Journal of Biometrics & Biostatistics
ISSN: 2155-6180

El Khiyari and Wechsler, J Biom Biostat 2016, 7:4
DOI: 10.4172/2155-6180.1000323

Research Article Open Access

Face Verification Subject to Varying (Age, Ethnicity, and Gender) Demographics Using Deep Learning

Hachim El Khiyari* and Harry Wechsler

Department of Computer Science, George Mason University, Fairfax, VA 22032, USA

Abstract

Human facial appearance is strongly influenced by demographical characteristics such as categorical age, ethnicity, and gender with each category further partitioned into classes-Black, White, Male, Female, Young (18-30), Middle Age (30-50), and Old (50-70)-and groups–mix of classes. Most subjects share a more similar appearance with their own demographic class than with other classes. We evaluate here the accuracy of automatic facial verification for subjects belonging to varying age, ethnicity, and gender categories. Towards that end, we use a convolutional neural network for feature extraction and show that our method yields better performance on individual demographics compared to a commercial face recognition engine. For one-class demographic groups, we corroborate empirical findings that biometric performance on verification is relatively lower for females, young subjects in the 18-30 age group, and blacks. We then expand the scope of our method and evaluate the accuracy of face verification for several multiclass demographic groups. We discuss the results and make suggestions for improving face verification across varying demographics.

Keywords: Age; Authentication; Biometrics; Convolutional neural networks; Deep learning; Demographics; Ethnicity; Face recognition; Gender; Face verification; Feature extraction

Introduction

The goal of automatic face authentication is to accurately verify the identity of subjects from captured images or video in unconstrained environments. Several factors affect accuracy and may be classified as external or internal. External factors are related to the environment in which the image or video is captured and include elements such as image quality, illumination, or camera viewpoint. Internal factors on the other hand are related to the subjects' appearance such as demographics, facial expression, or cosmetics. Both internal and external factors contribute significantly to the complexity of face authentication in unconstrained operational settings.

This paper focuses on the impact of multiclass demographics on face verification. We consider three demographic categories – age, ethnicity, and gender indexed along seven demographic classes – Black, White, Male, Female, Young (18-30), Middle Age (30-50), and Old (50-70). We partition our dataset into corresponding groups where each group belongs to one or more demographic classes. As an example, a one-class demographic group is the 18-30 age group which contains all male, female, black, and white subjects in the 18-30 age range. Klare et al. [1] reported that commercial and non-trainable algorithms (e.g. LBP, Gabor) have lower matching accuracies on females, blacks, and subjects in the 18-30 age group. Their study, confined to one-class demographics, did not consider performance on multiclass demographic groups which is addressed here.

Robust face verification across demographics requires a powerful face representation capable of capturing the subtle details of human facial appearance. We evaluate the use of deep learning to automatically capture robust feature representations using a convolutional neural network (CNN) and compare our method with a commercial of the shelf (COTS) face recognition engine. The use of deep learning for feature extraction was presented in the context of face recognition subject to aging (El Khiyari and Wechsler [2]). Here we expand the use of VGG-Face CNN (Parkhi et al. [3]) to assess the potential of deep learning for face verification across varying demographics. Experimental results show that the performance of our method is consistently higher across demographics than the COTS engine.

The outline of the paper is as follows. Section 2 provides a literature review of previous studies of demographical effects on face recognition. Section 3 discusses deep learning and provides more details on the convolutional neural network that is used in our experiments. The architecture, experimental design, and results are described in Sections 4-5. Section 6 discusses the results and makes recommendations for future R&D directions. Section 7 concludes the paper.

Demographics and related biometric work

A study by Klare et al. [1] on the role of demographic information on the performance of face recognition reported that commercial and non-trainable algorithms (e.g. LBP, Gabor) have lower matching accuracies on females, blacks, and subjects from the 18-30 age group. The authors considered only 1-class demographic groups and pointed out that training on a specific group improves testing performance on that same group. The use of VGG-Face mediates between learning the face space and enrollment for training. Klare et al. also found that biometric performance improves across all classes if training is evenly distributed across demographics.

O'Toole et al. [4] showed that the demographic distribution of the testing set affects performance and the choice of the decision threshold. The authors considered a testing dataset with mixed demographics and pointed out that this scenario tends to overestimate recognition accuracy since some part of performance is based on face categorization such as ethnicity or gender instead of the recognition of unique identities. The authors reported that verification accuracy

***Corresponding author:** El Khiyari H, Department of Computer Science, George Mason University, Fairfax, VA 22030, USA, Tel: (703) 931-5206; Fax: (703) 993-1710; E-mail: helkhiya@gmu.edu

Received October 24, 2016; **Accepted** November 26, 2016; **Published** November 30, 2016

Citation: El Khiyari H, Wechsler H (2016) Face Verification Subject to Varying (Age, Ethnicity, and Gender) Demographics Using Deep Learning. J Biom Biostat 7: 323. doi:10.4172/2155-6180.1000323

is lower for image pairs of the same demographic compared to image pairs from different demographics.

Large demographically diverse public datasets are scarce, so to overcome the lack of demographic variety, Riccio et al. [5] proposed combining multiple datasets spanning the range of age groups, ethnicities, and genders to form a large multi-demographic database. Phillips et al. [6] reported that a face recognition algorithm trained on a specific ethnicity performs best on that same demographic. Their results were analogous to findings from psychological research that indicate that humans are better at recognizing faces from their own ethnicity. Finally, Dantcheva et al. [7] suggested the use of demographic attributes as ancillary information to help in face recognition tasks especially in low-quality images.

The effects of population demographics on performance did not receive as much attention as other factors such as pose, illumination, and facial expression (PIE). This work builds on previous research and is first to consider both single class and multiclass demographic groups using the MORPH data set (Ricanek and Tesafaye [8]).

Deep learning using convolutional neural networks

Deep learning has gained considerable popularity in computer vision due to its high performance on complex tasks. Our face representation approach uses a deep convolutional neural networks (CNN) (Parkhi et al. [3]) to extract rich image descriptors suitable for multi-demographics face recognition. Inspired by biological vision, CNNs take advantage of the spatial correlation in natural images by restricting the connectivity of the network nodes to local receptive fields.

The recent surge in popularity of CNNs is due to the increase in processing power of modern computers and advances in parallel computing. CNNs are composed of millions of parameters and use a deep architecture capable of learning complex features for visual recognition tasks. To reduce training time, CNNs use regularization techniques such as parameter sharing, pooling, and dropout to minimize the number of parameters in the network. However, despite the efficient learning techniques used by CNNs, a very large training dataset is still required in order to avoid over-fitting.

Due to the limited size of our training dataset, we leverage the pre-trained VGG-Face CNN (Parkhi et al. [3]) described in Figure 1. The pre-trained CNN was learned from a large face dataset containing 982,803 web images of 2622 celebrities and public figures. While VGG-Face can only identify subjects in its training dataset, it can however be used as a feature extractor for any arbitrary face image by running the image through the entire network, then extracting the output of the first fully-connected layer, FC-1. The extracted feature vector serves as a highly discriminative, compact, and interoperable encoding of the input image. Once the features are extracted from the FC-1 layer of the VGG-Face CNN, they can be used for training and testing arbitrary face classifiers as we do in our experiments. We use the MatConvNet toolbox (Vedaldi and Lenc [9]) which consists of a library of MATLAB functions implementing CNN architectures for computer vision. MatConvNet provides the pre-trained implementation of the VGG-Face CNN that we use for feature extraction.

The network is composed of a sequence of convolutional, pool, two fully-connected (FC) layers, FC-1 and FC-2, and a final softmax layer. The convolutional layers use filters of dimension 3 while the pool layers perform subsampling with a factor of 2.

Architecture and Methodology

The architecture of the face recognition used in the experiments is shown in Figure 2. The core of the architecture is VGG-Face CNN that performs feature extraction. Image descriptors for enrollment purposes are extracted from the output of the first fully-connected FC-1 layer of the CNN. Verification is performed by comparing descriptors of pairs of images and computing their L_1 distance.

Image Preprocessing

All images were normalized using in-plane rotation to horizontally align the left and right eyes. The eye coordinates are available from the metadata provided with the MORPH dataset. The datasets contain RGB images which are fed to the convolutional neural network in their original color channels. Face images cropped and rescaled to a standard $224 \times 224 \times 3$ dimension are then fed to the VGG-Face CNN. The neurons of the first convolutional layer compute dot products for their receptive fields along the 3 color channels. A sample of preprocessed images for MORPH is shown in Figure 3.

Feature extraction

We use the VGG-Face network provided by the MatConvNet

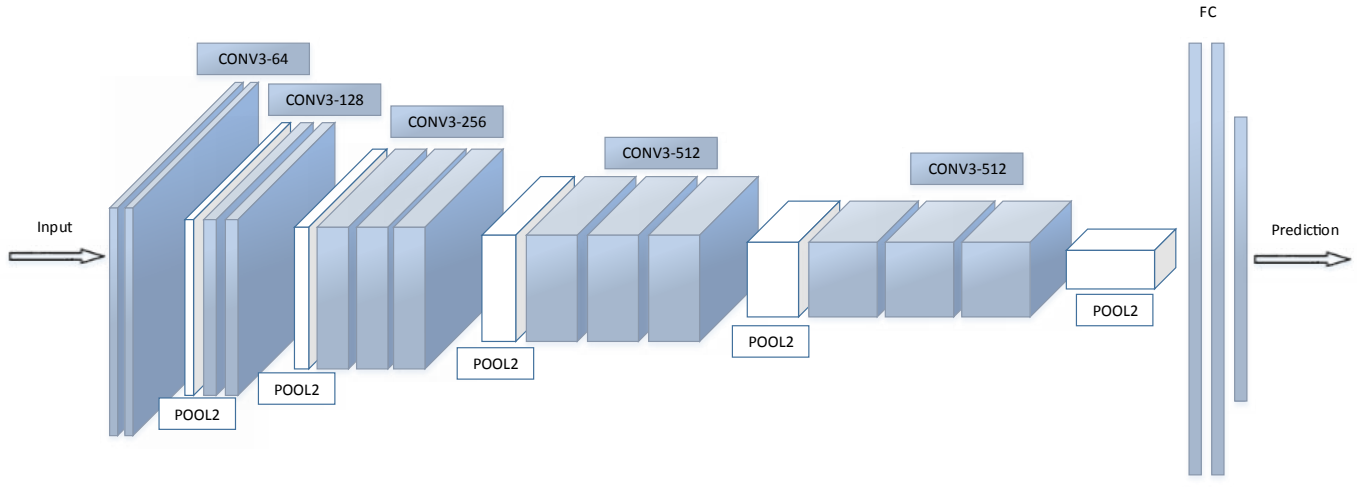

Figure 1: VGG-Face CNN Architecture (Parkhi et al. [3]).

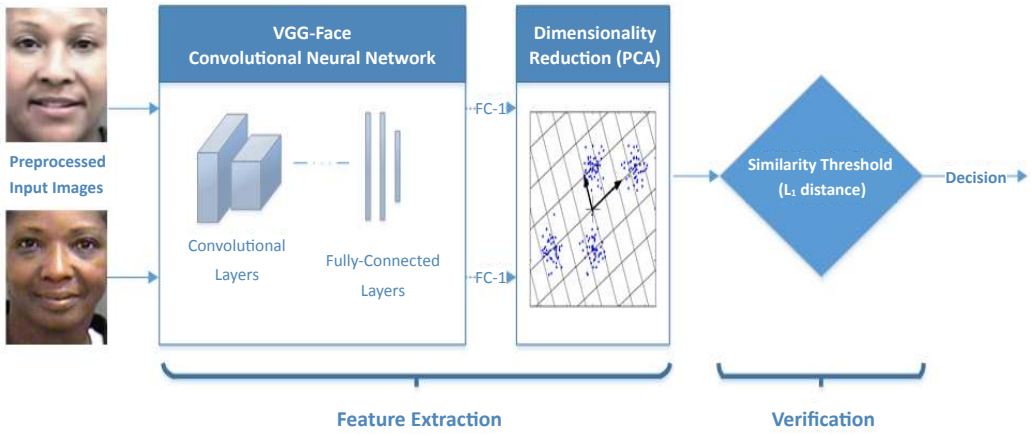

Figure 2: Authentication architecture using VGG-Face CNN feature descriptors.

Figure 3: Subjects from the MORPH dataset belonging to multiple demographic classes (Ricanek and Tesafaye [8]).

Category	Class	# Samples	# Subjects
Age	18-30	1376	344
	30-50	1376	344
	50-70	144	36
Ethnicity	White	1448	362
	Black	1448	362
Gender	Male	1448	362
	Female	1448	362

Table 1: Demographic composition.

toolbox (Vedaldi and Lenc [9]) for feature extraction. The VGG-Face network described in section 3 has a deep architecture and is composed of 3 × 3 convolution layers, 2 × 2 pooling layers, and 3 fully-connected layers. While the network was originally trained to perform classification rather than feature extraction, the output layer of the network is not used in our experiments and 4096-dimensional descriptors are instead extracted from the first fully-connected layer, FC-1. To extract features from an image in our dataset, the image is preprocessed and fed to the CNN as a 224 × 224 × 3 array of pixel intensities. Each convolutional layer performs a filtering operation on the preceding layer resulting in an activation volume which in turn becomes the input of the following layer. Pooling is used throughout the network to reduce the number of nodes by down sampling the activation maps using the max operation. The fully-connected layers of the network are used for learning the classification function. The extracted features from the output of the first fully-connected layers (FC-1) are L_2-normalized by dividing each component by the L_2-norm of the feature vector. The normalized features are then used for verification.

Dimensionality reduction

The performance and interoperability of our authentication ("verification") method is evaluated using L_1 distances measuring the similarity between pairs of feature vectors. To improve performance, we apply principal component analysis (PCA) to reduce the dimension of the extracted features (Jolliffe [10]).

Since the first convolutional layer of VGG-Face contains 4096 nodes, the FC-1 features are 4096-dimensional vectors. PCA is a linear projection such that the variance of the projected data is maximized. To find the lower dimensional PCA subspace of the CNN features X_1, X_2, …., X_n, we construct the matrix $X = [X_1, X_2, \ldots, X_n]$ where each vector X_i is centered by extracting the mean of X. We construct the covariance matrix of the data and compute its eigenvectors and corresponding eigenvalues. We select the top eigenvectors with the largest eigenvalues such that 95% of the variance of the original data is preserved. The top eigenvectors are called principal components and form the orthogonal basis on which the FC-1 features are projected.

Experimental Design and Results

Images were collected from the MORPH dataset (Ricanek and Tesafaye [8]) and partitioned into multiple demographic groups where each partition is composed entirely of subjects belonging to one or more demographic classes. Table 1 shows the demographic composition of the dataset including the number of subjects and samples. From the dataset, we constructed multiple subsets corresponding to one-class, two-class, or three-class demographic groups. One-class demographic groups included subjects that share a single common demographic characteristic (e.g. whites, females). Two-class demographic groups included subjects sharing 2 demographic characteristics (e.g. white females, young blacks). Three-class demographic groups included subjects sharing 3 demographic characteristics (e.g. Middle age black females, young white males). For each demographic group, the corresponding images were balanced so they contain the same proportions from the other demographics. For example, the 18-30 age group contained the same proportion of whites, blacks, males, and females, as the 30-50 group.

We performed the preprocessing, feature extraction, dimensionality reduction, and classification as described in section 4. The first experiment reports on the performance of one-class demographics groups, and compares our method with the 2015 release of a commercial off the shelf (COTS) face recognition system. The second and third experiments report the performance of two-class and three-class demographic groups respectively.

The performance and interoperability of our authentication

(using repeated verification) method was evaluated using L_1 distances measuring the similarity between pairs of feature vectors. To improve performance, we applied principal component analysis (PCA) (see Section 4) to reduce the dimension of the extracted features.

One-class demographic groups

We partitioned the dataset into 7 demographic groups: Female, Male, Black, White, Young (18-30), Middle Age (30-50), and Old (50-70). Figures 4-6 show the performance of our method and the COTS face recognition system. The receiver operating characteristic (ROC) curves represent the tradeoffs between false accept and false reject rates for different values of the similarity threshold. Table 2 provides a summary of the results.

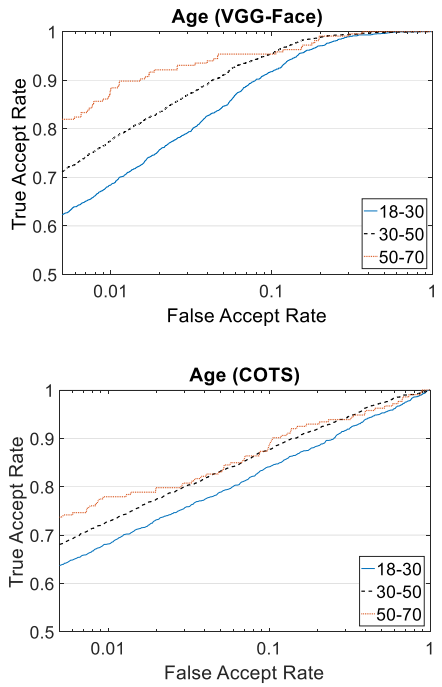

Figure 4: ROC of age group demographic (Left ROC shows our method. Right ROC shows the COTS system).

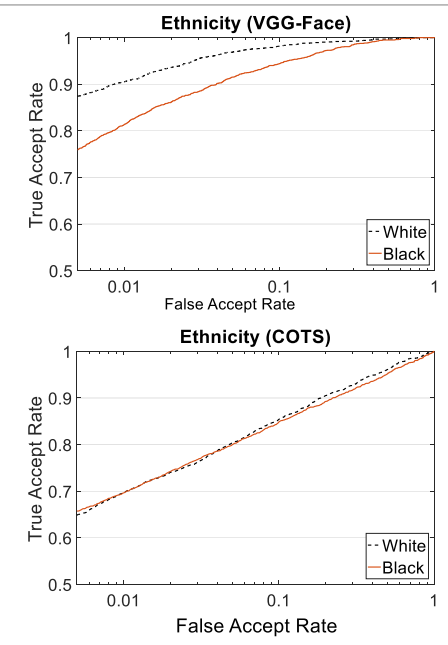

Figure 5: ROC of ethnicity demographic. Left ROC shows our method. Right ROC shows the COTS system.

Our method shows consistently higher performance than the COTS system. The groups that show the lowest relative performance for each demographic class are the Young (18-30), Females, and Blacks groups. These results are consistent with Klare et al. [1]. Performance using the COTS system was lower however, there was a smaller performance gap within each class especially between black and white.

Two-class demographic groups

Table 3 shows a summary of the results on verification accuracy for two-class demographic groups. Groups containing the 50-70 age group are not included in the table because they contain less subjects and therefore their performance cannot be directly compared with the other groups. Note that the number of subjects used in this experiment is smaller than the first experiment, so the performance cannot be directly compared. The results show that there can be a very significant gap in performance across demographics. For a 0.01 FAR, the TAR is 56% for young females and 94% for white males.

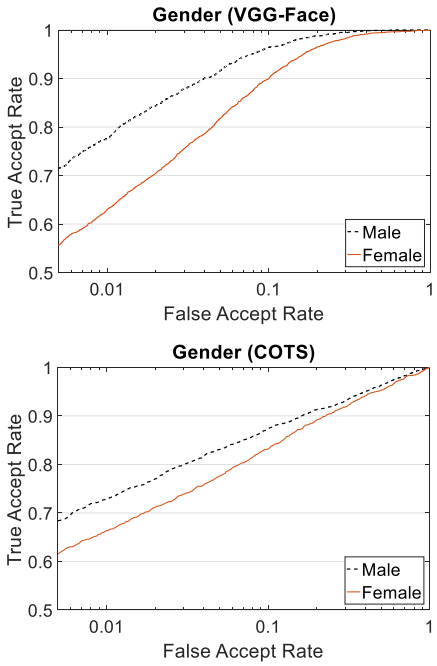

Figure 6: ROC of gender demographic. Left ROC shows our method. Right ROC shows the COTS system.

	CNN		COTS	
	TAR (FAR 0.01)	TAR (FAR 0.1)	TAR (FAR 0.01)	TAR (FAR 0.1)
Females	63.03	89.96	66.34	83.17
Young	68.41	91.76	68.18	84.14
Middle Age	77.42	95.49	72.9	87.69
Males	77.67	96.41	72.96	87.41
Black	81.4	94.43	69.63	84.93
Old	87.96	95.37	77.93	88.73
White	90.52	98.16	69.9	85.5

Table 2: Performance on one-class demographic groups.Values represent the true accept rate (TAR) for different false accept rates (FAR).

	TAR (FAR 0.01)	TAR (FAR 0.1)	# Samples	# Subjects
Young Females	0.56	0.85	688	172
Middle Age Females	0.65	0.91	688	172
Black Females	0.72	0.92	724	181
Young Males	0.72	0.94	688	172
Young Blacks	0.78	0.92	688	172
White Females	0.82	0.96	724	181
Middle Age Males	0.82	0.97	688	172
Black Males	0.83	0.95	724	181
Middle Age Blacks	0.85	0.95	688	172
Young Whites	0.88	0.96	688	172
Middle Age Whites	0.91	0.99	688	172
White Males	0.94	0.99	724	181

Table 3: Two-class Demographic group performance. Values represent the true accept rate (TAR) for different false accept rates (FAR).

	TAR (FAR 0.01)	TAR (FAR 0.1)	# Samples	# Subjects
Young Black Females	0.70	0.89	344	86
Young White Females	0.77	0.92	344	86
Middle Age Black Females	0.78	0.93	344	86
Young Black Males	0.81	0.93	344	86
Middle Age White Females	0.83	0.96	344	86
Middle Age Black Males	0.87	0.95	344	86
Young White Males	0.90	0.97	344	86
Middle Age White Males	0.94	0.99	344	86

Table 4: Three-class demographic group performance. Values represent the true accept rate (TAR) for different false accept rates (FAR).

Three-class demographic groups

In this experiment, the dataset was partitioned into 8 disjoint demographic groups. The TAR is reported in Table 4. Groups containing the 50-70 class were again excluded due to the low number of subjects in our dataset. The most challenging group is the young black females while the least challenging is the white middle age males.

Comparative performance evaluation and discussion

Demographics play an important role in the performance evaluation of face recognition systems. Datasets used in research often have an unbalanced distribution across demographics. For example, the complete MORPH dataset is composed mostly of black males: 85% of the images are of male subjects and 77% of the images are of black subjects. Performance evaluation using demographically unbalanced datasets can provide misleading results since the demographics in an operational and uncontrolled environment can be different than the demographics of the testing dataset. Our experiments reported here use balanced demographic groups.

Low performance could be traced to high intraclass or low interclass image variations. While our experiments do not provide evidence of inherently high interclass variations for any demographic group, some factors can affect intraclass variations. For example, in the female group, which shows consistently low performance, intraclass variation can be due to factors such as 1) use of cosmetics; 2) eyebrow plucking–ocular region is the most discriminative region of the face so change in the appearance of the eyebrows or eyelashes can have a significant effect on performance.; and 3) long hair–even though, we crop images to exclude hair, some women have long hair that sometimes occludes parts of the face.

Deep learning provides a powerful framework for learning robust face representations across demographics. However, the choice of subjects used to train the convolutional neural network can affect its ability to learn the most relevant features needed for recognition. One possible approach to improve accuracy across demographics is to train the CNN using a demographically-balanced training dataset. This can be challenging since CNNs require a very large training dataset to avoid overfitting. Another more plausible approach is to fine-tune a pre-trained CNN such as VGG-Face to optimize it for a particular demographic group, i.e. initialize the CNN parameters with the pre-trained values, then resume training with a different (small) dataset and update the weights using stochastic gradient descent and backpropagation. The small dataset used for fine tuning can be selected to match the demographics of the population during operation. In an uncontrolled environment, several CNNs would need to be trained (fine-tuned) for each demographic group. During testing, dynamic selection of an optimized CNN need to precede recognition.

Conclusions

This paper advances the use of deep learning for face verification across multiple demographic groups. Each demographic group considered contained the same proportions as the other demographic groups. The results reported showed that a deep learning approach using a pre-trained convolutional neural network outperforms a commercial face recognition system. For one-class demographic groups, we corroborated that verification accuracy is relatively lower for females, young subjects in the 18-30 age range, and blacks. For two-class and three-class demographic groups, we showed that performance can vary widely and identified which groups are the most challenging. This paper highlights the role demographics play on verification performance and provides direction for future research in face recognition under uncontrolled operational settings.

References

1. Klare BF, Burge MJ, Klontz JC, Vorder Bruegge RW, Jain AK (2012) Face Recognition Performance: Role of Demographic Information. IEEE Transactions on Information Forensics and Security 7: 1789-1801.
2. El Khiyari H, Wechsler H (2016) Face Recognition across Time Lapse Using Convolutional Neural Networks. Journal of Information Security 7: 141-151.
3. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. Proceedings of the British Machine Vision Conference (BMVC).
4. O'Toole AJ, Phillips PJ, An X, Dunlop J (2012) Demographic effects on estimates of automatic face recognition performance. Image and Vision Computing 30: 169-176.
5. Riccio D, Tortora G, De Marsico M, Wechsler H (2012) EGA - Ethnicity, gender and age, a pre-annotated face database. IEEE Workshop on Biometric Measurements and Systems for Security and Medical Applications (BIOMS), Salerno: 1-8.
6. Phillips PJ, Jiang F, Narvekar A, Ayyad J, O'Toole AJ (2011) An other-race effect for face recognition algorithms. ACM Trans Appl Percept.
7. Dantcheva A, Elia P, Ross A (2016) What Else Does Your Biometric Data Reveal? A Survey on Soft Biometrics. IEEE Transactions on Information Forensics and Security 11: 441-467.
8. Ricanek K, Tesafaye T (2006) MORPH: A Longitudinal Image Database of Normal Adult Age-Progression. IEEE 7th International Conference on Automatic Face and Gesture Recognition, Southampton UK.
9. Vedaldi A, Lenc K (2015) MatConvNet - Convolutional Neural Networks for MATLAB. Proc of the ACM Int Conf on Multimedia.
10. Jolliffe IT (2002) Principal Component Analysis. (2ndedn), Springer.