Regularizing Deep Learning Architecture for Face Recognition with Weight Variations

Shruti Nagpal*, Maneet Singh*, Mayank Vatsa, and Richa Singh
IIIT-Delhi, New Delhi, India
{shrutin,maneets,mayank,rsingh}@iiitd.ac.in

Abstract

Several mathematical models have been proposed for recognizing face images with age variations. However, effect of change in body-weight is also an interesting covariate that has not been much explored. This paper presents a novel approach to incorporate the weight variations during feature learning process. In a deep learning architecture, we propose incorporating the body-weight in terms of a regularization function which helps in learning the latent variables representative of different weight categories. The formulation has been proposed for both Autoencoder and Deep Boltzmann Machine. On extended WIT database of 200 subjects, the comparison with a commercial system and an existing algorithm show that the proposed algorithm outperforms them by more than 9% at rank-10 identification accuracy.

1. Introduction

It is well understood that body-weight loss or gain affects the facial appearance of a person. Recently, it has been established as a challenge to face recognition algorithms by Singh et al. [7] and Wen et al. [13]. With time, the changes in weight are inevitable and unpredictable. There is no definite pattern of these changes and drastic weight variations are common with an increase in awareness about fitness and health as well as several other factors, such as genetics, eating habits, and medical conditions. Weight variations lead to structural variations which affect facial appearance and lead to decrease in the performance of automated face recognition systems. Therefore, it is an important and challenging aspect of automatic face recognition research that requires significant efforts.

A brief summary of the recent related research is tabulated in Table 1. To the best of our knowledge, there are only two research threads on face recognition with body weight variations. Singh et al. [7] propose an algorithm using Gabor features and a combination of three neural networks for addressing the problem of face recognition with weight and age variations. They reported the rank-1 identification accuracy of 20.34% on the WhoIsIt (WIT) dataset comprising of atleast 10 images each, of 110 subjects. Wen et al. [13] performed verification experiments on two datasets: synthetic and real, containing 120 and 242 subjects respectively. The datasets have two images per subject with weight variations. They used a Partial Least Squares (PLS) based approach and reported a verification accuracy of 69.6% using SIFT and PLS on the real face dataset and 88.7% on the synthetic dataset.

Figure 1. Sample Images from the eWIT dataset showing weight variations. Each column corresponds to one subject.

It is well understood that weight variations are a factor of age as well. However, as shown in Figure 1, there is no relation between the two. With growing age, one can lose or gain any amount of weight. Therefore, in this research, we attempt to learn the facial features with weight variations. This research extends the first two research threads on face recognition with weight variations and proposes a deep-learning based architecture. The two-fold contributions of this paper are:

- Deep-learning based algorithm is proposed which significantly improves the recognition accuracy with respect to the current state-of-the-art for face recognition with weight variations. The algorithm utilizes a

*Equal contribution by the student authors.

Covariate	Author	Method	Dataset Used
Age	Bereta et al. [2] (2013)	Performance of various local descriptors with distance measures is evaluated for age progression. Local descriptors are applied to Gabor wavelet images	FG-NET Aging dataset
	Yang et al. [14] (2014)	Graph matching on feature encoded in the local Gabor binary pattern histogram sequence projected in the LDA subspace	FG-NET Aging dataset
Weight (and Age)	Singh et al. [7] (2014)	Deep learning architecture based on neural networks and Gabor filters.	WIT dataset: 110 subjects, 1109 images covering age and weight variations
	Wen et al. [13] (2014)	PLS method with LBP and SIFT	Synthetic Dataset: 120 subjects, 240 images Real Dataset: 242 subjects, 484 images. Both datasets contain weight variation
	Proposed (2015)	Deep Boltzmann Machine coupled with regularization incorporating weight variations	eWIT dataset: 200 subjects, 2036 images covering age and weight variations

Table 1. Literature Review of recent papers in face recognition with age and weight variations

body-weight based regularization approach to modify the loss function of deep-learning architecture such as Deep Boltzmann Machine [6] and Sparse-Stacked Denoising Autoencoder (SDAE) [11].

- We build upon the existing WhoIsIt (WIT) dataset [7] (1109 images of 110 subjects) and present extended-WIT (eWIT) dataset with 2036 images corresponding to 200 subjects.

In the next section, we explain the proposed algorithm, followed by details about the dataset used. Section 4 presents the experimental protocol and results.

2. Proposed Algorithm

Since there is no well defined pattern in which the weight increases or reduces, we propose a representation learning based algorithm for learning facial features. Deep learning algorithms have been utilized in encoding facial information and recognizing individuals with variations in pose, expression, and illumination as well as in video sequences [3], [8], [9], [10]. Applying these algorithms in a straightforward manner for age and/or weight variations may not yield good performance. In order to incorporate these variations in feature encoding, we propose a modification in the feature learning process via *regularization*.

The objective function of any deep-learning based architecture such as autoencoders and deep Boltzmann machine is to minimize a loss function $\mathcal{L}$. The representation is learnt based on learning the features (or weight matrix) that minimize the *loss*. Traditionally, these unsupervised feature learning algorithms are optimized using regularizers such as KL-divergence and l_p norm. In the proposed algorithm, we modify the loss function by adding a regularizer which is dependent on the body-weight labels of the images. Inspired from Singh *et al.* [7], three weight labels are utilized, namely *thin*, *moderate*, and *heavy*. Since these three categories are discrete *attributes*, we have quantified them into three (approximate) numerical values as 50, 75, and 100 respectively. For a given sample, S_{bw} with bw as the weight category, a body-weight parameter α_{bw} is defined as,

$$\alpha_{bw} = \frac{S_{bw}}{225} \tag{1}$$

where, bw = {*thin, moderate, heavy*}, and S_{bw} can take one of the three values, depending on the weight category. For a deep learning architecture, the loss function $\mathcal{L}$ with a network weight matrix W, is then modified by introducing a l_p norm regularization as,

$$\mathcal{L} = \mathcal{L} + \lambda_p \left\| \alpha_{bw} W \right\|_p^p \tag{2}$$

Here, λ_p is the regularization parameter which is learnt and body-weight parameter α_{bw} for each sample is calculated at the time of training. Based on the three weight categories, α_{bw} can take three values, thereby producing α_{thin}, α_{mod} and α_{heavy}. Using different regularization approaches, the loss function of the network can be modified as follows:

- with l_1 norm regularization:

$$\mathcal{L} = \mathcal{L} + \lambda_1 \left\| \alpha_{bw} W \right\|_1 \tag{3}$$

- with l_2 norm regularization:

$$\mathcal{L} = \mathcal{L} + \lambda_2 \left\| \alpha_{bw} W \right\|_2^2 \tag{4}$$

- with l_1 + l_2 norm regularization:

$$\mathcal{L} = \mathcal{L} + \lambda_1 \left\| \alpha_{bw} W \right\|_1 + \lambda_2 \left\| \alpha_{bw} W \right\|_2^2 \tag{5}$$

Regularization is used to prevent over-fitting, it helps the learner to converge faster, and prevents convergence at

the local minima. While the loss function drives the deep learning algorithm to be sensitive to the variations along with manifold of high density, the regularization influences the learner to be less sensitive to the input. This helps in encoding variations on the manifold but disregarding the orthogonal variations. This means that the regularization in the proposed modification helps in learning latent variables representative of different body-weight categories. In other words, by incorporating the body-weight information of each sample in the optimization function, the network is forced to modify the latent variables according to these variations. Therefore, the feature representation is expected to be robust towards body-weight variations. The proposed regularization approach is applied in two deep learning architectures, Deep Boltzmann Machine (DBM) [6] and Sparse-Stacked Denoising Autoencoder (SDAE) [11].

2.1. Regularized Deep Boltzmann Machine

Deep Boltzmann Machines are stacked Restricted Boltzmann Machines (RBM) having undirected edges between the layers [6]. They are extremely useful for unsupervised learning of feature representations from a given large unlabeled data. The energy function of a RBM can be formulated as follows:

$$E(x,h) = -a^T x - b^T h - x^T W h \tag{6}$$

where *x* and *h* represent the visible and hidden units, respectively. *W* is the weight matrix where weight w_{ij} signifies weight of connection between the hidden unit h_j and visible unit x_i. *a* represents the bias weights for visible units and *b* represents the bias weights for the hidden units. The probability distribution of a RBM, over the hidden and visible units is defined as:

$$P(x,h) = \frac{1}{Z} exp(-E(x,h)) \tag{7}$$

where, *Z* is the partition function, which is a normalization constant. This further leads to the formulation of marginal probability which is the sum of all possible combinations of the hidden unit configurations, i.e.,

$$P(x) = \sum_h P(x,h) = \frac{1}{Z} \sum_h exp(-E(x,h)) \tag{8}$$

Using the training data $\mathbf{X}$, RBMs are trained to minimize the negative log likelihood, i.e. the loss function $\mathcal{L}_{rbm}$ is defined as:

$$\mathcal{L}_{rbm} = -\sum_{x \epsilon X} log(P(x)) \tag{9}$$

As explained earlier, we modify this loss function by adding regularization terms to it. Equations (3), (4) and (5) show the updated loss functions of RBM:

- l_1 norm regularization:

$$\mathcal{L}_{rbm} = -\sum_{x \epsilon X} log(P(x)) + \lambda_1 \left\| \alpha_{bw} W \right\|_1 \tag{10}$$

- l_2 norm regularization:

$$\mathcal{L}_{rbm} = -\sum_{x \epsilon X} log(P(x)) + \lambda_2 \left\| \alpha_{bw} W \right\|_2^2 \tag{11}$$

- l_1 + l_2 norm regularization:

$$\mathcal{L}_{rbm} = -\sum_{x \epsilon X} log(P(x)) + \lambda_1 \left\| \alpha_{bw} W \right\|_1 + \lambda_2 \left\| \alpha_{bw} W \right\|_2^2 \tag{12}$$

A layer-by-layer greedy training [1] is used to stack RBMs and train a DBM.

2.2. Regularized Sparse-Stacked Denoising Autoencoder

Sparse denoising autoencoders are stacked to form a deep learning architecture and greedy layer-by-layer training is used to train the architecture[11]. The output layer of the first autoencoder is connected to the input layer of the second autoencoder and so on. An autoencoder consists of two components, the encoder and the decoder. The encoder transforms the input vector into a hidden representation, and the decoder tries to map it back to the input vector. For a given input vector, *x*, the hidden representation, *y*, is calculated as:

$$y = \phi(Wx + b) \tag{13}$$

where, *W* is the weight matrix, w_{ij} represents the weight of the connection from the i^{th} input node to the j^{th} hidden node. ϕ represents the activation function of the nodes and *b* represents the bias. The decoder maps the learnt features to the data space, using Equation 14.

$$z = \phi(W' y + b') \tag{14}$$

W' is the weight matrix, w'_{ij} represents the weight of the connection from the i^{th} hidden node to the j^{th} decoder output node. ϕ represents the activation function of the nodes and b' represents the bias. The loss function of an autoencoder is formulated as:

$$\mathcal{L}_{ae} = \left\| x - z \right\|_F^2 = \left\| x - \phi(W' \phi(Wx + b) + b') \right\|_F^2 \tag{15}$$

Similar to RBM (DBM) formulation, we modify the loss function of the SDAE by adding regularization terms to it. Following equations represent the updated loss functions of SDAE:

- l_1 norm regularization:

$$\mathcal{L}_{ae} = \left\| x - z \right\|_F^2 + \lambda_1 \left\| \alpha_{bw} W \right\|_1 \tag{16}$$

- l_2 norm regularization:

$$\mathcal{L}_{ae} = \|x - z\|_F^2 + \lambda_2 \|\alpha_{bw} W\|_2^2 \quad (17)$$

- l_1 + l_2 norm regularization:

$$\mathcal{L}_{ae} = \|x - z\|_F^2 + \lambda_1 \|\alpha_{bw} W\|_1 + \lambda_2 \|\alpha_{bw} W\|_2^2 \quad (18)$$

As mentioned previously, the proposed regularization based modifications in the loss functions help in incorporating body-weight variations in the learnt feature space (or latent space). Once the deep-learning based architectures are trained to represent the face images, Random Decision Forest (RDF) [4] based classification is used for recognition.

2.3. Random Decision Forest based Identification

Random Decision Forest is an ensemble of decision trees which is used to solve classification problems [4]. It can handle the non-linearity in the feature space, is robust towards outliers, and provides stable performance with increase in gallery size [4]. Input to the RDF classifier are the features extracted using deep learning architecture and output is the class label and probabilistic match score for each class. RDF training is performed separately using labeled training data and then used for classifying probe samples from the test data.

3. eWIT Dataset

The only publicly available dataset that contains weight as well as age variations for the subjects is the WhoIsIt (WIT) dataset [7]. It contains 1109 images for 110 subjects. We have extended this dataset and created the extendedWIT (referred to as eWIT). The eWIT database consists of frontal face images of public figures taken from the Internet. The extended database contains a total of 2036 images of 200 subjects, each subject having at least 10 and at most 14 images. Each face image has been labeled as either *thin*, *moderate* or *heavy*.

Out of the 2036 images in the database, 437 are labeled as *thin*, 1309 as *moderate*, and 290 as *heavy*. The age range of the entire dataset is between 1 to 96 years, with the average age being 34.29 years. The average age difference between the youngest and oldest image of each subject is 28.78 years. More details about the dataset are given in Table 2 which also provides a tabular representation of the number of images in each of the three weight categories. The extended dataset will be made publicly available to the research community. [1]

[1] http://iab-rubric.org/resources/whoisit.html

Attribute	Value
Number of Subjects	200
Number of Images	2036
Age Range	[1 - 96]
Average Age	34.29
Images per Subject	[10 - 14]
Weight category wise distribution of images	
Thin	437
Moderate	1309
Heavy	290

Table 2. Description of eWIT dataset

4. Experiments and Results

For all images in the eWIT database, faces are detected using Viola Jones detector [12], geometric normalization is performed, and the inter-eye distance is fixed to 90 pixels. eWIT is partitioned into two subsets, training and testing, such that 50% images of each subject are in training and the remaining are in testing. The identification experiments are performed with 200 classes.

Since the number of images in the eWIT database are not sufficient to train a DBM or SDAE, we use a transfer learning based approach for training. A DBM/SDAE is first trained with over 600,000 frontal face images from multiple datasets to learn the unsupervised feature representation of face images. These images and subjects are non-overlapping with the individuals in the eWIT dataset. As mentioned by Salakhutdinov and Hinton [6], "*high-level representations can be built from a large supply of unlabeled sensory inputs and very limited labeled data can then be used to only slightly fine-tune the model for a specific task at hand*". The proposed algorithm utilizes this property, learns the unsupervised features, and then fine tunes the trained algorithm on a smaller number of images from from the eWIT dataset. Using the trained deep-learning algorithm, features are extracted for the training set of eWIT. A Random Decision Forest is then trained on these features for identification. The testing partition comprising 50% of the images from every subject are used for testing.

The results of identification experiments are compared with the existing algorithm proposed by Singh et al. [7] and a commercial-off-the-shelf (COTS) system, Verilook [5]. Table 3 shows the rank-1 and rank-10 identification accuracies of all the algorithms. Figure 3 shows the CMC curves for the DBM architectures, Figure 4 shows the CMC curves for the SDAE architectures, and Figure 5 shows the comparison of the proposed algorithm with state-of-the-art algorithm and COTS. Key results of our experiments are:

- Rank-1 accuracy obtained using the existing algorithm [7] is 17.7% whereas, COTS yields 14.3%. Rank-

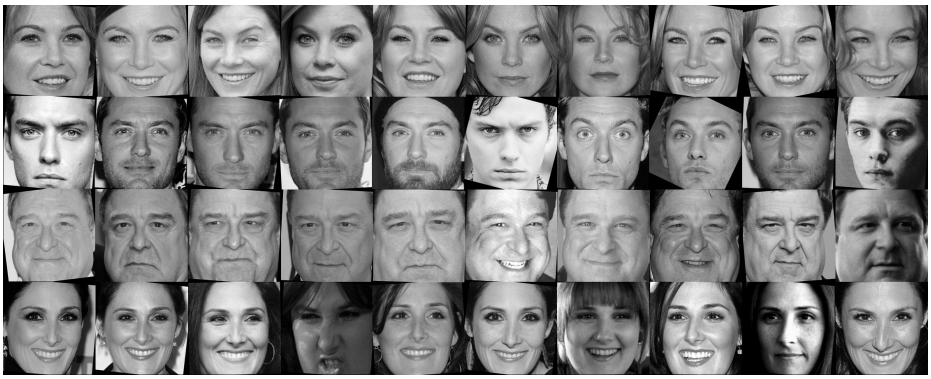

Figure 2. Sample images added to the WIT dataset.

Algorithms		Rank-1 Accuracy	Rank-10 Accuracy
COTS (VeriLook)		14.3	47.0
Singh et al. [7]		17.7	51.2
SDAE	KL Divergence	19.5	56.2
	l_1 norm	21.9	58.7
	l_2 norm	20.1	57.7
	l_1 norm + l_2 norm	23.0	60.3
DBM	No regularization	20.1	57.4
	l_1 norm	22.3	59.6
	l_2 norm	20.7	58.4
	l_1 norm + l_2 norm	23.4	61.9

Table 3. Identification accuracies obtained by the existing and proposed algorithms on the eWIT database.

10 accuracies obtained by these two approaches are 51.2% and 47.0% respectively. Rank-1 accuracy obtained using the SDAE with KL divergence is 19.5% and rank-10 accuracy is 56.2%, whereas, the deep Boltzmann machine yields the rank-1 accuracy of 20.1% and rank-10 accuracy of 57.4%

- Experiments using SDAE and DBM are performed with l_1 and l_2 norm. For SDAE with l_1 norm regularization, the rank-1 accuracy is improved to 21.9%, whereas for DBM, the accuracy is 22.3%. Similarly, when we apply l_2 norm regularization, the rank-1 accuracy for SDAE is 20.1% and DBM is 20.7%. This clearly demonstrates that l_1 norm yields better results.

- On applying l_1 norm + l_2 norm, the results are marginally better than the other techniques for both SDAE and DBM. Rank-1 accuracy obtained for SDAE and DBM are 23.0% and 23.4% respectively. Rank-10 accuracies obtained with l_1 and l_2 norm together for SDAE and DBM are 60.3% and 61.9% respectively. As shown in Figure 6, the proposed algorithm performs at least 10% better than the existing algorithm (10.7%) and COTS (14%).

- With l_1 norm + l_2 norm (elastic net), both grouping effect and sparsity promoting regularization are applied and hence improved results are observed.

- For rank-10 accuracy, the variants of the proposed algorithm have the standard deviation of less than 1% whereas the standard deviation of COTS is 8.13%. This shows that the proposed algorithm is more stable as compared to the commercial system.

- Computationally, on a 6C 2.4GHz workstation with 64GB RAM, the regularized DBM and regularized SDAE based feature extraction followed by RDF based classifier require less than 1 second for identification.

5. Conclusion and Future Work

The contributions of this research is two-folds: (1) proposing a novel algorithm to learn latent variables via body-weight attuned regularization approach, and (2) two deep learning based algorithms, one with DBM and another with SDAE are presented to address the problem of face recognition with weight variations. Results on the extended WIT database, on 200 subjects, show that the proposed regularization forces the feature learner to adapt the variations due to body-weight changes. Results also show that the regularized DBM and regularized SDAE both perform better than an existing algorithm and Verilook commercial matcher. Currently, we are exploring accommodating facial aging information along with weight variations in the feature learning process. We also plan to extend the proposed approach with different regularization and deep learning architectures.

References

[1] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, U. D. Montral, and M. Qubec. Greedy layer-wise training of deep networks. In *In NIPS*. MIT Press, 2007.

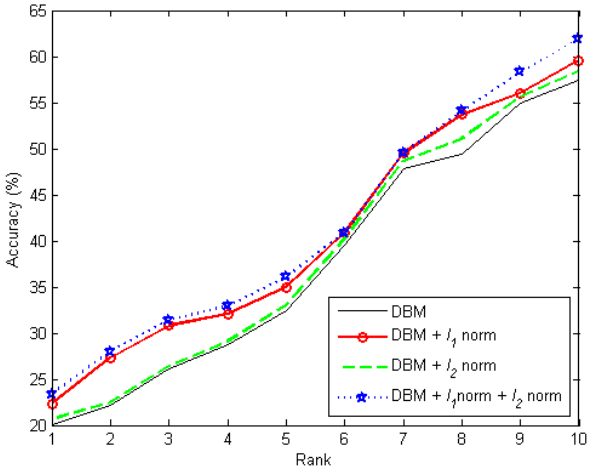

Figure 3. Identification results with DBM.

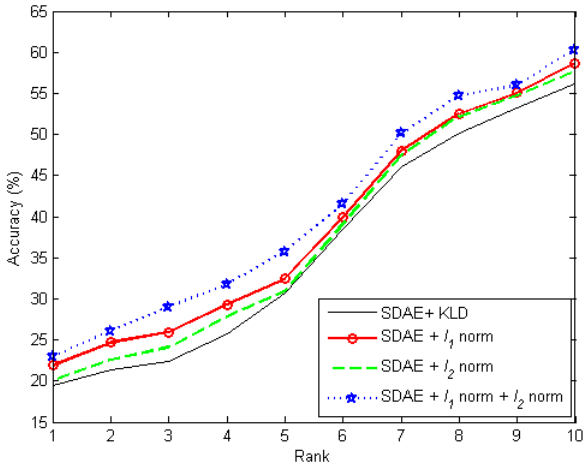

Figure 4. Identification results with SDAE.

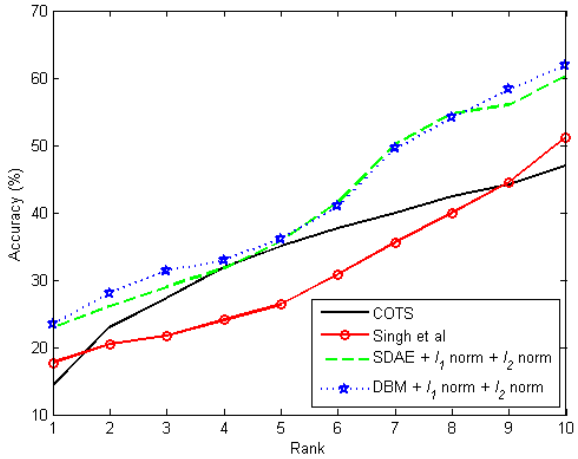

Figure 5. Identification results comparing COTS, existing algorithm [7], best of DBM and best of SDAE.

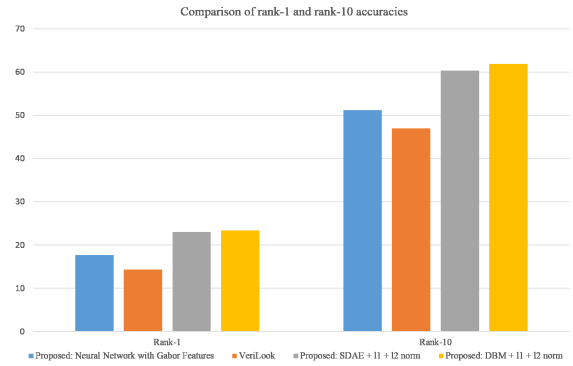

Figure 6. Bar Graph comparing COTS (VeriLook) [5], existing algorithm [7], best of DBM and best of SDAE.

[2] M. Bereta, P. Karczmarek, W. Pedrycz, and M. Reformat. Local descriptors in application to the aging problem in face recognition. *Pattern Recog.*, 46(10):2634–2646, Oct. 2013.

[3] G. Goswami, R. Bhardwaj, R. Singh, and M. Vatsa. MDL-Face: Memorability augmented deep learning for video face recognition. In *IEEE IJCB*, pages 1–7, 2014.

[4] T. K. Ho. Random decision forests. In *IEEE ICDAR*, volume 1, pages 278–282 vol.1, Aug 1995.

[5] http://www.neurotechnology.com/verilook.html. Verilook.

[6] R. Salakhutdinov and G. Hinton. Deep boltzmann machines. In *AISTATS*, volume 5, pages 448–455, 2009.

[7] M. Singh, S. Nagpal, R. Singh, and M. Vatsa. On recognizing face images with weight and age variations. *IEEE Access*, 2:822–830, 2014.

[8] Y. Sun, Y. Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, pages 1988–1996, 2014.

[9] Y. Sun, X. Wang, and X. Tang. Deep learning face representation from predicting 10,000 classes. In *IEEE CVPR*, pages 1891–1898, 2014.

[10] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE CVPR*, pages 1701–1708, 2014.

[11] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, 2010.

[12] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.

[13] L. Wen, G. Guo, and X. Li. A study on the influence of body weight changes on face recognition. In *IEEE IJCB*, pages 1–6, 2014.

[14] H. Yang, D. Huang, and Y. Wang. Age invariant face recognition based on texture embedded discriminative graph model. In *IEEE IJCB*, pages 1–8, 2014.