MULTIMODAL BIOMETRICS RECOGNITION FROM FACIAL VIDEO VIA DEEP LEARNING

Sayan Maity, Mohamed Abdel-Mottaleb, and Shihab S. As

University of Miami; 1251 Memorial Drive; Coral Gables; Florida 33146-0620
s.maity1@umail.miami.edu, mottaleb@miami.edu, sasfour@miami.edu

ABSTRACT

Biometrics identification using multiple modalities has attracted the attention of many researchers as it produces more robust and trustworthy results than single modality biometrics. In this paper, we present a novel multimodal recognition system that trains a Deep Learning Network to automatically learn features after extracting multiple biometric modalities from a single data source, i.e., facial video clips. Utilizing different modalities, i.e., left ear, left profile face, frontal face, right profile face, and right ear, present in the facial video clips, we train supervised denosing autoencoders to automatically extract robust and non-redundant features. The automatically learned features are then used to train modality specific sparse classifiers to perform the multimodal recognition. Experiments conducted on the constrained facial video dataset (WVU) and the unconstrained facial video dataset (HONDA/UCSD), resulted in a 99.17% and 97.14% rank-1 recognition rates, respectively. The multimodal recognition accuracy demonstrates the superiority and robustness of the proposed approach irrespective of the illumination, non-planar movement, and pose variations present in the video clips.

KEYWORDS

Multimodal Biometrics, Autoencoder, Deep Learning, Sparse Classification.

1. INTRODUCTION

There are several motivations for building robust multimodal biometric systems that extract multiple modalities from a single source of biometrics, i.e., facial video clips. Firstly, acquiring video clips of facial data is straight forward using conventional video cameras, which are ubiquitous. Secondly, the nature of data collection is non-intrusive and the ear, frontal, and profile face can appear in the same video. The proposed system, shown in Figure 1, consists of three distinct components to perform the task of efficient multimodal recognition from facial video clips. First, the object detection technique proposed by Viola and Jones [1], was adopted for the automatic detection of modality specific regions from the video frames. Unconstrained facial video clips contain significant head pose variations due to non-planar movements, and sudden changes in facial expressions. This results in an uneven number of detected modality specific video frames for the same subject in different video clips, and also a different number of modality

David C. Wyld et al. (Eds) : CSITA, ISPR, ARIN, DMAP, CCSIT, AISC, SIPP, PDCTA, SOEN - 2017
pp. 67– 75, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70107

specific images for different subject. From the aspect of building a robust and accurate model, it is always preferable to use the entire available training data. However, classification through sparse representation (SRC) is vulnerable in the presence of uneven number of modality specific training samples for different subjects. Thus, to overcome the vulnerability of SRC whilst using all of the detected modality specific regions, in the model building phase we train supervised denoising sparse autoencoder to construct a mapping function. This mapping function is used to automatically extract the discriminative features preserving the robustness to the possible variances using the uneven number of detected modality specific regions. Therefore, by applying Deep Learning Network as the second component in the pipeline results in an equal number of training sample features for the different subjects. Finally, using the modality specific recognition results, score level multimodal fusion is performed to obtain the multimodal recognition result.

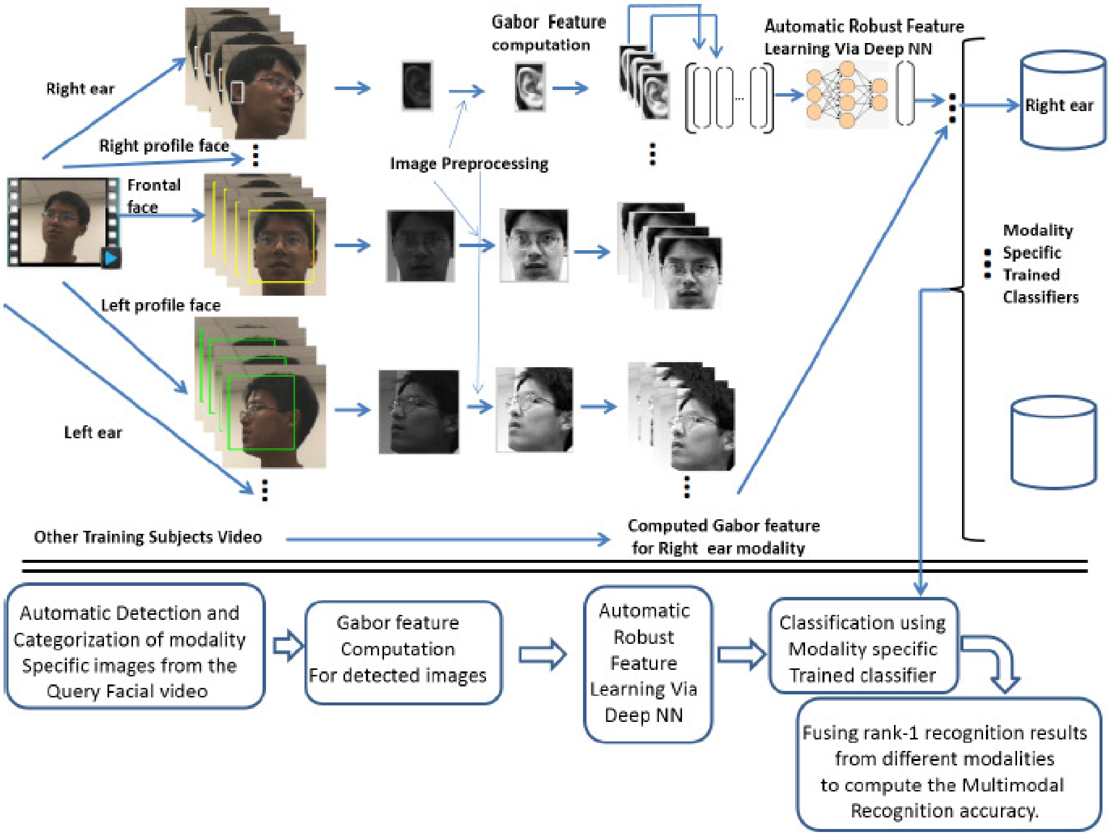

Right ear
Right profile face
Frontal face
Left profile face
Left ear
Image Preprocessing
Gabor Feature computation
Automatic Robust Feature Learning Via Deep NN
Right ear
Modality Specific Trained Classifiers
Other Training Subjects Video
Computed Gabor feature for Right ear modality
Automatic Detection and Categorization of modality Specific images from the Query Facial video
Gabor feature Computation For detected images
Automatic Robust Feature Learning Via Deep NN
Classification using Modality specific Trained classifier
Fusing rank-1 recognition results from different modalities to compute the Multimodal Recognition accuracy.

Fig. 1. System Block Diagram: Multimodal Biometrics Recognition from Facial Video

Due to the unavailability of proper datasets for multimodal recognition studies [2], often virtual multimodal databases are synthetically obtained by pairing modalities of different subjects from different databases. To the best of our knowledge, the proposed approach is the first study where multiple modalities are extracted from a single data source that belongs to the same subject. The main contributions of the proposed approach is the application of training a Deep Learning Network for automatic feature learning in multimodal biometrics recognition using a single source of biometrics i.e., facial video data, irrespective of the illumination, non-planar movement, and pose variations present in the face video clips.

The remainder of this paper is organized as follows: Section 2 details the modality specific frame detection from the facial video clips. Section 3 describes the automatic feature learning using supervised denoising sparse autoencoder (deep-learning). Section 4 presents the modality specific classification using sparse representation and multimodal fusion. Section 5 provides the experimental results on the constrained facial video dataset (WVU [3]) and the unconstrained facial video dataset (HONDA/UCSD [4]) to demonstrate the performance of the proposed framework. Finally, conclusions and future research directions are presented in Section 6.

2. MODALITY SPECIFIC IMAGE FRAME DETECTION

To perform multimodal biometric recognition, we first need to detect the images of the different modalities from the facial video. The facial video clips in the constrained dataset are collected in a controlled environment, where the camera rotates around the subject's head. The video sequences start with the left profile of each subject (0 degrees) and proceed to the right profile (180 degrees). Each of these video sequences contains image frames of different modalities, e.g., left ear, left profile face, frontal face, right profile face, and right ear, respectively. The video sequences in the unconstrained dataset contains uncontrolled and nonuniform head rotations and changing facial expressions. Thus, the appearance of a specific modality in a certain frame of the unconstrained video clip is random compared with the constrained video clips.

The algorithm was trained to detect the different modalities that appear in the facial video clips. To automate the detection process of the modality specific image frames, we adopt the Adaboost object detection technique, proposed by Viola and Jones [1]. The algorithm is trained to detect frontal and profile faces in the video frames, respectively, using manually cropped frontal face images from color FERET database, and profile face images from the University of Notre Dame Collection J2 database. Moreover, it is trained using cropped ear images from UND color ear database to detect ear images in the video frames. By using these modality specific trained detectors, we can detect faces and ears in the video frames. The modality specific trained detectors are applied to the entire video sequence to detect the face and the ear regions in the video frames.

Before using the detected modality specific regions from the video frames for extracting features, some preprocessing steps are performed. The facial video clips recorded in the unconstrained environment contain variations in illumination and low contrast. Histogram equalization is performed to enhance the contrast of the images. Finally, all detected modality specific regions from the facial video clips were resized; ear images were resized to 110 X 70 pixels and faces images (frontal and profile) were resized to 128 X 128 pixels.

3. AUTOMATIC FEATURE LEARNING USING DEEP NEURAL NETWORK

Even though the modalitiy specific sparse classifiers result in relatively high recognition accuracy on the constrained face video clips, the accuracy suffers in case of unconstrained video because the sparse classifier is vulnerable to the bias in the number of training images from different subjects. For example, subjects in the HONDA/UCSD dataset [4] randomly change their head pose. This results in a nonuniform number of detected modality specific video frames across different video clips, which is not ideal to perform classification through sparse representation.

In the subsequent sections we first describe the gabor feature extraction technique. Then, we describe the supervised denoising sparse autoencoders, which we use to automatically learn equal number of feature vectors for each subject from the uneven number of modality specific detected regions.

3.1 Feature Extraction

2D Gabor filters [5] are used in broad range of applications to extract scale and rotation invariant feature vectors. In our feature extraction step, uniform down-sampled Gabor wavelets are computed for the detected regions:

$$\psi_{\mu,\nu}(z) = \frac{||k_{\mu,\nu}||^2}{s^2} e^{(\frac{-||k_{\mu,\nu}||^2||z||^2}{2s^2})} [e^{ik_{\mu,\nu}z} - e^{\frac{-s^2}{2}}], \tag{1}$$

where $z = (x, y)$ represents each pixel in the 2D image, $k_{\mu,\nu}$ is the wave vector, which can be defined as $k_{\mu,\nu} = k_\nu e^{i\phi_\mu}$, $k_\nu = \frac{k_{max}}{f^\nu}$, k_{max} is the maximum frequency, and f is the spacing factor between kernels in the frequency domain, $\phi_\mu = \frac{\pi\mu}{2}$, and the value of s determines the ratio of the Gaussian window width to wavelength. Using equation 1, Gabor kernels can be generated from one filter using different scaling and rotation factors. In this paper, we used five scales, $\nu \in 0, ..., 4$ and eight orientations $\mu \in 0, ..., 7$. The other parameter values used are $s = 2\pi$,$k_{max} = \frac{\pi}{2}$, and $f = \sqrt{2}$.

Before computing the Gabor features, all detected ear regions are resized to the average size of all the ear images, *i.e.*, 110×70 pixels, and all face images (frontal and profile) are resized to the average size of all the face images, i.e., 128×128 pixels. Gabor features are computed by convolving each Gabor wavelet with the detected 2D region, as follows:

$$C_{\mu,\nu}(z) = T(z) * \psi_{\mu,\nu}(z), \tag{2}$$

where $T(z)$ is the detected 2D region, and $z = (x, y)$ represents the pixel location. The feature vector is constructed out of $C_{\mu,\nu}$ by concatenating its rows.

3.2 Supervised Stacked Denoising Auto-encoder

The application of neural networks to supervised learning [6] is well proven in different applications including computer vision and speech recognition. An autoencoder neural network is an unsupervised learning algorithm, one of the commonly used building blocks in deep neural networks, that applies backpropagation to set the target values to be equal to the inputs. The reconstruction error between the input and the output of the network is used to adjust the weights of each layer. An autoencoder tries to learn a function $x_i = \hat{x}_i$, where x_i belongs to unlabeled training examples set $\{x_{(1)}, x_{(2)}, x_{(3)}, ..., x_{(n)}\}$, and $x_i \in \mathbb{R}^n$.

In other words, it is trying to learn an approximation to the identity function, to produce an output $\hat{x}$ that is similar to x, in two subsequent stages: (i) An encoder that maps the input x to the hidden nodes through some deterministic mapping function $f : h = f(x)$, then (ii) A decoder that maps the hidden nodes back to the original input space through another deterministic mapping function $g : \hat{x} = g(h)$. For real-valued input, by minimizing the reconstruction error $||x - g(f(x))||_2^2$, the parameters of encoder and decoder can be learned.

To learn features, which are robust to illumination, viewing angle, pose etc., from modality specific image regions, we adopted the supervised autoencoder [7]. The supervised autoencoder is trained using features extracted from image regions ($\hat{x}_i$) containing variations in illumination, viewing angle and pose, whereas the features of selected image regions, (x_i), with similar illumination and without pose variations are utilized as the target. By minimizing the objective criterion given in Equation 3 (subject to, the modality-specific features of the same person are similar), the supervised autoencoders learn to capture the modality specific robust representation.

$$\underset{W,b_e,b_d}{min} \frac{1}{N} \sum_i \left(\|(x_i - g(f(\hat{x}_i))\|_2^2 + \lambda \|(f(x_i) - f(\hat{x}_i)\|_2^2\right); \qquad (3)$$

where the output of the hidden layer, h, is defined as $h = f(x) = tanh(Wx + b_e), g(h) = tanh(W^T h + b_d)$, N is the total number of training samples, and λ is the weight preservation term. The first term in Equation 3 minimize the the reconstruction error, *i.e.*, after passing through the encoder and the decoder, the variations (llumination, viewing angle and pose) of the features extracted from the unconstrained images will be repaired. The second term in Equation 3 enforces the simillarity of modality specific features corresponding to the same person.

After training a stack of encoders its highest level output representation can be used as input to a stand-alone supervised learning algorithm. A logistic regression (LR) layer was added on top of the encoders as the final output layer which enable the deep neural network to perform supervised learning. By performing gradient descent on a supervised cost function, the Supervised Stacked Denoising Auto-encoder (SDAE) automatically learned fine-tuned network weights. Thus, the parameters of the entire SDAE network are fine-tuned to minimize the error in predicting the supervised target (*e.g.*, class labels).

3.3 Training the Deep Learning Network

We adopt the two stage training of the Deep Learning Network, where we have a better initialization to begin with and a fine tuned network weights that lead us to a more accurate high-level representation of the dataset. The steps of two stage Deep Learning Network training are as follows:

*Step*1. Stacked Denoising Autoencoders are used to train the initial network weights one layer at a time in a greedy fashion using Deep Belief Network (DBN).

*Step*2. The weights of the Deep learning network are initialized using the learned parameters from DBN.

*Step*3. Labelled training data are used as input, and their predicted classification labels obtained using the Logistic regression layer along with the initial weights of the network used as an objective function to fine tune the entire network .

*Step*4. Finally, the learned network weights are used to extract image features to train the sparse classifier.

The network is illustrated in Figure 2, which shows a two-category classification problem (there are two output values), where the decoding part of SDAE is removed and the encoding part of SDAE is retained to produce the initial features. In addition, the output layer of the whole network, which is also called logistic regression layer, is added. The following sigmoid function is used as activation function of the logistic regression layer:

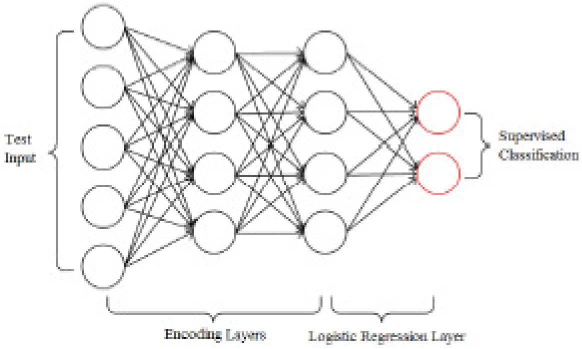


Fig. 2. Supervised Stacked Denoising Auto-encoder

$$h(x) = \frac{1}{e^{-Wx-b}}; \tag{4}$$

where x is the output of the last encoding layer y^l, in other words the features are pretrained by the SDAE network. The output of the sigmoid function is between 0 and 1, which denotes the classification results in case of two class classification problem. Therefore, we can use the errors between the predicted classification results and the true labels associated with the training data points to fine-tune the whole network weights.The cost function is defined as the following cross-entropy function:

$$Cost = -\frac{1}{m}\left[\sum_{i=1}^{m} l^{(i)} log(h(x^{(i)})) + (1 - l^{(i)}) log(1 - h(x^{(i)}))\right]; \tag{5}$$

where $l^{(i)}$ denotes the label of the sample $x^{(i)}$. By minimizing the cost function, we update the network weights.

4. MODALITY SPECIFIC AND MULITMODAL RECOGNITION

The modality specific sub-dictionaries (d_j^i) contain feature vectors generated by Deep Learning Network using the modality specific training data of each individual subject; where i represents the modality, $i \in 1, 2, ..., 5$; and j stands for the number of training video sequence.

Later, we concatenate the modality specific learned sub-dictionaries d_j^i of all the subjects in the dataset to obtain the modality specific (*i.e.*, left ear, left profile face, frontal face, right profile face, and right ear) dictionary D_i, as follows.

$$D_i = [d_1^i; d_2^i; ...; d_j^i]; \forall i \in 1, 2, ..., 5 \qquad (6)$$

4.1 Multimodal Recognition

The recognition results from the five modalities — left ear, left profile face, frontal face, right profile face, and right ear are combined using score level fusion. Score level fusion has the flexibility of fusing various modalities upon their availability. To prepare for fusion, the matching scores obtained from the different matchers are transformed into a common domain using a score normalization technique. Later, the weighted sum technique is used to fuse the results at the score level. We have adopted the *Tanh* score normalization technique [8], which is both robust and efficient. The normalized match scores are then fused using the weighted sum technique:

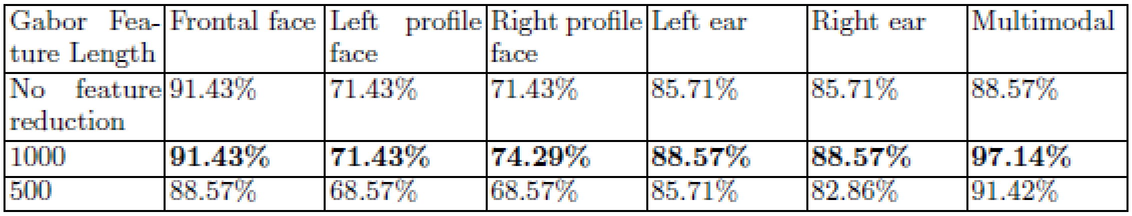
$$S_p = \sum_{i=1}^{M} w_i * s_i^n; \qquad (7)$$

where w_i and s_i^n are the weight and normalized match score of the i^{th} modality specific classifier, respectively, such that $\sum_{i=1}^{M} w_i = 1$. In this study, the weights $w_i, i = 1, 2, 3, 4, 5$; correspond for the left ear, left profile face, frontal face, right profile face, and right ear modalities, respectively. These weights can be obtained by exhaustive search or based on the individual performance of the classifiers [8]. Later, the weights for the modality specific classifiers in the score level fusion were determined by using a separate training set with the goal of maximizing the fused multimodal recognition accuracy.

5. EXPERIMENTAL RESULTS

In this section we describe the results of the modality specific and multi-modal recognition experiments on both datasets. The feature vectors automatically learned using the trained Deep Learning network resulted in length of 9600 for frontal and profile face; 4160 for ear. In order to decrease the computational complexity and to find out most effective feature vector length to maximize the recognition accuracy, the dimensionality of the feature vector is reduced to a lower dimension using Principal Component Analysis (PCA) [9]. Using PCA, the number of features is reduced to 500 and 1000. In Table- 1 the modality specific recognition accuracy obtained for the reduced feature vector of 500, 1000 is shown. Feature vectors of length 1000 resulted in best recognition accuracy for both modality specific and multimodal recognition.

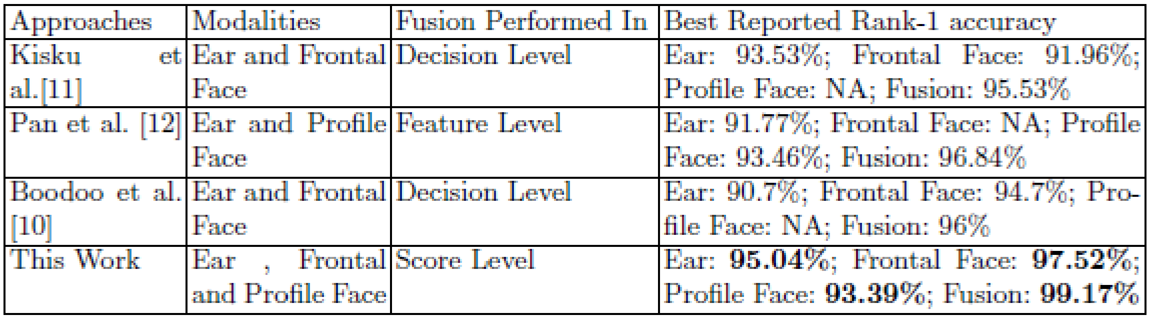
Table 1. Modality Specific and Multimodal Rank-1 Recognition Accuracy



Gabor Feature Length	Frontal face	Left profile face	Right profile face	Left ear	Right ear	Multimodal
No feature reduction	91.43%	71.43%	71.43%	85.71%	85.71%	88.57%
1000	**91.43%**	**71.43%**	**74.29%**	**88.57%**	**88.57%**	**97.14%**
500	88.57%	68.57%	68.57%	85.71%	82.86%	91.42%

The best rank-1 recognition rates, using ear, frontal and profile face modalities for multimodal recognition, compared with the results reported in [10{12] is shown in Table 2.

Table 2. Comparison of 2D multimodal (frontal face, profile face and ear) rank-1 recognition accuracy with the state-of-the-art techniques



Approaches	Modalities	Fusion Performed In	Best Reported Rank-1 accuracy
Kisku et al.[11]	Ear and Frontal Face	Decision Level	Ear: 93.53%; Frontal Face: 91.96%; Profile Face: NA; Fusion: 95.53%
Pan et al. [12]	Ear and Profile Face	Feature Level	Ear: 91.77%; Frontal Face: NA; Profile Face: 93.46%; Fusion: 96.84%
Boodoo et al. [10]	Ear and Frontal Face	Decision Level	Ear: 90.7%; Frontal Face: 94.7%; Profile Face: NA; Fusion: 96%
This Work	Ear , Frontal and Profile Face	Score Level	Ear: **95.04%**; Frontal Face: **97.52%**; Profile Face: **93.39%**; Fusion: **99.17%**

6. CONCLUSION

We proposed a system for multimodal recognition using a single biometrics data source, i.e., facial video clips. Using the Adaboost detector, we automatically detect modality specific regions. We use Gabor features extracted from the detected regions to automatically learn robust and non-redundant features by training a Supervised Stacked Denoising Auto-encoder (Deep Learning) network. Classification through sparse representation is used for each modality. Then, the multimodal recognition is obtained through the fusion of the results from the modality specific recognition.

REFERENCES

[1] Viola, P. and Jones, M.: Grid Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition, pp. 511{518.(2001).

[2] Zengxi Huang and Yiguang Liu and Chunguang Li and Menglong Yang and Liping Chen: A robust face and ear based multimodal biometric system using sparse representation. In: Pattern Recognition, pp.2156{2168.(2013).

[3] Gamal Fahmy and Ahmed El-sherbeeny and Susmita M and Mohamed Abdel-mottaleb and Hany Ammar: The effect of lighting direction/condition on the performance of face recognition algorithms. In:SPIE Conference on Biometrics for Human Identification, pp.188{200.(2006).

[4] K.C. Lee and J. Ho and M.H. Yang and D. Kriegman: Visual Tracking and Recognition Using Probabilistic Appearance Manifolds. In:Computer Vision and Image Understanding.(2005).

[5] Chengjun Liu and Wechsler, H.:Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. In: IEEE Transactions on Image Processing, pp.467{476.(2002).

[6] Rumelhart, David E. and McClelland, James L.: Parallel Distributed Processing: Explorations in the Microstructure of Cognition. In:MIT Press. Cambridge, MA, USA.(1986).

[7] Shenghua Gao and Yuting Zhang and Kui Jia and Jiwen Lu and Yingying Zhang: Single Sample Face Recognition via Learning Deep Supervised Autoencoders. In:IEEE Transactions on Information Forensics and Security, pp.2108{2118.(2015)

[8] Ross, A. A. and Nandakumar, K. and Jain, A. K.: Handbook of multibiometrics. In:Springer.(2006)

[9] Turk Matthew and Pentland Alex: Eigenfaces for recognition.In: J. Cognitive Neuroscience. MIT Press, pp.71{86.(1991)

[10] Nazmeen Bibi Boodoo and R. K. Subramanian: Robust Multi biometric Recognition Using Face and Ear Images. In:J. CoRR.(2009)

[11] Dakshina Ranjan Kisku and Jamuna Kanta Sing and Phalguni Gupta: Multibiometrics Belief Fusion. In:J. CoRR.(2010)

[12] Xiuqin Pan and Yongcun Cao and Xiaona Xu and Yong Lu and Yue Zhao: Ear and face based multimodal recognition based on KFDA. In:International Conference on Audio, Language and Image Processing. pp.965{969.(2008)