

Automatic Face Annotation in Personal Photo Collections Using Context-Based Unsupervised Clustering and Face Information Fusion

Jae Young Choi, *Student Member, IEEE*, Wesley De Neve, Yong Man Ro, *Senior Member, IEEE*, and Konstantinos N. Plataniotis, *Senior Member, IEEE*

Abstract—In this paper, a novel face annotation framework is proposed that systematically leverages context information such as situation awareness information with current face recognition (FR) solutions. In particular, unsupervised situation and subject clustering techniques have been developed that are aided by context information. Situation clustering groups together photos that are similar in terms of capture time and visual content, allowing for the reliable use of visual context information during subject clustering. The aim of subject clustering is to merge multiple face images that belong to the same individual. To take advantage of the availability of multiple face images for a particular individual, we propose effective FR methods that are based on face information fusion strategies. The performance of the proposed annotation method has been evaluated using a variety of photo sets. The photo sets were constructed using 1385 photos from the MPEG-7 Visual Core Experiment 3 (VCE-3) data set and approximately 20 000 photos collected from well-known photo-sharing websites. The reported experimental results show that the proposed face annotation method significantly outperforms traditional face annotation solutions at no additional computational cost, with accuracy gains of up to 25% for particular cases.

Index Terms—Clustering, context, face annotation, face information fusion, generic learning, personal photos.

I. INTRODUCTION

THE WIDESPREAD use of digital cameras and mobile phones, as well as the popularity of online photo sharing applications such as “Flickr” [1] and “Facebook” [2] has led to the creation of numerous collections of personal photos [6]. These collections of personal photos need to be managed by users. As such, a strong demand exists for automatic content

Manuscript received July 30, 2009; revised November 30, 2009; accepted March 3, 2010. Date of publication July 26, 2010; date of current version October 8, 2010. This work was supported by the National Research Foundation of Korea, under Grant KRF-2008-313-D01004, and by the IT Research and Development Program of MKE/KEIT, under Grant 2009-F-054-01, Development of Technology for Analysis and Filtering of Illegal and Objectionable Multimedia Content. This paper was recommended by Associate Editor S. Yan.

J. Y. Choi, W. De Neve, and Y. M. Ro are with the Image and Video Systems Laboratory, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 305-732, Korea (e-mail: jygchoi@kaist.ac.kr; wesley.deneve@kaist.ac.kr; ymro@ee.kaist.ac.kr).

K. N. Plataniotis is with the Multimedia Laboratory, Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S3G4, Canada (e-mail: kostas@comm.utoronto.ca).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TCSVT.2010.2058470

annotation techniques [12], [58], [63] that facilitate efficient and effective search in collections of personal photos [7], [15]. Personal photos are commonly annotated along the “who,” “where,” and “when” dimension in that order of importance [3]. Indeed, recent user studies [3]–[5] report that people prefer to organize their photos according to who appears in their photos (e.g., family members or friends). Consequently, the act of labeling faces on personal photos (termed “face annotation” [10] hereafter) is of crucial importance for subject-based organization of personal photo collections [21]–[23].

In general, manual face annotation is a time-consuming and labor-intensive task. To eliminate the need for a manual face annotation step, computer-based face detection [9] and face recognition (FR) [38] techniques should be integrated into an automatic face annotation system [10], [21], [27]. As stated in [9], automatic face detection has become a mature technique. However, traditional FR solutions [38] are still far from adequate in terms of face annotation accuracy for practical applications. This is mainly due to the fact that only appearance information (e.g., shape and texture) of a single face image is used in order to determine the identity of a subject [21]–[28]. This observation especially holds true when having to deal with uncontrolled photo acquisition circumstances. Such acquisition circumstances are frequently encountered in collections of personal photos.

In contrast to generic image sets, it is well-known that collections of personal photos contain rich and powerful context clues [14], [15]. These context clues include metadata such as timestamps and global positioning system tags. Thus, context information can be used as a complementary source of information in order to improve the face annotation accuracy of traditional FR solutions [21], [22], [27]. This paper proposes a novel framework that systematically leverages the benefits of context information such as situation awareness information with current FR techniques for the purpose of face annotation. In particular, we aim at developing an automatic face annotation system that is feasible for use in real-world personal photo collections, in terms of both face annotation accuracy and computational cost. The distinct characteristics of the proposed face annotation method are as follows.

- 1) *Unsupervised* clustering techniques, namely situation and subject clustering, have been designed in order to

group face images that belong to the same subject. The proposed clustering techniques effectively combine content (e.g., color and texture) and context-based information (e.g., photo capture time) of personal photos in order to achieve a reliable clustering performance.

- 2) In order to take advantage of the availability of multiple face images belonging to the same subject, we propose two effective *face information fusion* methods: weighted feature fusion and confidence-based majority voting. These two methods have been designed to take into account the *confidence* of each individual FR result (as obtained for each corresponding face image), thus exploiting a complementary effect that originates from the availability of multiple face images belonging to the same subject. We incorporate these face information fusion strategies into current FR techniques, aiming to improve the overall face annotation accuracy.

The performance of the proposed face annotation method has been evaluated using a variety of photo sets. These photo sets were constructed using 1385 photos from the MPEG-7 Visual Core Experiment 3 (VCE-3) data set and approximately 20 000 Web photos collected from well-known photo-sharing websites such as Flickr [1]. The experimental results show that the proposed face annotation method significantly improves face annotation accuracy by at least an order of magnitude compared to baseline FR solutions (making use of different feature extractors) that only use a single face feature [38], with accuracy gains of up to 25% for particular cases. In addition, our face annotation system is able to achieve a level of face annotation accuracy that meets the requirements of practical applications. Also, the proposed face annotation framework is straightforward to implement and has a low computational complexity.

The remainder of this paper is organized as follows. Section II reviews existing work on face annotation in personal photo collections. In addition, we discuss the differences between our work and already existing face annotation methods. Section III subsequently presents an overview of the proposed face annotation framework. Section IV first explains the definition of situation clustering in personal photo collections. This explanation is then followed by a detailed discussion of the proposed situation clustering method. Our subject clustering method is outlined in Section V. Section VI explains the FR methods that make use of the proposed face information fusion techniques. In Section VII, we present a series of experiments that investigate the effectiveness of the proposed face annotation method. Finally, conclusions are drawn in Section VIII.

II. RELATED WORK

Using face annotation for cost-effective management of personal photo collections is an area of current research interest and intense development [6], [10], [11]. In the past few years, considerable research efforts have been dedicated to the development of face annotation methods that facilitate photo management and search [16]–[29].

Early work on face annotation focused on the development of semi-automatic methods that make use of intelligent user interfaces and relevance feedback techniques [16]–[20]. In [16]–[18], users are first required to manually label the detected face and clothing images through a photo browsing interface. By comparing the similarities of already labeled and unlabeled face/clothing images, the unlabeled face images are sorted and grouped according to the identify information (i.e., the name information) provided for the already labeled face and clothing images. Using a nearest neighbor classification method, a list of candidate identities is subsequently proposed to the user. The user then has to take a decision whether one of the proposed names for a given query face is correct or not. In [19], face images belonging to the same subject are first grouped based on time and torso information. The user is then asked to manually label these grouped faces. In addition, in this paper, all clustering errors need to be manually corrected by the user through a browsing interface.

A major limitation of semi-automatic face annotation is the requirement that users have to confirm or correct the identity of each individual in order to achieve reliable face annotation results for each photo. As such, this approach may be too cumbersome and time-consuming for practical annotation systems that have to deal with a high number of personal photos.

To avoid the need for user intervention during face annotation, automatic face annotation solutions have been developed [21]–[29]. Already existing methods for automatic face annotation can be roughly divided into methods only relying on face features, methods also making use of clothing features, and methods also making use of social context information.

In face annotation approaches that only make use of face features [24], [28], traditional FR solutions are directly applied to annotate faces on personal photos. However, these methods are still suffering from low face annotation accuracy when personal photos were captured in challenging circumstances [21], [27]. Indeed, face images detected in real-life personal photos are often subject to severe variations in illumination, pose, and spatial resolution [62] (see Fig. 7).

In [25]–[27] and [29], it is demonstrated that clothing information can be used to assist in the identification of subjects by complementing the identity information derived from face features. The underlying idea for using clothing images is that subjects in sets of photos taken during a short period of time (e.g., a given day) usually do not change their clothing [27]. In [26], a 4-D feature vector is extracted from each clothing image. This feature vector consists of one relative vertical position and three red-green-blue (RGB) pixels. Next, a probability density model is created using the extracted clothing feature vectors. To recognize the identity of a particular subject, the visual distance between pairs of clothing images is measured by computing the distance between the corresponding probability density models. In [27], the authors construct a Markov random field (MRF) for the personal photos captured in a particular event. This approach allows combining face similarity information with pairwise clothing similarity information. In this paper, pairwise clothing similarity information is computed using both color histogram

and Gabor texture features extracted from the clothing images. The authors show that an MRF-based inference process can lead to improved face annotation accuracy when incorporating clothing features, compared to an approach only using face similarity information. In [29], a clothing feature vector composed of a 50-dimensional banded auto-correlogram and a 14-dimensional color texture moment is used to estimate the posterior probability of intra and extra-personal variations. In this paper, clothing features are integrated with face features using a Bayesian framework in order to estimate the identity of subjects in personal photo collections.

In previous methods using clothing images, two limitations can be identified. First, most previous work has been done under the assumption that discriminatory information, used to identify the subjects in clothing images, can only be preserved within a specific event. Thus, a classifier model taking clothing features as an input needs to be rebuilt for every event in order to guarantee a reliable face annotation accuracy [26], [27]. Consequently, the overall face annotation accuracy might decrease when new test photos, taken during events that are different from the events considered during the training process, need to be annotated. Second, due to the high cost of manually labeling training clothing images for the purpose of recognizing clothing [29], previous methods may often be ineffective in case of a shortage of labeled images.

In [21]–[23], social context information drawn from photo collections was found to be useful for improving face annotation accuracy. In [21], the authors investigate the manual tagging behavior of members of “Facebook” [2], a popular online social network. The authors observe that more than 99% of the individuals tagged are friends or family members of the photographers (for the photo collections investigated). Starting from this observation, the authors propose a face annotation method based on a conditional random field (CRF) model. A CRF is used to combine a FR result with social context information (e.g., the number of times that each subject appears in the entire collection of personal photos of a user) in order to enhance the overall face annotation accuracy. In [22], authors treat the annotation of personal photos as a stochastic process, using a time function that takes as domain the set of all people appearing in a photo collection. In this paper, the authors construct a language probability model for every photo in order to estimate the probability of occurrence of each subject in the photo collections considered. In [23], likelihood scores are computed by taking into account the appearance frequency of each subject and the co-occurrence of pairs of individuals. The aforementioned approaches demonstrate that likelihood scores can be used to produce a limited set of candidate names for subjects appearing in a particular photo.

However, face annotation methods using social context information may be difficult to implement. The implementation difficulties mainly stem from the use of a time-consuming and cumbersome manual labeling effort in order to reliably estimate social context information. More precisely, both the occurrence probability of each subject and the co-occurrence probability of each pair of subjects need to be computed in advance at photo or event-level, using the previously labeled training photos (as described in [22] and [23]).

The research presented in this paper differs in three major aspects from work already described in the scientific literature.

- 1) Our face annotation approach utilizes FR techniques that rely on *face information fusion*. By taking advantage of the availability of multiple face images for the same subject, we are able to significantly improve the face annotation accuracy. Thus far, already existing methods for information fusion have mostly been used in *multimodal* biometric systems that usually consolidate multiple sources of evidence at decision or confidence-level [31]–[35]. However, few studies have described the effectiveness of evidence fusion using multiple representations of the same biometric feature (e.g., a face). Hence, to the best of our knowledge, this paper is the first attempt to devise a systematic face information fusion method that allows improving FR performance.
- 2) Previous face annotation methods utilize clothing information for the purpose of *subject recognition*. Specifically, clothing features in previous face annotation methods are employed as complementary evidence when determining the identity of faces in photos (by combining clothing feature information with FR results). In our method, however, clothing features are utilized for the purpose of *clustering* face images. Our experiment results show that the use of clothing features for clustering face images in personal photo collections can significantly improve the clustering performance (compared to a clustering technique that only makes use of face features).
- 3) Our method does not require the manual creation of training images. Indeed, since clothing features are used by *unsupervised* clustering techniques, our face annotation solution does not require labeling of clothing and face images by hand. In addition, our face annotation framework incorporates a training scheme based on generic learning (GL) [30]. This approach solves the problem of having an insufficient number of training face images. Consequently, in contrast to previous approaches using clothing features [25]–[27], [29] and social context [21]–[23], the face annotation accuracy of our method does not suffer from a shortage of training images.

III. OVERVIEW OF THE PROPOSED FRAMEWORK

The proposed face annotation method largely consists of three sequential steps: situation clustering, subject clustering, and FR based on a GL-based training scheme and face information fusion.

Fig. 1 provides an overview of the proposed face annotation framework. Situation and subject clustering techniques make effective use of both content and context-based information. Three different types of context information are exploited by the two clustering techniques, namely temporal re-occurrence, spatial re-occurrence, and visual context information (see Table I for more details). In the situation clustering step, photos in the collection are divided into a number of situation clusters. Each situation cluster consists of multiple photos

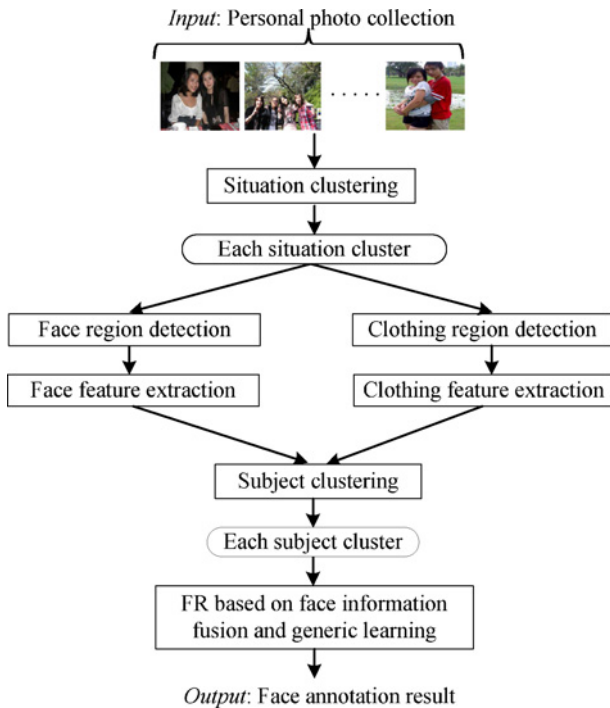


Fig. 1. Overview of the proposed face annotation framework.

that are similar in terms of both capture time and visual characteristics (explained in more detail in Section IV). The primary purpose of situation clustering is to be able to reliably apply visual context information when clustering the subjects that appear in photos belonging to a particular situation cluster. The goal of subject clustering is to group multiple face images that belong to the same subject. To this end, face and associated clothing regions are detected and segmented in all photos belonging to a particular situation cluster.

These regions are then properly normalized to have a pre-specified prototype size and rotation. Face and clothing features (e.g., color information [60]) are subsequently extracted, making it possible to utilize face and clothing information in a complementary way during subject clustering.

Multiple face images within each subject cluster are transformed into corresponding low-dimensional face features such as eigenfeatures [48]. This transformation is done by using a face feature extractor [12] constructed with a GL-based training scheme [30]. Multiple face features of the same subject are then combined using the proposed face information fusion strategies. This allows for improved matching with the face feature of a target subject (i.e., a subject with a known identity). Finally, based on the obtained FR results, the face images in each subject cluster are annotated with the identity of the correct target subject. Each of the three steps shown in Fig. 1 will be described in more detail in the following sections.

IV. SITUATION CLUSTERING

It is well-known that photographers tend to take *multiple* photos in the proximity of a single location during a short period of time [25], [27]. Thus, photos taken in a short

TABLE I
CONTEXT INFORMATION USED BY THE PROPOSED CLUSTERING
TECHNIQUES

Type	Description
Temporal re-occurrence context	It is likely that multiple face images of the same subject appear in photos that are taken sequentially within a short time interval.
Spatial re-occurrence context	It is likely that a given subject appears in multiple photos that have been grouped together based on location information.
Visual context	1) The same subject may not appear more than once in the same photo. 2) For a given subject, it is likely that appearance characteristics such as hair styling and clothing remain consistent in a sequence of photos collected over a short period of time.

period of time are likely to have visual characteristics that are stationary or similar. These characteristics reflect the temporal and spatial re-occurrence context described in Table I. In this section, we outline a situation clustering technique that takes advantage of temporal and spatial re-occurrence contextual information.

A. Definition of Situation Cluster

In [40] and [41], similarity in capture time and content are separately used to automatically cluster photos into different events. In [40], the authors create a time similarity matrix. The rows and columns of this matrix are filled out with the time similarity scores computed for each pair of adjacent photos, taking into account the capture time of the photos. Based on this time similarity matrix, the similarity scores between two photos are first obtained. In a next step, the photo collection is segmented into several events by comparing the computed similarity scores. In [41], a color histogram technique is used to represent photos. To merge photos into events, a block-based color histogram correlation method is used to compute the image similarity between two different photos.

In the event-based photo clustering methods described above, a single event cluster may still contain photos that are completely dissimilar in terms of visual information, although the photos considered belong to the same event. For example, personal photos taken during a single event may contain different objects (e.g., sea, rocks, and trees). Further, the visual appearance of a particular subject (e.g., in terms of clothing and hair) may be completely different for photos taken at different days. In that case, already existing event-based clustering methods may not be able to correctly group multiple face images belonging to the same subject.

To overcome the aforementioned drawback, we split a particular event into several *situation* clusters, taking into account both the capture time and the visual similarity of the photos. We define a “situation cluster” as a group of photos that have close capture times and similar visual characteristics. Fig. 2 illustrates a number of personal photos that have been clustered into two situations (all photos belong to the same event). As shown in Fig. 2, the visual appearance of the subject is consistent for the photos linked to a particular situation. Hence, the visual context can be readily utilized



Fig. 2. Illustration of situation clusters in a personal photo collection available on “Flickr.” The text enclosed in brackets represents the capture time of the above photos in the following format: *year:month:day:hour:minute*. Based on the similarity of the visual characteristics and the proximity of the capture time, the photos taken during the trip of the user can be divided into two situation clusters.

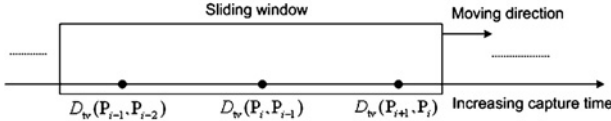


Fig. 3. Situation boundary detection using a sliding window. The sliding window moves from the oldest photo to the most recent photo. Each time the sliding window is moved, three consecutive dissimilarity values are computed.

during subject clustering. The proposed situation clustering method is explained in more detail in the following subsection.

B. Detecting Situation Boundaries

Let $\mathbf{P} = \{P_i\}_{i=1}^{N_p}$ be a collection of N_p personal photos that need to be annotated, where P_i denotes the i th photo. We assume that all of the photos in \mathbf{P} are arranged in ascending order according to their capture time. Let t_i be the capture time of P_i and let $\mathbf{V}_i = \{v_i^{(n)}\}_{n=1}^K$ be a set of K low-level visual feature vectors extracted from P_i . We assume $v_i^{(n)}$ is the n th low-level visual feature vector [39] of P_i , with $1 \leq n \leq K$. Low-level visual features may include texture and color. Note that the basic unit of t_i is minutes, as obtained by converting the “date” and “time” tags of an exchangeable image file format (EXIF) header into minutes [41].

In order to compute the difference in time between any two consecutive photos P_i and P_{i-1} , a time dissimilarity function is defined as follows:

$$D_t(t_i, t_{i-1}) = \frac{\log(t_i - t_{i-1} + c)}{\log(\nabla t_{\max})} \quad (1)$$

where c is a constant scalar to avoid a zero-valued input for the logarithmic scale function in the numerator term. In (1), ∇t_{\max} denotes the maximum value of all time differences computed between every pair of two adjacent photos. As such, since $t_{i-1} < t_i$, $\Delta t_{\max} = \max_{i, i-1} (t_i - t_{i-1})$ and $2 \leq i \leq N_p$. Note that in (1), a logarithmic function is used to properly scale the large variance of the capture time, which may range from a few hours to a few months. Thus, $D_t(t_i, t_{i-1})$ is less sensitive to the difference in time between P_i and P_{i-1} when assuming that both pictures are linked to the same situation. The central idea behind such insensitivity is that the duration of a single situation is usually short. To compare the visual characteristics between P_i and P_{i-1} , we define a visual dissimilarity function

as follows:

$$D_v(\mathbf{V}_i, \mathbf{V}_{i-1}) = \sum_n D_v^{(n)}(v_i^{(n)}, v_{i-1}^{(n)}) \quad (2)$$

where $D_v^{(n)}(v_i^{(n)}, v_{i-1}^{(n)})$ denotes a function that computes the dissimilarity between $v_i^{(n)}$ and $v_{i-1}^{(n)}$.

Compared with already existing event-based clustering methods, we consider both time and visual differences at the same time. As such, the final dissimilarity between P_i and P_{i-1} is computed using (1) and (2)

$$D_{tv}(P_i, P_{i-1}) = \exp(D_t(t_i, t_{i-1}) \times D_v(\mathbf{V}_i, \mathbf{V}_{i-1})). \quad (3)$$

In (3), an exponential function is used to emphasize both smaller time and visual differences. To be more specific, for smaller time and visual differences, the total difference $D_{tv}(P_i, P_{i-1})$ will also be small, whereas the total difference will be large for either large time or visual differences, or when both the time and visual difference are significant.

To divide a photo collection \mathbf{P} into a number of situation clusters, we need to detect the corresponding situation boundaries. The detection of situation boundaries in a photo collection rests on the following observation: photos adjacent to a boundary generally display a significant change in their capture time and visual content (as illustrated in Fig. 2). Based on this observation, three consecutive dissimilarity values $D_{tv}(P_{i-1}, P_{i-2})$, $D_{tv}(P_i, P_{i-1})$, and $D_{tv}(P_{i+1}, P_i)$ are computed using (3), forming a sliding window as depicted in Fig. 3.

The presence of a situation change boundary is checked at each window position, in the middle of the window, according to the following rule:

$$D_{tv}(P_i, P_{i-1}) > \kappa \cdot (|\Delta D_{tv}(P_i, P_{i-1})| + |\Delta D_{tv}(P_{i+1}, P_i)|) \\ \text{subject to } \Delta D_{tv}(P_i, P_{i-1}) > 0 \text{ and } \Delta D_{tv}(P_{i+1}, P_i) < 0 \quad (4)$$

where $\Delta D_{tv}(P_i, P_{i-1}) = D_{tv}(P_i, P_{i-1}) - D_{tv}(P_{i-1}, P_{i-2})$, ($\Delta D_{tv}(P_{i+1}, P_i)$ can be calculated in a similar way), κ controls the degree of merging ($0 < \kappa < 1$), and $|\cdot|$ denotes the absolute value function. It should be emphasized that, prior to (4), all $D_{tv}(P_i, P_{i-1})$ must first be normalized and rescaled in order to have the same range, with $1 \leq i \leq N_p$. In (4), P_i is considered a situation boundary if $D_{tv}(P_i, P_{i-1})$ is equal to the maximum of the three dissimilarity values mentioned above. The underlying idea behind this is that, if



Fig. 4. Examples of detected faces and associated clothing regions. The segmented face and clothing images are placed to the right side of each original photo. Note that each cropped face image is rescaled to a size of 86×86 pixels, while the corresponding clothing image has a rectangular resolution of 68×32 pixels.

P_i represents a situation boundary, then large differences will exist between both the capture time and visual features of P_{i-1} and P_i , much larger than the differences computed for each pair of adjacent photos included in the sliding window.

C. Determining an Optimal Clustering Resolution

Note that, by varying κ in (4), we are able to obtain situation clusters with a different granularity. For smaller κ , fine-grained situation clusters are obtained, while for larger κ , coarse-grained situation clusters are acquired. A simple but effective way of determining an optimal value for κ (i.e., a value that guarantees the best clustering resolution) is as follows. Let $\mathbf{S}_\kappa = \{\mathbf{S}_\kappa^{(n)}\}_{n=1}^M$ be a set of situation clusters $\mathbf{S}_\kappa^{(n)}$ generated for a particular value of κ , with M denoting the total number of situation clusters detected in \mathbf{P} . To determine the “confidence” (or “goodness”) of the set \mathbf{S}_κ for a given value of κ , we compute the average intra and inter-situation dissimilarities over \mathbf{S}_κ . The confidence score for a particular κ is then computed as follows:

$$C(\kappa) = \frac{\sum_{n=1}^{M-1} \sum_{P_i \in \mathbf{S}_\kappa^{(n)}} \sum_{P_j \in \mathbf{S}_\kappa^{(n+1)}} \frac{D_{\text{tv}}(P_i, P_j)}{|\mathbf{S}_\kappa^{(n)}| \cdot |\mathbf{S}_\kappa^{(n+1)}|} - \sum_{n=1}^M \sum_{P_i, P_j \in \mathbf{S}_\kappa^{(n)}} \frac{D_{\text{tv}}(P_i, P_j)}{|\mathbf{S}_\kappa^{(n)}|^2 - |\mathbf{S}_\kappa^{(n)}|}}{|\mathbf{S}_\kappa^{(n)}|^2 - |\mathbf{S}_\kappa^{(n)}|} \quad (5)$$

where $|\mathbf{S}_\kappa^{(n)}|$ denotes the number of photos included in a particular situation cluster $\mathbf{S}_\kappa^{(n)}$ with $1 \leq n \leq M$. Note that in (5), the first term on the right denotes the average inter-situation dissimilarity over \mathbf{S}_κ , calculated by summing the average dissimilarities between two adjacent situation clusters, while the second term denotes the average intra-situation dissimilarity over \mathbf{S}_κ . Using (5), the optimal κ is determined as follows:

$$\kappa_{\text{opt}} = \arg \max_{\kappa \in (0,1]} C(\kappa). \quad (6)$$

Note that in (6), the optimal value for κ , denoted as κ_{opt} , is determined by selecting the value of κ that maximizes $C(\kappa)$. This is realized by computing $C(\kappa)$ over the range $\kappa \in (0, 1]$, using a step size equal to “0.02.” In (6), the resulting set of situation clusters, generated at κ_{opt} , achieves a maximum inter-situation dissimilarity, while the intra-situation dissimilarity is minimal. As explained in the next section, subject clustering is subsequently applied to the individual situation clusters $\mathbf{S}_{\kappa_{\text{opt}}}^{(n)}$ ($n = 1, \dots, M$).

V. SUBJECT CLUSTERING

The ultimate goal of subject clustering is twofold: first, all face images of the same subject should be collected in a single cluster, and second, the face images of different subjects should be part of different clusters. This section provides a more detailed description of our subject clustering technique.

A. Extraction of Face and Clothing Regions

Based on the visual context described in Table I, it is reasonable to assume that the clothing of a particular subject typically remains invariant between photos acquired over a short period of time (i.e., the photos within a situation cluster). Hence, the features derived from clothing, as well as face features, can be used to differentiate one subject from other subjects for the purpose of clustering face images.

The face and clothing detection and segmentation methods used during the subject clustering step are as follows.

- 1) Given a single photo, using any state-of-the-art face detection technique, face regions are first detected and extracted. For normalization purposes, each of the detected face images is rotated and rescaled to 86×86 pixels, placing eye centers on fixed pixel locations (as recommended in [52]).
- 2) The associated clothing region is extracted using the position and relative scale of a bounding box that surrounds each detected face image (see Fig. 4). Based on extensive experiments, a simple but effective detection rule was devised: a clothing region is assumed to be located below the corresponding face region at a distance that is one-third of the height of the face region in terms of pixel rows. Further, the clothing region is segmented to have a size of 68×32 pixels.

Our experiments show that the chosen location and size of the clothing region allows for a sufficient amount of discriminative power in order to tell whether two clothing images belong to the same subject or not (see Section VII-A). Also, we have found that our clothing region detection rule is robust to the variation in spatial resolution of the original picture. Fig. 4 illustrates the detected faces and associated clothing regions from two real-world photos using our detection rule. The faces in these photos are located using the *Viola-Jones* face detection package [42], while the associated clothing regions are found using the proposed rule. As shown in Fig. 4, the use of a narrow clothing bounding box is helpful to find and extract reliable clothing regions when partial occlusions occur between the individuals appearing in a photo.

B. Subject Clustering Process

Since we do not have *a priori* information about possible identities or the nature of face or clothing feature observations, unsupervised clustering techniques are suitable for the purpose of subject clustering. In this paper, the average linkage-based hierarchical agglomerative clustering (HAC) [43], [55] technique is adopted for subject clustering. HAC procedures are among the best known unsupervised clustering solutions.

We now explain the proposed subject clustering method. Let FI_i be the i th face image and let CI_i be the corresponding clothing image detected in a photo that belongs to a particular $\mathcal{S}_{\kappa_{\text{opt}}}^{(n)}$, with $1 \leq i \leq N_s$, and where N_s denotes the total number of face (or clothing) images extracted from all photos in $\mathcal{S}_{\kappa_{\text{opt}}}^{(n)}$. Let f_i be a face feature of FI_i and let $c_i^{(n)}$ be the n th clothing feature of CI_i . Note that f_i can be obtained using any face feature extraction method (e.g., using global or local-based face features [59]). In addition, let $\{c_i^{(n)}\}_{n=1}^{N_c}$ be a set consisting of N_c clothing features, for instance representing color, texture, and edge information. Then, using f_i and $\{c_i^{(n)}\}_{n=1}^{N_c}$, a corresponding subject-identity feature can be defined as follows:

$$\mathbf{F}_i = \{f_i, c_i^{(1)}, \dots, c_i^{(N_c)}\}. \quad (7)$$

The authors of [32] show that a weighted sum of identity information obtained from multiple biometric modalities is effective for identity verification. Since \mathbf{F}_i includes distinct feature modalities consisting of face and multiple clothing features, we use a weighted sum of their dissimilarity scores when computing the dissimilarity between \mathbf{F}_i and \mathbf{F}_j ($i \neq j$). To this end, the dissimilarity function is defined as follows:

$$D_{\text{fc}}(\mathbf{F}_i, \mathbf{F}_j) = w_f \cdot D_f(f_i, f_j) + \sum_{n=1}^{N_c} w_c^{(n)} \cdot D_c^{(n)}(c_i^{(n)}, c_j^{(n)}) \quad (8)$$

where $D_f(f_i, f_j)$ and $D_c^{(n)}(c_i^{(n)}, c_j^{(n)})$ are metric functions that measure the dissimilarity between their two input arguments, w_f and $w_c^{(n)}$ denote user-defined weights to control the importance of the face and clothing features, and $w_f + \sum_{n=1}^{N_c} w_c^{(n)} = 1$. Appropriate weight values were experimentally determined by means of an exhaustive tuning process (see also Section VII-A). In (8), the weighted combination facilitates a *complementary effect* between its different components, positively affecting the classification performance. Indeed, the rationale behind this complementary effect is that a loss in discriminative classification power, caused by less reliable face or clothes features, can be compensated by other features with good discriminative capability. It is important to note that, prior to the computation of the weighted combination, $D_f(f_i, f_j)$ and $D_c^{(n)}(c_i^{(n)}, c_j^{(n)})$ must be normalized and rescaled in order to have the same range (from 0 to 1).

Using \mathbf{F}_i and $D_{\text{fc}}(\cdot)$, we summarize the proposed subject clustering algorithm in Table II. In addition, Fig. 5 visualizes the proposed subject clustering process. In Fig. 5, we assume that the number of subject clusters in the initial stage is seven. As such, $N_s = 7$. In the final iteration, three face images belonging to the same subject are grouped into \mathbf{C}_1 , while four face images belonging to another subject are assigned

TABLE II
PROPOSED ALGORITHM FOR SUBJECT CLUSTERING

- 1) Since a situation cluster $\mathcal{S}_{\kappa_{\text{opt}}}^{(n)}$ contains N_s subject-identity features, HAC begins with the creation of N_s singleton subject clusters. A singleton subject cluster is denoted as \mathbf{C}_i , where $i = 1, \dots, N_s$. Note that each \mathbf{C}_i consists of a single subject-identity feature \mathbf{F}_i .
- 2) Calculate the average dissimilarity (or distance) between \mathbf{C}_i and \mathbf{C}_j by summing the pairwise dissimilarities between the subject-identity features in the two selected subject clusters

$$D_{\text{cls}}(\mathbf{C}_i, \mathbf{C}_j) = \frac{1}{|\mathbf{C}_i| \cdot |\mathbf{C}_j|} \cdot \sum_{\mathbf{F}_m \in \mathbf{C}_i} \sum_{\mathbf{F}_n \in \mathbf{C}_j} D_{\text{fc}}(\mathbf{F}_m, \mathbf{F}_n) \quad (9)$$

where $|\mathbf{C}_i|$ and $|\mathbf{C}_j|$ represent the total number of subject-identity features observed in \mathbf{C}_i and \mathbf{C}_j , respectively, and $m \neq n$, and $1 \leq m, n \leq N_s$.

- 3) Find the two nearest clusters \mathbf{C}_{i^*} and \mathbf{C}_{j^*} by comparing all $D_{\text{cls}}(\mathbf{C}_i, \mathbf{C}_j)$ one by one in the following way:

$$(\mathbf{C}_{i^*}, \mathbf{C}_{j^*}) = \arg \min_{i,j} D_{\text{cls}}(\mathbf{C}_i, \mathbf{C}_j) \text{ for } i \neq j. \quad (10)$$

- 4) Merge the two nearest clusters into a single cluster $\mathbf{C}_i = \mathbf{C}_{i^*} \cup \mathbf{C}_{j^*}$ and subsequently remove \mathbf{C}_{i^*} and \mathbf{C}_{j^*} .

- 5) Repeat steps 2, 3, and 4 if any pair of subject clusters exists that satisfies $D_{\text{cls}}(\mathbf{C}_i, \mathbf{C}_j) < \zeta$, where ζ is a pre-determined stopping threshold value. Otherwise, when all $D_{\text{cls}}(\mathbf{C}_i, \mathbf{C}_j) \geq \zeta$, the subject clustering process is terminated.

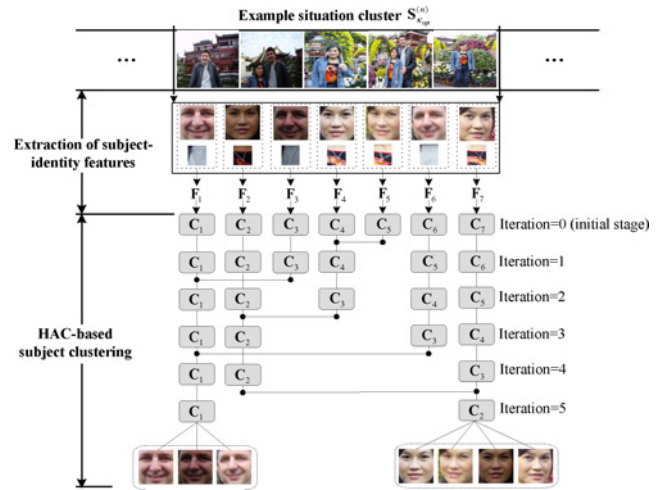


Fig. 5. Illustration of HAC-based subject clustering. During the initial stage, all seven subject-identity features lie in a singleton cluster. During each subsequent iteration, a selected pair of subject clusters is merged into a single subject cluster in a recursive manner. Finally, the subject clustering process terminates if all dissimilarities at hand exceed a pre-determined stopping threshold. Note that the pair of clusters chosen for merging consists of those two subject clusters that have the smallest inter-cluster dissimilarity [as described in (10)].

to \mathbf{C}_2 . In the subsequent FR step, the different face images available in the final subject clusters \mathbf{C}_1 and \mathbf{C}_2 are used for face recognition purposes, using face information fusion. The latter process will be discussed in more detail in Section VI.

C. Determining an Optimal Stopping Threshold

In general, selecting an appropriate value for the threshold ζ is critical for the clustering quality [43], [55]. The clustering quality depends upon how the samples are grouped into clusters and also on the number of clusters produced [46].

In a typical HAC-based clustering setting, the problem of finding an optimal stopping threshold is equal to that of deciding which hierarchical level actually represents a *natural* clustering.

In this section, we develop a robust stopping threshold selection criterion which aims at an optimal compromise between the *clustering error* and the *degree of merging*. In the proposed criterion, the clustering error is represented by the *cluster compactness*, which indicates to what extent all subject-identity features within a cluster are similar, based on the dissimilarities defined in (8). Further, the degree of merging reflects to what extent the number of generated clusters is close to the number of clusters in the context of natural grouping.

We now discuss how to determine an optimal stopping threshold. Suppose that HAC-based subject clustering generates a total number of N subject clusters when a pre-specified stopping threshold is ζ . The within-cluster error associated with ζ is then defined as follows:

$$e_w(\zeta) = \sum_{i=1}^N \sum_{\mathbf{F}_i^{(m)} \in C_i} D_{fc}(\mathbf{F}_i^{(m)}, \bar{\mathbf{F}}_i)^2. \quad (11)$$

In (11), $\mathbf{F}_i^{(m)}$ denotes the m th subject-identity feature assigned to C_i and $\bar{\mathbf{F}}_i$ denotes the mean identity feature of C_i such that $\bar{\mathbf{F}}_i = \{\bar{f}_i, \bar{c}_i^{(1)}, \dots, \bar{c}_i^{(N_c)}\}$, where $\bar{f}_i = \frac{1}{|C_i|} \sum_{\mathbf{f}_i^{(m)} \in C_i} f_i^{(m)}$, and $\bar{c}_i^{(k)} = \frac{1}{|C_i|} \sum_{c_i^{(k,m)} \in C_i} c_i^{(k,m)}$ with $f_i^{(m)}$ and $c_i^{(k,m)}$ denoting the face and the k th clothing feature vector of $\mathbf{F}_i^{(m)}$, respectively, and $1 \leq k \leq N_c$. Note that in (11), a sum-of-squared errors is used to represent the within-cluster error, a simple measure widely used in HAC [43]. Likewise, the between-cluster error with respect to ζ is defined as follows:

$$e_b(\zeta) = \sum_{i=1}^N D_{fc}(\bar{\mathbf{F}}_i, \bar{\mathbf{F}})^2 \quad (12)$$

where $\bar{\mathbf{F}} = \{\bar{f}, \bar{c}^{(1)}, \dots, \bar{c}^{(N_c)}\}$ denotes the global mean subject-identity feature, $\bar{f} = \frac{1}{N} \sum_{i=1}^N |C_i| \cdot \bar{f}_i$, and $\bar{c}^{(k)} = \frac{1}{N} \sum_{i=1}^N |C_i| \cdot \bar{c}_i^{(k)}$. At the beginning of HAC-based clustering, each cluster C_i has a single-subject identity feature (i.e., $\mathbf{F}_i^{(m)} = \bar{\mathbf{F}}_i$) so that $e_w(\zeta)$ is equal to zero. This means that, while HAC-based clustering proceeds, $e_w(\zeta)$ will be at least equal to or higher than the within-cluster error computed during the initial stage (which is zero). Thus, the minimum lower bound of $e_w(\zeta)$ is obtained during the initial stage of HAC-based clustering. On the other hand, $e_b(\zeta)$ achieves its maximum upper bound during the initial¹.

Based on these two observations, we now derive the *cluster compactness gain* to effectively measure the cluster compactness with respect to changes in the stopping threshold ζ . Let $\Delta e_w^{(i)}(\zeta)$ be the increase in within-cluster error caused by a subject cluster C_i during the last stage of clustering with a particular stopping threshold value ζ , compared to the within-cluster error computed during the initial stage. As such,

$\Delta e_w^{(i)}(\zeta)$ can be expressed as follows:

$$\Delta e_w^{(i)}(\zeta) = \sum_{\mathbf{F}_i^{(m)} \in C_i} D_{fc}(\mathbf{F}_i^{(m)}, \bar{\mathbf{F}})^2 - 0. \quad (13)$$

Note that in (13), the within-cluster error during the initial stage is equal to zero. Likewise, let $\Delta e_b^{(i)}(\zeta)$ be the decrease in between-cluster error caused by a subject cluster C_i during the last stage of clustering with a particular stopping threshold value ζ , compared to the between-cluster error computed during the initial stage. As such, $\Delta e_b^{(i)}(\zeta)$ can be written as follows:

$$\Delta e_b^{(i)}(\zeta) = \sum_{\mathbf{F}_i^{(m)} \in C_i} D_{fc}(\mathbf{F}_i^{(m)}, \bar{\mathbf{F}})^2 - D_{fc}(\bar{\mathbf{F}}_i, \bar{\mathbf{F}})^2 \quad (14)$$

where the first and second terms on the right-hand side of (14) denote the between-cluster error caused by C_i during the initial stage (the sum refers to all initial clusters that have resulted in the creation of C_i during the last stage) and the between-cluster error caused by C_i during the last stage for a particular ζ , respectively. Using (13) and (14), the cluster compactness gain parameterized by ζ for C_i can then be defined as follows:

$$\Lambda_i(\zeta) = \Delta e_b^{(i)}(\zeta) - \Delta e_w^{(i)}(\zeta). \quad (15)$$

In (15), $\Lambda_i(\zeta)$ measures the cluster compactness gain caused by C_i , given ζ , and this relative to the initial stage. Using (15), the cluster compactness gain for all subject clusters can simply be computed as follows:

$$\Lambda(\zeta) = \sum_{i=1}^N \Lambda_i(\zeta). \quad (16)$$

As previously discussed in this section, in addition to $\Lambda(\zeta)$, we take into account another important constraint when determining the final stopping threshold, aiming to further maximize the merging density of the resulting clusters. For this purpose, the merging degree factor $\gamma(\zeta)$ with respect to ζ is defined as follows:

$$\gamma(\zeta) = 1 - N/N_s \quad (17)$$

where N_s is the number of singleton clusters present during the initial stage. Note that at the beginning of HAC, the value of $\gamma(\zeta)$ is zero (i.e., the merging degree is the lowest), while $\gamma(\zeta)$ increases as the merging continues along with ζ . Note that $\gamma(\zeta)$ reaches its largest value when all the subject-identity features are merged into one cluster. We seek to fulfill an optimal balance between $\Lambda(\zeta)$ and $\gamma(\zeta)$ when determining the optimal stopping threshold ζ_{opt} . As such, ζ_{opt} is determined according to the following criterion:

$$\zeta_{\text{opt}} = \arg \max_{\zeta} (\Lambda(\zeta) + \gamma(\zeta)). \quad (18)$$

VI. FACE RECOGNITION BASED ON FACE INFORMATION FUSION

In this section, we propose a FR method based on the fusion of multiple face observations (instances) that all belong to a single identity. The primary characteristic of the proposed

¹Due to space limitations, the detailed proof for this observation can be found at http://ivylab.kaist.ac.kr/html/publication/paper/jy_tcvst_proof.pdf.

fusion methods is to account for the *confidence* (or belief) of the individual face features prior to their combination. The underlying idea is that stressing face images that lead to a higher discriminatory power may help eliminate noisy information to a certain degree via face information fusion, thus improving the FR performance. Hence, we expect that face images that have been the subject of large variations in appearance (e.g., in terms of illumination or pose) within a subject cluster are correctly annotated by virtue of a complementary effect.

Before explaining the proposed face information fusion, we first introduce a common notation. For the sake of conciseness, we denote a certain subject cluster by \mathbf{C} . Let $\text{FI}_q^{(m)}$ be the m th query face image (i.e., a face image to be annotated) in the set of all observations within \mathbf{C} and let $\text{FI}_t^{(n)}$ be the n th target image pre-enrolled in a target subject database. Also, let $\varphi(\cdot)$ be a face feature extractor [64] that returns a low-dimensional feature representation for a particular input face image. It is important to note that $\varphi(\cdot)$ is created with a training scheme based on GL [30]. In a typical GL-based FR system, the training process is performed using a generic database that consists of identities other than those to be recognized in testing operations. The use of GL allows avoiding an intensive manual labeling task, where manual labeling is required to create a large number of training face images. Please refer to [30] for more details regarding GL-based FR techniques. Finally, we define a function $l(\cdot)$ that returns an identity label for an input face image.

A. Face Recognition Using Weighted Feature Fusion

This section describes an FR method that makes use of a weighted combination, at feature level, of multiple face observations. We denote the low-dimensional feature vectors of $\text{FI}_q^{(m)}$ and $\text{FI}_t^{(n)}$ as $\mathbf{f}_q^{(m)}$ and $\mathbf{f}_t^{(n)}$, respectively. These feature vectors are formally defined as follows:

$$\mathbf{f}_q^{(m)} = \varphi(\text{FI}_q^{(m)}) \text{ and } \mathbf{f}_t^{(n)} = \varphi(\text{FI}_t^{(n)}) \quad (19)$$

where $1 \leq m \leq |\mathbf{C}|$, $1 \leq n \leq G$, $|\mathbf{C}|$ denotes the number of face images within the subject cluster \mathbf{C} , and G is the total number of subjects pre-enrolled in the target database.

Note that several defective face images (e.g., face images showing a strong variance in terms of illumination and viewpoint) may be part of a subject cluster. We regard such defective face images as outliers (see Fig. 6). The impact of outliers present in a subject cluster should be kept minimal. To this end, we associate a *weight* with each feature vector $\mathbf{f}_q^{(m)}$, representing the *distance* between the feature vector $\mathbf{f}_q^{(m)}$ and a corresponding prototype (i.e., a feature vector that is representative for the elements of the subject cluster). It is well-known that the median is more resilient to outliers than the mean [46]. Hence, we adopt a median feature vector, denoted by $\tilde{\mathbf{f}}_q$, as a prototype feature vector for each subject cluster. Note that each element of $\tilde{\mathbf{f}}_q$ is filled with the median value of all corresponding elements of $\mathbf{f}_q^{(m)}$ ($m = 1, \dots, |\mathbf{C}|$).

To diminish the influence of the elements of $\mathbf{f}_q^{(m)}$ that are far away from the prototype $\tilde{\mathbf{f}}_q$, the penalty-based Minkowski distance metric [44] is used. This distance metric is defined

as follows:

$$d_m = \left(\sum_k \left\{ \Phi(|\Gamma_k(\tilde{\mathbf{f}}_q) - \Gamma_k(\mathbf{f}_q^{(m)})|) \right\}^p \right)^{\frac{1}{p}} \quad (20)$$

where

$$\Phi(|x|) = \begin{cases} \tau \cdot |x| & \text{if } |x| > \delta \cdot \sigma_k \\ |x| & \text{otherwise} \end{cases} \quad (21)$$

and $\Gamma_k(\cdot)$ is a function that returns the k th element of the argument vector, σ_k stands for the standard deviation computed over the k th element samples of the feature vectors $\mathbf{f}_q^{(m)}$ that are part of \mathbf{C} , and δ and τ denote a user-specific threshold and a penalty constant, respectively. The parameters δ and τ are determined by means of a heuristic approach. Based on our experiments, 2.2 and 2.0 are found to be reasonable values for δ and τ , respectively. It should be emphasized that in (20), the distance d_m is forced to increase if the difference between each element in $\tilde{\mathbf{f}}_q$ and in $\mathbf{f}_q^{(m)}$ exceeds a certain $\delta \cdot \sigma_k$. The actual increase is controlled by the parameter τ [see (21)].

Using d_m provided by (20) and a soft-max function [46], we compute a weight that adjusts the influence of $\mathbf{f}_q^{(m)}$ on the fusion of face features

$$w_m = \frac{\exp(-d_m)}{\sum_{m=1}^{|\mathbf{C}|} \exp(-d_m)}. \quad (22)$$

Note that d_m should be normalized to have zero mean and unit standard deviation prior to the computation of w_m . In this paper, the widely used “z-score” technique is employed to normalize the distance scores. Other distance score normalization techniques are explained in detail in [32].

Using w_m , a single feature vector can be computed as a weighted average of the individual feature vectors $\mathbf{f}_q^{(m)}$ in \mathbf{C} as follows:

$$\mathbf{f}_q = \sum_{m=1}^{|\mathbf{C}|} w_m \cdot \mathbf{f}_q^{(m)}. \quad (23)$$

In (23), by assigning a higher weight to the reliable face features and a lower weight to the other face features (i.e., the outliers), the chance of assigning such outliers to the wrong subject class can be reduced.

The complementary effect of weighted feature fusion on the classification accuracy is visualized in Fig. 6. In Fig. 6, the first three most significant dimensions of the PCA feature subspace are plotted for each face image. It can be seen that two outliers ($\mathbf{f}_q^{(5)}$ and $\mathbf{f}_q^{(6)}$)—subject to a significant variation in pose and illumination—are located far from the feature vector $\mathbf{f}_t^{(1)}$ of the correct target subject, compared to the $\mathbf{f}_t^{(2)}$ of the incorrect target subject. As a result, the two outliers may be misclassified as the target identity $\mathbf{f}_t^{(2)}$ when performing FR in an independent way. However, the feature vector \mathbf{f}_q obtained using a weighted average of six individual feature vectors (including the two outliers) is much closer to $\mathbf{f}_t^{(1)}$ than $\mathbf{f}_t^{(2)}$. Consequently, the two outliers as well as the other query images are correctly identified through the use of a weighted face feature fusion.

To annotate $\text{FI}_q^{(m)}$, a nearest neighbor classifier is applied to determine the identity of $\text{FI}_q^{(m)}$, finding the smallest distance

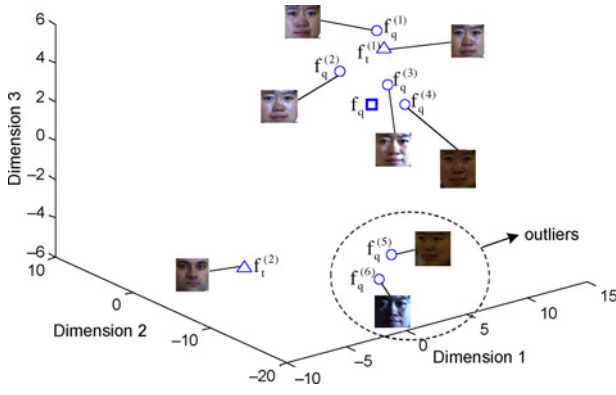


Fig. 6. Illustration of the complementary effect of weighted feature fusion on the classification accuracy. The six “circle” symbols represent the PCA feature vectors $f_q^{(m)}$ ($m = 1, \dots, 6$) of corresponding query images, which are all assumed to be part of a single subject cluster. The two “triangle” symbols represent the feature vectors $f_t^{(n)}$ ($n = 1, 2$) belonging to two different target subjects. In addition, a feature vector computed using a weighted average of the individual feature vectors $f_q^{(m)}$ is represented by a “square” symbol.

between f_q and $f_t^{(n)}$ ($n = 1, \dots, G$) in feature subspace as follows:

$$l(\text{FI}_q^{(m)}) = l(\text{FI}_t^{(n^*)}) \text{ and } n^* = \arg \min_{n=1}^G D_f(f_q, f_t^{(n)}) \quad (24)$$

where $D_f(\cdot)$ denotes a distance metric. Using (24), all $\text{FI}_q^{(m)}$ ($m = 1, \dots, |\mathbf{C}|$) contained in \mathbf{C} are annotated as subject identity $l(\text{FI}_t^{(n^*)})$ in a batch manner.

B. Face Recognition Using Confidence-Based Majority Voting

In confidence-based majority voting, the resulting identity (or subject) labels and corresponding confidence values are separately computed by matching each individual face feature $f_q^{(m)}$ against a set $\{f_t^{(n)}\}_{n=1}^G$ of G target face features. Let us denote the distance between $f_q^{(m)}$ and $f_t^{(n)}$ in the feature subspace as $d_{m,n}$. Note that $d_{m,n}$ can be computed using any distance metric (e.g., Euclidean distance). Based on $d_{m,n}$, we calculate the number of the votes for a particular identity label and an associated confidence value.

We now describe FR using confidence-based majority voting. Let $N_{\text{vote}}(n)$ be the total number of votes given to the n th target identity label, received from each individual $f_q^{(m)}$, that is

$$N_{\text{vote}}(n) = \sum_{m=1}^{|\mathbf{C}|} \delta(f_q^{(m)}, f_t^{(n)}) \quad (25)$$

where

$$\delta(f_q^{(m)}, f_t^{(n)}) = \begin{cases} 1 & \text{if } n = \arg \min_{k=1}^G d_{m,k} \\ 0 & \text{otherwise} \end{cases} \quad (26)$$

and $\delta(f_q^{(m)}, f_t^{(n)})$ is an indicator function that returns “one” when the minimum of the distance values computed between $f_q^{(m)}$ and $f_t^{(k)}$ ($k = 1, \dots, G$) is achieved at $k = n$, and “zero” otherwise.

In order to determine the confidence related to $N_{\text{vote}}(n)$, the $d_{m,n}$ are first normalized to have zero mean and unit standard

deviation in order to have the same scale. The normalized distances are then mapped onto values in the confidence domain using a sigmoid activation function [47]

$$c_{m,n} = \frac{1}{1 + \exp(d_{m,n})}. \quad (27)$$

Further, a sum normalization method [32] is used to obtain the normalized confidence of $c_{m,n}$, whose value ranges from 0 to 1 as follows:

$$c'_{m,n} = \frac{c_{m,n}}{\sum_{n=1}^G c_{m,n}}. \quad (28)$$

In (28), since $0 \leq c'_{m,n} \leq 1$ and $\sum_{n=1}^G c'_{m,n} = 1$, the confidence $c'_{m,n}$ can be regarded as the *a posteriori* probability that the identity label of $f_t^{(n)}$ is assigned to that of $f_q^{(m)}$, given $\text{FI}_q^{(m)}$. Using (26) and $c'_{m,n}$, the total confidence associated with $N_{\text{vote}}(n)$ is then determined as follows:

$$C_{\text{conf}}(n) = \sum_{m=1}^{|\mathbf{C}|} \delta(f_q^{(m)}, f_t^{(n)}) \cdot c'_{m,n}. \quad (29)$$

Note that in (29), $C_{\text{conf}}(n)$ is the sum of the confidence values of the n th target identity votes received from the individual $f_q^{(m)}$.

Finally, the target identity label that achieves the largest combined value of $N_{\text{vote}}(n)$ and $C_{\text{conf}}(n)$ is selected as the identity of $\text{FI}_q^{(m)}$ ($m = 1, \dots, |\mathbf{C}|$). This is done as follows:

$$l(\text{FI}_q^{(m)}) = l(\text{FI}_t^{(n^*)}) \text{ and } n^* = \arg \max_{n=1}^G (N_{\text{vote}}(n) \cdot C_{\text{conf}}(n)). \quad (30)$$

VII. EXPERIMENTS

In this section, we present a performance evaluation of the proposed face annotation method. To evaluate our annotation method, six different photo collections (see Table III for details) were created. Of all six photo collections, one photo set consisted of photos gathered from the MPEG-7 VCE-3 data set [36], while the remaining five photo sets were created using photos retrieved from popular photo sharing Web sites such as Flickr [1] and Picasa [37]. The MPEG-7 VCE-3 data set provides a total of 1385 personal photos, captured by a number of people participating in the MPEG-7 standardization effort. The remaining five data sets consist of photos posted on the weblogs of 18 different users. These users are members of Flickr or Picasso. The collected photos include real-life scenes such as a wedding, a trip, a birthday party, and so on. Note that the accurate capture times for most of the photographs used were extracted from the EXIF header stored in each image file. On the other hand, for photos with a missing capture time, the corresponding upload time was used as a substitute for the capture time.

To form a ground truth for each photo collection, the *Viola-Jones* face detection algorithm [42] was first applied to all photos used. The identities of all detected faces and corresponding clothing images were then manually labeled. Fig. 7 displays a number of face images used in our experiment. As shown

TABLE III
DETAILED DESCRIPTION OF THE PHOTO COLLECTIONS AND THE CORRESPONDING GROUND TRUTH DATA SETS USED IN OUR EXPERIMENTS

Name of the photo collection	P1	P2	P3	P4	P5	P6
Number of photos	1385	5441	3107	2215	4483	5679
Number of target subjects to be annotated	58	104	88	61	74	140
Number of photos containing target subjects	1120	4154	2652	1732	3241	4876
Number of detected face images belonging to target subjects	1345	4872	3012	2934	3652	5276
Average number of photos per target subject	23	47	32	41	45	40
Time span	1 years	4 years	3 years	2 years	3 years	5 years

Note that the “P1” photo collection is composed of photos taken from MPEG-7 VCE-3 data set, while the photo collections P2–P6 contain photos collected from the web.



Fig. 7. Example face images used in our experiment. Each row contains face images that belong to the same subject.

In Fig. 7, recognizing the face images used is significantly challenging due to severe illumination and pose variations, the use of heavy make-up, and the presence of occlusions. In order to focus on the FR accuracy, we have excluded face detection errors from our performance evaluation. It should be noted that users usually prefer to annotate known individuals, such as friends and family members [21]. In this context, we carried out manual labeling for individuals who appear at least ten times in the picture collections used, while ignoring individuals that appear less than ten times [22], [27].

Table III provides detailed information about the constructed ground truth data sets. It should be noted that, in our experiments, face recognition was performed using 529 target subjects (i.e., subjects with a known identity), distributed over six different target sets (one target set for each photo collection used). In particular, as shown in Table III, the number of target subjects used for the purpose of FR (i.e., the number of subjects with a known identity) is 58, 108, 88, 61, 74, and 140 for the P1, P2, P3, P4, P5, and P6 photo collections, respectively. Moreover, the target sets and the query sets are disjoint.

A. Evaluation of Clustering Performance

In this experiment, we assess the performance of the proposed situation and subject clustering methods. Recall that the final goal of situation and subject clustering is to group face images belonging to the same subject as correctly as possible. As such, this experiment focuses on assessing the accuracy of grouping face images by using both situation and subject clustering, rather than investigating the accuracy of situation detection alone. Local binary pattern (LBP) face descriptor [56] was adopted to represent face information,

while the MPEG-7 CS and EH descriptors were used to represent clothing information (i.e., color and texture). Further, we combined the MPEG-7 CS Descriptor with the MPEG “illumination invariant color descriptor” (IICD) in order to obtain a characterization of color features that is more robust to variations in illumination. In addition, to achieve an optimal fusion of face and clothing features, the weighting values defined in (8) were determined by means of an exhaustive tuning process. As such, the following weighting values were used: $w_f = 0.59$ (face), $w_c^{(1)} = 0.3$ (color), and $w_c^{(2)} = 0.11$ (texture).

In general, the following two issues need to be considered during the evaluation of the clustering performance: 1) each cluster should contain face images that belong to the same subject (to the extent possible), and 2) to facilitate the complementary effect that originates from the use of multiple face observations, face images belonging to the same subject, as many as possible, have to be merged in a single cluster.

In order to consider the aforementioned issues during the evaluation of the clustering performance, the *F*Score metric [55] is adopted to quantify the clustering performance. Suppose that a particular situation cluster contains R different subjects and N_s subject-identity features. Then let \mathbf{L}_r be the set of N_r subject-identity features all belonging to the same identity (i.e., the r th subject), where $1 \leq r \leq R$ and $N_s = \sum_{r=1}^R N_r$. It should be noted that \mathbf{L}_r can be obtained using the ground truth data sets described in Table III. Also, let us assume that a total of K subject clusters (\mathbf{C}_i , $i = 1, \dots, K$) are generated for the situation cluster under consideration and that N_i subject-identity features are grouped in each \mathbf{C}_i . Given that $N_i^{(r)}$ elements in \mathbf{C}_i (where $N_i = \sum_{r=1}^R N_i^{(r)}$) belong to \mathbf{L}_r , the “ F value” of \mathbf{L}_r and \mathbf{C}_i is then defined as

$$F(\mathbf{L}_r, \mathbf{C}_i) = \frac{2 \cdot R(\mathbf{L}_r, \mathbf{C}_i) \cdot P(\mathbf{L}_r, \mathbf{C}_i)}{R(\mathbf{L}_r, \mathbf{C}_i) + P(\mathbf{L}_r, \mathbf{C}_i)} \quad (31)$$

where $R(\mathbf{L}_r, \mathbf{C}_i) = N_i^{(r)}/N_r$ and $P(\mathbf{L}_r, \mathbf{C}_i) = N_i/N_i^{(r)}$ denote the clustering recall and precision for \mathbf{L}_r and \mathbf{C}_i , respectively. It should be noted that $R(\mathbf{L}_r, \mathbf{C}_i)$ represents the clustering performance related to the between-cluster error rate, while $P(\mathbf{L}_r, \mathbf{C}_i)$ reflects the within-cluster error rate. Based on (31), the *F*Score for the entire subject cluster [55] is defined as

$$F\text{Score} = \sum_{r=1}^R \frac{N_r}{N_s} F\text{Score}(\mathbf{L}_r) \quad (32)$$

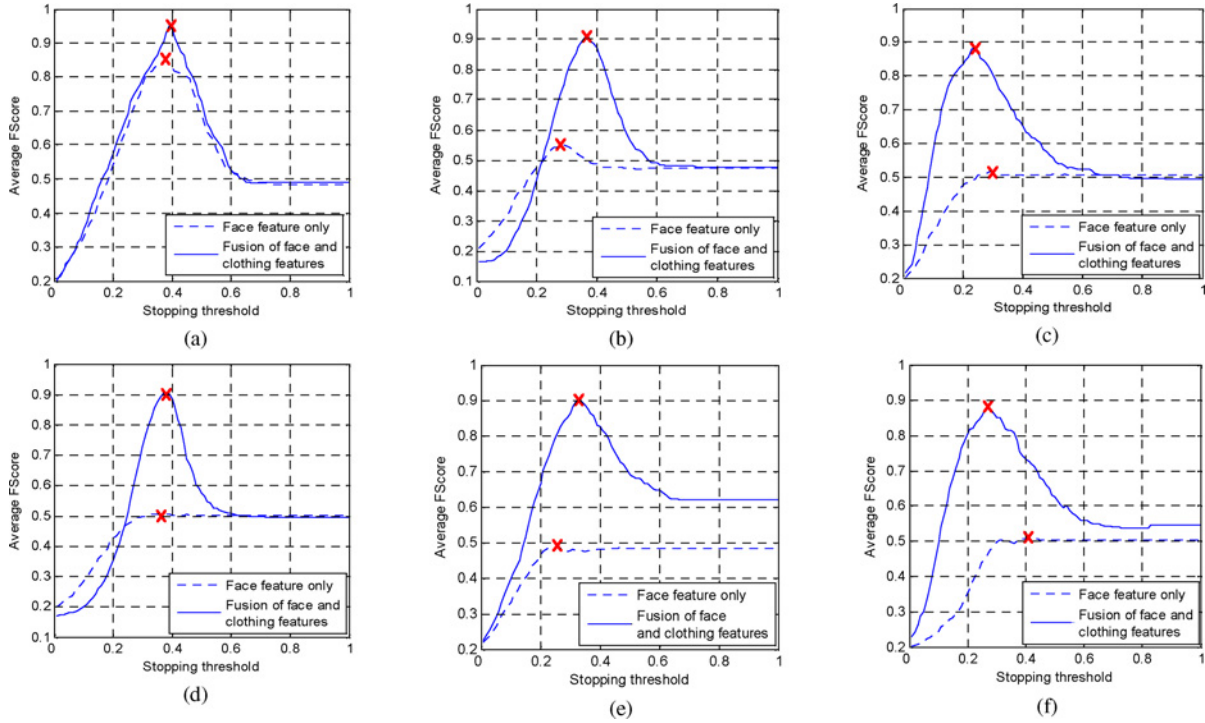


Fig. 8. *FScore* values averaged over all of the situation clusters generated for each photo collection. (a) “P1” photo set. (b) “P2” photo set. (c) “P3” photo set. (d) “P4” photo set. (e) “P5” photo set. (f) “P6” photo set. In each *FScore* plot, the corresponding “cross mark” represents the maximum *FScore*, as obtained for an optimal value of the stopping threshold (determined using the corresponding ground truth data set).

TABLE IV

AVERAGE SQUARED ERRORS AND CORRESPONDING STANDARD DEVIATIONS COMPUTED FOR OPTIMAL THRESHOLDS OBTAINED USING GROUND TRUTH INFORMATION AND THRESHOLDS OBTAINED USING THE PROPOSED METHOD

Name of the photo collection	P1	P2	P3	P4	P5	P6
Average squared error	0.0079	0.0047	0.0059	0.0061	0.0018	0.0072
Standard deviation	0.0017	0.0089	0.0065	0.011	0.0041	0.0086

Note that the range of the average squared error is between zero and one as the value of the thresholds ranges from zero to one.

TABLE V

PRECISION AND RECALL FOR THREE DIFFERENT FACE ANNOTATION METHODS WITH RESPECT TO SIX DIFFERENT PHOTO COLLECTIONS

Feature Extraction Algorithm	Photo Collection	Baseline		Clustering + Weighted Feature Fusion		Clustering + Confidence-Based Majority Voting	
		Precision	Recall	Precision	Recall	Precision	Recall
Bayesian	P1	0.67	0.7	0.95	0.95	0.93	0.92
	P2	0.48	0.51	0.73	0.71	0.68	0.68
	P3	0.41	0.44	0.78	0.79	0.79	0.77
	P4	0.52	0.57	0.83	0.85	0.82	0.84
	P5	0.46	0.49	0.82	0.84	0.79	0.81
	P6	0.42	0.47	0.68	0.67	0.65	0.64
RLDA	P1	0.70	0.72	0.92	0.88	0.88	0.90
	P2	0.58	0.59	0.74	0.72	0.69	0.69
	P3	0.54	0.58	0.74	0.76	0.71	0.72
	P4	0.62	0.64	0.77	0.77	0.72	0.73
	P5	0.57	0.58	0.70	0.71	0.68	0.68
	P6	0.51	0.54	0.69	0.68	0.66	0.65

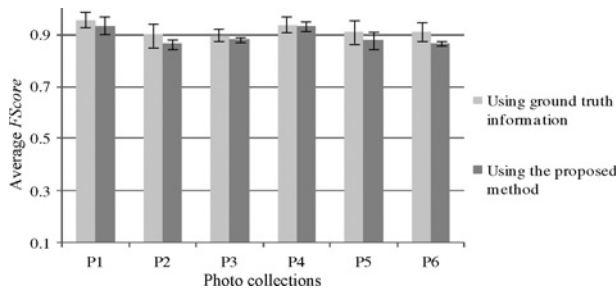


Fig. 9. Comparison of average $FScores$ and corresponding standard deviations, once for thresholds computed using ground truth information and once for thresholds computed using the proposed method. Note that $FScore$ values are averaged over all situation clusters generated for each photo set.

where

$$FScore(\mathbf{L}_r) = \max_{i=1}^K F(\mathbf{L}_r, \mathbf{C}_i). \quad (33)$$

In (33), the $FScore$ of \mathbf{L}_r , denoted as $FScore(\mathbf{L}_r)$, is the maximum of $F(\mathbf{L}_r, \mathbf{C}_i)$ attained at any subject cluster in the hierarchical tree. Note that the $FScore$ as defined in (32) will be one when every subject has a corresponding subject cluster that contains all face images belonging to that subject. Hence, the higher the $FScore$ value, the better the clustering result in the sense of natural grouping.

The resulting $FScores$ for the situation and subject clustering processes are shown in Fig. 8 with respect to the six different photo collections used. It is important to note that all $FScore$ curves shown in Fig. 8 are values averaged over all situation clusters produced for each photo collection. Looking into the results in Fig. 8, except for the “P1” photo set, the $FScores$ are relatively low (less than 0.55) when only making use of face information. However, when using a fusion of face and clothing information [as defined by (8)], the peak $FScore$ significantly increases for most photo sets. In particular, a peak $FScore$ of up to 0.9 is achieved for all of the photo sets. Based on the fact that the $FScore$ becomes one when perfect clustering results are achieved, we can demonstrate that the proposed clustering methods are able to attain a reliable clustering performance (for an appropriate stopping threshold).

B. Evaluation of Stopping Threshold Selection

As shown in Fig. 8, the $FScore$ curves vary along with the pre-determined stopping threshold. As such, selecting an optimal stopping threshold, at which a maximum $FScore$ is achieved, is of critical importance in order to achieve a feasible clustering performance. In this section, we evaluate the effectiveness of the proposed stopping threshold selection method described in Section V-C. Note that the following experimental results are obtained using a fusion of face and clothing features.

For each situation cluster, we compute the squared error between the optimal threshold values obtained using the ground truth and the threshold values determined using the proposed method. Table IV tabulates the squared errors averaged over all situation clusters created for each photo set. Also, the corresponding standard deviation is presented in order to demonstrate the stability of the results reported. Note that the

average squared errors range from zero to one. As can be seen in Table IV, the average squared errors for all photo sets are significantly small and close to zero. This indicates that the proposed method works well for estimating an optimal stopping threshold.

Further, Fig. 9 allows comparing the average $FScores$ and corresponding standard deviations for optimal threshold values, once computed using ground truth information and once computed using the proposed method. As expected, the $FScores$ are high (close to one) and nearly the same for all photo sets.

C. Evaluation of Face Annotation Performance

We tested the annotation performance of the proposed method over real-world personal photo collections. To construct a face feature extractor $\varphi(\cdot)$ (as defined in Section VI), principal component analysis (PCA) [48], Bayesian [49], fisher linear discriminant analysis (FLDA) [50], and regularized linear discriminant analysis (RLDA) [50] were adopted as feature extraction methods. PCA and FLDA are commonly used as a benchmark for evaluating the performance of FR algorithms [38]. The Bayesian approach shows the best overall performance in the FERET test [52], while RLDA is also a popular linear discriminant analysis FR technique. Note that grayscale face images were employed by the feature extraction algorithms considered (using the R channel of the red-green-blue (RGB) color space [12]). To measure similarity, the Euclidean distance was used for FLDA and RLDA, while the Mahalanobis distance and maximum *a posteriori* probability (MAP) were used for PCA and Bayesian, respectively [51].

As stated in Section VI, to train feature extractors using a GL-based scheme, we constructed a reasonably sized generic training set, consisting of a total of 6594 facial images of 726 subjects collected from three public face databases: CMU PIE [53], Color FERET [52], and AR [54]. During the collection phase, 1428 face images of 68 subjects (21 samples/subject) were selected from CMU PIE. As for Color FERET, 4480 face images of 560 subjects (eight samples/subject) were chosen from the “fa,” “fb,” “fc,” and “dup1” sets. As for AR DB, we selected face images with different expressions: neutral, smile, anger, and scream. As a result, 686 frontal-view images belonging to 98 subjects were chosen from two different sessions (as described in [54]).

In a typical face annotation system, performance results can be reported for the following two tasks.

- 1) *Subject identification (or classification)*: given a query face, the task of subject identification is to suggest a list of candidate target names.
- 2) *Subject-based photo retrieval*: when a user enters the name of a subject as a search term, the task is to retrieve a set of personal photos containing the subject corresponding to the given name.

In our experiments, the H -Hit rate, proposed in [26], was adopted to measure the accuracy of subject identification, while precision and recall were used to measure the performance of subject-based photo retrieval. When measuring the H -Hit rate, if the actual name of a given query face is in the

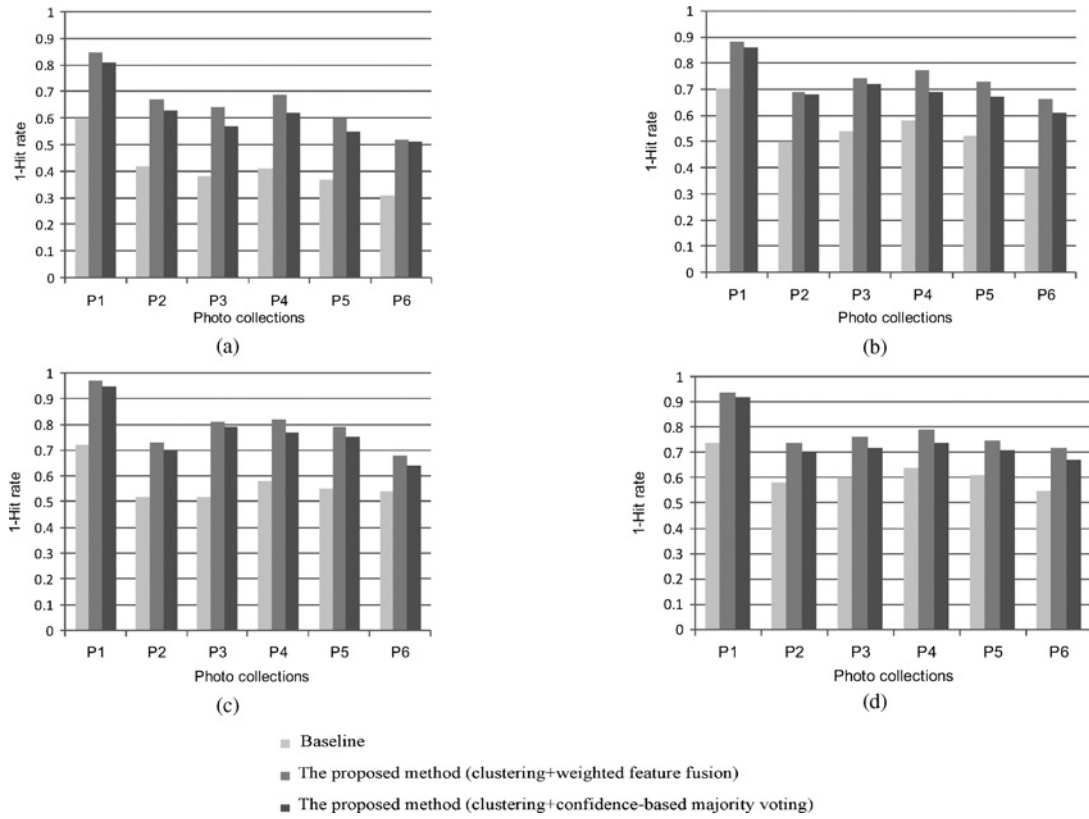


Fig. 10. Comparison of the 1-Hit rates obtained for three different annotation methods and four feature extraction methods. (a) PCA. (b) FLDA. (c) Bayesian. (d) RLDA. The subject clusters, used by the proposed FR methods, were generated using stopping threshold values determined by the method proposed in this paper.

list containing H names, then this query face is said to be hit by the name list. In addition, the precision and recall used in our experiment are defined as follows:

$$\text{precision} = \frac{1}{G} \sum_{n=1}^G \frac{N_{\text{correct}}^{(n)}}{N_{\text{retrieval}}^{(n)}} \text{ and } \text{recall} = \frac{1}{G} \sum_{n=1}^G \frac{N_{\text{correct}}^{(n)}}{N_{\text{ground}}^{(n)}} \quad (34)$$

where G is the total number of target subjects, $N_{\text{retrieval}}^{(n)}$ is the number of retrieved photos annotated with identity label n , $N_{\text{correct}}^{(n)}$ is the number of photos correctly annotated with identity label n , and $N_{\text{ground}}^{(n)}$ is the number of photos annotated with identity label n in the ground truth.

The performance of conventional appearance-based FR solutions [30] (only using a single face image) is referred to as baseline face annotation accuracy. Baseline FR also utilizes a training set that consists of training images corresponding to the target subjects. It should be noted that, when referring to the literature in the area of FR [10], [51], [59], the use of eight training images is usually sufficient to prevent a significant decrease in the FR performance caused by a shortage of training images. For this reason, the training set for baseline FR contained eight face images per target subject in our experiments. This guarantees fair and stable comparisons with the proposed FR technique that relies on a GL-based training scheme and face information fusion. That way, we are able to demonstrate that our face annotation method can achieve acceptable annotation accuracy, while not requiring training

face images for each target subject (which is in contrast to baseline FR).

Fig. 10 compares the 1-Hit rates of the proposed methods (“clustering + weighted feature fusion” and “clustering + confidence-based majority voting”) with the 1-Hit rates obtained for baseline FR. For FR using weighted feature fusion, the “ P ” value shown in (20) is set to 2. In Fig. 10, we can observe that the annotation task for personal photos collected from the Web is significantly challenging. Specifically, the 1-Hit rates obtained for baseline FR on the five Web photo collections (P2–P6) are noticeably low (less than 62 percent) for all feature extraction methods used. However, we can see that a substantial improvement in annotation performance can be achieved by the proposed face annotation methods, thanks to the use of face information fusion. In particular, in the case of weighted feature fusion, the 1-Hit rate, averaged over six photo sets, can be improved with 24.66%, 20.05%, 14.83%, and 22.83% for PCA, FLDA, RLDA, and Bayesian, respectively. It is also worth noting that the results of the weighted feature fusion method are better than those obtained for confidence-based majority voting. This result is consistent with previous reports [31], [32] that feature-level information fusion achieves better classification results than fusion methods working on other levels.

Table V shows the precision and recall annotation performance for three different face annotation methods, applied to six different photo sets. We only present the precision

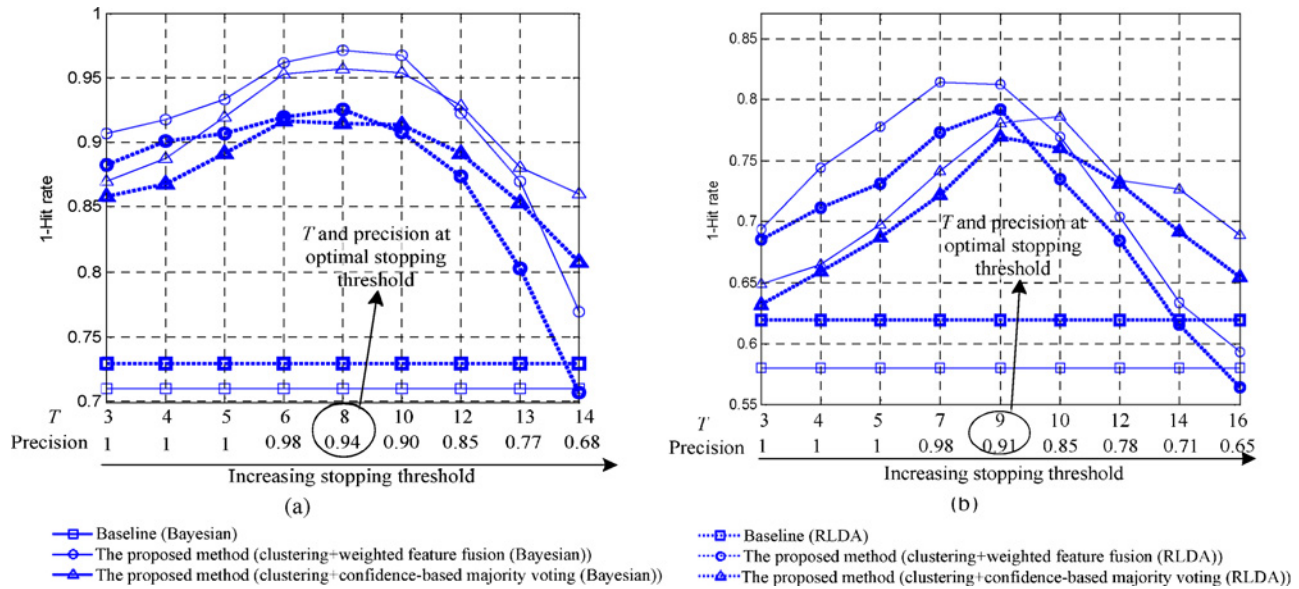


Fig. 11. Plot of the change in annotation accuracy according to the number of face images merged in a cluster and corresponding precision values. (a) “P1” photo set. (b) “P4” photo set. Note that “ T ” denotes the average number of face images in a cluster, where “ T ” is a function of the stopping threshold value. Also, the corresponding precision is presented below each “ T ” value, as computed by (31).

and recall for the Bayesian and RLDA approach, since these techniques achieve better 1-Hit rates than PCA and FLDA (as shown in Fig. 11). The experimental results in Table V confirm that the proposed annotation method is more effective. As shown in Table V, compared to baseline FR, the proposed annotation method can significantly improve the precision and recall for all feature extraction algorithms and photo sets used.

D. Effect of Clustering Performance on Face Annotation

As demonstrated in Sections VII-A and VII-B, the proposed clustering methods have proven to be effective: face images belonging to the same subject can be grouped together with a small clustering error rate. In practical applications, however, the clustering performance might be lower than its attainable optimum, dependent upon the face and clothing features chosen and the cluster parameters adopted (e.g., the dissimilarity metric used). In this sense, it is worth evaluating the robustness (or tolerance) of the proposed face annotation method against variations in clustering performance. Such an evaluation is important in order to ensure that our method can be readily extended to real-world applications. This motivated us to investigate how the face annotation accuracy is affected by two parameters related to clustering performance: the number of face images merged in a cluster and the precision given by (32). Note that the precision is adversely proportional to the within-cluster error rate (i.e., when the precision increases, the within-cluster error rate decreases).

Fig. 11 shows the variation in face annotation accuracy with respect to the number of face images merged in a cluster (denoted by T) and the corresponding precision values. We first observe that when the stopping threshold increases, T increases, whereas the precision decreases (i.e., the within-cluster error rate increases). Note that, for the “P1” photo set, for the optimal stopping threshold (i.e., computed by

making use of the ground truth), T is equal to 8 and the precision is equal to 0.94, while for the “P4” photo set, T is equal to 9 and the precision is equal to 0.91. As shown in Fig. 11, as T decreases, the annotation accuracy becomes worse than the annotation accuracy achieved for the optimal stopping threshold. However, we can observe that the annotation accuracy for weighted feature fusion is much better than the annotation accuracy for baseline FR, even when $T = 3$ in both photo sets (note that baseline FR only makes use of a single image). This indicates that weighted feature fusion is advantageous for the case where T is forced to be relatively small in an attempt to guarantee a high precision.

Looking into the robustness against precision, it can be observed that the annotation accuracy of weighted feature fusion is significantly influenced by the precision. In particular, the annotation accuracy drops rapidly at precision values less than 0.77 and 0.78 for the “P1” and “P4” photo sets, respectively. This can be attributed to the fact that false face images (i.e., face images whose identities differ from the identity comprising the majority of face images in a single subject cluster) may directly influence the fusion process at the level of features, although their effect might not be significant due to the assignment of small weighting values. On the other hand, confidence-based majority voting results in a slower decay in annotation accuracy compared to weighted feature fusion. In particular, confidence-based majority voting outperforms baseline FR by a significant margin, even when the precision is equal to 0.68 (“P1” photo set) and 0.71 (“P4” photo set).

E. Runtime Performance

We have measured the time required to annotate more than 5500 photos on an Intel Pentium IV 2.4 GHz CPU processor. The time needed to execute weighted feature fusion

in conjunction with clustering is about 345 s (this is about 60 ms per photo), while the time needed to execute confidence-based majority voting in conjunction with clustering is about 413 s (this is about 75 ms per photo). Note that the processing time needed for selecting an optimal stopping threshold value during subject clustering is included in the execution times measured. On the other hand, the preprocessing time required to create a feature extractor using a GL-based training scheme is not considered in the measurement of the execution times as this process can be executed off-line.

Regarding the overall computational complexity of the proposed face annotation framework, the most expensive step is found to be HAC-based subject clustering as the complexity of this stage is, in general, $O(n^3)$ [55], where n is the number of data points considered when performing HAC. The computational complexity is primarily due to the following two reasons: 1) the computation of pairwise similarity between all n data points, and 2) the repeated selection of a pair of clusters that is most similar. While HAC-based subject clustering is currently the main run-time performance bottleneck in our face annotation framework, it is interesting to know that efficient implementations exist of HAC-based clustering [45]. These implementations are able to considerably reduce both the time and memory complexity of HAC-based clustering [e.g., the time complexity can be readily reduced from $O(n^3)$ to $O(n^2)$].

VIII. CONCLUSION

In this paper, we proposed a new face annotation method that is particularly useful for large personal photo collections (usually consisting of thousands of photos). The proposed face annotation method systematically leverages contextual cues with current FR techniques in order to improve face annotation accuracy. We demonstrated that face images belonging to the same subject can be reliably merged in a cluster using the proposed situation and subject clustering techniques. In addition, to take advantage of the availability of multiple face images belonging to the same subject, we proposed a novel FR method using face information fusion. Further, to eliminate the need for training images, a training scheme based on generic learning was incorporated into the proposed FR method.

Our experimental results show that our face annotation method significantly outperforms conventional methods in terms of face annotation accuracy. In addition, our face annotation method is simple to implement, compared to already existing face annotation methods that utilize contextual information. Consequently, we believe that our face annotation method can be readily and effectively applied to real-world collections of personal photos, with a low implementation cost and feasible face annotation accuracy.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions. They would also like to thank the FERET Technical Agent of the U.S. National Institute of Standards and Technology for providing the FERET database.

REFERENCES

- [1] Flickr [Online]. Available: <http://www.flickr.com>
- [2] Facebook [Online]. Available: <http://www.facebook.com>
- [3] K. Rodden and K. R. Wood, "How do people manage their digital photographs?" in *Proc. ACM Hum. Factors Comput. Syst.*, 2003, pp. 409–416.
- [4] M. Ames and M. Naaman, "Why we tag: Motivations for annotation in mobile and online media," in *Proc. ACM Int. Conf. CHI*, 2007, pp. 971–980.
- [5] A. Matellanes, A. Evans, and B. Erdal, "Creating an application for automatic annotation of images and videos," in *Proc. Int. Conf. Florida Artif. Intell. Res. Soc.*, 2007, pp. 1–10.
- [6] M. S. Kankanhalli and Y. Rui, "Application potential of multimedia information retrieval," *Proc. IEEE*, vol. 96, no. 4, pp. 712–720, Apr. 2008.
- [7] R. Jain, "Multimedia information retrieval: Watershed events," in *Proc. ACM Int. Conf. MIR*, 2008, pp. 229–236.
- [8] M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 24, no. 1, pp. 34–58, Jan. 2002.
- [9] S. J. D. Prince, J. H. Elder, J. Warrell, and F. M. Felisberti, "Tied factor analysis for face recognition across large pose differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 970–984, Jun. 2008.
- [10] J. Zhu, S. C. H. Hoi, and M. R. Lyu, "Face annotation using transductive kernel Fisher discriminant," *IEEE Trans. Multimedia*, vol. 10, no. 1, pp. 86–96, Jan. 2008.
- [11] J. Y. Choi, Y. M. Ro, and K. N. Plataniotis, "Color face recognition for degraded face images," *IEEE Trans. Syst. Man Cybern.-Part B*, vol. 39, no. 5, pp. 1217–1230, Oct. 2009.
- [12] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 23, no. 1, pp. 1349–1380, Jan. 2001.
- [13] L. Cao, J. Luo, and T. S. Huang, "Annotating photo collections by label propagation according to multiple similarity cues," in *Proc. ACM Int. Conf. Multimedia*, 2008, pp. 121–130.
- [14] J. Yuan, J. Luo, H. Kautz, and Y. Wu, "Mining GPS traces and visual words for event classification," in *Proc. ACM Int. Conf. Multimedia Inform. Retrieval*, 2008, pp. 2–9.
- [15] Y. Tian, W. Liu, R. Xiao, F. Wen, and X. Tang, "A face annotation framework with partial clustering and interactive labeling," in *Proc. IEEE Int. Conf. CVPR*, 2007, pp. 1–8.
- [16] L. Zhang, Y. Hu, M. Li, W. Ma, and H. J. Zhang, "Efficient propagation for face annotation in family albums," in *Proc. ACM Int. Conf. Multimedia*, 2004, pp. 716–723.
- [17] B. Suh and B. B. Bederson, "Semi-automatic photo annotation strategies using event based clustering and clothing based subject recognition," *Int. J. Interacting Comput.*, vol. 19, no. 2, pp. 524–544, 2007.
- [18] B. Suh and B. B. Bederson, "Semi-automatic image annotation using event and torso identification," Comput. Sci. Dept., Univ. Maryland, College Park, Tech. Rep. HCIL-2004-15, 2004.
- [19] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang, "EasyAlbum: An interactive photo annotation system based on face clustering and re-ranking," in *Proc. Int. Conf. ACM CHI*, Jan Jose, CA, 2007, pp. 367–376.
- [20] Z. Stone, T. Zickler, and T. Darrell, "Autotagging Facebook: Social network context improves photo annotation," in *Proc. IEEE Int. Conf. CVPRW*, Jun. 2008, pp. 1–8.
- [21] N. O'Hare and A. F. Smeaton, "Context-aware person identification in personal photo collections," *IEEE Trans. Multimedia*, vol. 11, no. 2, pp. 220–228, Feb. 2009.
- [22] M. Naaman, R. B. Yeh, H. G. Molina, and A. Paepcke, "Leveraging context to resolve identity in photo albums," in *Proc. ACM Int. Conf. JCDL*, 2005, pp. 178–187.
- [23] A. Girgensohn, J. Adcock, and L. Wilcox, "Leveraging face recognition technology to find and organize photos," in *Proc. ACM Conf. Multimedia Inform. Retrieval*, 2004, pp. 99–106.
- [24] M. Zhao, Y. W. Teo, S. Liu, T. S. Chua, and R. Jain, "Automatic person annotation of family photo album," in *Proc. LNCS Int. Conf. CIVR*, 2006, pp. 163–172.
- [25] L. Chen, B. Hu, L. Zhang, M. Li, and H. J. Zhang, "Face annotation for family photo album management," *Int. J. Image Graph.*, vol. 3, no. 1, pp. 1–14, 2003.
- [26] D. Anguelov, L. Kuang-Chih, S. B. Gokturk, and B. Sumengen, "Contextual identity recognition in personal photo albums," in *Proc. IEEE Int. Conf. CVPR*, Jun. 2007, pp. 1–7.

- [27] S. Cooray, N. E. O'Connor, C. Gurrin, G. J. Jones, N. O'Hare, and A. F. Smeaton, "Identifying subject re-occurrences for personal photo management applications," in *Proc. IEEE Int. Conf. VIE*, 2006, pp. 144–149.
- [28] L. Zhang, L. Chen, and M. L. H. Zhang, "Automated annotation of human faces in family albums," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 355–358.
- [29] J. Wang, K. N. Plataniotis, J. Lu, and A. N. Venetsanopoulos, "On solving the face recognition problem with one training sample per subject," *Pattern Recognit.*, vol. 39, no. 6, pp. 1746–1762, 2006.
- [30] A. K. Jain, A. Ross, and S. Prabhaker, "An introduction to biometric recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 4–20, Jan. 2004.
- [31] A. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, 2005.
- [32] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
- [33] R. Brunelli and D. Falavigna, "Subject identification using multiple cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 10, pp. 955–966, Oct. 1995.
- [34] P. K. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhall, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, no. 3, pp. 1432–1882, 2010.
- [35] *Description of MPEG-7 Visual Core Experiments*, document N6905.doc, MPEG-7 Visual Group ISO/IEC JTC1/SC29/WG11, 2005.
- [36] Picasa [Online]. Available: <http://picasaweb.google.com>
- [37] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, no. 4, pp. 399–458, 2003.
- [38] S. Yang, S. K. Kim, and Y. M. Ro, "Semantic home photo categorization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 324–335, Mar. 2007.
- [39] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 364–373.
- [40] A. C. Loui and A. Savakis, "Automated event clustering and quality screening of consumer pictures for digital albuming," *IEEE Trans. Multimedia*, vol. 5, no. 3, pp. 390–401, Sep. 2003.
- [41] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Int. Conf. CIVR*, Dec. 2001, pp. 511–518.
- [42] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
- [43] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 2000.
- [44] X. Li, "Parallel algorithms for hierarchical clustering and cluster validity," *IEEE Trans. Pattern. Anal. Mach. Intell.*, vol. 12, no. 11, pp. 1088–1092, Nov. 1990.
- [45] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin, Germany: Springer, 2006.
- [46] C. L. Liu, "Classifier combination based on confidence transformation," *Pattern Recognit.*, vol. 38, no. 11, pp. 11–28, 2005.
- [47] M. A. Turk and A. P. Pentland, "Eigenfaces for recognition," *J. Cognitive Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.
- [48] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognit.*, vol. 33, no. 11, pp. 1771–1782, 2000.
- [49] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 7, pp. 711–720, Jul. 1997.
- [50] J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularized discriminant analysis for the small sample size problem in face recognition," *Pattern Recognit. Lett.*, vol. 24, no. 16, pp. 3079–3087, 2003.
- [51] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.
- [52] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1615–1618, Dec. 2003.
- [53] A. M. Martinez and R. Benavente, "The AR face database," Univ. Purdue, West Lafayette, IN, CVC Tech. Rep. 24, Jun. 1998.
- [54] Y. Zhao and G. Karypis, "Hierarchical clustering algorithms for document datasets," *Data Mining Knowl. Discovery*, vol. 10, no. 2, pp. 141–168, Mar. 2005.
- [55] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary pattern: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.
- [56] B. S. Manjunath, P. Salembier, and T. Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*. New York: Wiley, 2002.
- [57] E. Chang, K. Goh, G. Sychay, and G. Wu, "CBSA: Content-based soft annotation for multimodal image retrieval using Bayes point machines," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 1, pp. 26–38, Jan. 2003.
- [58] Y. Su, S. Shan, X. Chen, and W. Gao, "Hierarchical ensemble of global and local classifiers for face recognition," *IEEE Trans. Image Process.*, vol. 18, no. 8, pp. 1885–1886, Aug. 2009.
- [59] B. S. Manjunath, J. R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 6, pp. 703–715, Jun. 2001.
- [60] V. Perlibakas, "Distance measures for PCA-based face recognition," *Pattern Recognit. Lett.*, vol. 25, no. 12, pp. 1421–1430, 2004.
- [61] Y. Pang, Y. Yuan, and X. Li, "Gabor-based region covariance matrices for face recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 7, pp. 989–993, Jul. 2008.
- [62] G. J. Qi, X. S. Hua, Y. Rui, J. Tang, T. Mei, M. Wang, and H. J. Zhang, "Correlative multimedia video annotation with temporal kernels," *ACM TOMCCAP*, vol. 5, no. 1, article 3, Oct. 2008.
- [63] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of Gabor phase patterns (HGPP): A novel object representation approach for face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 1, pp. 57–68, Jan. 2007.



Jae Young Choi (S'08) received the B.S. degree from Kwangwoon University, Seoul, South Korea, in 2004, and the M.S. degree from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2008. He is currently pursuing the Ph.D. degree from the Image and Video Systems Laboratory, Department of Electrical Engineering, KAIST.

In 2007, he was an Intern Researcher with the Electronics and Telecommunications Research Institute, Daejeon. In 2008, he was a Visiting Student Researcher with the University of Toronto, Toronto, ON, Canada. His current research interests include face recognition/detection, image/video indexing, pattern recognition, machine learning, the social web, and personalized broadcasting technologies.



Wesley De Neve received the M.S. degree in computer science and the Ph.D. degree in computer science engineering, both from Ghent University, Ghent, Belgium, in 2002 and 2007, respectively.

He is currently working as a Senior Researcher with the Image and Video Systems Laboratory (IVY Lab), in the position of Assistant Research Professor. IVY Lab is part of the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea. Prior to joining KAIST, he was a Post-Doctoral Researcher with both Ghent University-IBBT, Ghent, and Information and Communications University, Daejeon. His current research interests and areas of publication include the coding, annotation, and adaptation of image and video content, GPU-based video processing, efficient XML processing, and the semantic and the social web.



Yong Man Ro (M'92–SM'98) received the B.S. degree from Yonsei University, Seoul, South Korea, and the M.S. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea.

In 1987, he was a Visiting Researcher with Columbia University, New York. From 1992 to 1995, he was a Visiting Researcher with the University of California, Irvine, and with KAIST. He was a Research Fellow with the University of California, Berkeley, and a Visiting Professor with the University of Toronto, Toronto, ON, Canada, in 1996 and 2007, respectively. He is currently a Full Professor with KAIST, where he is directing the Image and Video Systems Laboratory, Department of Electrical Engineering. He participated in the MPEG-7 and MPEG-21 international standardization efforts, contributing to the definition of the MPEG-7 texture descriptor, the

MPEG-21 DIA visual impairment descriptors, and modality conversion. His current research interests include image/video processing, multimedia adaptation, visual data mining, image/video indexing, and multimedia security.

Dr. Ro received the Young Investigator Finalist Award from ISMRM in 1992 and the Scientist Award in Korea in 2003. He served as a TPC member of international conferences such as IWDW, WIAMIS, AIRS, and CCNC, and was the Co-Program Chair of IWDW 2004.



Konstantinos N. Plataniotis (S'90–M'92–SM'03) received the B.Eng. degree in computer engineering from the University of Patras, Patras, Greece, in 1988, and the M.S and Ph.D degrees in electrical engineering from the Florida Institute of Technology, Melbourne, in 1992 and 1994, respectively.

He is currently a Professor with the Multimedia Laboratory, Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto (U of T), Toronto, ON, Canada, and is an Adjunct Professor with the School of Computer

Science, Ryerson University, Toronto. He is the Director of the Knowledge Media Design Institute, U of T, and the Director of Research at the U of T's Identity, Privacy and Security Institute. His current research interests include biometrics, communications systems, multimedia systems, and signal and image processing.

Dr. Plataniotis is the Editor-in-Chief for the IEEE SIGNAL PROCESSING LETTERS from 2009 to 2011, a Registered Professional Engineer in the province of Ontario, and a member of the Technical Chamber of Greece. He is the recipient of the 2005 IEEE Canada's Outstanding Engineering Educator Award "for contributions to engineering education and inspirational guidance of graduate students." He is also the co-recipient of the 2006 IEEE Transactions on Neural Networks Outstanding Paper Award for the 2003 paper titled, "Face Recognition Using Kernel Direct Discriminant Analysis Algorithms."