Automatic Photo Indexing Based on Person Identity

Seungji Yang[1], Kyong Sok Seo[1], Sang Kyun Kim[2], Yong Man Ro[1],
Ji-Yeon Kim[2], and Yang Suk Seo[2]

[1] Image and Video Systems Lab., Information and Communications University (ICU),
Munjiro 119, Yuseong, Daejeon, South Korea
{yangzeno, imksseo, yro}@icu.ac.kr

[2] Computing Lab., Digital Research Center, Samsung Advanced Institute of Technology
(SAIT), San 14-1, Giheung, Yongin, Kyunggi, South Korea
{skim, jikim, ysseo}@sait.samsung.co.kr

Abstract. In this paper, we propose a novel approach to automatically index digital home photos based on person identity. A person is identified by his/her face and clothes. The proposed method consists of two parts: clustering and indexing. In the clustering, a series of unlabeled photos is aligned in taken-time order, and is divided into several sub-groups by situation. The situation groups are decided by time and visual differences. In the indexing, SVMs are trained with features of pre-indexed faces to model target persons. The representative feature vector of the person group from the clustering is queried to the trained SVMs. Each SVM outputs a numeric confidence value about the query person group. The query person group is determined to the target person by the most confident SVM. The experimental results showed that the proposed method outperformed traditional person indexing method using only face feature and its performance increased to 93.56% from 72.31%.

1 Introduction

Recently, digital cameras are getting popular as providing a convenience for users to easily take a lot of photos. And it is gradually replacing traditional film camera, so the volume of digital photos is continually increased. The digitization of photos makes users easy to show their lives and experiences to their family or friends. But, people might not move their photos from digital cameras to their personal storages in PC since 'moving photos' sometimes means extra works such as manually sorting, selecting, and annotating pictures. Unless people are having enough time to do that or willing to take care of the photos for each event, they would rather leave the photos in the memory stick of digital camera.

Digital photo album [1] can be a useful tool for organizing these large amounts of digital home photos, and it basically works based on automatic clustering or indexing method. However, the traditional digital photo albums still need user's manual work in many parts. When users need to arrange the photos in the digital photo album, they often feel that it is nuisance since it is hard to browse their photos in some meaningful orders. Thus, the manual indexing is pretty time-consuming, tedious, erroneous, and inconsistent, so that it has been a big hurdle for users to use digital cameras.

Y.-S. Ho and H.J. Kim (Eds.): PCM 2005, Part II, LNCS 3768, pp. 877 – 888, 2005.

Under these circumstances, person-based photo indexing is strongly needed because people are most likely to browse photos based on persons who are taken in the photos.

Traditionally, the person-based indexing has been focused on detecting face position of person and representing it efficiently as a compact feature vector. And, many researchers have been worked for this face detection and recognition. But, the main interest of the face detection and recognition technology has been focused on the security system [2 - 3], e.g., intelligent surveillance system, automatic gate control, and face search system to find criminals. Since the security camera takes pictures in a fixed place, the picture usually contains static environment. And it makes the system easy to detect or recognize face from the picture. On the contrary, general home photos contain more complex background with big illumination variation because people bring their cameras anywhere and take pictures whenever they want [4]. Thus the face detection and recognition from the general home photos must be more difficult than those for traditional security system. And, if only face information is used for the face recognition, its performance should be quite low.

To solve the recognition problem for digital home photos, we propose a novel approach for automatic person indexing. The proposed method consists of two steps: clustering and indexing. In the clustering, firstly, a sequence of unlabeled photos is divided into several subsets by situation-based clustering. The situation groups are decided by time and visual differences. So, the face and clothes features can be considered as a person identity in a situation since ones tend not to change their clothes in a certain time range. After that, using the face and clothes information, the person groups are generated by person-identity-based clustering. In the indexing process, support vector person (SVP) for each target person is modeled with support vector machine (SVM). The representative feature vector of the generated person groups in clustering process is queried to the trained SVPs. The SVPs output a confidence value representing how much the query person group is related to a target person in the database. Using the confidence value, the query person group is determined to the target person by selecting the most confident SVP.

The paper is composed of 5 sections. Section 2 covers the details regarding the person-identity-based clustering, where situation clustering is also represented. In Section 3, person-identity-based indexing is described including the clustering process. Section 4 provides the experimental results, and conclusions are drawn in Section 5.

2 Person-Identity-Based Clustering

To be an effective and accurate clustering for digital home photos, our person-identity-based clustering involves two steps. First, a sequence of photos is aligned in order of taken-time and it is divided into several sub-groups by situation clustering. In each situation group, a person is identified by his/her face and clothes. These person identity features are extracted from the photos. Second, in every situation, several person groups are generated as merging similar person identity features. Note that any similarity matching between person identity features in different situation groups would not be happened in this stage. Fig.1. illustrates the overall procedure of the person identity-based clustering.

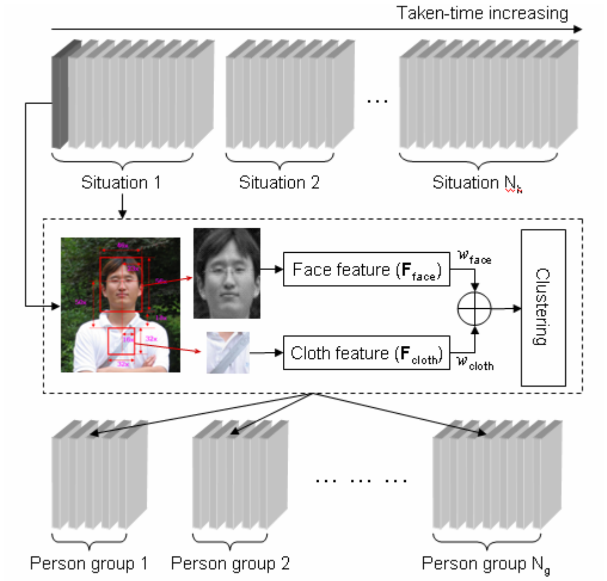

Fig. 1. Overall procedure of the person-identity-based clustering

2.1 Situation-Based Clustering

When clustering the large amounts of home photos based on human, a critical problem is that the photos are taken in a very complex background with variable illumination [4]. But, most people tend to take several pictures in the same place, and these people usually wear same clothes during a certain range of time. Then, a group of neighboring photos according to taken-time may have similar situation. Therefore, when clustering the large amount of photos, they should be divided into smaller amounts of subsets, and it is better to find the same person using helpful information in each subset. Under this observation, 'situation' is defined as the place in which photos would be taken. So, a subset of photos associated in the same situation often contains similar background. This situation can be useful as a fundamental cluster for photo clustering or indexing.

In this paper, given a sequence of photos, situation change boundary is detected by visual change and time gap between adjacent photos with time [5]. For the similarity matching to detect the situation change boundary, first, the photos are aligned in taken-time order. Then, the similarity matching is performed with two neighboring photos. The time dissimilarity (D_{time}) between the current $(i)^{\text{th}}$ photo and the previous $(i-1)^{\text{th}}$ photo are measured as follows,

$$D_{time}(i)=\frac{\log\{\mathbf{F}_{time}(i)-\mathbf{F}_{time}(i-1)+C_{time}\}}{D_{time\_\max}}, \tag{1}$$

where, the taken-time feature $\mathbf{F}_{\text{time}}$ is obtained from Exif header of photo data, log(•) is a time scale function, C_{time} is a constant to avoid zero for input of the scale function.

$D_{\text{time\_max}}$ is maximum time difference. The value of time dissimilarity is scaled, so that it can be less sensitive to the large time difference. The time dissimilarity at the same situation is insensitive [5].

Using the time dissimilarity, the content-based similarity matching between the $(i)^{\text{th}}$ and $(i-1)^{\text{th}}$ photos is performed with several visual features and their importance [5]. It can be written as,

$$D_{total}(i) = \exp\left[D_{time}(i) \times \sum_{f \in \mathbf{F}} \{ w_f(i) \times D_f(i) \} \right] \tag{2}$$

where, $D_f(i)$ is the dissimilarity defined as $D_f\{\mathbf{F}_f(i) - \mathbf{F}_f(i-1)\}$ in which $D(\bullet)$ is the similarity matching function for the feature f. Using the exponential function, the dissimilarity value at smaller value of feature difference is reduced while it is enlarged at higher value. w_f(i) is an importance value for the f feature. It can be adaptive to the visual semantics of photo.

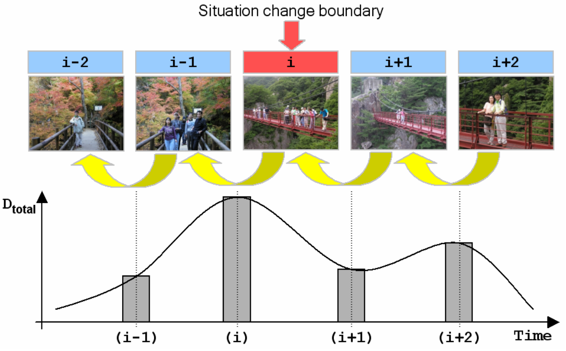

Fig. 2. Situation change detection (This is an example of situation change detection from 5 sequential photos, where the $(i)^{\text{th}}$ photo has the biggest time and visual differences from the previous photo and the difference is assumed as being over a threshold.)

To detect the change of situation between the $(i)^{\text{th}}$ photo and the $(i-1)^{\text{th}}$ photo, we compare $D_{\text{total}}(i-1)$, $D_{\text{total}}(i)$, and $D_{\text{total}}(i+1)$. If the $(i)^{\text{th}}$ photo is a situation change boundary, $D_{\text{total}}(i)$ would have a peak value among three dissimilarity values. Fig. 2 illustrates this overall situation change boundary detection in a sequence of 5 photos. Finally, the situation boundary is detected by (3).

$$D_{total}(i) > \beta \times \{\Delta D_{total}(i) + \Delta D_{total}(i+1)\}, \text{subject to } \Delta D_{total}(i) > 0, \Delta D_{total}(i+1) < 0, \tag{3}$$

where, $\Delta D_{\text{total}}(i) = D_{\text{total}}(i) - D_{\text{total}}(i-1)$ and β is a threshold to detect situation change.

2.2 Person Identity Extraction

To extract person identity, face detection is essential, and many researches have been studied for the face detection. One of the most popular methods is the AdaBoost method with Haar-like features [7 - 8]. The AdaBoost is a boosting method that creates a strong classifier from lots of weak classifiers [9], and the Haar-like features are the inputs of AdaBoost classifier dedicated to face detection. This face detection is another big issue, so it is not considered in this paper.

As mentioned before, face and clothes information is used to identify a person in a given situation. In order to extract the person identity from a photo, the region of the face is detected first, and clothes region of corresponding face is extracted using the detected face position. The size of facial images is normalized as 46 by 56 pixels, so that the center positions of the two eyes in the facial image is located on the 24th row and the 16th and 31st column for the right and left eye, respectively. When deciding the corresponding clothes region, determining the size and position of the clothes region is difficult. There should be a trade-off, i.e., the clothes region should be not only large enough to represent individual identity, but also small enough not to be interfered with each other. Under lots of observations, a heuristic rule is discovered; the clothes location is defined by 18 lines below the face region, and the size of clothes is 32 by 32 pixels. But it is assumed that a person has no clothes information if faces are located at the margin of the photo, or two faces are too closely located, which could affect the other's body. Fig. 3 shows the detection of the face region and the corresponding clothes region.

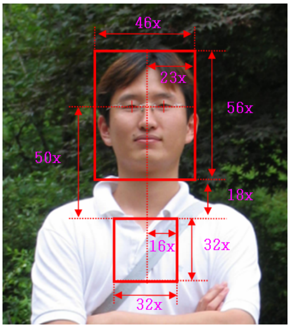

Fig. 3. Detection of face and clothes region

To describe the facial and clothes images, their visual features are extracted. When extracting features, any state-of-art technology can be applied for the face and clothes descriptions. Unlike the face feature, color and texture features seem to be effective for representing the clothes since ones would design the clothes as considering colors and shapes mostly. So the color and texture features are extracted from the defined clothes region. Furthermore, the information about people who are taken together is also helpful because people appearing in a photo means that they must be different. Using this information, a person can be more accurately recognized.

2.3 Person-Identity-Based Clustering

Person identity-based clustering merges similar person identities into a cluster so that all of person identities in a situation are clustered into several groups composed with a same person's images. First of all, feature dissimilarities of face and clothes need to be measured, which are person identities. The dissimilarity (D_{person}) between features of i^{th} and j^{th} person identity is measured as follows,

$$D_{person}(i,j) = w_{face} \cdot \tilde{D}_{face}(i,j) + \sum_{f \in \mathbf{F}_{clothes}} \left\{ w_f \cdot \tilde{D}_f(i,j) \right\} \tag{4}$$

where, $D_{\text{face}}(\bullet)$is a function measuring dissimilarity of the i^{th} and j^{th} face feature, and $D_{\text{f}}(\bullet)$ is a function measuring dissimilarity of the i^{th} and j^{th} clothes features where **f** is one of the clothes feature set $\mathbf{F}_{\text{clothes}}$. w_{face} and w_{f} are weighting values to represent importance of the face feature and the **f** clothes features, respectively. And, '~' is the notation for the normalized dissimilarity. Since the dissimilarities might not be normally distributed, their ranges should be normalized and rescaled to have the range from 0 to 1. This normalization is calculated by (5),

$$\tilde{\mathrm{D}}_f = \frac{\hat{\mathrm{D}}_f - \min(\hat{\mathrm{D}}_f)}{\max(\hat{\mathrm{D}}_f) - \min(\hat{\mathrm{D}}_f)}, \text{ where } \hat{\mathrm{D}}_f = \frac{\mathrm{D}_f - \mu_{\mathrm{f}}}{\sigma_f}, \tag{5}$$

where, μ_{f} and σ_{f} are the mean and variance of the dissimilarities of the **f** features, respectively.

After the dissimilarity measurement, person groups are generated as merging similar person identity features into a cluster so that all person identities in a situation are clustered into several groups composed with a same person's images. In this case, the revealed data is only the dissimilarity values. On this limited condition, several unsupervised clustering methods can be utilized, and one of them is Agglomerative Hierarchical clustering method [10]. This clustering method starts with all singleton clusters and forms the sequence by merging clusters. Major steps of the method are contained in the following procedures:

1. Start with all singleton clusters composed with single person identity, $C(g)$ where $g = 1, 2, \ldots, G$ and G is the number of the person identity clusters, in a situation.
2. Calculate the dissimilarity $D_{\text{cluster}}(a, b)$, between two clusters as follows,

$$D_{cluster}(a,b) = \frac{1}{n_{C(a)} \cdot n_{C(b)}} \cdot \sum_{i \in C(a)} \sum_{j \in C(b)} D_{person}(i,j) \tag{6}$$

   where $n_{C(a)}$ and $n_{C(b)}$ represent the number of persons in the two cluster $C(a)$ and $C(b)$, respectively.
3. Find the nearest two clusters, noted a' and b', as follows,

$$(a',b') = \underset{a,b,a \neq b}{\arg\min} D_{cluster}(a,b) \tag{7}$$

4. Merge the two nearest clusters into one cluster as being $C(a') = C(a') \cup C(b')$ and removing $C(b')$.
5. Repeat (2), (3), (4), until specified dissimilarity has been reached.

Finally, most clusters are composed with person identities of a same person.

3 Person-Identity-Based Indexing

The proposed person-identity-based indexing uses the results of the person-identity-based clustering. The clustered person groups based on situations are automatically indexed to target people in pre-stored database. In this paper, we build a SVM model for target people, called support vector person (SVP), instead of collecting feature instances to represent a person in the database. In the database where face information is only available, the clothes information is no longer valuable.

3.1 Support Vector Person Modeling

SVM [11] is a popular learning method that uses a hypothesis space of linear functions in a high dimensional feature space. Support vectors are trained with a learning algorithm from optimization theory implementing a learning bias derived from statistical learning theory [12]. The linear function given feature instance x can be written as,

$$f_{svm}(\mathrm{x}) = w_0 + \sum_{u=1}^{k} w_u K_u \langle \mathrm{x}, \mathrm{c}_u \rangle, \tag{8}$$

where, w is the parameter vector that control the function, and K_u is a kernel function.

In this paper, we use a radial basis function (RBF) [13] as the kernel, defined as follows.

$$K_u \langle \mathrm{x}, \mathrm{c} \rangle = \exp\{-\gamma_u \cdot (\mathrm{x} - \mathrm{c}_u)^T \cdot (\mathrm{x} - \mathrm{c}_u)\}, \tag{9}$$

where, the kernel K_u is a function that uses the distance between an input vector **x**, and trained the center c_u, where γ_u is a constant. This linear function is trained with labeled dataset composed with positive and negative data. To model SVP for each person, the SVPs are trained with positive and negative face features that are correctly classified by human in prior. The SVP of the person, u, is written as,

$$\mathbf{SVP}_{u \in \mathrm{U}} = \mathbf{SVM}_u \{\mathrm{u}^+(\mathbf{F}_{face}), \mathrm{u}^-(\mathbf{F}_{face})\}, \tag{10}$$

where, **U** is a person set indexed in the database, u^+ is a set of positive face samples that belong to the person, u^+, and u^- is a set of negative face samples that belong to the others except the u.

3.2 Person-Identity-Based Indexing

Fig. 4 shows a general flow of the proposed person-identity-based indexing. As shown in the figure, from the person groups clustered in the person clustering, a representative feature vector is constructed. The representative feature vector is average of each component of the face feature vectors in the person group. It is written as,

$$\mathbf{F}'_{face}\Big|_g = \{f'_{face}(1), f'_{face}(2), \cdots, f'_{face}(N_f)\}, \text{where } f'_{face}(i) = \frac{1}{n_{C(k)}} \cdot \sum_{j=1}^{n_{C(k)}} \{f_{face}(j)\big|_i\}, \tag{11}$$

where, $\mathbf{F}'_{face}|_g$ is an average face feature representing the g^{th} person group.

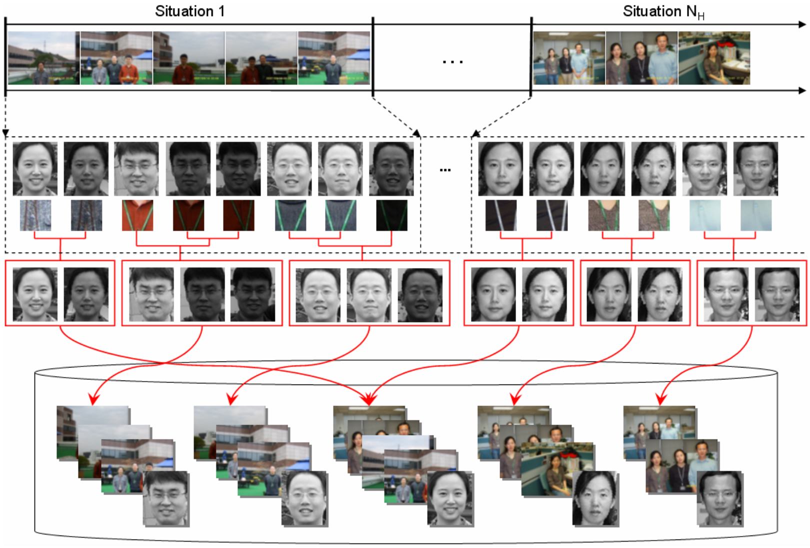

Fig. 4. A general flow of the proposed person indexing

The representative feature vector is used as input to the modeled SVPs. Then, the SVP outputs confidence values which indicates how much the face features is likely to the person. The confidence value of the g^{th} person group for the person, u, v_g(u), is obtained as follows,

$$v_g(u) = K_u\left(\mathbf{F}'_{face}\Big|_g, \mathbf{SVP}_u\right), \tag{12}$$

where, the kernel, K_u can be a distance measurement for the SVPs, and it outputs confidence value, v_g(u) of the face feature $\mathbf{F}'_{\text{face}}$ of the g^{th} person group about the SVP for the person, u.

Finally, using the confidence values from all SVPs, a target person of the person group g, g_{target}, is determined by selecting the most confident SVP, as follows,

$$g_{target} = u = \underset{u\in\mathbf{U}, g\in\mathbf{G}}{\arg\max}\{v_g(u)\}, \tag{13}$$

4 Experiments

To verify the proposed method, experiments were performed with MPEG-7 official dataset for visual core experiment part 3 (VCE-3) [15]. This dataset contains 1385 home photos taken by general users. Using an automatic face detection tool developed by SAIT, 1819 faces were detected and proposed to the VCE-3 [16]. These 1819 facial images were extracted from 1120 photos of the dataset, and 72 people were appeared in the 1120 photos. The experiments were composed of two parts: one was for the demonstration of the usefulness of person-identity-based clustering, and the other was for the verification of the person-identity-based indexing.

4.1 Experiment 1: Person-Identity-Based Clustering

First of all, the 1819 facial images were divided into 31 situation groups by the situation-based clustering. Since the dataset is divided into many situations, every person can be appeared in any situation. To settle the truth number of clusters, we used a notation of facial image groups to represent one person's facial images in a situation. As examining all situations, total 195 facial image groups were detected. After the situation-based clustering, the person identity-based clustering got started with the 31 situations and 1819 singleton clusters. These 1819 singleton clusters would be merged into 195 facial image groups if no error was occurred.

All of the clusters are merged into smaller number of clusters based on dissimilarity values using the weighted combination of the face and clothes features. For the face description, MPEG-7 advanced face recognition descriptor (AFRD) was used, and MPEG-7 color structure descriptor (CSD) with illumination invariant color descriptor (IICD) and edge histogram descriptor (EHD) were used for the color feature and texture feature. These were combined with weight values of $w_{\text{color}} = 0.272$, $w_{\text{texture}} = 0.181$, and $w_{\text{face}} = 0.546$. As an exceptional case, $w_{\text{face}} = 1.0$ was used if the clothes features were not available.

When the clusters were being merged, evaluation method of clustering performance was needed to see how many errors are increased in each step of the proposed method. The error rate is the ratio of the number of minor facial images to the number of major facial images in the cluster. The error rate, $e(g)$, of the person group, g, is computed as changing dissimilarity threshold as follows,

$$e(g)\big|_{threshold} = \frac{n(g) - n_{major}(g)}{n(g)}, \tag{14}$$

where, $n(g)$ is the number of facial images in the g person group, and $n_{\text{major}}(g)$ is the number of major facial images. As easily realized, the bigger dissimilarity threshold leads the smaller number of person groups but increases error rate.

In order to examine the effect using the feature combination, the proposed method was compared with the method using face features only. As Fig. 5 shows the result, 195 clusters were generated with 23.03% of error rate using the face only while 14.02% of error rate was obtained using the proposed method.

4.2 Experiment 2: Person-Identity-Based Indexing

To verify the person-identity-based indexing, some of the original data in the previous experiment was selected, which was containing facial images of 27 persons who had more than 30 facial images. Total number of facial images was 1135. These 1135 facial images were extracted from 764 photos. About 60% of them, 684 facial images were used for training data and the remainders, 451 facial images, were used for testing data. Correspondingly, the proposed SVPs were modeled for the 27 persons.

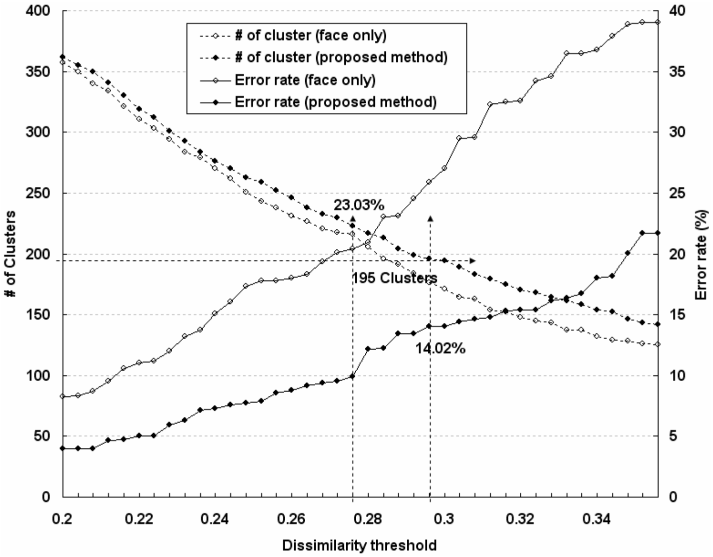

Fig. 5. Person-identity-based clustering results: with 1819 facial images

The test dataset was divided into 13 situation groups by the situation-based clustering, and the truth number of facial image groups was 36. Fig. 6 shows the person-identity-based clustering results of the test dataset. As shown in the results, the error rate was 5.987% when the person groups were merged into 36 groups.

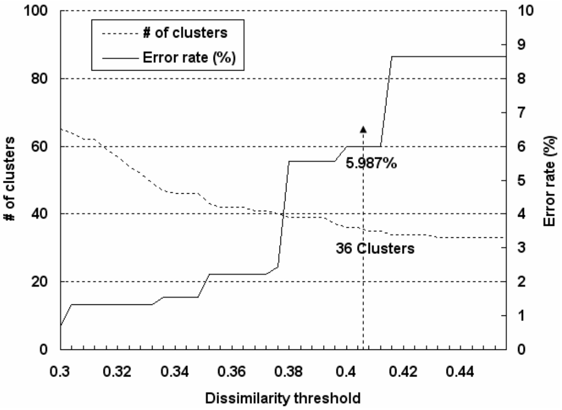

Fig. 6. Person-identity-based clustering results: with only 451 facial images

The stopping criterion for the person-identity-based clustering was heuristically determined: until the number of clusters reached to 10% of the initial number. With the criterion, the person groups were merged into 46 groups, and 1.552% of the error rate was obtained. These merged person groups were used for the next process of the person-identity-based indexing.

Recall and precision were used to evaluate indexing results. The recall is defined as G/T where G is the number of true-positive facial images and T is the number of the true facial images. The precision is defined as G/N where N is the number of positive facial images. Average performance over the all 27 persons was 93.56% (recall: 93.11% and precision: 94.01%) using the proposed method while 72.31% was obtained in the case of using only face features.

5 Conclusions

In this paper, we propose a novel approach for automatic person indexing for digital home photos. The proposed method is composed of two steps: clustering and indexing. In the clustering, a series of unlabeled photos is aligned in order of taken-time and is divided into sub-sets by situation-based clustering. The situation groups are decided by time and visual differences. In every situation group, the face and clothes features are considered as a person identity. They are used to generate the person clusters. In the indexing, pre-indexed face dataset is trained with SVMs in order to model target persons. The representative feature vector of the person clusters is queried to the trained SVMs. Each SVM outputs a numeric confidence value about the query person group. A target person for the query person cluster is determined by the most confident SVM. The experiment results showed that the proposed method outperformed the traditional indexing method using only face features and its performance increased to 93.56% from 72.31%.

References

1. Yagawa, Y., Iwai, N., Yanagi, K., Kojima, K. Matsumoto, K.: The Digital Album: a personal file-tainment system, Proc. of IEEE Intl' Conference on MCS (1996) 433-439.
2. Yang, M. H., Kriegman, D., Ahuja, N.: Detection Faces in Images: A Survey, IEEE Trans. on PAMI (2002) 34-58.
3. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey, ACM Computing Surveys (2003) 399-458.
4. Ahang, L., Chen, L., Li, M., Zhang, H.: Automated Annotation of Human Faces in Family Albums, Proc. of the 11th ACM intl. conf. on Multimedia (2003) 355-358.
5. Yang, S., Kim, S.K., Seo, K.S., Ro, Y.M.: Automated Situation Clustering of Home Photos for Digital Albuming, SPIE (2005)
6. Manjunath, B. S., Salembier, P., Sikora, T.: Introduction to MPEG-7, John Wiley & Sons, LTD (2002).
7. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features, CVPR (2001) 511-518.
8. Wu, B., Ai, H., Huang, C., Lao, S.: Fast Rotation Invariant Multi-View Face Detection Based on Real Adaboost, AFGR (2004) 79-84.
9. Freund, Y., Schapire, R. E.: A decision-theoretic generalization of on-line learning and an application to boosting, Computational Learning Theory: Eurocolt '95 (1995) 23-37.
10. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification second ed., John Wiley & Sons, LTD, (2001)

11. Vapnik, V. N.: The Nature of Statistical Learning Theory, second ed. Springer (1999)
12. Cristianini, N., Shawe-Taylor, J.: An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press (2000).
13. Moody, J., Darken, C. J.: Fast learning in networks of locally tuned processing units, Neural Computation (1989) 281-294.
14. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification second ed., John Wiley & Sons, LTD (2001)
15. NIST Database, /M7photo/Photo_DB/Face-CE_testset_org, MPEG-7 Visual Group, ISO/IEC JTC1/SX29/WG11 (2005)
16. Bober, M., Kim, S.K.: Description of MPEG-7 Visual Core Experiments, MPEG-7 Visual Group ISO/IEC JTC1/SC29/WG11 N6905 (2005)