2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Discriminant Auto Encoders for Face Recognition with Expression and Pose Variations

Chathurdara Sri Nadith Pathirage
Department of Computing
Curtin University
Perth, Western Australia
Email: c.nadithpa@postgrad.curtin.edu.au

Ling Li
Department of Computing
Curtin University
Perth, Western Australia
Email: L.Li@curtin.edu.au

Wanquan Liu
Department of Computing
Curtin University
Perth, Western Australia
Email: W.Liu@curtin.edu.au

***Abstract*—The key challenge of face recognition is to develop effective feature representations for reducing intra-personal variations while enlarging inter-personal differences. This paper presents a novel non-linear discriminant error criterion which can be used in effective feature learning from raw pixels. Unlike many existing methods which assume the problem to be linear in nature, the proposed method utilizes a novel deep learning (DL) framework which makes no prior assumptions thus exploiting the full potential of learning a highly non-linear transformation. High level representations learnt via the proposed model are highly supervised and can help to boost the performance of subsequent classifiers such as LDA. This study clearly shows the value of using non-linear discriminant error criterion as a tractable objective to guide the learning of useful high level features in various face related problems. The extracted features are learnt from local face regions and the results of the experiments performed on 3 different face image databases demonstrate the superiority and the generalizability of our method compared to existing work, as well as the applicability of the concept onto many different deep learning models of the same nature.**

I. Introduction

Face recognition (FR) has been broadly investigated and used successfully in various computer vision related areas during the past decades. While many face recognition methods were proven to achieve impressively good performance in the constrained environments, their performance is still unsatisfactory in unconstrained environments mainly due to the large variations in face images caused by pose, expression, lighting, occlusion and etc. In fact, a face with such variations can be regarded as a neutral face exposed to different kind of face noises. De-noising such face noises while extracting robust and discriminative features to enlarge the inter-personal variations and shrink the intra-personal margins at the same time remains a central and challenging problem in face recognition.

In order to obtain effective features, many promising works have been conducted in the recent years. They can be mainly classified into linear [17][3][1][8] and non-linear feature based methods. Regardless of the complexity involved in non-linear techniques, Deep Learning (DL) methods have shown an impressive progress over conventional linear techniques due to the immense power of learning data adaptive features for problems that possess highly non-linear characteristics. These methods are mainly based on Convolutional Neural Network (CNN) [9], Deep Belief Networks (DBN) [6] and Deep Auto Encoder (DAE) [4].

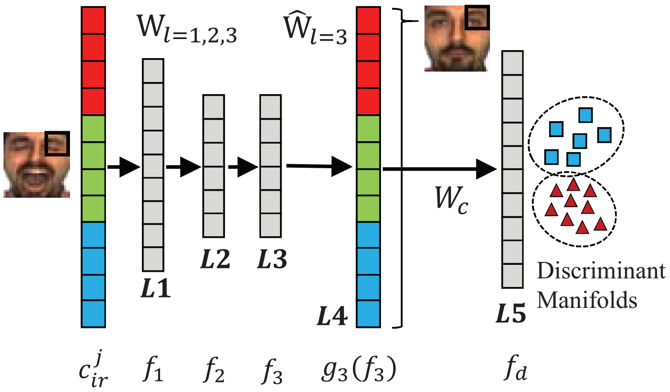

Fig. 1. DDA Net where $c_{ir}^{j} \in \mathbb{R}^{(36*3)}$, $f_1 \in \mathbb{R}^{75}$, $f_2 \in \mathbb{R}^{50}$ denote the combined patch feature and the low dimensional noisy feature learned at Layer 1 (L1) and Layer 2 (L2) respectively while $f_3 \in \mathbb{R}^{50}$ denotes the noise-less feature learned at Layer 3 (de-noising layer) in the observed low dimensional space. $g_3(.)$ represents the decoder function. Hence the discriminant layer where $f_d \in \mathbb{R}^{class\ count-1}$ is shown as the right most layer.

In general, CNNs that consist of very high number of parameters (weights) require huge amounts of data to pre-train and to fine-tune the model accurately. This is due to the complexity involved in such architectures compared to auto encoder models. Hence CNNs were used with large datasets for training involving different illumination and slight pose variations. In contrast, DAE models provide simpler design to reduce the complexity involved in CNN. In [10], a decent DAE model is proposed to handle the pose problem in FR, where each shallow Auto Encoder (AE) is trained to achieve simple but tractable goals required to address the global non-linear objective. This model requires pose information in the training phase hence limits the optimization with the number training samples with the pose. Irrespective of the models above, the structure of a dataset that can be described with the inter-class and intra-class evaluation criteria was generally not incorporated into the objective of learning discriminative features.

In this paper, we propose a DAE model that utilizes a novel non-linear discriminant error criterion. In contrast to the hand engineered features (LBP [1], SIFT [8]) which do not follow a learning process, the proposed model can learn dynamic data-adaptive features directly from the raw pixels. These features are hence highly coherent and able to effectively characterize the higher order discriminant information for various problem

978-1-5090-4847-2/16/$31.00 ©2016 IEEE

domains such as face pose and expressions. More importantly, our method requires no prior information about the pose or expression and can be trained with any database of interest in pre-training and fine-tuning stages. Different face regions usually have different structures and it is beneficial to learn more region-specific features for a better discriminant face representation. We perform discriminant feature learning for each patch via stacking feature learning modules in a deep architecture to exploit hierarchical and complementary information to enhance the recognition performance. The main contributions of the proposed method are listed as follows:

- Problem formulation for non-linear discriminant analysis to yield effective data-adaptable features via DL Nets;
- The patch based framework is proposed along with other learning constraints to breakdown the non-linearity. Thus facial expressions and face poses are formulated as tractable goals while achieving a high level of generalization.

The rest of the paper is organized as follows. Section 2 details the novel non-linear criterion and the Deep Auto Encoder Model that utilizes this criteria (Deep Discriminant Analysis Nets); Section 3 evaluates the performances of the proposed model on AR, Curtin and MultiPIE databases against expression and pose problems followed by the conclusion in Section 4.

II. Non-Linear Discriminant Analysis

This section includes discussions on the novel non-linear discriminant cost model and its optimization.

A. *Non-Linear Discriminant Analysis*

We propose a novel discriminant cost formulation in learning the non-linear discriminant subspace where the inter-class distance is maximized while minimizing the intra-class distance.

Assume that training data is given as $X = [x_1, x_2, ..., x_N]$ where $x_j \in \mathbb{R}^d (j = 1, 2, ..., N)$ represents the image of a person as a column vector and N is the total number of images. Let N_i denote the number of images in class $i (i = 1, 2, ..., C)$ such that $N = \sum_i N_i$. We define the between-class cost C_b and within-class cost C_w as follows:

$$C_b(W) = \sum_{i=1}^{C} N_i \; \vartheta(\; \Phi(Wm_i) - \Phi(Wm) \;) \tag{1}$$

$$C_w(W, x_j) = \sum_{i=1}^{C} \sum_{x_j \in C_i} \vartheta(\; \Phi(Wx_j) - \Phi(Wm_i) \;) \tag{2}$$

where $m = (1/N) \sum_{i=1}^{N} x_i$ is the mean of the dataset, m_i represents the mean of the class i, $\Phi(\mathbf{x}) = sigmoid(\mathbf{x}) = 1/(1 + e^{-x})$ which is the non-linear squashing function and $\vartheta(.)$ denotes a distance metric which in our case, the $L2$ norm. The objective of the non-linear discriminant analysis is to find a set of directions on a manifold in which the between-class distance is maximized while minimizing the within-class distance as:

$$C(W, x_j) = (C_w(W, x_j)/N) + \lambda(N/C_b(W)) \tag{3}$$

where λ is the regularzing parameter in the cost function between the within-class and between-class distances and it can be learnt via validation dataset. Our aim is to learn the projection W via optimization.

B. *Deep Discriminant Analysis (DDA) Nets*

In this section we describe the motivations behind the proposed model and the constitution of the novel framework.

1) Evolution of DDA Nets: A typical DL cost model learns a feature representation that is not necessarily discriminative itself. AE model could fail to learn effective features in reconstructing the original input from a noisy input due to the absence of discriminant information in the commonly used squared error function that utilized to train the model in an unsupervised manner. In our approach, we perform metric learning over different classes while shrinking the inner-class distance and expanding the between-class distance via utilizing the non-linear distance measurement explained above. Similar models such as [19][10][14] can be further improved by incorporating such class discriminant information. In general, DAE models are fairly easy to train and can be used to subdivide the global non-linearity involved in a certain problem domain into a series of sub-objectives where each sub-objective is modelled via a simple AE [10]. Therefore the proposed model is built based on DAE model with an aim to learn a projection of the input into a deep feature space (Fig. 1). The manifold distances are calculated based on the single neutral face representative instance.

2) DDA Net Framework: We utilize a novel deep architecture where each shallow AE is trained to achieve simple but tractable goals required to address the global non-linear objective as a whole. The framework follows a patch based approach (Fig. 2) to further refine the global non-linear objective into simpler tasks. We choose non-overlapping patches of the face image of size and stride 6x6 respectively. Furthermore, it limits the number of parameters of the model that need to be learnt while training each DDA Net in a parallel environment.

The proposed DDA Net model consists of three interconnected learning process: the progressive non-linear dimension reduction process, de-noising process, and a discrimination process. The first two layers of the framework perform non-linear dimension reduction and yield a low dimensional feature whose effective dimension is half the dimension of the original RGB features. The 3^{rd} layer will perform the de-noising based on a strong supervisory signal which is the neutral frontal face in our case. The last layer will perform the discriminant analysis based on a single representative face image thus ensures the features observed in the reconstruction layer are highly discriminative. The stacked R, G, B channel patch features are combined and used as the input to the framework. The combined patch features can be represented as:

$$c_{ir}^{j} = [p_{ir}^{jR}; \; p_{ir}^{jG}; \; p_{ir}^{jB}] \tag{4}$$

where $p_{ir}^{jR}, p_{ir}^{jG}, p_{ir}^{jB} \in \mathbb{R}^{36}$ are the corresponding colour channels RGB for the j^{th} patch of the r^{th} image that belongs to the i^{th} class. The combined patch feature (R, G, B) of the

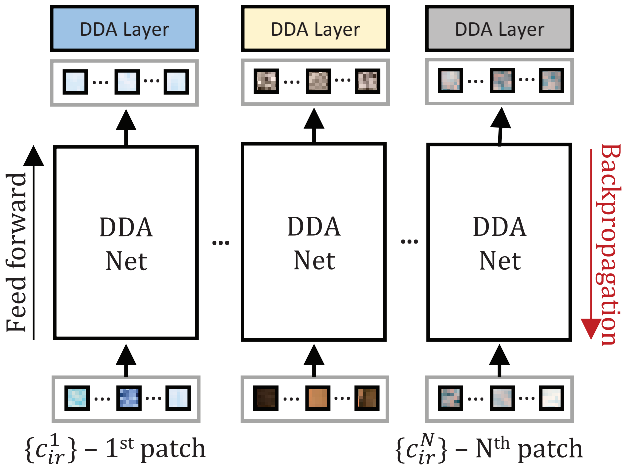

Fig. 2. Patch based DDA Framework that converts each patch of a face image to its corresponding frontal face patch followed by the non-linear discriminant analysis process (DDA layer).

r^{th} image in the i^{th} class is denoted by $c_{ir}^{j} \in \mathbb{R}^{108}$. In order to reduce the effects of illumination and colour contrasting, histogram equalization is performed on the V-channel after converting the RGB image to the HSV colour space. The pixel intensities are then normalized from the range 0-255 to 0-1 to be compatible with the DDA Net's operating range. This setup ensures that the proposed model learns the optimal combinations of weights for pixel based on the global objective function. We perform the dimensionality reduction in steps (75% energy, 50% energy) as shown in Fig. 1. Formally, Layer 1 and Layer 2 are progressive non-linear dimension reduction layers and Layer 1 objective function is,

$$\left[W_{l=1}^{*}, b_{l=1}^{*}, \widehat{W}_{l=1}^{*}, \widehat{b}_{l=1}^{*}\right]_{j} = argmin_{W,b,\widehat{W},\widehat{b}} \sum_{i=1}^{S}\sum_{r=1}^{N_i}\left\|c_{ir}^{j} - g_1(f_1(c_{ir}^{j}))\right\|_2^2 \tag{5}$$

where S is the total number of identities, N_i is the number of images that belongs to the i^{th} class, $c_{ir}^{j} \in \mathbb{R}^{108}$ is as described above and $g_1(.), f_1(.)$ are the decoder and encoder functions respectively. $g_1(x) = f_1(x) = sigmoid(\mathbf{x}) = \frac{1}{1+e^{-x}}$. The Layer 2 objective function is as follows,

$$\left[W_{l=2}^{*}, b_{l=2}^{*}, \widehat{W}_{l=2}^{*}, \widehat{b}_{l=2}^{*}\right]_{j} = argmin_{W,b,\widehat{W},\widehat{b}} \sum_{i=1}^{S}\sum_{r=1}^{N_i}\left\|\left\{h_{ir}^{j}\right\}_1 - g_2(f_2(\{h_{ir}^{j}\}_1))\right\|_2^2 \tag{6}$$

where $\{h_{ir}^{j}\}_1 \in \mathbb{R}^{75}$ is the learnt representation of Layer 1 for the j^{th} patch of the r^{th} image that belongs to the i^{th} class and all other parameters are the same as above. The nodes in the higher layer learn the statistical dependencies among the nodes in the adjacent lower layer so that the higher layer can discover more complex patterns (abstract) in the input by eliminating the noise and reducing dimensions. Next, a de-noising layer is employed to perform the de-noising against the frontal face.

Algorithm 1 Training Algorithm:

```
$X^j = [c_{11}^j ... c_{1N_1}^j c_{21}^j ... c_{2N_2}^j ... c_{S1}^j ... c_{SN_S}^j]$;
$V^j = validation\ patch\ set$;
$\{W^j\}_0 = weights\ obtained\ after\ pretraining\ stage$;
$\{ve^j\}_0 = +\infty \quad //temporary\ variables$;
$count_{epoch} = 0$;
$error_{validation} = +\infty$;
while $(error_{validation} > 0) AND (count_{epoch} < t)$ do
    $F^j = DDA_4(X^j, \{W^j\}_{idx})$
    $e^j = C(\{W_c^j\}_{idx},\ F^j)$
    $\nabla\{W_c^j\}_{idx} = \frac{\partial e^j}{\partial\{W_c^j\}_{idx}}$
    $\nabla F^j = \frac{\partial e^j}{\partial F^j}$
    $\nabla\{W^j\}_{idx} = \nabla F^j \frac{\partial DDA_4(X^j, \{W^j\}_{idx})}{\partial\{W^j\}_{idx}}$
    $\{W_c^j\}_{idx+1} = \{W_c^j\}_{idx} - \ell\ (\nabla\{W_c^j\}_{idx})$
    $\{W^j\}_{idx+1} = \{W^j\}_{idx} - \ell\ (\nabla\{W^j\}_{idx})$
    $VF^j = DDA_4(V^j, \{W^j\}_{idx+1})$
    $\{ve^j\}_{idx+1} = C(\{W_c^j\}_{idx+1},\ VF^j)$
    $error_{validation} = \{ve^j\}_{idx} - \{ve^j\}_{idx+1}$
    $count_{epoch} = count_{epoch} + 1$
end while
{where $idx$ is the iteration index, $\ell$ denotes the learning
    rate, $t$ is a large positive integer to denote the maximum epoch count and $DDA_4(.)$ represents the features
    obtained at the reconstruction layer (Layer 4) and $W_c$
    is the projection matrix to the DDA space. In training,
    the breaking condition that mostly occurs is the validation
    error criterion.}
```

The Layer 3 objective function can be written as:

$$\left[W_{l=3}^{*}, b_{l=3}^{*}, \widehat{W}_{l=3}^{*}, \widehat{b}_{l=3}^{*}\right]_{j} = argmin_{W,b,\widehat{W},\widehat{b}} \sum_{i=1}^{S}\sum_{r=1}^{N_i}\left\|\left\{c_{ir}^{j}\right\}_F - g_3(f_3(\{h_{ir}^{j}\}_2))\right\|_2^2 \tag{7}$$

where $\{h_{ir}^{j}\}_2 \in \mathbb{R}^{50}$ is the learnt representation of Layer 2 for the j^{th} patch of the r^{th} image that belongs to the i^{th} class, and $\{c_{ir}^{j}\}_F \in \mathbb{R}^{108}$ is the corresponding combined feature of the corresponding frontal face of the input face image. The denoising layer acts as a decoder as well as a strong regularizer where all the lower layers can be treated as the encoding layers. Finally, we utilize the DDA layer on the reconstructed image layer (Layer 4) to learn the initial projection matrix W_c during the pre-training process. The Layer 5 (DDA layer) objective function can be written as:

$$[W_c^{*}]_j = argmin_{W_c}\ \ C(W_c,\ g_3(\{h_{ir}^{j}\}_3)) \tag{8}$$

where $\{h_{ir}^{j}\}_3 \in \mathbb{R}^{50}$ is the learnt representation of Layer 3 for the j^{th} patch of the r^{th} image that belongs to the i^{th} class and $C(.)$ is the cost model described in Eq. 3. We use the neutral face image of each class as the representative image instead of using the actual mean image. This will ensure that

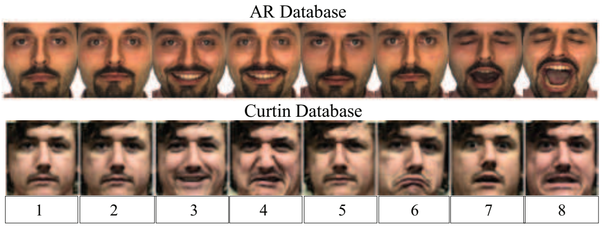

Fig. 3. Different expressions and the corresponding indices.

the proposed model will learn a forceful projection for each noisy face image to approximate its frontal neutral face.

After constructing a DDA framework with initialized DDA Nets, each image patch can be transformed into its corresponding neutral face patch at the reconstruction layer while establishing hierarchical feature representations at different layers. In order to refine the features further, we perform training on the database of interest by stacking the pre-initialized layers one after another to optimize the objective as below:

$$\begin{aligned}\left[W_l^*|_{l=1}^L,\ b_l^*|_{l=1}^L,\ \widehat{W}_L^*,\ \widehat{b}_L^*,\ W_c^*\right] = \\ argmin_{W_l|_{l=1}^L, b_l|_{l=1}^L, \widehat{W}_L, \widehat{b}_L, W_c}\ \ C(W_c,\ p(c_{ij}^r))\end{aligned} \tag{9}$$

where $p(c_{ir}^j) = g_L(f_L(f_{L-1}(f_{L-2}(c_{ir}^j))))$ with $L = 3$; and $W_l|_{l=1}^L$ denotes the encoders weights, $\widehat{W}_L$ denotes the decoder weights and W_c is the projection matrix learnt via optimizing the non-linear discriminant error criteria $C(.)$.

We employ the full batch gradient descent algorithm and the backpropagation mechanism on the cost functions to find the optimal parameters. After the fine-tuning phase of the DDA Nets, the feature representations learnt at the de-noising layer f_3 and the reconstruction layer f_4 will ensure that they are highly discriminative and suitable for recognition purpose. Since the low dimensional feature (f_3) and the features at the reconstruction layer (f_4) are obtained via a strong supervised non-linear discriminant criterion, these features are highly favourable with LDA analysis followed by Nearest Neighbour (NN) classifier for recognition. In addition, we utilize the DDA Net features followed by PCA and SRC separately to observe the properties of the learnt features. We report the experimental results in the following section.

III. Experiments

Face images in AR [11], Curtin [12] and MultiPIE [5] databases are used for experiments. All images are cropped, aligned and resized to 33x33 resolution. Experiments were conducted on expressions and poses separately to evaluate the effectiveness of proposed approach. Furthermore, a validation dataset is used in every test case to select the optimal parameters for the objective functions.

A. Facial Expression Experiments

Three distinct experiments were conducted in regard to the expression problem, each consisting of 8 test cases (TC) to evaluate the performance of the proposed model. Fig. 3 shows the images used with different expressions including extremely opened mouth.

1) Same Identity Experiment: In this setup, experiments were conducted on the AR database in isolation, and training and testing were performed on different images of the same subject. One image out of the 8 images from each identity was taken for testing while the other 7 images from the same identities were used for training. The test cases were formed as shown below:

- For each test case i: Test on the i^{th} image of each identity and train on the remaining 7 images of each identity.

The data split for the training, validation and testing process of the 100 identities in the AR database is described as follows. One of the 8 images from 75 identities were used for testing in each test case. The remaining 7 images from the 75 identities were used for training. Images of the same indices from the other 25 identities were used for validation. Results of face recognition on the same identity experiment are shown in Table I.

TABLE I. Results of the same identity experiments on AR database.

Method	1	2	3	4	5	6	7	8
DDA-SRC	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
DDA-LDA	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**99.2**
DDA-PCA	**98.7**	**98.7**	**96**	**96**	**94.7**	**98.7**	**84**	**84**
SFDAE-SRC [14]	100	100	100	100	97.3	98.7	98.7	96
SFDAE-LDA [14]	100	100	100	100	100	100	96	93
SFDAE-PCA [14]	98.7	98.7	94.7	97.3	94.7	97.3	80	79
SRC [18]	100	98.7	98.7	97.3	96	97.3	94.7	96
KDA (Gaussian)	100	100	100	100	98.7	100	97.3	97.3
KPCA (Gaussian)	94.6	94.6	90.6	93.3	86.6	96	77.3	81.3
LDA	100	100	100	98.7	97.3	100	93.3	92
PCA	94	93.3	88	92	84	94.7	74.7	78.7

2) Cross Identity Experiment: In this Experiment, training and testing are performed on mutually exclusive datasets from the AR database. This setting exploits the model's generalization ability on mutually exclusive datasets that were observed under the same environmental conditions.

All 8 images of 50 identities from the AR database were taken for training and another 25 identities were used for validation. Images of the remaining 25 identities were split into gallery and test image. The i^{th} test case follow the same format as described in the previous setup, i.e. in Test Case i ($i = 1, ..., 8$), Image i was used for testing and the other 7 images formed the gallery. Results are shown in Table II:

TABLE II. Results of the cross identity experiments on AR database.

Method	1	2	3	4	5	6	7	8
DDA-SRC	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
DDA-LDA	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
DDA-PCA	**100**	**97**	**89**	**95.7**	**95.7**	**93**	**85**	**89**
SFDAE-SRC [14]	100	100	100	100	100	100	98	97
SFDAE-LDA [14]	100	100	100	100	100	100	96	93
SFDAE-PCA [14]	100	95.7	87	95.7	95.7	91.3	82	87
SRC [18]	100	100	95.6	100	95.7	95.7	95.7	91.3
KDA (Gaussian)	100	100	100	100	95.7	95.7	81.3	89.3
KPCA (Gaussian)	82.6	95.6	82.6	91.3	82.6	91.3	70	82.6
LDA	100	100	95.7	100	95.7	95.7	78.3	87
PCA	78.3	91.3	70	87	78.3	91.3	69.6	74

3) Cross database experiment: In this experiment, training and testing were performed on different databases that are captured under different environmental conditions. This is a challenging task for the proposed framework with an aim to test the generalization ability for face images captured under the various environmental conditions. The i^{th} test case follows the same formats described in the previous setup. The training set consist of 75 subjects from the AR database and another 25 subjects from AR were taken for validation. 50 subjects from the Curtin database were used for testing. Experimental results in terms of face recognition rates are shown in Table III:

TABLE III. RESULTS OF THE CROSS DATABASE EXPERIMENTS ON AR AND CURTIN DATABASE.

Method	1	2	3	4	5	6	7	8
DDA-SRC	**100**	**100**	**100**	**100**	**100**	**100**	**92**	**100**
DDA-LDA	**100**	**100**	**100**	**100**	**100**	**100**	**94**	**100**
DDA-PCA	**98**	**100**	**98**	**100**	**98**	**100**	**92**	**98**
SFDAE-SRC [14]	100	100	100	100	100	100	90	97
SFDAE-LDA [14]	100	100	100	100	100	100	90	93
SFDAE-PCA [14]	98	100	98	96	96	98	86	88
SRC [18]	100	100	100	100	96	100	88	100
KDA (Gaussian)	98	100	100	98	96	98	82	90
KPCA (Gaussian)	98	100	96	100	98	98	88	94
LDA	96	100	100	96	90	94	76	88
PCA	96	96	88	98	94	94	82	92

Discussion. As shown in Table I, Table II and Table III, DDA Net performs consistently better compared to other methods for face images with different types of facial expression including extremely opened mouth. The differences in DDA-PCA vs SFDAE-PCA and (K)PCA, DDA-LDA vs SFDAE-LDA and (K)LDA, DDA-SRC vs SFDAE-SRC and SRC clearly show the improvement by utilizing the non-linear discriminant error criterion. Comparing DDA-LDA and DDA-PCA, it is conceivable that the opened mouth faces lie in a LDA structure (Gaussian) in the observed low dimensional space. It also shows that learnt features favour linear approximation of the face images in the observed space (refer DDA-SRC performance). Hence DDA Net outperforms the other linear dimensionality reduction techniques and the shallow models (kernals). Irrespective of the identities at training and the environmental conditions in which the images are acquired (Table II and Table III), DDA Net is highly invariant to the opened mouth noise thus claims its immense generalization ability with better low dimensional features for robust face recognition.

B. Face Pose Testing

The pose related experiments were carried out in the same fashion as mentioned in the previous section. The gray scale images showing various face pose from the popular MultiPIE database [10] as well as the Curtin database are used in these experiments. There are 6 poses per subject in the MultiPIE database, as shown in Fig. 4. Test cases are defined as in the previous experiments, except that there are now only 6 of them (excluding the frontal pose) which are labelled by the pose angles. We follow the usual test setups of other pose related methods where the same identity experiment is excluded since the training needs to contain the pose to be recognized. The other two experiments were conducted instead, as detailed below.

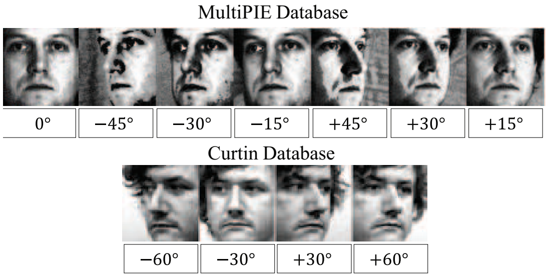

Fig. 4. Images with different poses and the indices.

1) Cross identity experiment: In this setup, training and testing was performed on mutually exclusive datasets from the MultiPIE database. 6 different test cases (excluding the frontal case) on different face pose of varying degree 45° – 15° are conducted. We follow the same dataset configuration used in [10] to compare and contrast our method with the existing state of the arts 2D and 3D methods presented in [10]. The images of all 337 subjects from MultiPIE at 7 poses with neural expression and frontal illumination are used. We use images from the first 200 subjects (subject ID 001 to 200) for training, which consisted of 4,207 images in total. The images from the remaining 137 subjects are used for testing, with 1,879 images in total. Out of them, the frontal face images from the earliest session for the 137 subjects are used as gallery images (137 in total), and images from the other poses are used as probe images (1,742 in total). SRC, LDA, PCA, KDA, KPCA results were excluded since these techniques failed (or not significant) and not robust under large pose variations (beyond the range of -15° to +15°). Experimental results in terms of face recognition rates are shown in Table IV:

TABLE IV. RESULTS OF THE CROSS IDENTITY EXPERIMENTS ON MULTIPIE.

Method	-45°	-30°	-15°	+15°	+30°	+45°	Average
DDA-SRC	**73**	**92**	**98.5**	**98.5**	**93**	**74.1**	**88.2**
DDA-LDA	**63**	**90**	**99.2**	**99.2**	**92**	**64.4**	**84.6**
DDA-PCA	**60**	**86**	**98.5**	**98.5**	**88**	**60.7**	**82**
SPAE [10]	84.9	92.6	96.3	95.7	94.3	84.4	91.4
DAE [4]	69.9	81.2	91	91.9	86.5	74.3	82.5
GMA [16]	75	74.5	82.7	92.6	87.5	65.2	79.6
CCA [7]	53.3	74.2	90	90	85.5	48.2	73.5
PLS [15]	51.1	76.9	88.3	88.3	78.5	56.5	73.3
MDF [13]	78.7	94	99	98.7	92.2	81.8	90.7
Asthana11 [2]	74.1	91	95.7	95.7	89.5	74.8	86.8

2) Cross database experiment: This setup aims to evaluate the proposed model against different databases captured under different environmental conditions. This is a novel experiment setting to assess the generalization ability of the proposed model with respect to the pose problem. None, but one, of the previous related work has reported their ability to work in this kind of setting, hence only that work will be compared. The images of all 337 subjects at pose variation (+45° to -45°) in the MultiPIE database were taken as the training set while all images of the 50 subjects at the same pose range from the Curtin database were used as the test set. Note that the Curtin database only consists of face images at pose angles +/-60°,

+/-30° and 0° (Fig. 4). The frontal face images from the 50 subjects in the Curtin database were used as gallery images and images of the other poses as probe images (each pose consists of 3 images). SRC, LDA, PCA results were excluded as mentioned above. Experimental results are shown in Table V:

TABLE V. RESULTS OF THE CROSS DATABASE EXPERIMENTS ON MULTIPIE AND CURTIN DATABASE.

Method	-60°	-30°	+30°	+60°
DDA-SRC	**33**	**70**	**71**	**31**
DDA-LDA	**28**	**77**	**77.8**	**27.6**
DDA-PCA	**24**	**69.2**	**71**	**23**
SFDAE-SRC [14]	20	60	62	19
SFDAE-LDA [14]	19	68	69.7	16
SFDAE-PCA [14]	17	53.8	54.4	15

Discussion. As shown in Table IV, the proposed model outperforms many 2D methods such as DAE [4], GMA [16], CCA [7], PLS [15] and 3D method (Asthana11 [2]) as displayed in the average column. It outperforms the best state of the art method SPAE [10] in small pose variations (+/-15°). Note that in this case, DDA Net learns the same transformation from +/-15° to 0° like SPAE, but with a non-linear discriminant criteria which leads to better performance. Although it does not perform as well as SPAE in large poses (beyond 30°), unlike SPAE, the proposed model can be used with various kinds of machine learning problems to learn effective features while preserving the class structure of data to facilitate recognition. More importantly, the proposed model can be trained with no prior information about the pose or expression while such information is necessary for training the SPAE model. Moreover, the non-linear discriminant error criterion combined with the deep structure can successfully be applied in complex problem domains where the regular LDA criteria fail. As shown by Table V, the proposed model demonstrates a decent level of tolerance for recognizing faces captured under different environmental conditions within the pose range (-30° to +30°), while other existing methods fail in such a challenging setting. Although the results for larger pose variations are unsatisfactory, the results clearly show the ability of the proposed model to learn effective features across databases irrespective of the environmental conditions, thus unveils a new direction for further improvements.

IV. CONCLUSION

We proposed a novel deep AE model that can progressively perform the non-linear dimension reduction while de-noising the input face image in the observed low dimensional space by a non-linear discriminant error criteria. Experiments show that the proposed model is able to learn effective features in various face related problem domains and show a good level of generalizability.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *Computer vision-eccv 2004*, pages 469–481. Springer, 2004.

[2] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3d pose normalization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 937–944. IEEE, 2011.

[3] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[4] Y. Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[5] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. The cmu multi-pose, illumination, and expression (multi-pie) face database. Technical report, Technical report, Carnegie Mellon University Robotics Institute. TR-07-08, 2007.

[6] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[7] H. Hotelling. Relations between two sets of variates. *Biometrika*, pages 321–377, 1936.

[8] J. Hu, J. Lu, and Y.-P. Tan. Discriminative deep metric learning for face verification in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1875–1882. IEEE, 2014.

[9] G. B. Huang, H. Lee, and E. Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2518–2525. IEEE, 2012.

[10] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1883–1890. IEEE, 2014.

[11] B. Y. Li, A. Mian, W. Liu, and A. Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 186–192. IEEE, 2013.

[12] B. Y. Li, A. Mian, W. Liu, and A. Krishna. Using kinect for face recognition under varying poses, expressions, illumination and disguise. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 186–192. IEEE, 2013.

[13] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan. Morphable displacement field based image matching for face recognition across pose. In *Computer Vision–ECCV 2012*, pages 102–115. Springer, 2012.

[14] C. S. N. Pathirage, L. Li, W. Liu, and Z. Min. Stacked face de-noising auto encoders for expression-robust face recognition. In *Digital Image Computing: Techniques and Applications (DICTA), 2015.* APRS, 2015.

[15] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 593–600. IEEE, 2011.

[16] A. Sharma, A. Kumar, H. Daume III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2160–2167. IEEE, 2012.

[17] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[18] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(2):210–227, 2009.

[19] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*, 2014.