COMBINING METADATA AND CONTEXT INFORMATION IN ANNOTATING PERSONAL MEDIA

Ming Zhao[1]*, Tat-Seng Chua*[1]*, Ramesh Jain*[2]

[1] Department of Computer Science, National University of Singapore
21 Lower Kent Ridge Road, Singapore 119077
E-mail: {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
E-mail: jain@ics.uci.edu

ABSTRACT

The proliferation of digital cameras, including those with mobile phones, has resulted in enormous volumes of photographs and video in personal media. To make these resources retrievable, we need to be able to annotate $4W$'s – when, where, who and what. While time and rough location via GPS are generally available, the annotation of who, what and exact where remains challenging. In addition to visual analysis, we must make use of the available metadata and contextual-level information to extract such information. This paper focuses on the "who" annotation problem. By assuming that personal media is mostly about the chronicles of a person, we can make use of patterns of person's social activities to supplement the results of face detection and recognition. By assuming that the same groups of people tend to appear in same events, and wear the same cloths within a short duration and nearby places, we can bring in the body context and social network information to estimate the person's presence in the photos and other examples of the same recognized persons. The combination of face detection/recognition, body context and social network has resulted in an effective system for person annotation in personal media. Experiments on a photo album containing over 1500 photos demonstrate that our approach is effective.

1. INTRODUCTION

Personal digital media are replacing tradition analog media in our daily life and the quantity is exploding. This stimulates the strong need for efficient management tools. Ideally, to achieve the aims of such an efficient management tool, personal media should be annotated with the basic information of $4W$: "when", "where", "who" and "what". Of the $4W$'s, "when" can be acquired easily with the time information recorded by the camera. Part of "where" can be provided by GPS, which gives the location information of the camera when the media is taken. However, the actual objects being taken can be far away from the camera as the main subject such as a building could be several kilometers away. GPS can be noisy because it is not reliable in some situations, especially within the cities and indoors. Moreover, GPS only provides rough location information within the granularity of tens of meters. It thus provides only "where" at the granularity of city name or town name, but not the type of "where" such as home, office, school, park etc. On the annotation of "who", face recognition is usually applied. However, the state-of-art face detection and recognition algorithms are not good enough to cope with the complex situation in photos, such as uncontrolled pose, lighting condition and expression. The annotation of "what" is highly related to the image annotation technique. However, the visual content based image annotation is still quite hard.

Although we can rely on content information to resolve the $4W$ problems, the use of content information alone is inadequate as it is not even easy for humans to resolve "when" and "where". To tackle this problem in a realistic way, it is very important to make use of the available metadata and context information accompanying the personal media.

The information provided by the metadata is very important for the annotation. Various kinds of metadata can be available from Exif (Exchangeable image file format). Exif is a specification for storing interchange information in the image file formats used by digital cameras. This standard was written by the Japan Electronic Industry Development Association (JEIDA) to encourage interoperability between imaging devices, and it is commonly used in digital cameras today. The metadata defined in the Exif specification covers an extensive spectrum including date and time information, camera settings, location information (which could come from a GPS receiver connected to the camera), and descriptions and copyright information. With these metadata, some of the problems of $4W$ can be made easier. For example, we can infer $sunrise$ and $sunset$ from time and location. With the focal length and F-number, we can infer $landscape$ and $outdoor$. By combining weather information with time and location, it will be easier to infer "rain"and "snow" etc. With the altitude from GPS, we can infer "mountain", "beach" and "ocean" etc.

Context information also plays an important role in the annotation, because extrinsic information is often as significant as intrinsic, if not more so. With the extensive information contained in personal media, a specific kind of infor-

mation usually does not come alone. Instead, it often occurs with other related information. So, personal media are usually organized in events. Most personal media are taken to capture events. Meanwhile, the personal-media-taking itself is also an event. The context information within a photos and an event is very useful for the annotation. For example, a small face or a profile face is difficult to be recognized. However, its identity can be inferred from social context such as its popularity in the photo album, other recognized persons in the same photo or in the same event according to their relationship with it, or its occurrence in other photos of the same event. Its identity can also be inferred from the visual context such as body, which is very similar to the recognition process by humans. "Outdoor" can be inferred from the presence of "mountain", "sunrise", "bird" etc. in the same photo or in other photos of the same event.

Realizing the importance of metadata and context in personal media annotation, the paper tries to propose a framework to combine the content, metadata and context. In particular, we apply the proposed framework to the annotation of "who" in family photo album. Face recognition is applied to the content information. However, its performance is limited by the uncontrolled condition of family photos. With the metadata of time and location, photos are clustered into events. Based on the fact that the same groups of people tend to appear in similar events and they tend to wear the same clothes within an event, two types of context information is used: social context information and visual context information. In each event, the social context information is used to estimate the probability of the persons' presence. The visual context information is used to identify other examples of the same recognized persons. Within each event, the visual context 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 a specific person.

The rest of the paper is organized as follows. Section 2 presents related work while Section 3 designing the whole framework. Event clustering and social context estimation is discussed in Section 4. Visual information, including face and visual context, is discussed in Section 5. Section 6 presents the experiments and results before the conclusions are drawn in Section 7.

2. RELATED WORK

The annotation of "who" can be approached with face recognition. While current face recognition systems perform well under relatively controlled environments [1], they tend to suffer when variations in pose, illumination or facial expressions are present. As real life family photographs tend to exhibit large variance in illuminations, poses and expressions of face images, automatic annotation of family photos is hard to 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 person'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].

3. THE FRAMEWORK FOR AUTOMATIC PERSON ANNOTATION

This section discusses the overall framework of combining content (face), metadata (time and location) and context (social context and visual context) for automatic family album annotation. Generally, this framework works with five

steps: (1) Event clustering with content and metadata. (2) Recognition based on content and metadata. (3) Social context estimation based on event clustering and recognition results. (4) Visual context clustering and combined with recognition. (5) Combine recognition and social context estimation. For the annotation of "who", we only use metadata for event clustering and only use face for recognition. The visual context is the person's body.

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. The metadata 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 the metadata of 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

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

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

4. EVENT CLUSTERING AND SOCIAL CONTEXT ESTIMATION

4.1. Event Clustering with Metadata

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 bringing 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 across events. Event is also important for constructing contextual estimators for estimating the probability of the presence of a person in a photo.

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

$$ATD(C_k,p) \leq F(C_k) \quad (2)$$

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

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

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.

If two events happens very close or they are taken by different persons, they can not be clustered into different events by time alone. In these cases, location is applied to split the events. Each photo p is resolved with a top three nearest location names according to a world location gazetteer. For each cluster C_k, all photos should be in close physical proximity. Thus, at least one of the top three location names previously resolved should match. If there are photos without a match in their top three location names, group these photos together and split the cluster C_k into different clusters.

4.2. 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 $P_Q(p,i)$:

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

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

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

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:

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

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].

5. VISUAL INFORMATION

5.1. Face Recognition in Photos

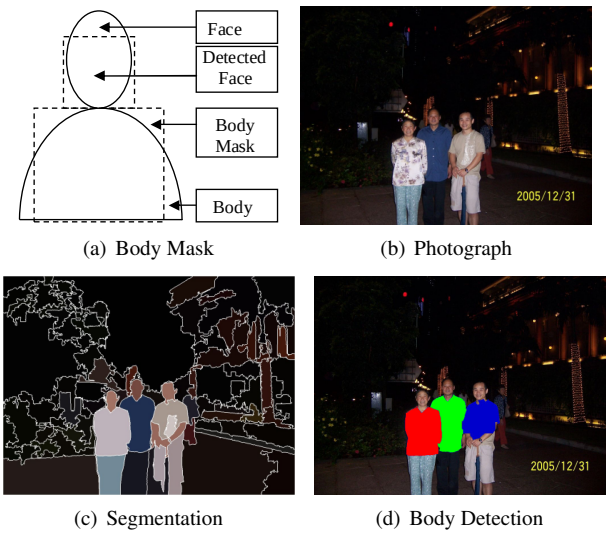
Face recognition [1] is not effective with photos having no restriction on the pose, illumination and expression. So, measures must be taken to circumvent these problems. Face and eye detection is applied with AdaBoost [12] to get the accurate face position. And the generalized quotient image [13] is used for delighting. Models of three views, *i.e.* left-view, front-view and right-view, are built for each person with 2DHMM [14]. 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

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

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

5.2. 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 1(c). With the help of the detected face region in a training data set, a body mask, shown in Figure 1(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 1.



(a) Body Mask (b) Photograph

(c) Segmentation (d) Body Detection

Fig. 1. Body Detection

The detected bodies in an event are then grouped into clusters using the constrained clustering method [17]. The "Must-Link" body regions come from the affine image matching and feature points matching, which can identify body regions that are highly similar and should be clustered together. Let B_1 and B_1 denote two body regions. The affine image matching minimize the difference between B_1 and

B_2, *i.e.*

$$E = \sum_{x,y} |B_1(x,y) - B_2(x',y')|^2 \quad (8)$$

where $[x', y']^T = \mathbf{A}[x, y]^T$ and $\mathbf{A}$ is an affine transformation matrix. SIFT features [18] are extracted for B_1 and B_2. Then, RANSAC [19] is applied to match the feature points in B_1 and B_2.

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 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 5.3.

5.3. Graphical Model for Combining Face and Body

A graphical model is proposed for combined face and body information, which is shown in Figure 2. 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: $c_k \rightarrow b_k$: the body clustering provides a set of clusters with relatively small variations for body information; $c_k \rightarrow r_k$: 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; $r_k \rightarrow f_k$: a cluster is encouraged to be split into several clusters if there are several face recognition clusters in it.

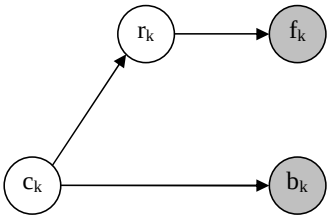


Fig. 2. Graphical Model for Combined Clustering

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

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

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

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

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

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

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

6. 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 3. 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 3 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 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.

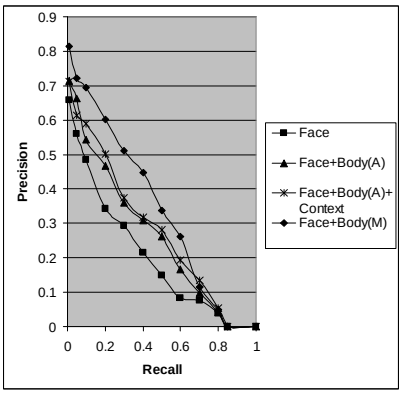


Fig. 3. Recall vs. Precision Performance

7. CONCLUSIONS

The annotation of $4W$'s (when, where, who and what) is essential for the efficient management of the large-scale data in personal media. In this paper, we propose a framework to tackle these annotation problems by making use of content, metadata and context, and this framework is applied to the annotation of "who" in family photo album. First, face recognition is performed with the content. Then, photos are clustered into events with the metadata of time and location. Within each event, the social context information is used to estimate the probability of the persons' presence in each event, and the visual context information is clustered, and then combined with face recognition results using a graphical model. Finally, the clusters with high face recognition confidence and social context estimation probabilities are identified as belonging to a specific person. The experiments show that the proposed framework works effectively. With the help of metadata for event clustering, the visual context can play a key role in improving the performance of person annotation, and the social context can also contribute to estimate the person's presence and improve the recall. Future work includes applying the proposed framework to the annotation of "where" and "what", improving the performance of visual context clustering, better use of social context information, and making use of more metadata for the annotation.

8. REFERENCES

[1] Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.: Face recognition: A literature survey. ACM Computing Surveys **35** (2003) 399–458

[2] Yip, A.W., Sinha, P.: Contribution of color to face recognition. Perception **31** (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** (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** (2004) 91–110

[19] Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Comm. of the ACM **24** (1981) 381–395