Beyond Principal Components: Deep Boltzmann Machines for Face Modeling

Chi Nhan Duong [1], Khoa Luu [2], Kha Gia Quach [1], Tien D. Bui [1]
[1] Concordia University, Computer Science and Software Engineering, Montréal, Québec, Canada
[2] Carnegie Mellon University, CyLab Biometrics Center, Pittsburgh, PA, USA
[1] {c_duon, k_q, bui}@encs.concordia.ca, [2] kluu@andrew.cmu.edu

Abstract

The "interpretation through synthesis", i.e. Active Appearance Models (AAMs) method, has received considerable attention over the past decades. It aims at "explaining" face images by synthesizing them via a parameterized model of appearance. It is quite challenging due to appearance variations of human face images, e.g. facial poses, occlusions, lighting, low resolution, etc. Since these variations are mostly non-linear, it is impossible to represent them in a linear model, such as Principal Component Analysis (PCA). This paper presents a novel Deep Appearance Models (DAMs) approach, an efficient replacement for AAMs, to accurately capture both shape and texture of face images under large variations. In this approach, three crucial components represented in hierarchical layers are modeled using the Deep Boltzmann Machines (DBM) to robustly capture the variations of facial shapes and appearances. DAMs are therefore superior to AAMs in inferring a representation for new face images under various challenging conditions. In addition, DAMs have ability to generate a compact set of parameters in higher level representation that can be used for classification, e.g. face recognition and facial age estimation. The proposed approach is evaluated in facial image reconstruction, facial super-resolution on two databases, i.e. LFPW and Helen. It is also evaluated on FG-NET database for the problem of age estimation.

1. Introduction

The "interpretation through synthesis" approach has become one of the most successful and popular face modeling approaches over the last two decades. Given a new face image, the purpose is to "describe" that image by generating a new synthesized image that is similar to it as much as possible. This aim can be achieved by an optimization process on the appearance parameters of the model based *apriori* on constrained solutions. Therefore, a subspace model to be suitable and practical must also provide a basis for a broad range of variations that are usually unseen.

Figure 1. A comparison of facial interpretation in real world images between our DAMs approach and the AAMs. The first row: original images; The second row: shape free images; The third row: facial interpretation using PCA-based AAMs; The fourth row: facial interpretation using our proposed DAMs approach.

Active Appearance Models (AAMs), one of the most successful face interpretation methods, were first introduced by Cootes et al. in 1998 [4]. Since then, there have been numerous improvements and adaptations based on the original approach [2, 5, 11, 15]. However, Gross et al. [8] showed that AAMs perform well in person-specific cases rather than generic ones. In these AAMs approaches, the capability of facial generalization and reconstruction are highly dependent on the characteristics of training databases. This is because at the heart of AAMs, Principal Component Analysis (PCA) is used to provide a subspace to model variations in training data. The limitation of PCA to generalize to illumination and poses, particularly for faces, is very well known. This is the reason why AAMs have difficulties in generalizing to new faces under these challenging conditions. On the other hand, the variations in data are not only large but also non-linear. For example, the variations in different facial expressions or poses are non-linear. It apparently violates the linear assumptions of PCA-based models. Thus, single PCA model is unable to interpret the facial variations well.

Recently, Deep Boltzmann Machines (DBM) [19] have

gained significant attention as one of the emerging research topics in both the higher-level representation of data and the distribution of observations. In DBM, non-linear latent variables are organized in multiple connected layers in a way that variables in one layer can simultaneously contribute to the probabilities or states of variables in the next layers. Each layer learns a different factor to represent the variations in a given data. Thanks to the nonlinear structure of DBM and the strength of latent variables organized in hidden layers, it efficiently captures variations and structures in complex data that could be higher than second order.

Moreover, DBM is shown to be more robust with ambiguous input data [19]. There are some recent works using DBM as shape prior model [6, 22, 26]. Far apart from these methods, the higher-level relationships of both shape and texture are exploited in our proposed DAMs so that the reconstruction of one can benefit from the information on the other. This paper proposes a novel Deep Appearance Models (DAMs[1]) approach to find a set of parameters in both shape and texture to characterize the identity, facial poses, facial expressions, lighting conditions of a given face. In addition, our proposed approach also has ability to generate a compact set of parameters in a robust model that can later be used for classification. Specifically, the DBMs-based shape and texture models are first independently constructed. Then the interactions between these shapes and textures are further modeled using a deeper hidden layer. By this way, after fitting the model to new images, these interactions can be used as a compact set of parameters that represent both shape and appearance of faces for further discriminative problems.

2. Related Work

This section reviews previous AAMs-based approaches for face modeling and model fitting. In the steps of the appearance modeling, let $\mathcal{I} \subset \mathbb{R}^2$ be the image domain and $\mathcal{D} \subset \mathbb{R}^2$ be the texture domain. A shape $\mathbf{s} = (r^1_{\mathcal{D}}, ..., r^n_{\mathcal{D}})$ and texture $\mathbf{g}$ in AAMs are represented in two linear PCA models as follows:

$$\begin{aligned} \mathbf{s}(b_s) &= \mathbf{s}_0 + \mathbf{P}_s b_s \\ \mathbf{g}(r_{\mathcal{D}}; b_g) &= \mathbf{g}_0(r_{\mathcal{D}}) + \mathbf{P}_g(r_{\mathcal{D}}) b_g \end{aligned} \tag{1}$$

where $\{\mathbf{s}_0, \mathbf{P}_s\}$ and $\{\mathbf{g}_0, \mathbf{P}_g\}$ are learned from a given training set. In order to fit this model to a new testing image, a warping operator $W(r_{\mathcal{D}}; \mathbf{s})$ and a similarity transformation $N(r_{\mathcal{I}}; \mathbf{q})$ defined in Eqn. (2) will be employed on that testing image.

$$\begin{aligned} W(r_{\mathcal{D}}; \mathbf{s}) &= r_{\mathcal{I}} \\ N(r_{\mathcal{I}}; \mathbf{q}) &= \begin{pmatrix} 1+\rho_1 & -\rho_2 \\ \rho_2 & 1+\rho_1 \end{pmatrix} r_{\mathcal{I}} + \tau \end{aligned} \tag{2}$$

[1]Noted that the term DAM is also used for "Direct Appearance Models" in [9].

where $q = \{\rho, \tau\}$ composes of the global rotation ρ and the translation τ. The parameters of shapes and textures are optimized so that the sum of squared errors between that testing image and the model texture instance are minimized:

$$[b^*_s, b^*_g] = \arg\min_{b_s, b_g} \|[I \circ N \circ W](b_s; \mathbf{q}) - \mathbf{g}(b_g)\|^2_{\mathcal{D}} \tag{3}$$

where $[I \circ N \circ W](r_{\mathcal{D}}, b_s; \mathbf{q}) = I(N(W(r_{\mathcal{D}}; b_s); \mathbf{q}))$ is the normalized shape-free image warped from the input image I using W and N operators defined in Eqn. (2).

However, Maaten et al. [24] showed that simply using single low-rank Gaussian function we are unable to capture the texture distributions of facial images in real world containing numerous factors, e.g. facial expressions, facial poses, lighting conditions, etc. Therefore, they presented a mixture of K probabilistic PCA to model the texture variations and employed the Expectation-Maximization (EM) algorithm to train the appearance model. Joan et al. [11] also used the probabilistic PCA to model the appearance. In the fitting steps, a test image is linearized and projected to a latent texture space before the shape parameters are optimized using the gradient descent algorithm. Their method is prominent to detect the facial features. However, the assumption of multivariate Gaussian distribution is a prerequisite condition in these methods.

Fitting steps in AAMs are an iterative optimization process. It measures the cost between a new testing image and a model texture in the coordinate of a reference frame. Generally, previous fitting techniques can be divided into two categories, i.e. discriminative and generative approaches. In the first category, the optimizing process is updated using a trained parameter-updating model. There are several ways to train a model in this approach, e.g. perturbing the parameters and recording the residuals [4], directly using texture information to predict the shape [9], linear regression technique [5] and non-linear regression method [20], etc. These techniques usually require low computational costs. However, since the mapping function is fixed and independent of current model parameters, their performance in term of fitting quality is still limited.

In the second category, the fitting steps are formulated as an image alignment problem and iteratively solved using the Gaussian-Newton optimization. Matthews et al. [15] presented a project out inverse algorithm to work on the orthogonal complement of the texture subspace. Although the algorithm runs very fast since most of the terms can be precomputed, it can not perform well in generic AAMs when testing faces are from untrained subjects. Other methods find the shape and texture increments either simultaneously [8] or alternatively [17]. Amberg et al. [2] presented the compositional framework. Recently, Tzimiropoulos et al. [23] presented a fitting algorithm that works effectively in both forward and inverse cases. However, their method is also limited due to the assumption of the PCA-based model.

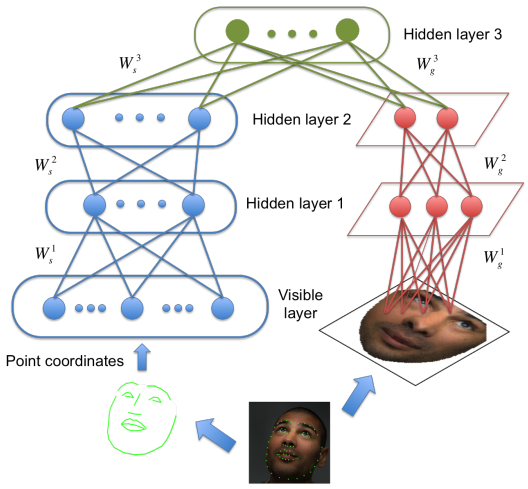

Figure 2. Deep Appearance Models that consists of shape model (left), texture model (right) and the joint representation of shape and texture.

3. Deep Appearance Models (DAMs)

There are three main steps to construct the DAMs as presented in Figure 2. Firstly, two appropriate prior models are separately formulated for shapes and textures. Then, the interactions between these two models will be interpreted in an efficient way. Finally, a fitting algorithm will be presented in order to synthesize any given new face image.

Facial shape structures and texture variations in DAMs are mathematically modeled using the Deep Boltzmann Machines. DBM is capable to model high-order correlations among input data. Its undirected connections provide both bottom-up and top-down passes to efficiently send updates between the texture model and the shape model. These modeling shape and texture parameters are then embedded in a higher-level layer that can be learned by clamping both shapes and textures as observations for the models.

3.1. Shape Modeling

In practical applications, facial shapes usually vary drastically due to many factors, e.g. facial expressions, facial poses, etc., as shown in Figure 1. In this approach, the coordinates of landmark points are considered as observations. Then a two-layer DBM is used to learn their distributions in order to generalize all possible patterns of facial shapes.

Let $\mathbf{s} = [x_1, y_1, ..., x_n, y_n]^T$ be a shape with n landmark points $\{x_i, y_i\}, x_i \in \mathbb{R}, y_i \in \mathbb{R}$; $\mathbf{h}_s^{(1)}, \mathbf{h}_s^{(2)}$ be the binary hidden variables of the first and second layers respectively; $\theta_s = \{\mathbf{W}_s^{(1)}, \mathbf{W}_s^{(2)}\}$ be the model parameters. Since $\{x_i, y_i\}$ are the coordinates of the i^{th}-landmark point, the interactions between visible and hidden units $\mathbf{h}_s^{(1)}$ in the first layer are formulated using Gaussian Restricted Boltzmann Machines. Then, in the second and subsequent layers, binary-binary RBMs will be efficiently employed.

Given a shape $\mathbf{s}$, the energy of the state $\{\mathbf{s}, \mathbf{h}_s^{(1)}, \mathbf{h}_s^{(2)}\}$ in facial shape modeling is formulated as follows:

$$E(\mathbf{s}, \mathbf{h}_s^{(1)}, \mathbf{h}_s^{(2)}; \theta_s) = \sum_i \frac{(s_i - b_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{s_i}{\sigma_i} W_{sij}^{(1)} h_{sj}^{(1)} - \sum_{j,l} h_{sj}^{(1)} W_{sjl}^{(2)} h_{sl}^{(2)} \quad (4)$$

In Eqn. (4), the bias terms of hidden units in these two layers are ignored to simplify the equation. Its corresponding probability is then computed as follows:

$$P(\mathbf{s}; \theta_s) = \sum_{\mathbf{h}_s^{(1)}, \mathbf{h}_s^{(2)}} P(\mathbf{s}, \mathbf{h}_s^{(1)}, \mathbf{h}_s^{(2)}; \theta_s) = \frac{1}{Z(\theta_s)} \sum_{\mathbf{h}_s^{(1)}, \mathbf{h}_s^{(2)}} \exp^{-E(\mathbf{s}, \mathbf{h}_s^{(1)}, \mathbf{h}_s^{(2)}; \theta_s)} \quad (5)$$

where $Z(\theta_s)$ is the partition function.

3.2. Texture Modeling

Far apart from the shape modeling, the process of texture modeling is more complicated due to numerous factors, such as: lighting conditions, facial occlusions, facial expressions, image resolutions, etc. These factors can significantly change pixel values presented in these textures. In addition, compared to facial shapes, facial textures consist of much higher non-linear variations. Therefore, the texture model approach has to be sophisticated enough to represent these variations.

Given a new image in the image domain $\mathcal{I}$, its corresponding shape-free image will be computed by warping it based on a reference candidate in the texture domain $\mathcal{D}$. This step aims at modeling facial textures without further using shape factors. Then a two-layer DBM is employed to model these shape-free images. Since the observations in the texture domain $\mathcal{D}$ are real values, the Gaussian Restricted Boltzmann Machines are used in the bottom layer. The interactions between hidden units in higher layers are formulated by a binary RBM.

Similar to the shape modeling in Eqn. (4), given a shape-free image $\mathbf{g}$, the energy of the state $\{\mathbf{g}, \mathbf{h}_g^{(1)}, \mathbf{h}_g^{(2)}\}$ in facial texture modeling is formulated as follows:

$$E(\mathbf{g}, \mathbf{h}_g^{(1)}, \mathbf{h}_g^{(2)}; \theta_g) = \sum_i \frac{(g_i - b_i)^2}{2\sigma_i^2} - \sum_{i,j} \frac{g_i}{\sigma_i} W_{gij}^{(1)} h_{gj}^{(1)} - \sum_{j,l} h_{gj}^{(1)} W_{gjl}^{(2)} h_{gl}^{(2)} \quad (6)$$

and its probability is then computed as follows:

$$P(\mathbf{g}; \theta_g) = \sum_{\mathbf{h}_g^{(1)}, \mathbf{h}_g^{(2)}} P(\mathbf{g}, \mathbf{h}_g^{(1)}, \mathbf{h}_g^{(2)}; \theta_g) = \frac{1}{Z(\theta_g)} \sum_{\mathbf{h}_g^{(1)}, \mathbf{h}_g^{(2)}} \exp^{-E(\mathbf{g}, \mathbf{h}_g^{(1)}, \mathbf{h}_g^{(2)}; \theta_g)} \quad (7)$$

3.3. Appearance Modeling

A straightforward way to extract model parameters for both shape and texture is to do a weighted concatenation and apply a dimensional reduction method such as PCA. However, this is not an optimal solution since these parameters are presented in different domains, i.e. shape parameters b_s determine the coordinates of landmark points while texture parameters b_g present facial appearance in the texture domain $\mathcal{D}$. Therefore, the gaps between them still exist in the final model parameters although weight values are employed to balance the combined features.

Meanwhile, our Deep Appearance Models also aim to produce a robust facial shape and texture representation. It, however, can be considered as the problem of data learning from multiple sources. In this problem, the information learned from multiple input channels can complement each other and boost the overall performance of the whole model. Particularly, captions and tags can be used to improve the classification accuracy [10, 16, 21].

In order to generate a robust feature in DAMs, one should notice that the hidden units are powerful in term of increasing the flexibility of deep model. Beside the ability of capturing different factors from the observations, the higher layer these hidden units are in, the more independent of the specific correlations of an input source [21]. Therefore, we can use them as a source-free representation. From that reason, we construct one more high-level layer to interpret the connections between face shape and its texture. Since $\mathbf{h}_s^{(2)}$ and $\mathbf{h}_g^{(2)}$ are independent of the spaces where the coordinates and appearance are in, the new layer can encode the shape and texture information more effectively.

Let $\mathbf{h}^{(3)}$ be the connection layer and $\theta = \{\theta_s, \theta_g\}$ be the model parameters, the joint distribution over the multimodal input can be written as follows:

$$P(\mathbf{s}, \mathbf{g}; \theta) = \sum_{\mathbf{h}_s^{(2)}, \mathbf{h}_g^{(2)}, \mathbf{h}^{(3)}} P(\mathbf{h}_s^{(2)}, \mathbf{h}_g^{(2)}, \mathbf{h}^{(3)}) \left(\sum_{\mathbf{h}_s^{(1)}} P(\mathbf{s}, \mathbf{h}_s^{(1)}, \mathbf{h}_s^{(2)}) \right) \left(\sum_{\mathbf{h}_g^{(1)}} P(\mathbf{g}, \mathbf{h}_g^{(1)}, \mathbf{h}_g^{(2)}) \right) \tag{8}$$

Model Learning: The parameters in the model are optimized in order to maximize the log likelihood

$$\theta^* = \arg\max_{\theta} \log P(\mathbf{s}, \mathbf{g}; \theta) \tag{9}$$

Then the optimal parameter values can be obtained in a gradient descent fashion given by

$$\frac{\partial}{\partial \theta} \mathbb{E}\left[\log P(\mathbf{s}, \mathbf{g}; \theta)\right] = \mathbb{E}_{\text{data}}\left[\frac{\partial E}{\partial \theta}\right] - \mathbb{E}_{\text{model}}\left[\frac{\partial E}{\partial \theta}\right] \tag{10}$$

where $\mathbb{E}_{\text{data}}[\cdot]$ and $\mathbb{E}_{\text{model}}[\cdot]$ are the expectations with respect to data distribution and distribution estimated by Deep Appearance Models. The former term can be approximated by mean-field inference while the latter term can be estimated using Markov chain Monte Carlo based stochastic approximation.

3.4. Properties of Deep Appearance Models

Deep Appearance Models provide the capability of generating facial shapes using texture information and vice versa. For example, one can predict a facial shape from the appearance using DAMs as follows: (1) clamping the texture information $\mathbf{g}$ as observations for the texture model and initializing hidden units with random states; (2) performing standard Gibbs sampling as a posterior inference step; and (3) obtaining the reconstructed shape from $P(\mathbf{s}|\mathbf{g}; \theta)$. To generate the appearance from a given shape, one can apply the same way with reversed pathways after clamping the shape information to the shape model.

In addition, it is more natural to interpret both shapes and textures using higher hidden layers. In order to obtain this representation, one can clamp both observed shape $\mathbf{s}$ and texture $\mathbf{g}$ together before applying the Gibbs sampling procedure to estimate $P(\mathbf{h}^{(3)}|\mathbf{s}, \mathbf{g}; \theta)$. Eventually, probabilities of these hidden layers can be used as features. Notice that, beside the advantage of better features for discriminative tasks, one can easily see that even when one of two inputs is missing (i.e. shape), $P(\mathbf{h}^{(3)}|\mathbf{g}; \theta)$ is still able to approximate. Hence, DAMs can be considered as a more generative model compared to other appearance models.

The proposed method can also deal with facial reconstruction in various challenging conditions, such as: facial occlusions, facial expressions, facial off-angles, etc. These advantages of this method will be shown in Section 5.

4. Fitting in Deep Appearance Models

Given a testing face I, the fitting process in DAMs can be formulated as finding an optimal shape $\mathbf{s}$ that maximizes the probability of the shape-free image as in Eqn. (11).

$$\mathbf{s}^* = \arg\max_{\mathbf{s}} P(I(W(r_{\mathcal{D}}, \mathbf{s}))|\mathbf{s}; \theta) \tag{11}$$

Since the connections between textures and hidden units $\mathbf{h}_g^{(1)}$ are modeled by a Gaussian Restricted Boltzmann Machines, the probability of texture $\mathbf{g}$ given hidden units $\mathbf{h}_g^{(1)}$ is computed as follows:

$$P(\mathbf{g}|\mathbf{h}_g^{(1)}; \mathbf{s}, \theta) = \mathcal{N}(\sigma \mathbf{W}_{\mathbf{g}}^{(\mathbf{1})} \mathbf{h}_g^{(1)} + \mathbf{c}, \sigma^2 \mathbf{A}) \tag{12}$$

where $\mathbf{A}$ is the identity matrix and σ is the standard-deviation of visible units in the texture model; $\mathbf{W}_{\mathbf{g}}^{(\mathbf{1})}$ are learned weights of the visible-hidden texture and $\mathbf{c}$ is a bias of this hidden layer $\mathbf{h}_g^{(1)}$.

During the fitting steps, the states of hidden units $\mathbf{h}_g^{(1)}$ are estimated by clamping both the current shape $\mathbf{s}$ and the

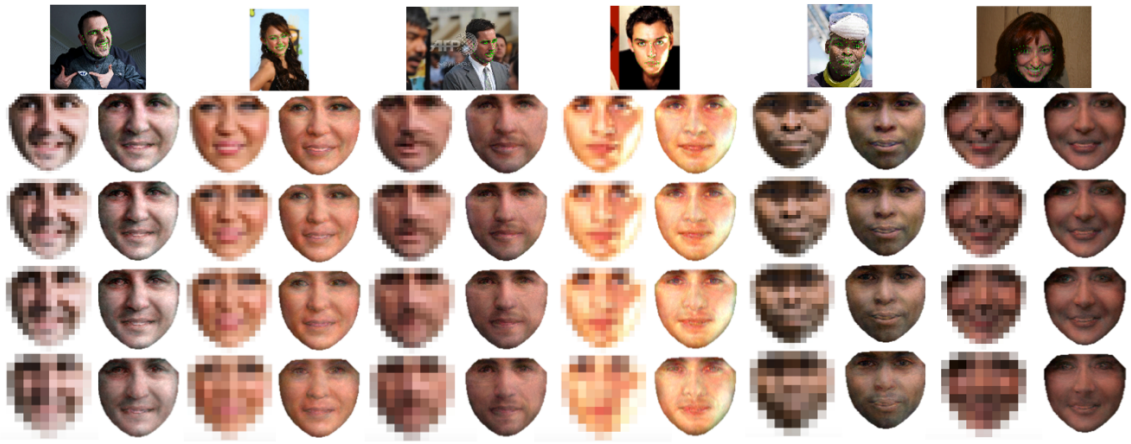

Figure 3. Facial image super-resolution reconstruction at different scales of down-sampling. The 1st row: original image, the 2nd row to the 5th row: down-scaled images with factors of 4, 6, 8, 12 (left) and reconstructed facial images using DAMs (right).

texture $\mathbf{g}$ to the model. The Gibbs sampling method is then applied to find the optimal estimated texture of the testing face given a current shape $\mathbf{s}$. By this way, the hidden units in DAMs can take into account both shape and texture information in order to reconstruct a better texture for further refinement.

Let $\mathbf{m} = \sigma \mathbf{W}_{\mathbf{g}}^{(1)} \mathbf{h}_g^{(1)} + \mathbf{c}$ be the mean of the Gaussian distribution, we have the following approximation:

$$P(I(W(r_{\mathcal{D}}, \mathbf{s}))|\mathbf{h}_g^{(1)}; \theta) = \mathcal{N}(\mathbf{m}, \sigma^2 \mathbf{A}) \tag{13}$$

The maximum likelihood can be then estimated as follows:

$$\begin{aligned} \mathbf{s}^* &= \arg\max_{\mathbf{s}} (P(I(W(r_{\mathcal{D}}, \mathbf{s}))|\mathbf{s}; \theta)) \\ &= \arg\max_{\mathbf{s}} \mathcal{N}(I(W(r_{\mathcal{D}}, \mathbf{s}))|\mathbf{m}, \sigma^2 \mathbf{A})) \\ &= \arg\min_{\mathbf{s}} \frac{1}{\sigma^2} \sum (I(W(r_{\mathcal{D}}, \mathbf{s})) - \mathbf{m})^2 \end{aligned} \tag{14}$$

Then the forward compositional algorithm can be use to solve the problem (14) by finding the updating parameter $\Delta\mathbf{s}$ that increases the likelihood:

$$\Delta\mathbf{s} = \arg\min_{\Delta\mathbf{s}} \|I(W(W(r_{\mathcal{D}}, \Delta\mathbf{s}), \mathbf{s})) - \mathbf{m}\|^2 \tag{15}$$

The linearization is taken place of the test image coordinate using first order Taylor expansion $I(W(W(r_{\mathcal{D}}, \Delta\mathbf{s}), \mathbf{s})) = I(W(r_{\mathcal{D}}, \mathbf{s})) + \mathbf{J}_I \Delta\mathbf{s}$ and the update parameter is given as:

$$\Delta\mathbf{s} = -(\mathbf{J}_I^T \mathbf{J}_I)^{-1} \mathbf{J}_I^T \left[I(W(r_{\mathcal{D}}, \mathbf{s})) - \mathbf{m}\right] \tag{16}$$

where $\mathbf{J}_I = \nabla I \frac{\partial W}{\partial \mathbf{s}}$ is the Jacobian.

5. Experimental Results

In this section, we evaluate our proposed Deep Appearance Models in both facial representation and reconstruction in four applications, i.e. face super-resolution; face off-angle reconstruction; occlusion removal and facial age estimation.

5.1. Databases

We aim to build a model that can represent face texture in-the-wild. Therefore, in the first three applications, we evaluate DAMs on two face databases in-the-wild, i.e. Labeled Face Parts in the Wild (LFPW) [3] and Helen [12]. These databases contain unconstrained facial images collected from various multimedia resources. These facial images have considerable resolutions and contain numerous variations such as poses, occlusions and expressions. For the age estimation application, FG-NET face aging database [1] is used to evaluate the method.

The LFPW database contains 1400 images in total with 1100 training and 300 testing images. However, a part of it is no longer accessible. Therefore, in our experiments, we only use 811 training and 224 testing images, the available remaining. Each facial image is annotated with 68 landmark points provided by 300-W competition [18].

The Helen database provides a high-resolution dataset with 2000 images used for training and 330 images for testing. The variations consist of pose changing from $-30°$ to $30°$; several types of expression such as neutral, surprise, smile, scream; and occlusions. Similar to LFPW, all faces in Helen are also annotated with 68 landmark points.

FG-NET is a popular face aging database. There are 1002 face images of 82 subjects with age ranges from 0 to 69 years. The annotations in FG-NET are also 68 landmarks in same format as LFPW and Helen databases.

5.2. Face Super-resolution Reconstruction

The proposed DAMs method is evaluated in its capability to recover high-resolution face images given their very low-resolution versions. Moreover, since LFPW and Helen databases also include numerous variations in poses, expressions and occlusions, the experiment becomes more

challenging. Therefore, our proposed method is very potential in dealing with the problem of super-resolution in various conditions of facial poses and occlusions.

In order to train the DAMs model, we combined 811 training images from LFPW and 2000 images from Helen database into one training set. The coordinates of facial landmarks were normalized to zero mean before setting as observations to train the shape model. In the texture modeling, shape-free images were first extracted by warping faces into the texture domain $\mathcal{D}$. The size of the shape-free image was set to 117×120 pixels based on the mean shape of the training data. Then texture model was trained to learn the facial variations represented in these shape-free images.

During the testing phase, since the number of visible units in the texture model are fixed, the testing low-scale facial shape-free image was first resized to 117×120 using bicubic interpolation method. Then both the shape and the shape-free image were clamped to DAMs. After 50 epoches in the alternating Gibbs updates, the face texture was reconstructed based on the current states of hidden unit $\mathbf{h}_g^{(1)}$. Different magnification factors α were used for evaluating the quality of DAMs reconstructions. Testing images were down-sampled in different magnification levels ranging from 4 to 12. They were then used as the inputs to the reconstruct module using our approach. Figure 3 shows the reconstruction results using the DAMs approach. Remarkable results are achieved using DAMs with very low-resolution input images, i.e. 10×10 pixels with the magnification factor $\alpha = 12$.

Our proposed approach is also compared with two baseline methods, i.e. bicubic interpolation method and PCA-based AAMs [23]. Root Mean Square Error (RMSE) is used as a performance measurement. RMSE is a common metric that is usually used for evaluating image recovery task. Although this metric is not always reliable for rating image quality visually [25], it could provide a qualitative view for comparing DAMs and other methods.

Although our method gives better reconstruction results in visualization than the others as shown in Figure 4, the RMSE results are not much better as shown in Table 1. This is because RMSE cannot fully evaluate the quality of reconstructed images in the task of image super-resolution [27]. Especially, we don't have the ground-truth for RMSE evaluation in these databases. For example, in the cases of occlusions and poses in those databases, although the reconstructed images obtained using PCA-based AAMs and bicubic methods are very blurry, their RMSEs are still low. This is because the reconstructed images still contains occlusion components or pose features which are quite similar to the original ones.

Figure 8 illustrates further reconstruction results obtained using bicubic method, PCA-based AAMs method and our DAMs approach. The PCA-based AAMs method is trained using the same dataset as DAMs and the length of texture parameter vector is 200, the highest level used in [23]).

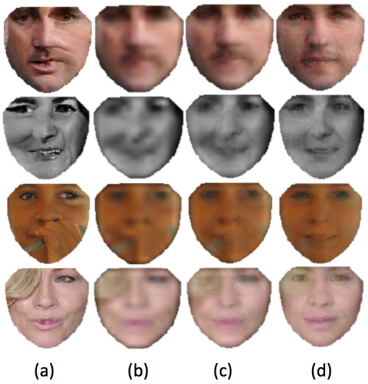

Figure 4. Results of average RMSEs over 4 images: (a) texture image; (b) Bicubic interpolation (RMSE = 14.95); (c) PCA-based AAMs reconstruction (RMSE = 15.18); (d) Deep Appearance Models reconstruction (RMSE = 16.25).

Table 1. The average RMSEs of reconstructed images using different methods against LFPW and Helen databases with $\alpha = 16$.

Methods	LFPW	Helen
Bicubic	19.53	22.13
AAMs [23]	19.74	22.3
DAMs (Ours)	**19.24**	**21.24**

Figure 5. Facial off-angle reconstruction. The 1st row: original image, the 2nd row: shape-free image, the 3rd row: PCA-based AAMs reconstruction [23], the 4th-row: DAMs reconstruction.

5.3. Facial off-angle Reconstruction and Occlusion Removal

This section illustrates the ability of DAMs to deal with facial poses and occlusions. Using the same trained model as in the previous experiment, facial images with different poses are represented in Figure 5.

Comparing to AAMs, our DAMs achieves better reconstructions especially in the invisible regions of extreme poses. These regions in shape-free images are blurry and noisy due to the non-linear warping operator. Therefore, the errors are spread out in the reconstructions of PCA-based

Figure 6. Occlusion removal: the 1st row: original image, the 2nd row: shape-free image, the 3rd row: PCA-based AAMs reconstruction still remains with occlusion and blurring effects, and the 4th-row: DAMs reconstruction can help to remove the occlusion.

Table 2. The MAEs (years) of different methods against impulsive noise.

Methods	No noise	Noise range			
		25	50	100	150
AAMs [23]	6.14	6.15	6.11	**6.13**	6.47
DAMs	**5.67**	**5.81**	**5.56**	6.14	**6.18**

AAMs approaches. Meanwhile, the generative capability of our proposed DAMs method can solve those challenging cases. From the results, it is easy to see that the blurry effects are effectively removed in DAMs reconstructions. Similarly, DAMs also shows its capability in the problem of facial occlusion removal. In Figure 6, the occlusions, e.g. hands, glasses, hair, etc., can be removed successfully without blurring effects. More interestingly, the occlusions are removed from faces without loosing facial features. For example, glasses are totally removed without making beard blurred as in the PCA-based AAMs reconstruction.

Using occluded faces as references and measuring the reconstruction quality by RMSE cannot illustrate the modeling capabilities of DAMs. To get a better evaluation protocol, we select a subset of 174 occluded faces of the first 29 subjects, i.e. 15 males and 14 females, from AR database [14]. We employ DAMs to reconstruct these occluded faces and then use their corresponding neutral faces, i.e. frontal face without occlusions, as references to compute the RMSE. In this testing set, each subject includes two faces with scarf and four other faces with both illumination and scarf. The average RMSE of DAMs is 45.08 while that of PCA-based AAMs is 47.36. The trained models in DAMs and AAMs use LFPW and Helen databases presented in Section 5.1. This experiment shows that DAMs achieve better reconstructions, i.e. closer to the neutral faces, compared to AAMs.

5.4. Facial Age Estimation

Besides some other previous age estimation approaches [7, 13], we will employ our proposed DAMs into this problem to further demonstrate its robustness and effectiveness.

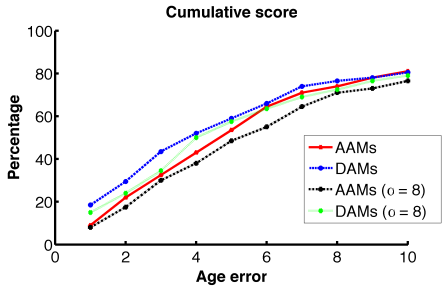

Figure 7. Cumulative scores of using reconstructed images from original-scaled and down-sampled images with a factor of 8.

Table 3. The MAEs (years) of different methods against low-resolution testing faces.

Methods	Magnification factor α			
	2	4	6	8
Bicubic	5.96	6.95	7.15	7.21
AAMs [23]	6.13	6.33	6.44	6.69
DAMs	**5.91**	**6.00**	**6.11**	**6.21**

Evaluations on reconstructed images: Since the texture is an important factor to predict a person's age given his facial image, this experiment will evaluate how good the reconstructed image is as well as how much aging information is retained by the model.

To make this task more challenging, we add noise to the testing facial image and then predict the age of that person using "clean" reconstructed face from DAMs. For the evaluation system, we re-implemented the age estimation system presented in [13] and trained it with 802 images from FG-NET. The remaining 200 images were used for testing. To generate noisy testing images, all pixels of facial images were mixed with uniform noise ranged within $[-r, r]$.

A similar experiment is also set up as follows: given the low-resolution testing face, the system will predict the age of that person using his high-resolution reconstructed face. The Mean Absolute Errors (MAEs) of different methods against noise and low-resolution testing faces are represented in Table 2 and Table 3, respectively. The performance in terms of Cumulative Scores (CS) is illustrated in Figure 7. From these results, in both cases, the smallest error is achieved with DAMs model. Therefore, our proposed model produces better reconstructed results under the effects of noise and low-resolution factors.

Evaluation on model features: Beside the ability of generalizing the faces, DAMs can produce a higher level representation for both facial shape and texture. Therefore, instead of using pixel values, we extracted the model parameters as described in Section 3.4 and evaluated them with the age estimation system. For the AAMs features, the number of features for shape and texture was chosen so that 93% of

Figure 8. Facial image superresolution. The original images (first row) are warped to shape-free images in texture domain (second row); then they are down-sampled by a factor of 8 from 117×120 to 15×15 (third row) The next three rows are the high-resolution reconstructed by bicubic method (fourth row), PCA-based AAMs (fifth row) and Deep Appearance Models (sixth row).

Table 4. Comparison of age estimation results on FG-NET database with four different features.

Inputs	MAEs (years)
DAMs-Mod	**5.28**
AAMs-Mod	5.35
DAMs-Rec	5.67
AAMs-Rec	6.14

Table 5. Computational time (hrs - hours, s - seconds) of DAMs and AAMs in three stages.

Stages	DAMs	AAMs [23]
Training	12.87 hrs	564.06 s
Fitting (per image)	19.17 s	2.28 s
Reconstruction (per image)	0.53 s	0.023 s

variations are retained. Table 4 lists the MAEs of four different inputs: reconstructed image of DAMs (**DAMs-Rec**) and AAMs (**AAMs-Rec**), model parameters extracted from AAMs (**AAMs-Mod**) and DAMs (**DAMs-Mod**). Not surprisingly, our DAMs feature achieves the lowest MAEs as compared with AAMs features.

6. Computational Costs

The computational costs of DAMs, i.e. training, fitting and reconstruction stages are discussed in this section. Both LFPW and Helen databases are combined to use in this evaluation. The numbers of training and testing images are 2811 and 554, respectively. The method is implemented in Matlab environment and runs in a system of Core i7-2600 @3.4GHz CPU, 8.00 GB RAM. The shape contains 68 landmarks and the appearance is represented in a vector of 9652 dimensions. Each layer was trained using Contrastive Divergence learning in 600 epochs. It is notice that the current version is implemented without using parallel processing. The computational costs of both DAMs and AAMs are shown in Table 5.

7. Conclusions

This paper has introduced novel Deep Appearance Models that have abilities of generalizing and representing faces in large variations. With the deep structured models for shapes and textures, the proposed approach was shown to achieve remarkable improvements in both facial reconstruction and facial age estimation tasks compared with PCA-based AAMs model. Moreover, the new model can produce a more robust face shape and texture representation based on their high-level relationships. Experimental results in several applications such as facial super-resolution, face off-angle reconstruction, occlusion removal and facial age estimation have shown the potential of the model in dealing with large variations.

References

[1] *FG-NET Aging Database*. http://www.fgnet.rsunit.com.

[2] B. Amberg, A. Blake, and T. Vetter. On compositional image alignment, with an application to active appearance models. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1714–1721. IEEE, 2009.

[3] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 545–552. IEEE, 2011.

[4] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Interprettting Face Images using Active Appearance Models. In *Proc. of the 3^{rd} Intl. Conf. on Automatic Face and Gesture Recognition*, pages 300–305, 1998.

[5] R. Donner, M. Reiter, G. Langs, P. Peloschek, and H. Bischof. Fast active appearance model search using canonical correlation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1690, 2006.

[6] S. A. Eslami, N. Heess, C. K. Williams, and J. Winn. The shape boltzmann machine: a strong model of object shape. *International Journal of Computer Vision*, 107(2):155–176, 2014.

[7] Y. Fu and T. S. Huang. Human age estimation with regression on discriminative aging manifold. *Multimedia, IEEE Transactions on*, 10(4):578–584, 2008.

[8] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12):1080–1093, 2005.

[9] X. Hou, S. Z. Li, H. Zhang, and Q. Cheng. Direct appearance models. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–828. IEEE, 2001.

[10] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In *Proceedings of the international conference on Multimedia information retrieval*, pages 527–536. ACM, 2010.

[11] S. Z. Joan Alabort-i medina. Bayesian active appearance models. In *Computer Vision and Pattern Recognition, 2014. CVPR 2014. IEEE Conference on*, pages 3438–3445. IEEE, 2014.

[12] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *Computer Vision–ECCV 2012*, pages 679–692. Springer, 2012.

[13] K. Luu, K. Ricanek, T. D. Bui, and C. Y. Suen. Age estimation using active appearance models and support vector machine regression. In *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*, pages 1–5. IEEE, 2009.

[14] A. Martınez and R. Benavente. The ar face database. *Rapport technique*, 24, 1998.

[15] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011.

[17] G. Papandreou and P. Maragos. Adaptive and constrained algorithms for inverse compositional active appearance model fitting. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[18] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 896–903. IEEE, 2013.

[19] R. Salakhutdinov and G. E. Hinton. Deep boltzmann machines. In *Intl. Conf. on Artificial Intell. and Statistics*, pages 448–455, 2009.

[20] J. Saragih and R. Goecke. A nonlinear discriminative approach to aam fitting. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[21] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[22] G. W. Taylor, L. Sigal, D. J. Fleet, and G. E. Hinton. Dynamical binary latent variable models for 3d human pose tracking. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 631–638. IEEE, 2010.

[23] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 593–600. IEEE, 2013.

[24] L. Van Der Maaten and E. Hendriks. Capturing appearance variation in active appearance models. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 34–41. IEEE, 2010.

[25] Z. Wang and A. C. Bovik. Mean squared error: love it or leave it? a new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, 26(1):98–117, 2009.

[26] Y. Wu, Z. Wang, and Q. Ji. Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3452–3459. IEEE, 2013.

[27] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image superresolution via sparse representation. *Image Processing, IEEE Transactions on*, 19(11):2861–2873, 2010.