wilcox, L.: Temporal event clustering for digital photo collections. In: Proceedings of the Eleventh ACM Internationl Conference on Multimedia. (2003)
11. Naaman, M., Song, Y.J., Paepcke, A., Garcia-Molina, H.: Automatic organization for digital photographs with geographic coordinates. In: ACM/IEEEE-CS Joint Conference on Digital Libraries. (2004) 53–62
12. Viola, P., Jones, M.: Robust real time object detection. In: IEEE ICCV Workshop on Statistical and Computational Theories of Vision, Vancouver, Canada (2001)
13. Wang, H., Li, S.Z., Wang, Y.: Generalized quotient image. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 2. (2004) 498–505
14. Cardinaux, F., Sanderson, C., Bengio, S.: Face verification using adapted generative models. In: The 6th International Conference on Automatic Face and Gesture Recognition, Seoul, Korea, IEEE (2004) 825–830
15. Zhao, M., Neo, S.Y., Goh, H.K., Chua, T.S.: Multi-faceted contextual model for person identification in news video. In: Multimedia Modeling. (2006)
16. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell **24**(5) (2002)
17. Wagstaff, K., Cardie, C., Rogers, S., Schroedl, S.: Constrained K-means clustering with background knowledge. In: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2001) 577–584
18. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110