Image and Vision Computing 30 (2012) 1052–1061

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

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

Face verification of age separated images under the influence of internal and external factors ☆

Gayathri Mahalingam *, Chandra Kambhamettu

Video/Image Modeling and Synthesis Laboratory, Dept. of Computer and Information Sciences, University of Delaware, Newark, DE 19716, United States

ARTICLE INFO

Article history:
Received 24 February 2012
Received in revised form 12 August 2012
Accepted 17 October 2012

Keywords:
Age progression
Face recognition
Face verification
AdaBoost
Local binary patterns

ABSTRACT

In this paper we study the task of face verification of age-separated images with the presence of various internal and external factors. We propose a hierarchical local binary pattern (HLBP) feature descriptor for robust face representation across age. The effective representation by HLBP across minimal age, illumination, and expression variations combined with its hierarchical computation provides a discriminative representation of the face image. The proposed face descriptor is combined with an AdaBoost classification framework to model the face verification task as a two-class problem. Experimental results on the FG-NET and MORPH aging datasets indicate that the performance of the proposed framework is robust with respect to images of both adults and children. A detailed empirical analysis on the effects of internal (age gap, gender, and ethnicity) and external (pose, expressions, facial hair, and glasses) factors in the face verification performance is also studied. The results indicate that the verification accuracy reduces as the age gap between the image pair increases. A quantitative comparison on the effects of gender on verification performance by both humans and the proposed machine learning approach is provided. The analysis indicate that the cues aid humans in verifying image pairs with large age gaps, while it aids machines for all age gaps. However, the cues mislead humans in the case of images of children and extra-personal pairs with large age gaps. Our analyses indicate that the pose and expression variations affect the performance, despite training with such variations, while facial hair and glasses act as discriminative cues. A study on the effects of ethnicity indicate that non-linear algorithms have insignificant effect in performance with the use of both generalized and individual ethnicity models when compared with linear algorithms.

1. Introduction

The human face has been an important modality in biometrics, and face recognition has been an active research area for the past several decades. Face recognition is important due to the breadth of applications such as crowd surveillance, security systems, border control, access control to buildings and other secured areas, identifying missing children, law enforcement, verification of duplicate enrollments, etc. Face verification is a challenging task due to the facial appearance changes, which are mainly caused by age progression besides other internal and external factors. The appearance changes of a face are attributed to shape (e.g. weight loss/gain) and texture changes (e.g. wrinkles, scar, etc.), as age progresses. Besides biological factors, factors such as ethnicity, habits, etc., and external factors such as eyeglasses, facial hair, pose and expression changes, etc. often contribute to the physical changes of the face. A detailed survey of contributions from both psychologists and computer scientists can be found in [1,2].

Current verification systems face challenges due to an inadequate set of images available for a subject across age. This is evident in the case of applications such as identification of missing children, verification of duplicate enrollments, etc. Hence, a verification system should take into account the variations caused by age in order to provide better verification performance on age separated images.

1.1. Background

Face verification across age has not been explored much in the past in spite of its importance in real-world applications. A detailed survey of the effects of aging on face verification tasks can be found in [3,4]. Earlier approaches [5–10] perform recognition by transforming one image to have the same age as the other, or by transforming both the images to reduce the aging effects. Ramanathan and Chellappa [11] proposed an aging model to perform face verification of images under the age 18. Park and Jain [12] also proposed a 3D aging model to perform recognition across age. The authors use a 3D aging model to perform age transformation of the images.

One of the shortcomings of the above mentioned approaches is that the information about the age of the probe image is required in order to perform the age transformation. This information is usually

☆ This paper has been recommended for acceptance by Rama Chellappa.

* Corresponding author.
E-mail addresses: mahaling@udel.edu (G. Mahalingam), chandrak@udel.edu (C. Kambhamettu).

0262-8856/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2012.10.003

not available in real-world applications. Also, the accuracy in age transformation relies on the accuracy of the aging model. Such inaccuracies may result in inappropriate age transformations causing instabilities to these approaches. Hence, we propose a discriminative approach to perform face verification across age progression.

Discriminative approaches proposed in the past [13–17] follow a non-generative approach to perform face verification across age progression. Ramanathan and Chellappa [13] proposed a discriminative approach for face verification across age progression. The authors adapted the probabilistic eigenspace framework and a Bayesian model to learn the differences between intra-personal pairs and extra-personal pairs. Ling et al. [14] also used a discriminative approach for face verification of age separated images. The authors proposed a face representation called *gradient orientation pyramid*, in which the image gradients are computed hierarchically to represent a face image. SVM based classification framework is then used for classification of the image pairs. Zhang et al. [15] and Wang et al. [17] used variants of LBP for extracting facial features and used them in their classification framework to classify image pairs of the same age. Kumar et al. [18] proposed *attribute* and *simile* classifiers which utilize the information from visual cues and perform face categorization in order to perform face verification. Li et al. [19] proposed a Q-stack model to perform face verification across age and head pose variations. Our work differs from the above mentioned approaches in the face representation and the classification framework.

We propose a discriminative approach for the task of face verification of age separated images. A discriminative feature based face representation coupled with a classification framework is proposed. The proposed framework has been applied to two aging databases, which include both internal and external variations in the face images. A detailed analysis on the performance of the proposed approach in comparison with other state-of-the-art approaches under these variations has been discussed in Section 5.

The rest of the paper is organized as follows. The problem formulation and our contributions are discussed in Section 2. The face verification framework is discussed in Section 3. Then, we introduce the hierarchical face representation in Section 4 and also provide a detailed analysis on the hierarchical face representation for the extraction of age invariant patterns. The experimental setup, datasets used, etc. are discussed in Section 5. Experiments to study the effects of internal and external factors are discussed in Sections 6 and 7, respectively. A detailed analysis on the performance of humans as well as the proposed approach is discussed in detail in Section 8. A statistical analysis on the human verification performance is also provided. Finally, Section 9 presents the conclusions and further discussions on this work.

2. Problem formulation

2.1. Face verification framework

Face recognition involves identifying the identity of an individual in the given probe image by comparing it with a gallery of individuals. But, the task of face verification involves identifying whether two images from an image pair belong to the same person or not. This method of verification is suitable for applications such as access control, border control, verification of photo-ID documents, etc. where the validation is performed by verifying the new photo with the old one. Earlier research works [20–22,14] have studied the face verification task as a two-class classification problem, where an image is either classified as intra-personal (belonging to the same individual) or extra-personal (belonging to different individuals).

The aim of this paper is to address the problem of the performance of a face verification system in classifying age separated image pairs under the influence of both internal and external factors. To address this problem effectively, we propose a discriminative face representation and a classification framework. Face verification is performed on two publicly available aging datasets, FG-NET [23] and MORPH [24] which involve more than 3000 and 70,000 intra-personal pairs, respectively.

In our proposed face representation, the uniform local binary pattern (LBP) operator is applied hierarchically to extract the features of a face image. The weak classifiers for every pair of images are then obtained by mapping the extracted features to the LBP feature space. Then we apply the AdaBoost learning algorithm proposed by Yoav et al. [25] to obtain the most discriminant features (strong classifiers) to represent the image pair. The final strong classifier, which combines a few hundreds of weak classifiers, can evaluate similarity between the two images. The entire framework is shown in Fig. 1. Table 1 provides a comprehensive list of various approaches proposed for face verification and their accuracies.

2.2. Contribution

First, we propose an effective face description in which we extend the LBP operator to a hierarchical LBP (HLBP). HLBP for each image is constructed by computing the uniform LBP at every level of the image pyramid and concatenating them together to form the HLBP descriptor. In our approach, we show that the face descriptor provides better performance in verifying age separated image pairs. We obtain near equal performance in verifying images of both adults and children when compared with other approaches. The use of LBP for face description is motivated due to its effective representation across illumination, minimal age, and expression variations as shown in [26,27]. Thus, LBP can be utilized to provide a robust face representation of images with aging effects. Also, the face is comprised of micro-patterns that are well captured by LBP. The effectiveness of the face representation is improved by computing the LBP hierarchically.

Second, we present a detailed study on the verification performance for two publicly available benchmark datasets. We evaluated the performance of the proposed approach with that of a variant of the proposed approach and with the state-of-the-art approach proposed by Ling et al. [14]. In addition, we study the effects of factors such as age gaps between the images, facial hair, pose, expressions, glasses, and ethnicity on the verification performance. A detailed analysis on the feasibility of using discriminant functions for face verification of age-separated images across ethnicity is studied. A comparison on the analysis of the effects of gender from both cognitive psychology and machine learning approach is discussed. The observations from the study are discussed in detail in Section 5.

Third, we present an evaluation of the performance by humans and machines for the task of face verification across various age gaps. We also present a statistical analysis of the performance results of humans in order to illustrate its significance in drawing important conclusions. The motivation behind this study is to analyze the performance of humans in verifying image pairs across age and gender, since humans use many visual cues to recognize faces from images. The face verification task involves images of both children and adults with various age gaps between the image pairs. The performance of

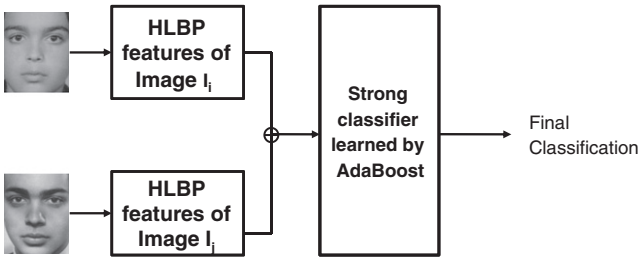

Fig. 1. Flowchart of the proposed face verification framework.

Table 1
A comparison of face verification methods across aging. The results for MORPH are from our experiments.

	Approach	Database (# of subjects, # of images)	EER reported (%)
Ramanathan and Chellappa [13]	Point five faces with Bayesian framework	Private database (200, unknown)	8.5%
Ling et al. [14]	Gradient orientation pyramid with SVM	Passport I (private) (200, unknown)	8.9%
		Passport II (private) (unknown, unknown)	11.2%
		FG-NET (62, 272) (only a subset)	24.1%
		MORPH (5060, 20,140)	29.38%
Our approach	Hierarchical LBP with AdaBoost	FG-NET (82,1002) (entire database)	24.08%
		MORPH (5060, 20,140)	16.49%

both humans and the machine learning approach is analyzed under two scenarios: 1. when the gender information of the subject in the image is provided, and 2. when the gender information of the subject in the image is not provided.

3. Classification framework

As in [20–22,14], we model the face verification task as a two-class classification problem. Face verification is a multi-class problem, which can be converted to a two-class problem by classifying the image pairs as intra-personal and extra-personal. Given two images I_i and I_j, the task is reduced to classifying this image pair as either intra-personal or extra-personal. A feature vector is obtained by mapping the image pair into a feature space, and is given as follows.

$$x = S\left(I_i, I_j\right), \tag{1}$$

where $x \in \Re^d$ is the feature vector from the d-dimensional feature space and the mapping function S is defined as

$$S : I \times I \rightarrow \Re^d \tag{2}$$

where I is the set of all images.

The AdaBoost algorithm is used to classify the feature vectors as belonging to intra-personal pairs and extra-personal pairs. AdaBoost introduced by Yoav et al. [25] is a strong tool to solve a two-class classification problem. The AdaBoost classifier learns a strong classifier by selecting a set of weak classifiers for every iteration from the training data. The final strong classifier is given by,

$$H(x) = \text{sign}(\textstyle\sum \alpha_t h_t(x)), \tag{3}$$

where $H(x)$ is the strong classifier, $\alpha_t \in \Re$ and $\alpha_t = \frac{1}{2} ln\frac{1-\varepsilon_t}{\varepsilon_t}$ where ε_t is the weighted error rate of weak classifier h_t. The final strong classifier represents the boundary in the feature space that separates the intra-personal and extra-personal pairs. In our experiments, we use the GMLAdaBoost library [28].

4. Hierarchical face description

Each face is described by constructing an image pyramid from the face image and computing LBP descriptors from each level of the pyramid. The final LBP descriptor is obtained by concatenating the LBP descriptors at each level of the pyramid.

The original LBP operator proposed by Ojala et al. [29] is a simple but very efficient and powerful operator for texture description. The operator labels the pixels of an image by thresholding the $n \times n$ neighborhood of each pixel with the value of the center pixel and considering the result value as a binary number. Fig. 2 shows an example of the basic LBP operator. The calculation of the LBP labels can be easily done in a single scan of the image. The histogram of the labels of the pixels of the image can be used as a texture descriptor. The gray-scale invariance is achieved by considering a local neighborhood for each pixel and by considering just the sign of the differences in the pixel values instead of their exact values. The LBP operator was then extended by Ojala et al. [30] in which the labels of each pixel are obtained using circular neighborhood having different sizes. The bi-linear interpolation of the pixel values from circular neighborhood allows the usage of any radius and number of pixels in the neighborhood. The LBP operator with P sampling points on a circle of radius R is given by,

$$\text{LBP}_{P,R} = \sum_{p=0}^{P-1} s\left(g_p - g_c\right) 2^p \tag{4}$$

where

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \tag{5}$$

where g_c corresponds to the gray value of the center pixel of the local neighborhood pixels with gray values g_p, $p = 0, \ldots, P-1$. Fig. 3 shows the (4,1) and (8,2) neighborhood circular LBP operator with 4 and 8 sampling points and radii 1 and 2, respectively.

Ojala et al. [31] also introduced another extension to the original operator which uses the property called *uniform patterns* according to which an LBP is called uniform if there exist at most two bitwise transitions from 0 to 1 or vice versa. Uniform patterns represent local micro-patterns of the image such as edges, spots and flat areas. In addition to this, uniform patterns can reduce the dimension of the LBP significantly, which is advantageous for face verification. Invariance to rotation of the face image can be achieved using the idea of rotation invariance proposed by Ojala et al. [29,31] as an extension to the LBP operator. The idea is to rotate the gray values of the neighboring pixels of an image pixel so as to obtain the least binary value for the operator. In our experiments, we use the $\text{LBP}^{u2}_{P,R}$ which is the uniform LBP operator with a window size of 5×5 around each pixel. In addition, we collect the LBP features in a hierarchical way, which has been shown to retain the most visual information as in [32,33].

Given an image $I(x,y)$, where (x,y) indicates pixel locations, we first define the pyramid of I as

$$G_k(I) = I(x, y, k) : k = 0, \ldots, s \tag{6}$$

with

$$G_0(I) = I(x, y, 0) \tag{7}$$

$$G_k(I) = [I(x, y, k-1) \otimes \Phi(x, y)] \tag{8}$$

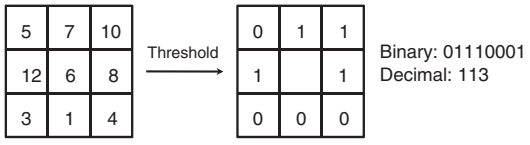

Fig. 2. The basic LBP operator.

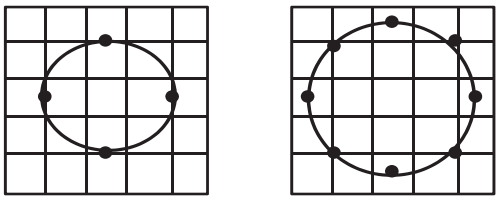

Fig. 3. $LBP_{4,1}$ and $LBP_{8,2}$ circular LBP operators.

where $\Phi(x,y)$ is the Gaussian kernel and s is the number of pyramid levels. The image at level k of the pyramid is obtained by convolving the Gaussian kernel with the image at level $k-1$. The image at each level of the pyramid is divided into blocks of size 8 and the LBP operator is applied to each block of the image at each level to obtain a LBP histogram for each block. The cumulative LBP histogram is obtained by concatenating the LBP histograms of each block of the image. The LBP operator is applied to all the images in the pyramid and a cumulative LBP histogram is obtained for the images at each level of the pyramid. The LBP pyramid of the image I is defined as follows;

$$L(I_0) = [\mathrm{LBP}(G_0(I)); \mathrm{LBP}(G_1(I)); \ldots; \mathrm{LBP}(G_s(I))] \tag{9}$$

where $L(I) \in \Re^{d \times s}$ which maps the image I into a $d \times s$ representation, where d is the length of the cumulative LBP histogram obtained from the image at each level of the pyramid. Fig. 4 illustrates the computation of a LBP pyramid from an image.

4.1. Kernels between HLBPs

Given an image pair (I_i, I_j) and corresponding LBP pyramids $L(I_i)$ and $L(I_j)$ (representing the LBP patterns from all scales of the pyramid), the feature vector x is given by

$$x = S\left(I_i, I_j\right) \tag{10}$$

where S is defined as the inner product between the LBP histograms of all the image blocks at all levels of the pyramid and given by

$$x = S\left(I_i, I_j\right) = \left(L(I_i) \cdot L\left(I_j\right)\right) \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{s \times 1} \tag{11}$$

where $\cdot$ is the Hadamard product between matrices. It is to be noted that $L(I_i)$ and $L(I_j)$ are $d \times s$ matrices, whose dot product gives a $d \times s$ representation whose inner product with an identity vector of dimension $s \times 1$ produces a d dimensional vector (i.e. x).

4.2. Age invariant patterns using hierarchical LBP

In order to establish similarity (or dissimilarity in case of extra-personal pairs) between two age separated images, it is important to extract age invariant texture patterns from both the images. It has been shown that discriminatory information such as the distribution of the edge direction in the face image is an age invariant feature [34]. A dense sampling using LBP local descriptors allows the extraction of such discriminatory information. Computing the LBP from a multi-scale image allows for the dense sampling of the texture patterns from an image. The extracted multi-scale LBP features are well suited for age invariant face verification as supported by our experimental results.

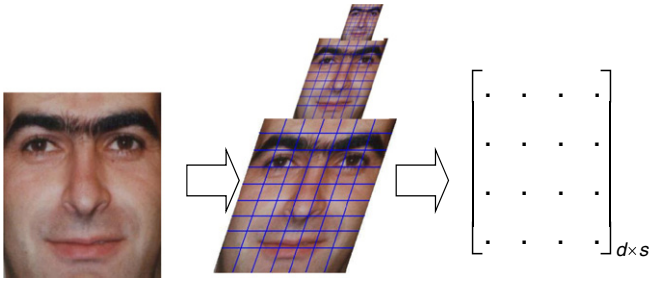

Fig. 4. Computation of LBP pyramid from an image.

There are two approaches using which a multi-scale LBP can be computed. In the first approach, the LBP codes are extracted for each pixel with different radii. The second approach involves extracting LBP codes from different image scales. However, the former approach has its own shortcomings due to the use of conventional LBP and is detailed as follows.

First, the conventional LBP methods extract only the micro structures (edges, corners, spots, etc.) of the images, while the HLBP extracts both micro and macro structures [35]. Texture classification algorithms [36,37] have shown that texture patterns cannot be discriminated with only micro patterns, but requires both micro and macro structures. Second, the stability of the LBP values decreases with the increase in the neighborhood radii (first method of multi-scale LBP extraction). This is due to the minimal correlation of the sampling points with the center pixel. Further, from a signal processing point of view, the sparse sampling adapted by the LBP operators from large neighborhood radii may not result in an adequate representation of the two-dimensional image signal. Also, aliasing effects are an obvious problem.

The second method of computing the multi-scale LBP generates multi-scale image structure which has been used in the past for effective texture analysis and matching which requires both the micro and macro structures to be extracted. Thus the multi-scale approach allows capturing both micro and macro texture patterns by computing LBP at different scales. These observations are verified from the equal error rates (EER) (shown in Table 2) from the 5-fold cross validation face verification experiments on FG-NET and MORPH datasets. The EER is defined as the error rate when the true acceptance rate (TAR) and true rejection rate (TRR) are equal, where TAR and TRR are given by,

$$\mathrm{TAR} = \frac{\#\ \text{of truly accepted intra} - \text{personal pairs}}{\#\ \text{of total intra} - \text{personal pairs}} \tag{12}$$

$$\mathrm{TRR} = \frac{\#\ \text{of truly rejected extra} - \text{personal pairs}}{\#\ \text{of total extra} - \text{personal pairs}} \tag{13}$$

where *accept* indicates that the images are from the same subject, and *reject* indicates that the images are from different subjects. The EERs from Table 2 indicate the effectiveness of HLBP when compared with LBP having large radii.

4.3. Performance analysis of hierarchical LBP

In order to show the effectiveness of the proposed hierarchical LBP face representation, we performed face verification experiments on the FG-NET database using two frameworks. The first variant uses HLBP with AdaBoost classification framework and the second variant uses HLBP with random forest (RF) classification framework. The motivation behind this experiment is to determine the robustness of the

Table 2
Experimental comparison on the performance of LBP (with various P and R values) and HLBP on FG-NET dataset.

LBP ($P=8, R=2$)	LBP ($P=16, R=2$)	LBP ($P=16, R=4$)	HLBP ($P=8, R=1$)
31.29%	35.78%	41.22%	24.08%
20.34%	26.34%	27.53%	16.49%

HLBP with different classification frameworks. Table 3 shows the EER obtained from the above mentioned approaches.

There are several observations made from the experimental results. First, it can be seen that the performance of both the classifiers is nearly equal which indicates the effectiveness of HLBP in face representation under different classification frameworks. The advantage of the HLBP is that it allows a better description of the features due to the feature extraction at different image scales. The extraction of LBP at each level of the hierarchy proves useful when images with different resolutions need to be verified.

Unlike the generative methods (see Section 1.1), the proposed framework does not require prior information (age, etc.) about the images. Also, the proposed framework is a discriminative approach which does not involve age estimation or age simulation in order to verify the image pairs. This potentially avoids the instabilities introduced by these processes in the face verification task.

5. Experiments and results

5.1. Experimental setup

5.1.1. Datasets

The face verification experiments are conducted on the FG-NET [23] and the MORPH [24] aging databases. Both FG-NET and MORPH are publicly available databases, which include images across ethnicity (mostly Caucasians and African Americans), age, gender, pose, illumination, expression, occlusions, facial hair, etc.

FG-NET includes 1002 images from 82 subjects with an age range of 0–69 years and an average of 12 images per subject. FG-NET includes real-world images from limited number of subjects (82) taken under uncontrolled conditions. Also, the largest age gap between an image pair of a subject is 45 years. These properties of FG-NET make it a challenging database for our experiments. Hence, we use the entire database (1002 images) in our experiments. Table 4 shows the statistics of the FG-NET database.

The MORPH aging database [24] is a publicly available database and consists of two *Albums* of images. Album1 consists of 1690 digitally scanned images of 631 subjects with an average of 4 images per subject. Album 2 consists of more than 20,000 digital images of more than 4000 individuals with an average of 4 images per subject.

5.1.2. Preprocessing

The images undergo a preprocessing stage cropping the face region and resizing it to 128 pixels. An image pyramid consisting of 3 scales (levels) is constructed and the LBP is computed at all the three levels. From our experiments, we deduce that effective, age invariant texture patterns can be extracted from the three levels of the image pyramid for images of both adults and children. No other preprocessing such as pose correction, normalization, etc. is performed.

5.1.3. Approaches

The face verification performance of the proposed approach is compared with LBP + AdaBoost, and SVM + GOP [14] approaches. The LBP + AdaBoost is a variant of the proposed approach where LBP is computed only at the finest scale.

Table 3
Performance analyses on the hierarchical LBP face descriptor. The table shows the average equal error rates (EER) from a 5-fold cross validation face verification experiment using different classification frameworks on the FG-NET database.

	HLBP + AdaBoost	HLBP + RF
FG-NET	24.08%	25.59%

Table 4
Statistics on the FG-NET database used in face verification task. "Std. age" represents the standard deviation of the age.

# of subjects	# of images	# of intra pairs	Mean age	Std. age
82	1002	5805	15.8	12.8

5.1.4. Experimental evaluation

We conducted experiments to analyze the effects of both internal (age gap, images of children and adults, and ethnicity) and external factors (pose, expressions, facial hair, and glasses) that affect the face verification performance, with the main focus on age separated images. A protocol similar to the one in [14] is followed for experiments on FG-NET. A 5-fold cross validation experiment is conducted for each database. The performance of the approaches is evaluated using the TAR–TRR curves (ROC curves). The various points on the ROC curve are computed by varying the classifier parameters. For AdaBoost, the ROC curves are generated by varying the *tetta*, a threshold parameter which varies the confidence score of the classifier. The average of the ROC curves from the 5-fold cross validation experiments is considered as the average performance of the approach. The average EER is computed in a similar way for each experiment.

6. Effects of internal factors

6.1. Effects of aging

We perform 5-fold cross validation face verification experiments on the FG-NET and the MORPH datasets. The effects of aging (both children and adults) and the effects of age gap between the images in the image pair are analyzed through these experiments. For the face verification experiment, we generated 5800 and 78,735 intra-personal pairs from the FG-NET and the MORPH datasets, respectively. An equal number of extra-personal pairs are randomly generated from images of different subjects. The training and testing pairs in each fold of the cross-validation are mutually exclusive. That is, the images from the same subject do not appear in both training and testing image pairs. The number of intra-personal and extra-personal pairs is equally divided among folds.

Figs. 5 and 6 show the TAR–TRR curves obtained from all the approaches mentioned in Section 5.1. It is evident from the results that the proposed approach outperforms all others. It is also evident that the hierarchical representation is effective for face verification tasks. This can be observed from the ROC curves of HLBP + AdaBoost and LBP + AdaBoost approaches. Since the entire database is used, it can be seen that the system effectively handles images of children and adults. Also, the system effectively handles image pairs with large age gaps. It is to be noted that preprocessing in terms of pose correction and normalization is not performed with these images. Table 5 shows the average EERs obtained for various approaches on the FG-NET and the MORPH dataset.

6.2. Effects of aging in children and adults

Facial changes due to aging are mainly manifested in terms of shape and texture variations [11,13]. However, these changes are manifested at different rates for different age groups. A child's face undergoes major shape changes, but minimal texture changes, while an adult's face undergo minimal shape changes than texture changes [11]. Hence, it is interesting to evaluate the above mentioned approaches on images of children and adults separately. This section discusses the face verification experiments on both images of children and adults from the FG-NET database.

The FG-NET database is divided into two subsets. The first subset (*children*) included 640 images from 67 subjects with age less than

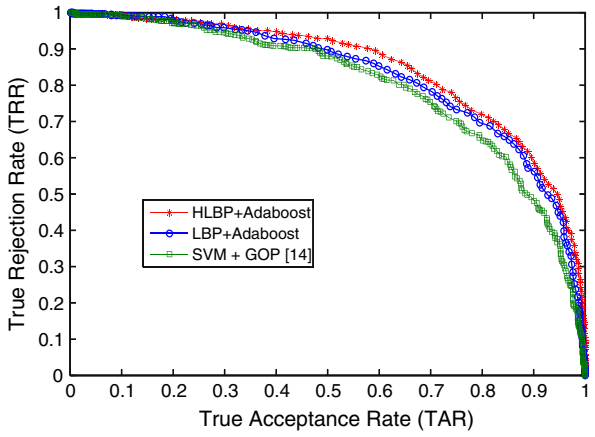

Fig. 5. TAR–TRR curves for face verification under the effects of aging on the entire FG-NET database.

Table 5
Effects of internal factors. The table shows the average EER obtained from the 5-fold cross validation experiments on the FG-NET and MORPH datasets.

Internal factor	Approach		FG-NET	MORPH
Aging	SVM + GOP		26.8%	29.38%
	LBP + AdaBoost		25.9%	24.52%
	HLBP + AdaBoost		24.08%	16.49%
Aging in children	SVM + GOP		34.27%	–
	LBP + AdaBoost		25.68%	–
	HLBP + AdaBoost		23.20%	–
Aging in adults	SVM + GOP		24.20%	29.38%
	LBP + AdaBoost		23.98%	24.52%
	HLBP + AdaBoost		20.72%	16.49%
Effect of age gaps	LBP + AdaBoost	0–2 years	37.54%	20.47%
		3–5 years	40.72%	27.88%
		6–8 years	41.64%	44.73%
		9–11 years	42.45%	39.68%
	HLBP + AdaBoost	0–2 years	30.83%	17.54%
		3–5 years	34.76%	18.40%
		6–8 years	35.08%	26.41%
		9–11 years	37.52%	30.26%

18 years and the second subset (*adults*) included 362 images from 61 subjects with age greater than 18 years. We generated 5718 intra-personal pairs for the first set and 2328 intra-personal pairs for the second subset of images. An equal number of extra-personal pairs is randomly generated for each subset to avoid the bias in training the classifiers. Face verification task is performed using a 5-fold cross validation and the average EERs and TAR–TRR curves are reported in Table 5 and Figs. 7 and 8.

The following are the observations made from our experiments. First, the performance of the approaches is lower for childrens' images when compared with the performance for adult images. It is a common observation that the task of face verification is harder for children than adults. This is due to the fact that a face profile undergoes large variations in shape before age 18 (as stated in [14]).

It can be seen that the performance of the proposed approach and its variant does not vary significantly between the two experiments (children and adults). This suggests that LBP provides an effective representation across age groups, which is improved further by computing the LBP at different scales.

6.3. Effects of age gaps

The motivation behind this study is to analyze the effect of age gaps between images in verification performance. Both the FG-NET and the MORPH databases are used for this experiment. The intra-personal image pairs are categorized into four categories based on the age gap between the images. We follow the categorization similar to [14]. The four categories include age gaps from 0 to 2 years, 3 to 5 years, 6 to 8 years, and 9 to 11 years. Face verification experiments are conducted for each category using a 5-fold cross validation. All the intra-personal pairs and an equal number of randomly selected extra-personal pairs are used to generate the folds. Table 6 shows the number of intra-personal pairs generated for each category for the FG-NET database.

The experiment is conducted using the approaches, HLBP + AdaBoost and the LBP + AdaBoost. The average of the equal error rates from each fold is used to evaluate the performance of the approaches. Table 5 shows the average EER obtained for various approaches on both FG-NET and MORPH databases for various age gaps between the images. Figs. 9 and 10 show the performance of the experiments on all four groups on FG-NET and MORPH, respectively. From the plots, it can be seen that the difficulty in verification increases with an increase in the age gap between the images. However, the rate of increase in the equal error rate reduces with an increase in the age gap. This can be observed from the equal error rates of the two different representations of the proposed approach. In addition, we observed that this rate of increase in the EER varies across datasets. This is due to the variation in the number of

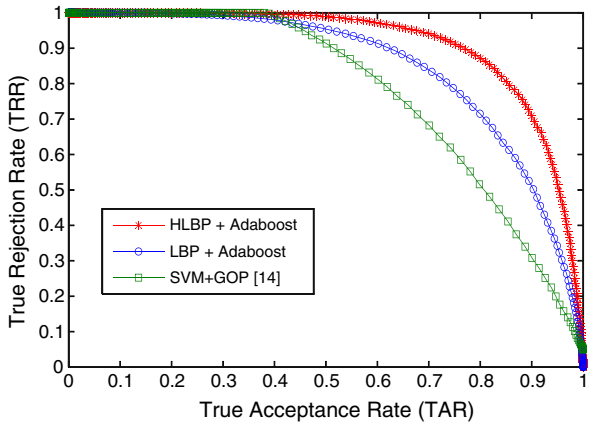

Fig. 6. TAR–TRR curves for face verification under effects of aging on the entire MORPH database.

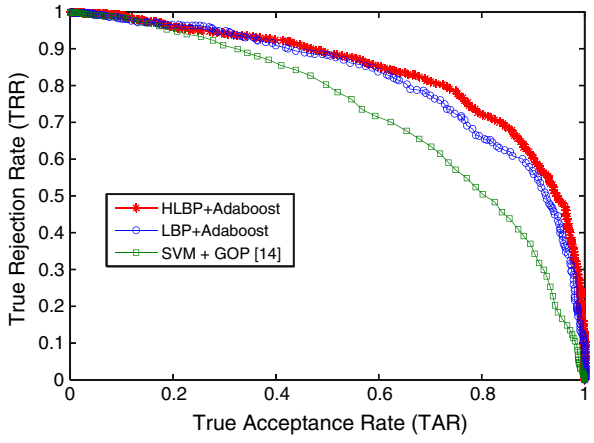

Fig. 7. TAR–TRR curves for face verification experiments showing the effects of aging in children from the FG-NET database.

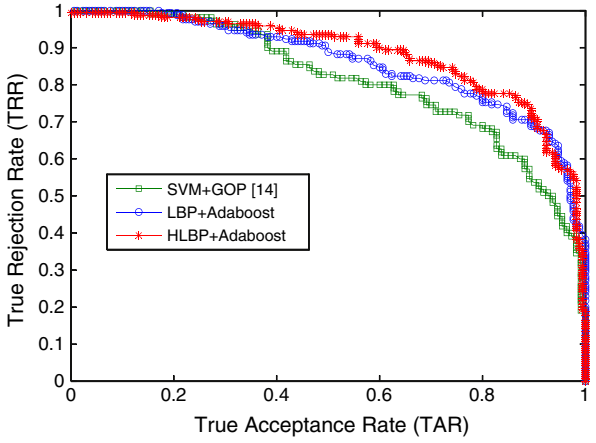

Fig. 8. TAR–TRR curves for face verification experiments showing the effects of aging in adults from the FG-NET database.

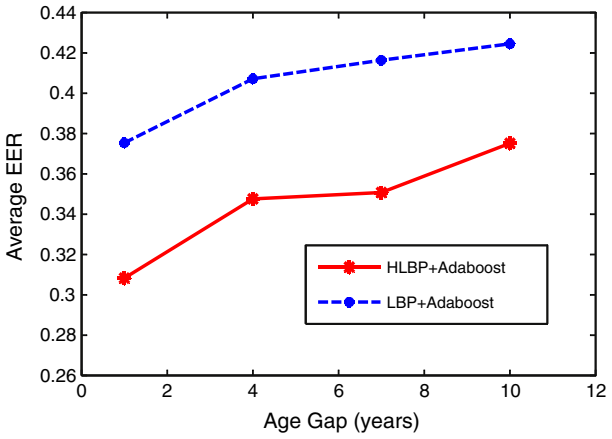

Fig. 9. Effects of age gaps on verification performance for the FG-NET database. The curves are shifted slightly along the x-axis for better illustration.

intra-personal and extra-personal pairs in each age group for a database.

6.4. Effect of ethnicity

Ethnicity is one other internal factor that can have a significant influence on the performance of the verification system. This is due to various anthropometry changes in the face across ethnicities. To evaluate the effects of ethnicity, we performed face verification experiments using the following ethnicities; 1. Caucasian (FG-NET [23]), 2. African American (MORPH [24]), 3. Hispanic (MORPH), 4. Indian (Indian face DB [38]), and 5. Japanese (JAFFE [39]).

The verification performances of both linear (PCA) and non-linear (HLBP + AdaBoost and LBP + AdaBoost) algorithms are evaluated under the following scenarios.

- *Exp1*: Training is performed with one ethnicity and tested on unknown ethnicities. Five classifiers, one for each ethnicity are trained and are used for face verification.
- *Exp2*: Training is performed on a mixture of ethnicities (cross-race) and tested on known and unknown ethnicities. Each classifier is trained with 4 ethnicities randomly selected in a leave-one-out fashion.

The motivation behind the design of these experiments is to evaluate the significance of *cross-race effect* [40] in face verification with both linear and non-linear algorithms. We generated 16,944 intra-personal pairs and an equal number of randomly selected extra-personal pairs. Table 7 shows the average EERs from all the approaches for both the experiments. The following observations are deduced from the EERs;

- The error rates from *Exp2* when compared with *Exp1* indicate a decrease in performance when the classifier is trained with image pairs from every ethnic group. This is due to the lack of an accurate generalization of the facial features across ethnicity. This also suggests that it is important to focus on individual models instead of generalized models for face verification/recognition.
- It can be seen that the non-linear algorithms provide better performance when compared with the linear algorithms in the case of both individual and generalized ethnicity models. This is consistent with the observation that non-linear functions perform better when compared with linear discriminant ones with sufficient training data [41].

Table 6
# of intra-personal pairs generated for age gap categories on FG-NET and MORPH database.

Age gap (years)	FG-NET	MORPH
0–2	846	123,444
3–5	1160	11,848
6–8	937	792
9–11	2862	956

7. Effect of external factors

Besides age related variations, the presence of external factors on the face image of a subject can affect the verification performance. In order to evaluate the effects of these external factors, we conducted face verification experiments on four subsets of the FG-NET database. Each subset includes images with pose, expression, eyeglasses, and facial hair variations, respectively. Two kinds of experiments were conducted in which the classifier is trained with (*Experiment1*) and without (*Experiment2*) these variations. Table 8 shows the average EER for both *Experiment1* and *Experiment2* under the influence of the external factors.

It is evident from the table that the performance is improved when the classifier is trained with images involving these variations. In particular, we can see that the facial hair and glasses act as discriminative cues thus providing a better performance. This also indicates

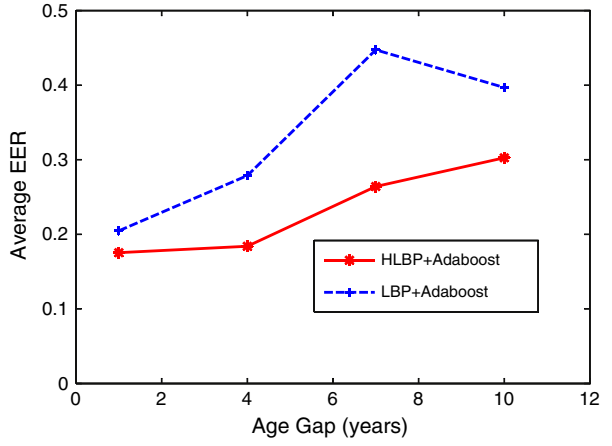

Fig. 10. Effects of age gaps on verification performance for MORPH database. The curves are shifted slightly along the x-axis for better illustration.

Table 7
Effects of ethnicity. Average EER from face verification experiments showing the effects of individual and generalized ethnicity training models.

Approach	*Exp1*	*Exp2*	
		Test on known ethnicity	Test on unknown ethnicity
PCA	35.08%	44.64%	47.08%
LBP + AdaBoost	29.73%	30.77%	34.83%
HLBP + AdaBoost	23.12%	24.76%	27.57%

that the learning process of the classifier is influenced by these cues, which can be a disadvantage if the testing images do not include glasses or facial hair. However, the effects of expressions and pose variations are significant when compared with the effects of facial hair and glasses. This can be observed from the respective EERs of *Experiment1*. This is due to the lack of all possible expressions or poses for training purposes. Hence the classifier may not be accurately trained for all possible expressions or pose variations. The results from *Experiment2* show that the verification task becomes extremely difficult when the classifier is not trained with images including these variations.

8. Human vs. machine learning evaluation of face verification

Humans are backed with knowledge base allowing them to use and interpret multiple information in recognizing faces. Human perception studies suggest that humans utilize visual cues more than discriminative cues to recognize face images [42]. Hence, the use of such cues in recognizing/verifying face images by humans can provide an insight on the use of these cues in the face verification task. In order to evaluate the performance of humans, with and without the presence of discriminative cue (gender is used in our experiments), we conducted cognitive psychology (CP) experiments involving 86 human subjects. The human subjects are psychology students (Caucasians and African Americans) with no knowledge about the datasets and the problem of face verification. This allows for cross-race training and testing. The gender information is provided as a hint which can be availed by the subject, if needed.

A similar face verification experiment is performed using the proposed machine learning approach (with and without gender information). The gender information is used to prune the search space for the classifier to perform verification. In order to provide a fair comparison between the machine learning approach and the human evaluation approach, 188 image pairs were selected from all age groups (baby, child, adult, and senior) and all age gap categories from both FG-NET and MORPH, and used for testing.

Table 9 shows the error rates in face verification from both the human experiment and the proposed machine learning approach with a 95% confidence level. The image pairs are categorized into five groups based on the age gap between the images. The following observations are made from the verification results.

Table 8
Effects of external factors. The table shows the average EER from a 5-fold cross validation face verification experiment on subsets of the FG-NET dataset.

External factor	Avg. EER when trained on images with external factors	Avg. EER when trained on images without external factors
Expression	23.40%	48.23%
Pose	24.57%	49.70%
Facial hair	12.09%	46.54%
Eyeglasses	10.14%	48.59%

Table 9
Face verification error rates from the cognitive psychology and the proposed approach with and without the use of discriminative cue (gender). The error rates are computed at 95% confidence level.

Age gap (years)	Human experiment		Proposed approach	
	Error rate (no hints)	Error rate (with hints)	Error rate (no hints)	Error rate (with hints)
0–2	15.28%	19.05%	24.18%	22.41%
3–5	34.17%	39.37%	26.58%	23.72%
6–8	37.48%	42.18%	30.75%	26.37%
9–11	30.74%	27.32%	33.89%	29.43%
>11	30.91%	29.31%	34.42%	30.28%

- The use of discriminative cues provides improved performance by the proposed approach when compared with the performance of the proposed approach without the use of these cues. The improvement in performance is achieved in terms of accuracy, lower time requirements, and graceful degradation of the search space, which allows for matching with only a subset of images from the gallery, where the subset of images is selected based on the discriminative cue of the probe image. This eventually reduces the execution time and also the computational cost, since the probe is compared with only a subset of the gallery.
- The cues aid humans in verifying image pairs with an age gap greater than 8 years. However, humans tend to mis-classify image pairs with an age gap less than 8 years. Also, we observed that the misclassified image pairs included images of subjects with an age gap less than 16 years. This indicates that the task of verifying images of children is difficult for humans as well. Also, humans tend to mis-classify images from different subjects as the age gap increases even with the presence of discriminative cues. Fig. 11 shows some sample intra-personal and extra-personal image pairs that were misclassified by humans.
- The performance of the machine learning approach is comparable with the performance of humans in verification of image pairs with large age gaps. It can be seen that the performance of the machine learning approach is significantly better with the use of gender information when compared with the performance by humans with the use of the same gender information.
- From Table 9, we note that the error rate achieved from human verification performance for the age gap range of 6–8 years is high even with the use of hints, when compared with the error rates obtained without the use of hints for the same age range. An observation of the image pairs for that age gap range shows that humans were misled in classifying extra-personal image pairs. This indicates that gender information did not prove useful for humans in verifying extra-personal image pairs. Also, the results indicate the presence of *cross-race effect*, which shows that humans find it difficult to identify faces from a race different from their own. Fig. 11 shows some sample image pairs that were misclassified by humans.

8.1. Statistical analysis of human evaluation

In order to evaluate the data obtained from the CP experiment, the data is analyzed using the hierarchical linear modeling (HLM) [43] also known as multi-level modeling. HLMs are suitable for data with many levels and allows for more accurate estimation of effects in situations where some of the data is "nested". For example, in the case of the CP experiment, it accounts for how 100 evaluations (say) of faces by one human subject are likely more related than 100 evaluations across multiple participants. Also, the images are collected across multiple levels such as gender, ethnicity, pose and expression variations.

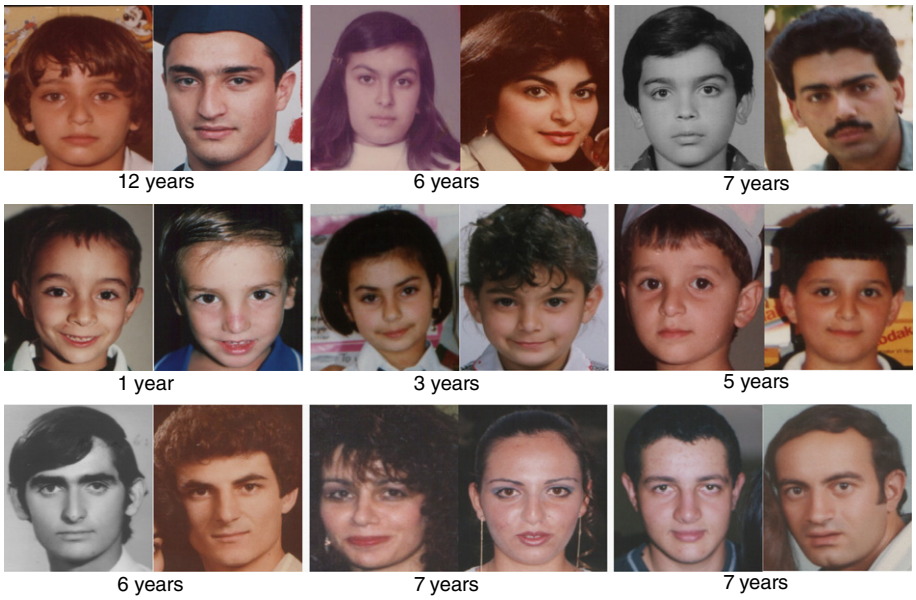

Fig. 11. Sample set of image pairs used in the cognitive psychology experiment. First row shows intra-personal pairs. Second and third rows show extra-personal pairs. The age gap between the images is listed below each pair. All the image pairs shown are mis-classified by humans.

A level 2 regression is followed in order to facilitate the analysis based on gender and ethnicity effects. A random intercepts model is followed where the regression equations are given by,

$$\begin{aligned}\beta_{0j} &= \pi_{00} + \pi_{01} W_j + u_{0j} \\ \beta_{1j} &= \pi_{10} + u_{1j}\end{aligned} \tag{14}$$

where, β is the intercept, π is the slope between the dependent variable and the corresponding level predictor, W refers to the level 2 predictor, and u refers to the standard error component for deviation of the intercept of a group from overall intercept (π_{00}). π_{01} refers to the age gap between the images, and π_{10} refers to the various age groups (children, adult, etc.).

The motivation behind the analysis is to determine whether the gender information helped humans in verifying image pairs across various age gaps and also to verify the statistical significance of the results from the CP experiment.

Table 10 shows the parameter values obtained from regression on the model. It can be observed from the parameter values (standard error in particular) that people are more likely to classify the image pairs correctly with the help of gender information as the age gap increased. This can be observed from the low standard error for π_{01} when compared with π_{00} (less age gap).

9. Conclusion and discussion

In this paper, we studied the problem of face verification with age variations using discriminative methods. Face image is holistically represented using the hierarchical local binary pattern feature descriptor. The LBP provides an effective representation across minimal age variations, illumination, and minimal pose variations, which makes it a suitable descriptor for description of images across age. The spatial information is incorporated by combining the LBP at each level of the Gaussian pyramid constructed for each face image. We presented an AdaBoost classifier that identifies the intra-personal and extra-personal image pairs across age gaps. Experiments on the FG-NET and MORPH database provided an insight on several factors that affect the performance of a face verification system.

From our experiments, we deduce that it is crucial for a face verification system to accurately learn the facial changes due to age progression for robust performance. There are several factors, both internal and external that affect the performance of a face verification system. The primary changes in the face of an individual are caused by shape and texture changes. While these changes attribute to internal factors, external factors such as pose, expressions, facial hair, glasses, etc. also affect the performance of a verification system. Our results indicate that variations in pose and expressions have a significant effect in the performance in spite of learning these variations. Our experiments on the effects of ethnicity suggest that the performance reduces when trained with multi-ethnic groups. The reduction in performance is observed when compared with the performance on training with an ethnic group and testing with the same/different ethnic group.

Besides these internal and external factors, biological factors such as gender, ethnicity, weight gain/loss, etc. also have a significant effect on the performance. The evaluation of performance by humans and machine learning approach under the presence of gender information suggests that factors such as gender, age group, ethnicity, etc. can act as discriminative cues and thus provide an improved performance in terms of accuracy, lower time requirements, and graceful degradation of the search space.

Table 10
HLM statistical analysis on the CG face verification experiment results. The parameter values from HLM are presented in the table.

Parameters	Coefficient (π)	Standard error (u)
Intercept β_{00} and π_{00} (overall intercept)	−0.9528	0.0529
Intercept β_{10} and π_{01} (age gap)	0.0169	0.0022

References

[1] N. Ramanathan, R. Chellappa, S. Biswas, Computational methods for modeling facial aging: a survey, J. Vis. Lang. Comput. 20 (3) (2009) 131–144.
[2] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Comput. Surv. 35 (4) (2003) 399–458.

[3] A. Lanitis, A survey of the effects of aging on biometric identity verification, Int. J. Biometrics 2 (1) (2010) 34–52.
[4] J. Zeng, H. Ling, J. Latecki, S. Figzhugh, G.-D. Guo, Analysis of facial images across age progression by humans, ISRN Machine Vision, (2012).
[5] J. Suo, X. Chen, S. Shan, W. Gao, Learning long term face aging patterns from partially dense aging databases, in: Proc. of the Intl. Conf. on Computer Vision, 2009, pp. 622–629.
[6] J. Suo, S. Zhu, S. Shan, X. Chen, A compositional and dynamic model for face aging, TPAMI 9 (5) (2009) 385–401.
[7] J. Suo, F. Min, S. Zhu, S. Shan, X. Chen, A multi-resolution dynamic model for face aging simulation, Comput. Vis. Pattern Recognit. (2007) 17–22.
[8] A. Lanitis, C.J. Taylor, T.F. Cootes, Toward automatic simulation of aging effects on face images, TPAMI 24 (4) (2002) 442–455.
[9] M. Tiddeman, B. Burt, D. Perrett, Prototyping and transforming facial texture for perception research, Comput. Graph. Appl. (2001) 42–50.
[10] R. Singh, M. Vatsa, A. Noore, S.K. Singh, in: Age Transformation for Improving Face Recognition Performance, Springer-Verlag, 2007, pp. 576–583.
[11] N. Ramanathan, R. Chellappa, Modeling age progression in young faces, Comput. Vis. Pattern Recognit. (2006) 387–394.
[12] U. Park, Y. Tong, A.K. Jain, Face recognition with temporal invariance: a 3d aging model, in: Proc. of Intl. Conf. on Automatic Face and Gesture Recognition, 2008, pp. 1–7.
[13] N. Ramanathan, R. Chellappa, Face verification across age progression, IEEE Trans. Image Process. 15 (11) (2006) 3349–3362.
[14] H. Ling, S. Soatto, N. Ramanathan, D. Jacobs, Face verification across age progression using discriminative methods, IEEE Trans. Inf. Forensics Secur. 5 (1) (2010) 82–91.
[15] G. Zhang, X. Huang, S.Z. Li, Y. Wang, X. Wu, Boosting local binary pattern (LBP)-based face recognition, in: Advances in Biometric Person Authentication, 2004, pp. 179–186.
[16] G.-D. 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, in: Intl. Conf. on Pattern Recognition, 2010, pp. 3392–3395.
[17] X. Wang, C. Zhang, Z. Zhang, Boosted multi-task learning for face verification with applications to web image and video search, Comput. Vis. Pattern Recognit. (2009) 142–149.
[18] N. Kumar, A. Berg, P. Belhumeur, S. Nayar, Describable visual attributes for face verification and image search, TPAMI 33 (10) (2011) 1962–1977.
[19] W. Li, A. Drygajlo, H. Qiu, Combination of age and head pose for adult face verification, in: Automatic Face and Gesture Recognition, 2011, pp. 77–82.
[20] B. Moghaddam, W. Wahid, A. Pentland, in: Beyond Eigenfaces: Probabilistic Matching for Face Recognition, 1998, pp. 30–35.
[21] K. Jonsson, J. Kittler, Y. Li, J. Matas, Support vector machines for face authentication, Image Vis. Comput. 20 (5–6) (2002) 369–375.
[22] P.J. Phillips, Support vector machines applied to face recognition, in: Advances in Neural Information Processing Systems 16 (NIPS), 2, 1999, pp. 803–809.
[23] Face, gesture recognition working group, FG-NET aging database, (2002).
[24] K. Ricanek, T. Tesafaye, MORPH: a longitudinal image database of normal adult age-progression, in: IEEE Intl. Conf. on Automatic Face and Gesture Recognition, 2006, pp. 341–345.
[25] F. Yoav, R.E. Schapire, A short introduction to boosting, J. Jpn. Soc. Artif. Intell. (1999) 771–780.
[26] T. Ahonen, A. Hadid, M. Pietikainen, Face recognition with local binary patterns, in: European Conference on Computer Vision, 2004, pp. 469–481.
[27] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, TPAMI 28 (12) (2006) 2037–2041.
[28] Y. Freund, R.E. Schapire, Game theory, on-line prediction and boosting, in: Proc. of the 9th Annual Conference on Computational Learning Theory, 1996, pp. 324–332.
[29] T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on feature distributions, Pattern Recognit. (1996) 51–59.
[30] T. Ojala, M. Pietikainen, T. Maenpaa, Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, TPAMI (2002) 971–987.
[31] T. Ojala, M. Pietikainen, T. Maenpaa, A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification, in: Second Intl. Conf. on Advances in Pattern recognition, 2001, pp. 397–406.
[32] L. Alvarez, F. Guichard, P.L. Lions, J.M. Morel, Axioms and fundamental equations of image processing, Arch. Ration. Mech. Anal. (1993) 199–257.
[33] D.K. Hammond, E.P. Simoncelli, Nonlinear image representation via local multiscale orientation, in: Tech. Rep., Courant Institute Technical Report, New York University, 2005.
[34] Z. Li, U. Park, A.K. Jain, A discriminative model for age invariant face recognition, IEEE Trans. Inf. Forensics Secur. (2011) 1028–1037.
[35] S.W. Lee, S.Z. Li, Learning multi-scale block local binary patterns for face recognition, in: LNCS 4642, ICB '07, 2007, pp. 828–837.
[36] Y. He, N. Sang, C. Gao, Pyramid-based multi-structure local binary pattern for texture classification, in: Proc. 10th Asian Conference on Computer Vision — Volume Part III, ACCV '10, Springer-Verlag, Berlin, Heidelberg, 2011, pp. 133–144, URL http://dl.acm.org/citation.cfm?id=1966049.1966061.
[37] T. Mäenpää, M. Pietikäinen, Multi-scale binary patterns for texture analysis, in: Proc. 13th Scandinavian Conference on Image analysis, SCIA '03, 2003, pp. 885–892.
[38] Indian face database, vis-www.cs.umass.edu/vidit/AI/dbase.html.
[39] Japanese female facial expression database, www.kasrl.org/jaffe.html.
[40] C.A. Meissner, J.C. Brigham, Thirty years of investigating the own-race bias in memory for faces: a meta-analytic review, Psychol. Public Policy Law 7 (2001) 3–35.
[41] J. Bekios-Calfa, J.M. Buenaposada, L. Baumela, Revisiting linear discriminant techniques in gender recognition, TPAMI 33 (4) (2011) 858–864.
[42] V. Bruce, A. Young, In the Eye of the Beholder: The Science of Face Perception, 2000.
[43] S. Raudenbush, A. Bryk, Hierarchical Linear Models (second edition), Thousand Oaks: Sage Publications, (2001).