Image and Vision Computing 58 (2017) 61–75

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Leveraging multiple cues for recognizing family photos ☆

Xiaolong Wang[a,*], Guodong Guo[b], Michele Merler[c], Noel C. F. Codella[c], Rohith MV[a], John R. Smith[c], Chandra Kambhamettu[a]

[a] CIS, University of Delaware, Newark, DE 19716, USA
[b] LCSEE, West Virginia University, Morgantown, WV 26506, USA
[c] IBM T. J.Watson Research Center, Yorktown Heights, NY 10598, USA

ARTICLE INFO

Article history:
Received 8 October 2015
Received in revised form 23 June 2016
Accepted 16 July 2016
Available online 25 July 2016

Keywords:
Family photo recognition
Social media
Semantics
Group photo analysis

ABSTRACT

Social relation analysis via images is a new research area that has attracted much interest recently. As social media usage increases, a wide variety of information can be extracted from the growing number of consumer photos shared online, such as the category of events captured or the relationships between individuals in a given picture. Family is one of the most important units in our society, thus categorizing family photos constitutes an essential step toward image-based social analysis and content-based retrieval of consumer photos. We propose an approach that combines multiple unique and complimentary cues for recognizing family photos. The first cue analyzes the geometric arrangement of people in the photograph, which characterizes scene-level information with efficient yet discriminative capability. The second cue models facial appearance similarities to capture and quantify relevant pairwise relations between individuals in a given photo. The last cue investigates the semantics of the context in which the photo was taken. Experiments on a dataset containing thousands of family and non-family pictures collected from social media indicate that each individual model produces good recognition results. Furthermore, a combined approach incorporating appearance, geometric and semantic features significantly outperforms the state of the art in this domain, achieving 96.7% classification accuracy.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Sharing images in social media (Facebook, Flickr, Instagram, etc.) is gaining rapid popularity in our modern lives. With universal access to mobile cameras and internet access almost everywhere in the world [1], millions of photographs are being constantly captured and shared to social media. As such, there is great demand for automatic annotation of photos to navigate and perform retrieval on such social collections at scale. Classification of personal photographs spans a broad spectrum of concepts and entities, from people to activities, from scenes to food or abstract concepts.

We focus our attention on one particular important aspect, which involves analyzing the social relationships among the subjects of a photograph. People seldom pose with strangers [2], and recognizing who they take pictures which constitutes the first step in analyzing the social relation (kinship, friends, etc.) in a given group photo. A quick search on Flickr and Instagram using the keyword "family" confirms that people not only possess family photographs but are keen on labeling them as such. This has further inspired us to create an automatic method for classifying images containing groups of people into family and non-family photographs.

As illustrated in Fig. 1, there are several common differences that can be observed between family and non-family photos. First, people with a kinship relation usually share similar facial attributes and their facial appearance is much more similar than that of people in non-family photos. An automated method to incorporate physiological similarities of faces can be expected to have a direct advantage toward the task from such a capability. Second, the spatial distribution and physical characteristics of individuals in two photo categories are distinctive. For example, people in family photos usually stand in a cascaded pattern and exhibit irregular height distribution. The age distribution of people in family photos is usually wider than people in non-family photos. There also exist unique observable phenomena related to composition of posing patterns in family photos. For example children usually stand in front of adults, and elder people tend to stand in the center. On the other hand, friends in non-family photos stand in random patterns and their ages vary in

☆ This paper has been recommended for acceptance by Maja Pantic.
* Corresponding author.
E-mail addresses: xiaolong@udel.edu (X. Wang), guodong.guo@mail.wvu.edu (G. Guo), mimerler@us.ibm.com (M. Merler), nccodell@us.ibm.com (N. C. F. Codella), rohithmv@udel.edu (R. MV), jsmith@us.ibm.com (J. Smith), chandrak@udel.edu (C. Kambhamettu).

http://dx.doi.org/10.1016/j.imavis.2016.07.006
0262-8856/© 2016 Elsevier B.V. All rights reserved.

Fig. 1. Examples of family photos (first row) and non-family photos (second row). The similarity of appearance, and the arrangement of people in a group photo reveal the type of the photo. In general, family members usually stand in a cascaded way and are more similar in their appearance.

a smaller range, since people tend to make friends within their age group. Third, the choice of environment, event and general context of a given photo provides additional cues as to what group of people in represented in it. For example, it is highly unlikely for family photos to be taken in a bar or club, whereas it is common for friends to snap pictures together in such environments. Based on this type of observations, we propose a model for recognizing family photos which fuses different kinds of discriminant features in a unified framework.

The contributions of this work include the following: (1) We propose a new geometry feature which captures people's standing pattern at the scene level. The proposed geometry feature is purely based on the relative position of people in the image. This is different from previous works [2,3] where age, appearance feature, and gender are used to analyze the photo. Our geometry feature is used to represent the relative position of individuals in the photo. As a result this proposed geometry feature is very robust and efficient to extract. The classification accuracy obtained by using only the geometry model is more than 87.0%. (2) We propose an appearance feature which can capture facial similarities of people. Compared to our previous framework [4], the facial appearance descriptor is extracted using a convolutional neural network trained on the FaceScrub dataset [5]. We demonstrate that this yields improved recognition performance in comparison to using hand-crafted features. (3) In addition to geometry and appearance features, we also use semantic information to discriminant two categories. We study the performance of semantic model system in this context. Finally, we propose a simple fusion scheme to combine all three approaches together, yielding even further improvements to performance.

We conducted experiments on a dataset containing thousands of group photos obtained from Flickr. While each proposed representation proved valid for the task, they demonstrated to carry complimentary information. The experimental results demonstrate that fusing geometry, appearance model, and semantic context information yields an improvement of 3.3% over the current state-of-the-art for family photo classification.

2. Related work

Photos with groups of people can provide many meaningful social context information in the photo [6]. These contextual features can help interpret demographic information, such as people's age and gender. Their experimental results demonstrated that using the context information can help improve the performance of event recognition. This work showed that context information extracted from the group photo can also aid demographic information perception.

Instead of estimating general demographic information (age and gender), other works also attempted to estimate the pairwise relation between individuals in a given photo. Singla et al. [7] used rule-based Markov Logic Network to detect and identify possible social relation of different individuals. Based on the general knowledge, such as parents are older than their children, and the gender of spouses are different. They used different constraints in the proposed framework. MC-SAT algorithm [8] was applied to combine hard and soft constraints to predict relationship between different individuals. Wang et al. [2] utilized pairwise facial features calculated from individuals within the group photo to identity person and estimate social relations. These pairwise features are extracted from each face pair. Pairwise height difference, age difference, and closeness are used as the social context feature. Their experimental results illustrated that social feature can help improve the recognition performance. Chen et al. [3] proposed a sub-graph learning based approach for group photo classification. They tried to classify the group photos into two general categories: family photo or non-family photo. In their work, different subgraphs were built to characterize social relationships. Age, gender, and pairwise distance between individuals were used to construct the social subgraphs' set. The feature of a given group photo is constructed by calculating the distribution of extracted subgraphs corresponding to the pre-trained subgraph set. SVMs [9] was used as the classifier to determine the category of the given photo [10]. However, their proposed framework is mainly built from age and gender information; facial similarities between people are not measured. As illustrated in their experimental results, when people have similar age and gender, their framework may not work well, such as the classification between photos with siblings and the photos with the classmates. Their experimental results also demonstrated that using general feature extraction (e.g., PHoG [11]) from the global photo did not work. In our previous work [4] which targets the family photo and non-family photo categorization problem, we proposed one fusion model with geometry and appearance feature. The geometry is built purely based on people's face potions. It is used to capture the global standing pattern of individuals in a photo. This is different from measuring individuals' pairwise distance. Since our geometry model only uses the location information, the scheme is

very efficient and simple to implement. Dense SIFT [12] feature with Modified Hausdorff distance (MHD) [13] is used to measure the facial similarities between people in the appearance model.

In this work, we aim to classify the photo into family photo and non-family categories same as previous works [3,4]. The geometry model advocated in our previous work [4] is also adopted in this paper. Meanwhile, we also advocate a new mid-level appearance feature for representing facial similarities in the group photo. In the recent years, feature representation learning has demonstrated superior performance in a variety of visual recognition tasks, such as image classification [14–16], object detection [17], and image segmentation [18,19]. Compared to our previous advocated model [4], we adopt Convolutional Neural Networks (CNNs) as the basis to measure the facial similarities instead of using manually-crafted feature (SIFT [12] or SIFT after processing [20,21]). To improve the matching efficiency, instead of using Modified Hausdorff distance, L1 norm is applied to measure the similarity between faces. The performance obtained using single appearance model is more than 90.0%. Furthermore, not limited to the low level and middle level information, we also use high level semantic features to discriminate different image categories. As far as we know, this is the first time that semantic information is applied in the family photo analysis task. We also fuse these different models together to improve the recognition performance. Experimental results demonstrate that fusing multiple cues extracted from the group photo can help improve the recognition performance. In general, compared to our previous work [4], in this paper, we advocate a new appearance model based on deep neural network and fuse the semantic information in the final classification model.

The structure of the whole paper is organized as follows: the proposed framework is presented in Section 3, where three different models are introduced. Section 3.1 presents the geometry model. The appearance model is described in Section 3.2. Section 3.3 reports the details of the general framework for extracting contextual semantic information. The experimental setting and results are presented in Section 4, and finally we give conclusion remarks.

3. Our approach

In this work, three different cues are extracted from a group photo to recognize family category. Afterwards, a fusion scheme is applied to fuse these cues together. The features aim to characterize the photo in different aspects. Fusion results demonstrate that these features are complementary to each other. These models also help us have a further understanding of the group photo classification problem.

3.1. Geometry model

Unique spatial distributions and patterns of people are typically observed that correlate with the relationship of the individuals captured in a photograph. These patterns can be in part described in terms of the physical distance difference between individuals [3], variations in height [2], or the physical proximities in the geometry of the people standing in a photo (such as pairwise distance between people). As indicated in Fig. 1, the height differences and the relative position vary between categories. These height differences are often associated with human relationship information, such as there is usually a distinct height difference between the parents and their children. The standing position in the photo can also reveal their social roles: for example, grandparents of the family usually stand in the group center, whereas the parents stand in the back of the children in most situations. In non-family photos, the height difference of people is muted, as the age gap is typically reduced. The physical proximity of people within the non-family photo is often different from family photos. Measuring pairwise distance and the height difference between people is the most commonly used approach. Wang et al. [2] calculated the Euclidean distance between face pair to construct social features. In Ref. [6], the distance between faces' centroid and each particular face is used as one of the contextual features to estimate the demographic information from the group photo. Chen et al. [3] counted the number of people between one people pair as the pairwise distance measurement to build the social subgraphs. Compared to these frameworks, instead of measuring the pairwise distance, our geometry feature is extracted at the global scene level to capture the overall standing pattern of the group people. The extracted feature can be directly used in the photo categorization problem. This scheme is much more efficient in characterizing the global standing pattern than other solutions, such as Minimal Spanning Tree (MST) which was used in [6]. While it is true that MST could be used to link all the faces in a picture, it not clear how such global representation would be computed. In Ref. [6], MST was used as a mean to compute the degree of each individual vertex (face), and not as a global descriptor for the group of all faces in the image. One limitation in using MST as a descriptor for our problem is that, given a set of vertices in a graph (our face points), there could be more than one MST if not all the edges have unique weights. Any representations based on MST would suffer from such lack of a deterministic structure.

3.1.1. Polygon construction

Our geometry feature is based on a polygon formed by the lines connecting the locations of people in the photo. Given a set of points, many types of polygons can be constructed. In our framework, our goal is to capture the contour shape of the group people's standing patterns in the photo. Also, the contour should capture people's standing difference between two different categories (family and non-family) as listed in Fig. 4. Meanwhile, our descriptor should have the following properties: (a) invariant to shift in the order of polygon vertices and the rotation of the image,(b) robust to variations in photo resolution,(c) unique and deterministic given an input image (unlike the MST which could produce different representations given the same input image), and (d) efficiency. The time complexity of our proposed model is O(n log(n)).

The entire extraction pipeline is illustrated in Fig. 2. Face locations of individuals in the group photo are used as the vertices to formulate the polygon. Faces are detected based on the algorithm proposed by Viola and Jones [22]. Then, a convex hull [23] is applied to all vertices to construct the primary polygon. Our observations indicate that fortunately, the convex hull of these points, in general, can approximate the shape of contour of the standing pattern. In some case, it is exactly the shape of the contour as illustrated in Fig. 3. Therefore, our solution (motivated by the above key observations) is given on the top of convex hull. It is known that convex hulls have been useful for many other applications, such as in computer visualization [24], path planning [25], shape matching [26], crystallography [27], and cartography [28]. In this work, convex hull is used as the basis to extract the geometry feature.

Given a set of points S in the Euclidean space, the convex hull is the smallest convex set that contains S. Each point s_i in S is associated with a non-negative weight w_i, then all non-negative weights sum to one. This can be calculated as follows:

$$\left\{\sum_{i=1}^{n} w_i s_i | (\forall i \ : w_i \geq 0) \ \wedge \sum_{i=1}^{n} w_i = 1\right\}, \tag{1}$$

where we assume that there are n different points in set S. Because the photos used in our work are collected from the unconstrained environment, the distribution of people within a group photo will

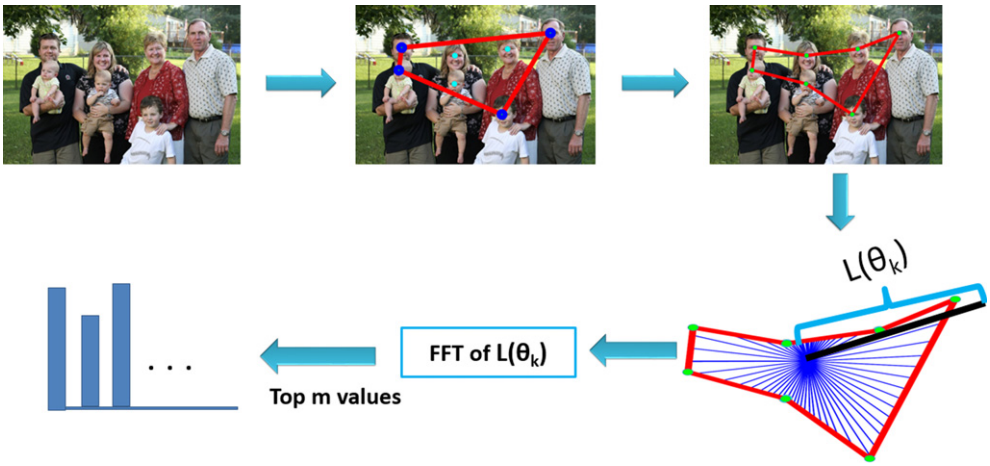

Fig. 2. An illustration of extracting geometry feature of the given group photo. In our geometry feature extraction pipeline, each vertex on the polygon corresponds to one face in the group photo.

have a high degree of variation. The constructed polygon using convex hull cannot guarantee that all faces are located on the polygon. In fact, the polygon is built by the convex hull, which minimizes the perimeter, leaving the possibility that points are lying on or within the constructed polygon. Let us assume that there are n individuals standing in one photo. After applying the convex hull, we can build one polygon which has m sequential vertices enclosing these n points, where $m <= n$. Since we need to characterize the geometry shape of standing pattern of all people, our next step is to add remaining $n - m$ vertices to the constructed polygon.

Let us assume that the coordinates of sequential vertices of the polygon built by the convex hull are $S = [(u_1, v_1), (u_2, v_2), \ldots, (u_m, v_m)]$. The $n - m$ remaining vertices are $S' = [(u'_1, v'_1), (u'_2, v'_2), \ldots, (u'_{n-m}, v'_{n-m})]$. To include these remaining vertices to the current polygon without abruptly changing the structure of the constructed polygon, for each point in S', we locate its two closest sequential vertices in set S. Afterwards, the new polygon is built based on their relative positions. Each point (u'_i, v'_i) can be considered as the node lying between two adjacent points (u_j, v_j) and (u_{j+1}, v_{j+1}) in set S. It means that any point (u_g, v_g) in the 2D space can be represented by

$$u_g = \alpha \cdot u_j + (1 - \alpha) \cdot u_{j+1},$$
$$v_g = \alpha \cdot v_j + (1 - \alpha) \cdot v_{j+1}, \tag{2}$$

where $\alpha \in [0, 1]$. Given the point (u'_i, v'_i), to locate its two nearest vertices in set S, we measure the Euclidean distance. The distance between (u'_i, v'_i) in $S'$and (u_g, v_g) can be calculated by

$$D = \sqrt{(u'_i - u_g)^2 + (v'_i - v_g)^2}. \tag{3}$$

To find its two nearest neighbors lying on the constructed polygon for the given point (u'_i, v'_i), the derivative of D with respect to α is calculated, which is $\frac{\partial D}{\partial \alpha} = 0$. Then we can obtain α, which is represented by

$$\alpha = \frac{(u_j - u_{j+1})(u_g - u_{j+1}) + (v_j - v_{j+1})(v_g - v_{j+1})}{(u_j - u_{j+1})^2 + (v_j - v_{j+1})^2}. \tag{4}$$

Obtaining α, we can measure the distance D between the (u'_i, v'_i) and the sequential vertices in S based on Eq. (3). Based on the calculated distances D, we can get the minimum distance. This measurement can help locate two sequential corresponding vertices in S. We name this procedure "convex hull post-processing". This post-processing step can help build the comprehensive polygon where all faces are located as vertices on the polygon. To illustrate the whole procedure, we have listed several examples in Fig. 4. These examples also can illustrate people's standing difference between two categories truly exists.

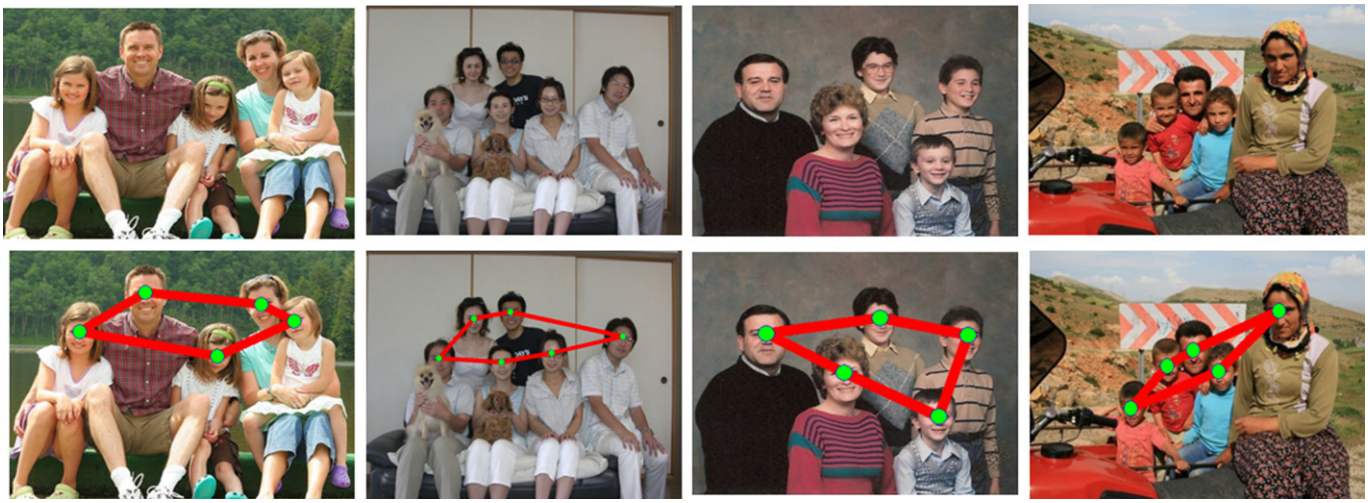

Fig. 3. Examples of group photo with the contour generated by applying convex hull directly.

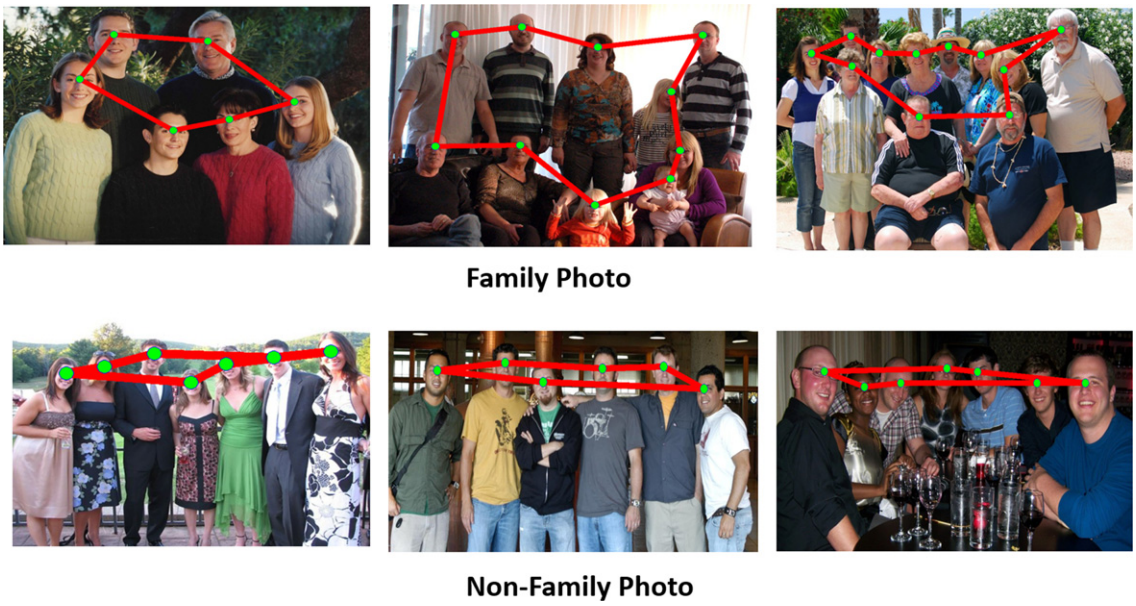

Fig. 4. Illustrations of constructed polygon based on individual arrangement in a photo based on our approach. From the constructed polygons, we can see that there is a significant difference between people arrangement of family photo and non-family photo in most cases.

3.1.2. Geometry feature extraction

The constructed polygon is used as the basis to extract geometry feature. The extracted geometry feature is directly used to describe people's standing pattern. In this work, we propose a mid-level geometry feature based on Fast Fourier Transform (FFT) [29]. The whole feature extraction framework proceeds as follows:

(a) Calculate the center of the 2*D* polygon.

$$u_{cen} = \frac{1}{n}\sum_{i=1}^{n} u_i,$$
$$v_{cen} = \frac{1}{n}\sum_{i=1}^{n} v_i, \tag{5}$$

where (u_{cen}, v_{cen}) indicates the centroid of the polygon. (u_i, v_i) is the coordinate of vertices on the polygon. n is the number of vertices (number of individuals).

(b) After obtaining the centroid, we divide the angle around the polygon center into K folds evenly. This can be represented as

$$\theta_k = \frac{2\pi}{K} * k, k \in \{1, 2 \ldots, K\}. \tag{6}$$

θ_k is the angle corresponding to the kth ray going through the center to the edge of the polygon. K is the total number of rays casting from the centroid. In our work, K is set to 64.

(c) Given the origin (u_{cen}, v_{cen}) and angle θ_k, K rays are casted radially from the origin with an equal angular interval. Afterwards, the distance between the perimeter of the polygon and the center is calculated along the specified direction θ_k. As illustrated in Fig. 2, $L(\theta_k)$ denotes the distance between the original centroid (u_{cen}, v_{cen}) and the perimeter along the ray oriented along θ_k.

(d) After obtaining the ray vector L, we calculate Fast Fourier Transform (FFT) for each ray. Fourier transform is used to convert the vector from its original space domain to a representation in the frequency domain. An FFT rapidly computes the Fourier transformation by factorizing the Discrete Fourier Transform matrix into a product of sparse factors [30,31]. We adopt FFT method in this paper due to its fast computation speed, effective representation capability, and desired satisfactory performance. These advantages have been demonstrated in our experiments. For calculation details, please refer to work [30]. FFT has been used previously in several shape representations [32,33].

Then the amplitude spectrum is extracted and normalized. Top T amplitude spectrum values are used as the geometry feature to represent the geometry information for a given group photo. Based on our experimental investigation, T can be set to a value around 50 typically. Using amplitude as the feature is invariant to many influences, such as the shift in the order of polygon vertices and the rotation of the image. Normalization can also deal with variations in photos' resolution. These calculations are very helpful for characterizing the geometry of similar spatial distribution of people across photos with different resolutions and orientations. Our geometry feature can capture the similarity in their arrangement effectively.

In general, the idea of our geometry feature is different from previous approach of geometry model used in object recognition [34] and face recognition [35,36] where shape is mainly used to describe the geometry information of one single object. Whereas, our advocated geometry feature aims to encapsulate the global view of people in the group photo. Meanwhile, we think there is certainly more than one model to formulate the geometry descriptor for recognizing family photo. Our proposed geometry model is based on considerations of the efficiency in computation and the accuracy in performance.

3.2. Appearance model

To measure the similarity of people in the group photo, we adopt deep learning feature scheme to extract facial features. One advantage of this deep learning model-based approach is its improved capacity for face representation in comparison to hand-crafted features (LBP [37], SIFT [12], etc.). As demonstrated in previous works in facial image analysis [38–41], convolutional neural network(CNN) [42] obtains a very good result on popular benchmarks, such as in the Labeled Faces in the Wild (LFW) [43] and Youtube Faces (YTF) [44]. The good performance of convolutional neural network is

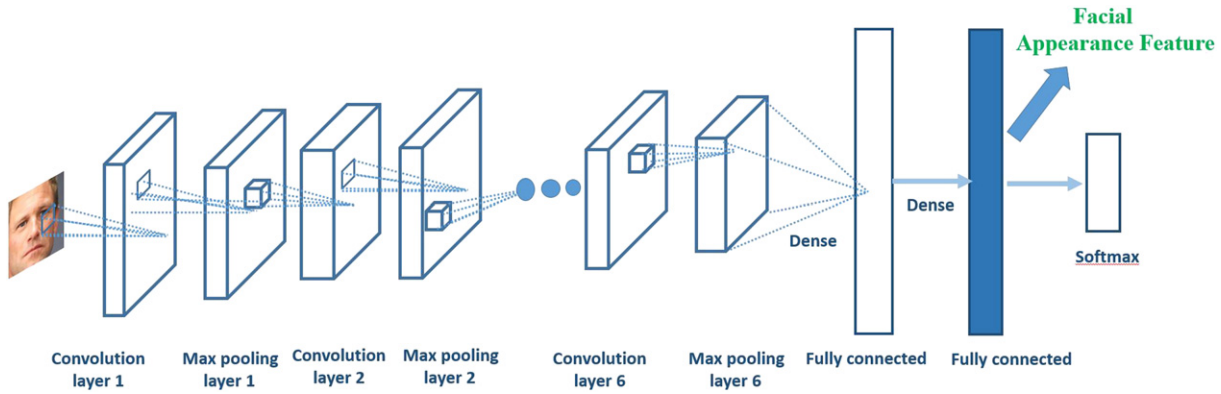

Fig. 5. The proposed architecture of CNN for learning face representation. To be concise, only three convolution layers are shown here. As indicated in the figure, feature maps extracted from the second fully connected are used the facial appearance feature representation.

likely due to several reasons. One reason is because of the availability of search engine for crawling large amounts of photos. Another is the availability of scalable computation resources, especially GPU technology.

Based on existing studies in face identification and related works in measuring the similarities between people with kinship relation, we attempt to apply deep models in extracting appearance features in group photo. Particularly, we apply CNN as the basis to extract middle-level appearance feature to characterize facial similarities in group photo. In our proposed pipeline, CNN is trained for face identification task. During the training phase, samples from different individuals are labeled with different labels that distinguish personal identity. As studied in previous works [45–47], age plays an important role in measuring the facial similarity, especially in face representation in kinship [48]. In our framework, the influence due to age is also considered in the proposed appearance model. The whole framework is illustrated in Fig. 6. The whole appearance model mainly includes three different steps: Pairwise Feature (PF) Extraction, codebook construction and DOGF feature extraction.

3.2.1. Pairwise Feature (PF) Extraction

To measure the appearance similarities among people in the group photo, CNN is applied as the basis to build the framework. In the beginning, faces and fiducial points are detected [22,49]. In this work, all facial images are aligned based on eyes' coordinates. The basic structure of CNN used in our framework includes five convolutional layers, followed by two fully connected layers and a softmax layer. All face images are resized into the same size 227 × 227 with three channels (RGB). The first convolutional layer is calculated by the convolution between the input RGB image with 96 different kernels with a stride of 4. The size of these 96 kernels is 11 × 11 × 3. The size of the sequential convolutional layer filters are 5 × 5 × 256, 3 × 3 × 384, 3 × 3 × 384 and 3 × 3 × 256. Along with these convolutional layers, there are two fully-connected layers. The output neurons obtained from these two fully connected layers are 4096 and 530 respectively. In the proposed framework, max-pooling layers are applied after each convolutional layer. The structure of each max-pooling layer is set as 3 × 3 with a stride of 2. ReLU function [50] is applied as the activation function of all convolution layers. The

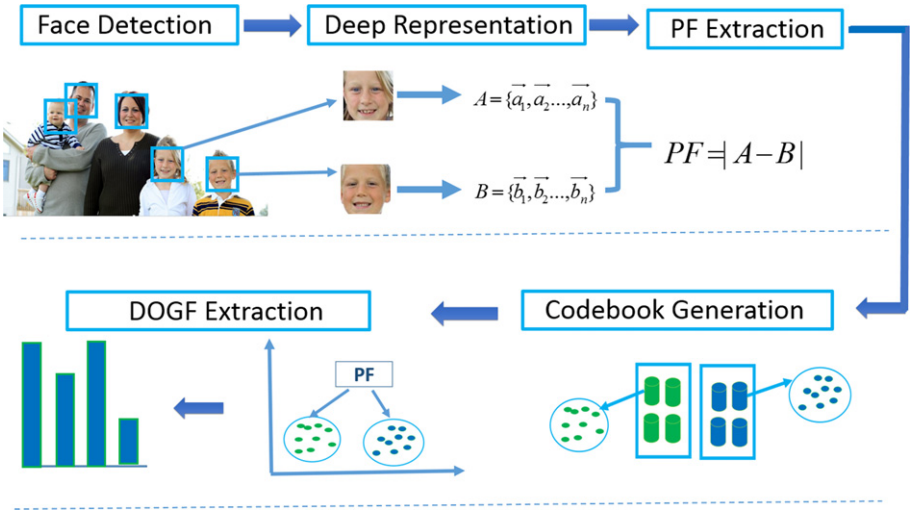

Fig. 6. An illustration of the proposed DOGF feature extraction. To be concise, only one face pair is listed here. Other facial pairs all follow the same processing procedure. In the codebook generation step, two colors represent different classes (family and non-family), which are represented by green and blue. In each category, four codebooks are learned from corresponding pairwise feature set divided by different age gaps.

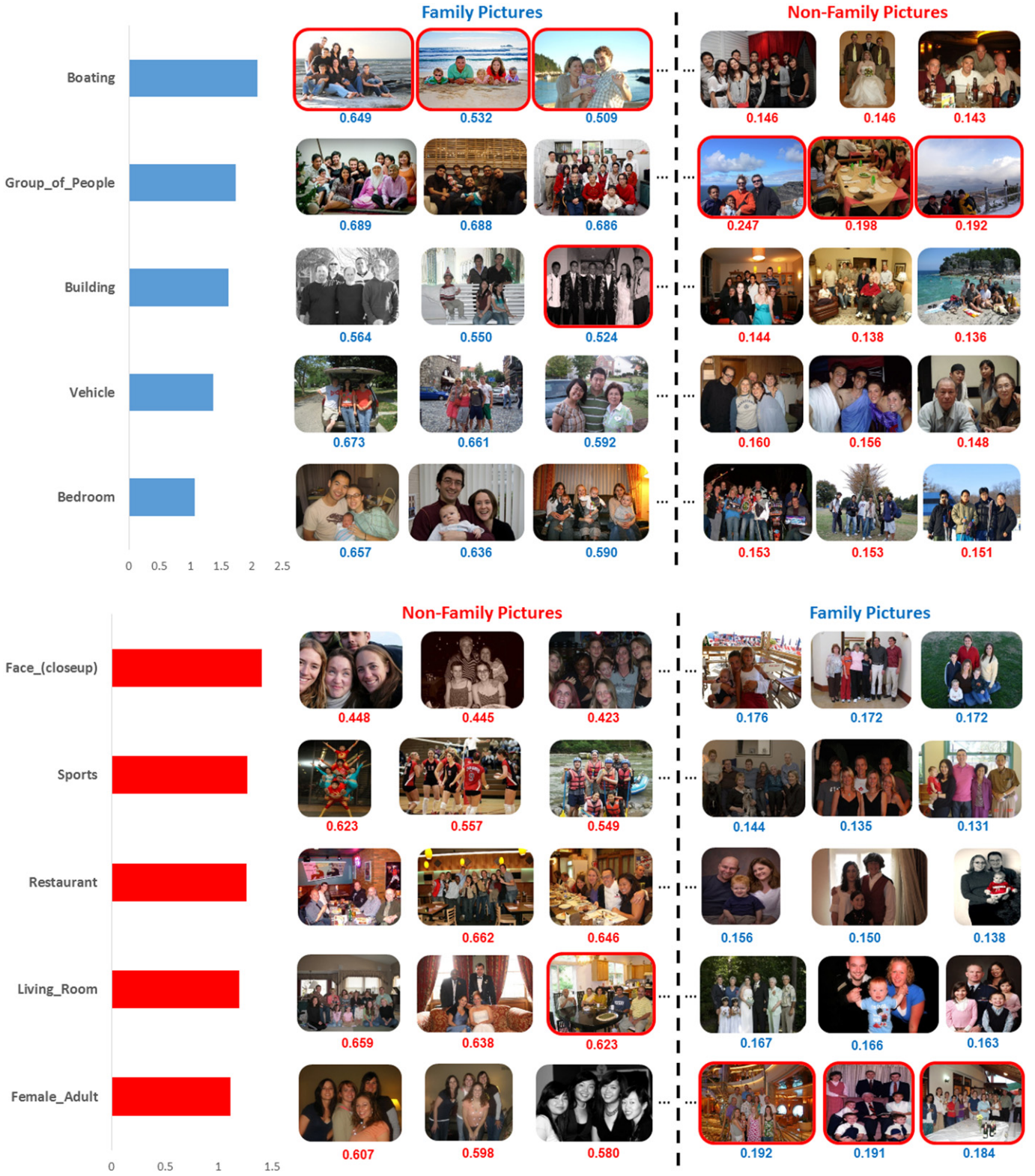

Fig. 7. Top five most discriminative semantic concepts associated with family (top) and non-family (bottom) photos. Under each photo we report the score of the corresponding semantic model. Images with red border represent mislabelings of the classifier.

whole framework is trained using back-propagation along with the softmax function as indicated in Fig. 5. CNN generally includes convolution and pooling operations. In the following, we introduce these operations briefly.

Convolution layers. Each neuron in the convolution layer is calculated by the convolution between the local receptive field in the preceding layer and the learned kernels (weights). In general, neurons in the same feature map share the same weights but are calculated from

different input receptive fields. An activation function is applied at the end of each layer as follows:

$$e_j^l = \phi\left(\sum_{i=1}^{N} e_i^{l-1} * w_{ji} + b_j^l\right), \tag{7}$$

where e_i^{l-1} is the input neuron from $l-1$ layer, N is the total number of input neurons, andb_j^l denotes the bias. $*$ denotes the convolutional operation. ϕ is the activation function where ReLU function is used in this work.

Pooling layers. In general, pooling layer downsamples the input feature maps. Pooling layer only changes the size of the input maps while not altering the number of input feature maps. The schemes implemented in the pooling layer usually include averaging, calculating the maximum, or using learned combinations of the neurons within the given block [51]. In our proposed framework, max pooling is applied. This operation is used to maintain specificity and is very efficient in characterizing the feature for specific topics. Its mechanism is similar as mammalian visual cortex [52].

In this work, Caffe [53] is applied to build the deep neural network to learn deep facial representation. For weights initialization, we use a Gaussian distribution with zero mean and a standard deviation of 0.01. In the beginning, zeros are used as the initial value for the biases. In each iteration, all the weights are updated based on a batch size of 128. The momentum is set as 0.9 and the weight decay is set as 0.005 for all layers. The FaceScrub dataset [5] is used to train the whole network.

In this work, feature maps extracted from the second fully connected layer are used to represent the facial characteristics. Assuming the representation for two facial images are $A = \{\vec{a_1}, \cdots, \vec{a_e}\}$ and $B = \{\vec{b_1}, \cdots, \vec{b_e}\}$, where e is the number of feature dimension. $PF = |A - B|$ is defined as the pairwise feature (PF) to measure the facial similarity between a pair of face images. For a group photo including n different individuals, there are $J = C_n^2$ different face pairs to compare. The pairwise feature set for a group photo including n faces can be represented as $PF_{photo} = \{PF_1, PF_2 \ldots, PF_J\}$.

3.2.2. Facial codebook construction

There may be many different kinship relations in a given photo. Age progression usually restricts the performance of measuring facial similarity with age gaps. Considering the age influence, we propose one codebook scheme using the extracted pairwise feature and age label information. There are seven age categories labeled in the dataset, which are $[1, 5, 10, 16, 28, 51, 75]$. These age labels represent infant, kid, school-age child, teenager, youth, middle-aged adult and elder. Our appearance model is used to measure the similarity between facial pairs. Age gap is calculated between two compared individuals. We have defined four different age gaps, which are $G_{age} < 10, 10 \leq G_{age} \leq 20, 20 < G_{age} \leq 40, G_{age} > 40$. In the testing phase, the age of each person is estimated using the scheme proposed in Guo et al. [54]. For age estimation, we follow the experimental setting as discussed in [3] where cross-validation is used. There are several reasons why we use bio-inspired features: a) an intuitive but important descriptor to capture the difference in age estimations; b) convincing performance in many real-world face datasets (as demonstrated in experiment results listed in previous works [54–56]). Our age estimation pipeline can obtain the estimation accuracy around 98.0% for seven age categories classification.

Based on the age gap, in the training stage, pairwise feature set PF_{photo} can be divided into four different groups. Within each divided pairwise feature set, K-means [57] is used to learn facial similarity codebooks. Each codebook has H different codewords. In this work, H is set to 9. The analysis of different values' setting is discussed in Experiments section. We have two different categories, which are family and non-family. Codebooks for each category are learned separately. These two codebook sets are represented as $\vec{G}_{mh}$ and $\vec{G}'_{mh}(h = 1, 2 \ldots, H)(m = 1, 2 \ldots, M)$, where m is the number of age groups and h denotes the number of codebooks. In this work, M is set to 4 since we have four different defined age gaps. For different group photos with different relations, age gaps calculated from different face pairs can vary a lot, construction of codebook groups based on age information is therefore necessary.

3.2.3. DOGF feature extraction

For jth face pair in the group photo, we calculate pairwise feature PF_j and estimate the specified age gap group m. Based on the estimated age gap, we can find its two corresponding codebooks, $\vec{G}_{mh}$ and $\vec{G}'_{mh}$. Then we calculate the pairwise cosine similarities between PF_j and $(\vec{G}_{mh}, \vec{G}'_{mh})$. This can be represented as $\vec{d}_j = (d(PF_j, \vec{G}_{mh}), d(PF_j, \vec{G}'_{mh}))$. Because age codebooks used for each facial pair with different kinship relations are not the same, the similarity feature calculated using our proposed scheme can also capture the co-occurrence of different relations in these photos.

For a group photo with n different individuals, C_n^2 different pairwise cosine similarity facial features can be extracted. Our appearance model aims to advocate a appearance descriptor to measure the facial similarity of people in group photo. Our appearance feature for representing facial similarities in the given group photo are named as Degree Of Group similarity Feature (DOGF). Based on the obtained C_n^2 pairwise similarity features, DOGF for one group photo can be calculated as $\vec{F} = \frac{1}{J}\sum_{j=1}^{J}\vec{d}_j$, where $J = C_n^2$. $\vec{F}$ is directly used as appearance feature for classification. In the testing phase, the age of the given face image is estimated using the scheme proposed in [54].

3.3. Semantic model

We observe that the context in which a picture is taken can also represent a valid cue to predict whether it is a family photo or not. Intuitively, the places people go to and the activities they perform tend to be different when in the company of family members as opposed to with their friends. For example, it is more likely for people to perform a sport activity with friends rather than family, while lunches at someone's house would seem more likely to be family gatherings. Based on this observation, we investigated whether the semantics of the context in which pictures are taken hold any correlation with the group of people represented in the picture (family or not family). We therefore trained a set of 764 visual semantic models from a set of half a million images downloaded from the web, following the framework introduced by the IBM Multimedia Analysis and Retrieval System (IMARS) [58]. The training dataset of images was manually annotated and organized in a hierarchical faceted taxonomy, which includes concepts related to "objects", "scenes", " people", "activities" and "events". Each model SC_i is an ensemble of SVMs with linearly approximated χ^2 kernel, learned on top of bags of examples randomly sampled from the set of manually labeled web images. Each individual SVM uses one of several different visual descriptors including color histogram, color correlogram, wavelet texture, edge histogram, gist, and lbp histogram, extracted at multiple different regions of the image, in a similar fashion to the spatial pyramid framework. The score for a Semantic Concept i on in a new image x is then

$$SC_i(x) = \sum_{k=1}^{N_i} w_k b_k(x) \tag{8}$$

$SC_i(x)$ is the weighted sum of the scores on x of the individual SVMs, which we define base models b_k in the ensemble. The weights w_k are learned via cross-validation during training. Finally, the score is

Fig. 8. Examples of group photos used in our experiment. (We can see the diversities of the family.)

normalized to the [0, 1] range by fitting a sigmoid on the prediction scores of the validation set.

For our domain of family pictures, we map each photo x to the semantic space by concatenating all the models scores into a N-dimensional Semantic Model Vector [59], that is, a vector in which each dimension has a semantic meaning corresponding to the prediction of a visual classifier to the given picture

$$SMV(x) = [SC_1(x), \ldots, SC_i(x), \ldots, SC_N(x)]$$

We then use this concatenated vector as a feature on top of which we train an SVM model to distinguish between family and non-family pictures, as explained in the following Section. The experimental results described in Section 4 confirmed that the visual semantics of a picture, even if not as strong as other features, indeed can be used to predict whether a picture is a family photo or not. Furthermore, it proved to be a complementary cue which can be combined with the other descriptors to boost recognition performance.

In order to qualitatively evaluate our choice of visual classifiers and determine the most discriminative ones for family pictures, we trained two linear SVMs on top of the Semantic Model Vector representation: one using family pictures as positives and non-family ones as negatives, and the other inverting the roles. In Fig. 7 are reported the top five weights of the SMVs, family in blue and non-family in red. The larger the weight, the higher the association of a visual concept with a class of pictures versus the other. For each category, we report in the Figure the top three and bottom three pictures from our dataset, ranked according to their visual classifier scores (shown below each photo). The visual classifiers are not always perfect and sometimes might make errors, which are highlighted with a red border in the Figure. We notice however that even when the exact semantic is lost, the discriminative power of the classifier can still contribute to the classification of a picture. For example the first three pictures of the first row do not contain a boat, but the classifier is picking up a correlation between family photos and settings close to bodies of water. It is interesting to observe how larger groups and

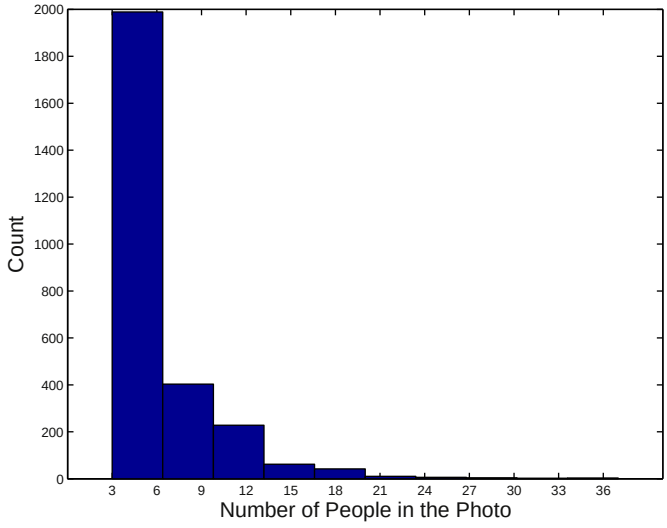

Fig. 9. Histogram illustration of number of people included in group photos used in the whole experiment.

Table 1
Classification results using different schemes on Dataset One used in [3]. Fusion Scheme I indicates the fusion scheme based on geometry and appearance information. Fusion Scheme II indicates the fusion model using appearance, geometry and semantic information together. Bold text indicates the results obtained based on the newly proposed scheme.

Method	Accuracy
Chen et al. [3]	90.3%
Wang et al. [4]	93.9%
Geometry (Ours)	86.3%
Appearance (Ours)	**91.0%**
Semantic (Ours)	**75.0%**
Fusion Scheme I (Ours)	**95.3%**
Fusion Scheme II (Ours)	**96.0%**

Table 2
Classification results using different schemes on Dataset Two. Fusion Scheme I indicates the fusion scheme based on geometry and appearance information. Fusion Scheme II indicates the fusion model using appearance, geometry and semantic information together. Chen et al.* indicates the assumption that approach [3] could classify the newly added images on the collected dataset with 100% classification accuracy. Bold text indicates the results obtained using the newly proposed scheme.

Method	Accuracy
Chen et al.* [3]	91.7%
Wang et al. [4]	93.4%
Geometry (Ours)	87.3%
Appearance (Ours)	**90.2%**
Semantic (Ours)	**76.3%**
Fusion Scheme I (Ours)	**95.1%**
Fusion Scheme II (Ours)	**96.7%**

vehicles are more likely to relate to family pictures, perhaps to commemorate large family gatherings and family road trips. On the other hand, it would seem that people tend to eat out at restaurants and go to sporting events with friends rather than family. Even the rooms of the house can become an indicator of who we are taking pictures with: in the living room with friends, in the bedroom with our family.

3.4. Classification

Our goal is to classify a given group photo into two different categories, family or non-family photo. This problem can be regarded as a binary classification problem. We use SVMs [60] with RBF kernel as the classification method for our system, applied on top of each feature separately. Following the optimization scheme advocated in [61], we employed a five-fold cross validation approach and grid search over the parameters space to estimate C and γ for the RBF kernel.

There are three different features in our approach, each feature can be considered as a weak learner toward the final task. To fuse different features together, many approaches can be applied, such as feature level early fusion and score-level late fusion [62,63]. In this work, we apply weighted fusion based on the output of each RBF kernel obtained in each individual model. The fusion is formulated as follows:

$$R_c(x,y) = \sum_m w_m \, e^{-\frac{||x-y||^2}{2\sigma_m^2}}, \tag{9}$$

Fig. 10. (a) Illustrations of correctly classified photos using the geometry model. (b) Examples of misclassified photos using geometry model.

where $R_c(x, y)$ is the combined kernel value for samples x and y, and σ_m is the RBF parameter of kernel m. w_m is the weight associated with the model (appearance, geometry, semantic information). In this work, w_m is obtained via cross-validation on the training data, following the scheme of Ayache et al. [64]. x and y are the feature vectors associated with the model. In our problem, there are three different models. We refer the reader to [64] for specific details.

4. Experiments

In this Section, we describe the details of the dataset used to evaluate the proposed approach for family photo classification and illustrate the performance obtained by different models. The experimental results and analysis are also demonstrated. In order to provide a deeper understanding of the proposed model, we report classification results obtained from different models individually,

Fig. 11. (a) Illustrations of correctly classified group photos based on appearance model (DOGF). (b) Illustrations of the photos misclassified using geometry model and correctly classified by the fusion model based on geometry and appearance cues.

e.g., geometry, appearance and semantic model. We finally compare our proposed fusion framework with previous works [3,4] under the same experimental setting.

4.1. Dataset

The first dataset used for family photo classification was collected by Chen et al. [3]. This dataset includes 1167 family photos and 1263 non-family photos. We refer to this as "Dataset One". Because of the unbalance in the number of examples between the two different categories in Dataset One, we enhanced the dataset by adding more photos to each image class. The expanded dataset we collected includes 1420 group photos for each of the two categories (family and non-family). We name the expanded dataset as "Dataset Two". In order to collect the dataset, we followed the same collection scheme as Chen et al. [3]. All images came from a public dataset collected by Gallagher and Chen [6]. This public dataset was mainly collected from social media (e.g., Flickr) using keywords, such as "family portrait" and "group photo". The original public dataset provides some initial labels, such as family, group, and wedding. However, from the perspective of our classification problem of interest, the dataset presented some labeling errors. For example, some family photos appeared in the group category, and several non-family photos were included in the non-family photo category. We corrected such mislabelings in our newly extended and organized dataset with the help of human annotators. Five people were involved in labeling the new extended dataset. A photo was labeled as family photo or non-family photo only if all members agreed on its label. Otherwise the photo was not used in the experiments. In this paper, we don't consider group photos where family members and friends are mixed. We also don't consider the family with adopted children. Our dataset does not include such samples. As illustrated in Fig. 1, the collected dataset presents a wide variety of subjects in different poses including sitting, standing or laying. It is also a very challenging dataset for measuring facial similarities because of face occlusions, changes of facial expressions, and even faces with sun-glasses. Although the images are collected using English tags, they are still very representative due to its high diversities, i.e, Asian, Caucasian, African American, etc.(illustrated in Fig. 8). Moreover, the dataset has a very wide range of the number of people included in the group photo as illustrated in Fig. 9. The number is from 3 to 37.

The group photos used in the experiment all include three or more people. For photos with two people, standard kinship verification approaches [65–68] can be applied to determine the pairwise relation. Our work aims at determining the category of group photos, which is different from a strict kinship verification problem or from the work of Fang et al. [69] where a corresponding family is predicted when given one probe face image. In general, the scope of this work is to recognize the category of group photo (family or non-family), not that of working on individual facial images.

4.2. Experimental results and discussion

To evaluate the performance of the proposed framework, five-fold cross validation is applied. For each fold, the accuracy is calculated as

$$Accuracy = \frac{N_{correct}}{N_{total}} \times 100\%, \tag{10}$$

where $N_{correct}$ represents the number of correctly classified photos and N_{total} is the total number of samples in the testing set. The final performance is calculated by averaging the accuracy obtained in different folds. We evaluate the performance of the proposed scheme on Dataset One and Two. The experimental results are illustrated in Tables 1 and 2. We also compare the performance of different schemes on Dataset One used by Chen et al. [3] and Dataset Two. We also draw the ROC curve for comparison as listed in Figs. 13 and 14. The experimental results demonstrate that the proposed scheme outperforms previous works [3,4] significantly. As for these misclassified samples listed in [3], our algorithms can correctly classify them as demonstrated in Fig. 17.

From the result, we can find that each separate model performs well. Experimental results demonstrate that the discriminative pattern between two different categories truly exists and our geometry model can capture these discriminatory information from

Fig. 12. Illustrations of the photos correctly classified using the semantic model.

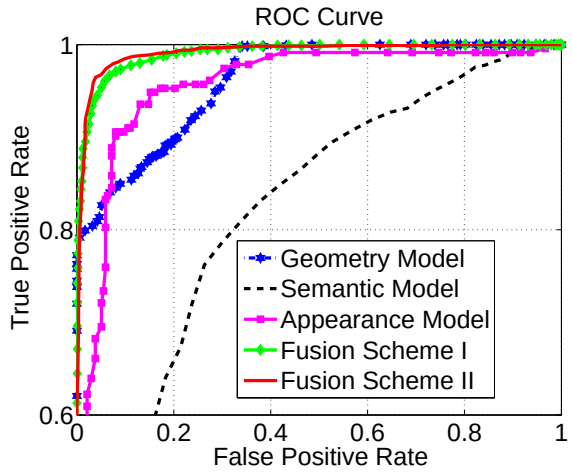

Fig. 13. Performance comparison of different models on Dataset One which is used in work [3]. Fusion Scheme I indicates the fusion model using geometry and appearance information. Fusion Scheme II indicates the fusion model using appearance, geometry and semantic information together.

the given group photo. Our appearance feature DOGF also works well in the unconstrained dataset. Although the images from the dataset used in our experiment are collected from unconstrained environments, the obtained performance is quite promising. The appearance model alone can achieve a classification accuracy up to 90.2% on the expanded dataset. This has improved the performance compared to our previous model [4] based on SIFT feature with Modified Hausdorff measure scheme. Our experimental results also demonstrate that supervised deep learning model is efficient in representing facial characteristics. Although contextual semantic information by itself does not obtain a comparable performance compared to the geometry and appearance models, it provides complimentary information. In fact the Fusion Scheme II, which includes such semantic information, further boosts the classification performance. For example, in Dataset Two, the final accuracy after fusing all different models is 96.7%, which is 3.3% higher than our previous model [4]. We have compared with the model proposed by Chen et al. [3] on two datasets. Our model obtains 96.0% compared to 90.3% reported in [3] on Dataset One. Since Chen et al. [3] did not

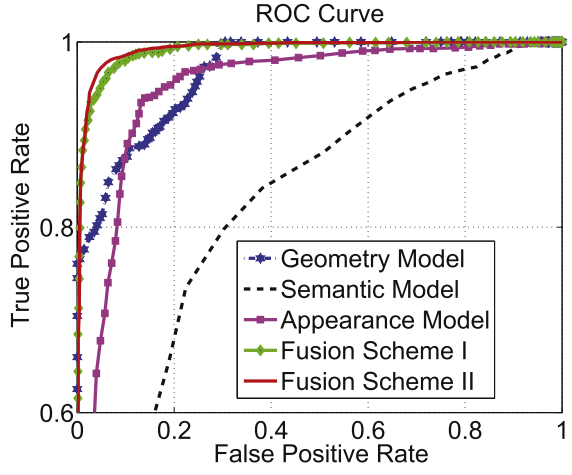

Fig. 14. Performance comparison of different models on Dataset Two. Fusion Scheme I indicates the fusion model using geometry and appearance information. Fusion Scheme II indicates the fusion model using appearance, geometry and semantic information together.

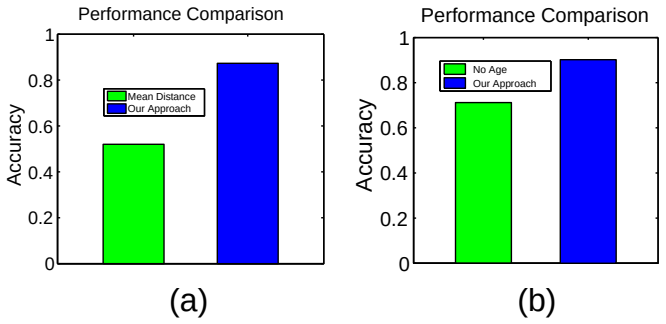

Fig. 15. (a) Performance comparison between mean distance and the proposed geometric feature scheme. Mean distance indicates that we calculate the center position of the polygon, then calculate the distance between the center to all the vertices in the polygon. Afterwards, the calculated mean distance is used as the feature to classify the photo category. (b) Performance comparison for different appearance models. No age indicates that there is no age information used in appearance model. Our approach indicates the proposed DOGF based appearance scheme.

release their source code, we assume that their approach can classify the images added to the expanded Dataset (Dataset Two) with 100% accuracy. With such assumption, their model would achieve a classification accuracy 91.7%, almost 5.0% lower than our proposed framework. Cross-validation in the training stage is also applied to select the optimal parameters. We have listed the accuracy comparisons between different K values as illustrated in Fig. 16 (a). The performance comparison with different numbers of codebook setting used in the appearance model is illustrated in Fig. 16 (b).

One advantage of the proposed geometry model is that our geometry feature extraction is purely based on the position pattern of the people in the group. It performs very well in typical family photos where people stand in a typical pattern, e.g., parents stand in the back of the children and the elders stand in the center. However, our geometry model does not work well in situations where individuals are positioned in an atypical way, e.g. standing in a row as shown in Fig. 10 (b). As analyzed in previous works [46,70], it is difficult to measure the facial similarities of people with large variations in pose, illumination, and age gaps. In Fig. 11 (a) and Table 2, we illustrate how the proposed DOGF feature achieves a very satisfactory result in discriminating the two different categories. Our appearance model can also compensate with the failure cases produced by the geometry model where family members stand in a less traditional way, such as standing in a row. It also works very well in other difficult situations, for example, when people sit around the table (e.g., dinner). Meanwhile, in cases where the appearance feature does not work, we can rely on the geometry model. Our experimental results demonstrate that these two information are complementary to each other.

To measure the performance of the advocated geometry model, we also compare the proposed geometry feature with another baseline where we calculate the distance between the centroid of the

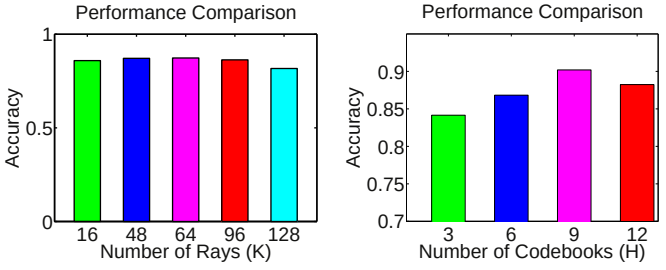

Fig. 16. (a) Performance comparison for different k values setting in the geometry model. (b) Performance comparison between different numbers of codebooks used in the appearance model.

Fig. 17. Illustration of correctly classified samples by our scheme but mis-classified using [3].

polygon and all vertices lying on the polygon. We then get the mean value of all these distances and use the mean distance as the feature for classification. The comparison result is listed in Fig. 15 (a) and demonstrates that the proposed geometry feature performs much better than the baseline in characterizing the standing pattern of the group photo.

For appearance model, we also compare the performance using age information vs without using age information. As illustrated in Fig. 15 (a), experimental results show that age information plays an important role in the final classification result. Without employing age information, the performance is very low. This result is consistent with the findings of previous studies [3]. These results show that the DOGF feature is not only very efficient but also quite effective in measuring the facial similarity in group photo.

As demonstrated with examples in Figs. 7 and 12, we find that the semantic information associated with the context (places, objects, people and activities) in which the two categories of photos are taken provides some insights on whether or not a picture is a family photo. Family photos usually include larger groups and are taken inside homes or, when outdoors, more likely within natural environments. On the other hand, we find that pictures belonging to the non-family category are more likely to be at sporting events or restaurants. From the experimental results, we can observe that while the visual semantic information by itself constitutes the weakest cue to distinguish family photos from other ones, it provides complimentary information with respect to other features and can help improve the recognition performance if integrated in an appropriate fusion scheme. In this paper, we did not consider the family with adopted children. When the family includes adopted children, the appearance model is not reliable for such kinds of photos. However, the geometry and semantic information would still help classify the photo correctly.

5. Conclusion

We have developed a novel framework to automatically classify family photos and non-family photos. Our work introduced multiple contributions: First, we have proposed a novel geometry feature to characterize the social relationship in a group photo. Second, a face descriptor based on a deep neural network architecture is proposed to measure similarities of individuals in a group photo with the goal of estimating their relation. Third, semantic information about the context in which the picture was taken is incorporated into our model to further improve the recognition performance. Furthermore, we have combined our multiple cues in a fusion scheme that can increase the recognition performance by more than 6% compared to each single model, demonstrating that the proposed features can complement to each other. Our method achieves 96.7% accuracy on a dataset, expanded over an existing one, containing thousands of family and non-family pictures collected from social media.

References

[1] D. Li, M.C. Chuah, EMOD: an efficient on-device mobile visual search system, Proceedings of the 6th ACM Multimedia Systems Conference, 2015.
[2] G. Wang, A. Gallagher, J. Luo, D. Forsyth, Seeing people in social context: recognizing people and social relationships, ECCV, 2010. pp. 169–182.
[3] Y.-Y. Chen, W.H. Hsu, H.-Y.M. Liao, Discovering informative social subgraphs and predicting pairwise relationships from group photos, Proceedings of the 20th ACM International Conference on Multimedia, 2012. pp. 669–678.
[4] X. Wang, G. Guo, M. Rohith, C. Kambhamettu, Leveraging geometry and appearance cues for recognizing family photos, 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015. pp. 1–8.
[5] H.-W. Ng, S. Winkler, A data-driven approach to cleaning large face datasets, 2014 IEEE International Conference on Image Processing (ICIP),, 2014. pp. 343–347.
[6] A.C. Gallagher, T. Chen, Understanding images of groups of people, CVPR, 2009.
[7] P. Singla, H. Kautz, J. Luo, A. Gallagher, Discovery of social relationships in consumer photo collections using Markov logic, CVPRW, 2008. pp. 1–7.
[8] H. Poon, P. Domingos, Sound and efficient inference with probabilistic and deterministic dependencies, AAAI, 2006.
[9] C.J. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Disc. 2 (2) (1998) 121–167.
[10] G. Lu, Y. Yan, N. Sebe, C. Kambhamettu, Knowing where I am: exploiting multi-task learning for multi-view indoor image-based localization., British Machine Vision Conference (BMVC), 2014.
[11] A. Bosch, A. Zisserman, X. Munoz, Representing shape with a spatial pyramid kernel, Proceedings of the 6th ACM international Conference on Image and Video Retrieval, ACM. 2007, pp. 401–408.
[12] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[13] D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge, Comparing images using the Hausdorff distance, PAMI 15 (9) (1993) 850–863.
[14] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, 2012. pp. 1097–1105.
[15] H. Chang, Y. Zhou, P. Spellman, B. Parvin, Stacked predictive sparse coding for classification of distinct regions in tumor histopathology, Proceedings of the IEEE International Conference on Computer Vision, 2013.
[16] Y. Zhou, H. Chang, K. Barner, P. Spellman, B. Parvin, Classification of histology sections via multispectral convolutional sparse coding, 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[17] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Advances in Neural Information Processing Systems, 2015. pp. 91–99.
[18] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. pp. 3431–3440.
[19] Y. Zhou, H. Chang, K.E. Barner, B. Parvin, Nuclei segmentation via sparsity constrained convolutional regression, 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), 2015.
[20] Lu, G. Sebe, N. Xu, C. Kambhamettu, C. Memory efficient large-scale image-based localization, Multimedia Tools and Applications, 2015.
[21] G. Lu, V. Ly, C. Kambhamettu, Structure-from-motion reconstruction based on weighted hamming descriptors, 2014 International Joint Conference on Neural Networks (IJCNN), 2014.
[22] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, CVPR, 1, 2001. I–511.
[23] R.L. Graham, F. Frances Yao, Finding the convex hull of a simple polygon, J. Algorithms 4 (4) (1983) 324–331.

[24] T. Horvath, G. Marton, P. Risztics, L. Szirmay-Kalos, Ray coherence between a sphere and a convex polyhedron, Computer Graphics Forum, 11, Wiley Online Library. 1992, pp. 163–172.
[25] S. Meeran, A. Share, Optimum path planning using convex hull and local search heuristic algorithms, Mechatronics 7 (8) (1997) 737–756.
[26] J. Corney, H. Rea, D. Clark, J. Pritchard, M. Breaks, R. MacLeod, Coarse filters for shape matching, IEEE Comput. Graph. Appl. 22 (3) (2002) 65–74.
[27] A. Łukaszewski, A. Szczepkowicz, Computer simulation of FIM images—the convex hull model, Vacuum 54 (1) (1999) 67–71.
[28] J. Hershberger, J. Snoeyink, Cartographic line simplification and polygon CSG formulae in O (n log n) time, Comput. Geom. 11 (3) (1998) 175–185.
[29] C. Burrus, T.W. Parks, DFT/FFT and Convolution Algorithms: Theory and Implementation, John Wiley & Sons, Inc.. 1991.
[30] C. Van Loan, Computational Frameworks for the Fast Fourier Transform, 10. Siam. 1992.
[31] X. Peng, S. Zhang, Y. Yang, D.N. Metaxas, Piefa: personalized incremental and ensemble face alignment, Proceedings of the IEEE International Conference on Computer Vision, 2015. pp. 3880–3888.
[32] Y. Rui, A.C. She, T.S. Huang, Modified Fourier descriptors for shape representation—a practical approach, Proc of First International Workshop on Image Databases and Multi Media Search, Citeseer. 1996, pp. 22–23.
[33] C.T. Zahn, R.Z. Roskies, Fourier descriptors for plane closed curves, IEEE Trans. Comput. 100 (3) (1972) 269–281.
[34] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (4) (2002) 509–522.
[35] C. Xu, Y. Wang, T. Tan, L. Quan, Automatic 3D face recognition combining global geometric features with local shape variation information, Proceedings. Sixth IEEE international Conference on Automatic Face and Gesture Recognition, 2004, 2004. pp. 308–313.
[36] E. Vezzetti, F. Marcolin, 3D human face description: landmarks measures and geometrical features, Image Vis. Comput. 30 (10) (2012) 698–712.
[37] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 2037–2041.
[38] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: closing the gap to human-level performance in face verification, 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. pp. 1701–1708.
[39] Y. Sun, Y. Chen, X. Wang, X. Tang, Deep learning face representation by joint identification-verification, Advances in Neural Information Processing Systems, 2014. pp. 1988–1996.
[40] Schroff, F. Kalenichenko, D. Philbin, J. Facenet: a unified embedding for face recognition and clustering, CVPR, 2015.
[41] X. Wang, R. Guo, C. Kambhamettu, Deeply-learned feature for age estimation, 2015 IEEE Winter Conference on Applications of Computer Vision, 2015.
[42] B.B. Le Cun, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Handwritten digit recognition with a back-propagation network, Advances in Neural Information Processing Systems, Citeseer. 1990,
[43] G.B. Huang, M. Ramesh, T. Berg, E. Learned-Miller, Labeled faces in the wild: a database for studying face recognition in unconstrained environments, Technical Report 07-49, University of Massachusetts, Amherst. 2007.
[44] L. Wolf, T. Hassner, I. Maoz, Face recognition in unconstrained videos with matched background similarity, 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011. pp. 529–534.
[45] G. Guo, G. Mu, K. Ricanek, Cross-age face recognition on a very large database: the performance versus age intervals and improvement using soft biometric traits, 2010 20th International Conference on Pattern Recognition (ICPR), 2010. pp. 3392–3395.
[46] Z. Li, U. Park, A.K. Jain, A discriminative model for age invariant face recognition, IEEE Trans. Inf. Forensics Secur. 6 (3) (2011) 1028–1037.
[47] R. Guo, L. Liu, W. Wang, A. Taalimi, C. Zhang, H. Qi, Deep tree-structured face: a unified representation for multi-task facial biometrics, 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016.
[48] S. Xia, M. Shao, J. Luo, Y. Fu, Understanding kin relationships in a photo, IEEE Trans. Multimedia 14 (4) (2012) 1046–1056.
[49] S. Milborrow, F. Nicolls, Locating facial features with an extended active shape model, ECCV 2008. pp. 504–513.
[50] V. Nair, G.E. Hinton, Rectified linear units improve restricted Boltzmann machines, Proceedings of the 27th International Conference on Machine Learning (ICML-10), 2010. pp. 807–814.
[51] D. Scherer, A. Müller, S. Behnke, Evaluation of pooling operations in convolutional architectures for object recognition, ICANN 2010, Springer. 2010, pp. 92–101.
[52] T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization, Proc. Natl. Acad. Sci. 104 (15) (2007) 6424–6429.
[53] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: convolutional architecture for fast feature embedding, Proceedings of the ACM International Conference on Multimedia, ACM. 2014, pp. 675–678.
[54] G. Guo, G. Mu, Y. Fu, T.S. Huang, Human age estimation using bio-inspired features, CVPR, 2009. pp. 112–119.
[55] G. Guo, X. Wang, A study on human age estimation under facial expression changes, CVPR, 2012.
[56] X. Wang, V. Ly, G. Lu, C. Kambhamettu, Can we minimize the influence due to gender and race in age estimation? 2013 12th International Conference on Machine Learning and Applications (ICMLA), 2, 2013. pp. 309–314.
[57] K. Alsabti, S. Ranka, V. Singh, An efficient parallel algorithm for high dimensional similarity join, Proceedings of the First Merged International and Symposium on Parallel and Distributed Processing 1998, Parallel Processing Symposium, 1998. IPPS/SPDP 1998., 1998. pp. 556–560.
[58] J. Smith, L. Cao, N. Codella, M. Hill, M. Merler, Q. Nguyen, E. Pring, R. Uceda-Sosa, Massive-scale learning of image and video semantic concepts, IBM Journal of Research and Development 59 (2/3) (2015) 7:1–7:13.
[59] M. Merler, B. Huang, L. Xie, G. Hua, A. Natsev, Semantic model vectors for complex video event recognition, Multimedia, IEEE Transactions on 14 (1) (2012) 88–101.
[60] V. Vapnik, The Nature of Statistical Learning Theory, 1999.
[61] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, ACM Trans. Intell. Syst. Technol. (TIST) 2 (3) (2011) 27.
[62] A. Jain, A. Ross, K. Nandakumar, Introduction to Biometrics, Springer. 2011.
[63] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, Fusion strategies in multimodal biometric verification, ICME, 3, 2003. III–5.
[64] S. Ayache, G. Quénot, J. Gensel, Classifier fusion for SVM-based multimedia semantic indexing, Advances in Information Retrieval Springer. 2007, pp. 494–504.
[65] R. Fang, K.D. Tang, N. Snavely, T. Chen, Towards computational models of kinship verification, ICIP, 2010. pp. 1577–1580.
[66] G. Guo, X. Wang, Kinship measurement on salient facial features, IEEE Trans. Instrum. Meas. 61 (8) (2012) 2322–2325.
[67] J. Lu, J. Hu, X. Zhou, Y. Shang, Y.-P. Tan, G. Wang, Neighborhood repulsed metric learning for kinship verification, CVPR, 2012. pp. 2594–2601.
[68] X. Wang, C. Kambhamettu, Leveraging appearance and geometry for kinship verification, 2014 IEEE International Conference on Image Processing (ICIP), 2014.
[69] R. Fang, A.C. Gallagher, T. Chen, A. Loui, Kinship classification by modeling facial feature heredity, ICIP, 2013.
[70] X. Peng, J. Huang, Q. Hu, S. Zhang, A. Elgammal, D. Metaxas, From circle to 3-sphere: head pose estimation by instance parameterization, Comput. Vis. Image Underst. (2015) 92–102.