Stacked Face De-noising Auto Encoders For Expression-Robust Face Recognition

Chathurdara Sri Nadith Pathirage*, Ling Li*, Wanquan Liu*, Min Zhang†
*Department of Computing, Curtin University, Perth, Western Australia
†Department of Mathematics, School of Science, Tianjin University, Tianjin 300072, PR China
Email: c.nadithpa@postgrad.curtin.edu.au

***Abstract*—Recent advancement in unsupervised and transfer learning methods of deep learning networks has seen a complete paradigm shift in machine learning. Inspired by the recent evolution of deep learning (DL) networks that demonstrates a proven pathway of addressing challenging dilemmas in various problem domains, we propose a novel DL framework for expression-robust feature acquisition. The framework exploits the contributions of different colour components in different local face regions by recovering the neutral expression from various expressions. Furthermore, the framework rigorously de-noises a face with dynamic expressions in a progressive way thus it is termed as stacked face de-noising auto-encoders (SFDAE). The high-level expression-robust representations that are learnt via this framework will not only yield better reconstruction of neutral expression faces but also boost the performance of the subsequent LDA [1] classifier. The experimental results reveal the superiority of the proposed method to the existing works in terms of its generalization ability and the high recognition accuracy.**

I. Introduction

Face recognition (FR) under varying expressions has been a big challenge in automatic FR systems due to the non-linear characteristics that are observed in various facial expressions. Despite the fact of its importance, there is only very little attention given to 2D image based approaches compared to 3D model based approaches [2] [3] [4] in the recent history. The facial deformations that appear in some expressions, such as closed eyes, teeth, etc., will introduce additional noise, thus degrade the performance of recognition systems. There have been some attempts in the related field in the past as discussed in [5]. The existing approaches can be divided into linear and non-linear methods.

Commonly used linear approaches include subspace model analysis (PCA, FLD) [6], and their extensions such as Enhanced Fisher Linear Discriminant (EFLD) [7] and Exponential Discriminant Analysis (EDA) [8]. In these methods, a linear subspace is learnt to extract facial features followed by a classifier. The techniques such as Global/Local Linear Regression [9], Linear Regression Classification (LRC) [10] etc. model the FR as a linear regression problem. A comprehensive review on all of the above methods were rigorously carried out by evaluating their strengths and weaknesses in performing robust face recognition under varying expressions in [11].

Recently, a decent attempt that outperforms all linear methods was proposed in [12] based on sparse representation coding (SRC) that exploits the tremendous potential of Compressive Sensing (CS) theory for problems in pattern recognition domain. In the SRC framework, it is assumed that the whole set of training samples from a dictionary (each image is a base atom) can approximate a test image of a given class by discriminatively finding sparse coefficients in which they form a linear combination of the atoms in the dictionary. While the SRC-based methods depict the power of constraining the sparsity in FR problems via L1 minimization, they have some disadvantages due to the fundamental design of the SRC framework. Firstly, for accurate recognition, a sufficiently large training image set for each subject is needed to construct a good over-complete dictionary. In practice it may not be possible to acquire a large set of images per identity. Secondly, if the model is to be applied on thousands of real world identities, the size of the dictionary would be really large and thereby pose practical challenges in speed and performance. Furthermore, the SRC framework also suffers from the linear nature due to its fundamental assumption of having linear combinations of dictionary atoms to approximate a given test image.

All of the above methods are well understood for their fundamental linear nature and in a sense have reached their maturity. Researchers are now looking for non-linear models to address problems that consist of non-linear characteristics. To overcome the bottlenecks of the models of linear nature, some approaches such as [13] [14] were suggested to incorporate non-linearity in the learning domain of the problem. These methods are based on recently-emerged Deep Learning (DL) framework and showcase a tremendous power of learning highly non-linear transformations at the cost of boosted complexity and added vulnerability to the effects of over-fitting due to their immense power and flexibility in fitting a task. In [13] [14] a novel DL model based on Deep Convolutional Networks (DCN) was introduced to convert a random face of an identity to its frontal representation and proven to yield promising results in the pertinent field. One of the drawbacks of employing convolutional networks is having a very high number of parameters (weights) to be trained thus requiring big amounts of data to pre-train and to fine-tune the model accurately. Also this model consists of eight layers imposing practical challenges in employing such kind of a system without a tremendous computing power. Such systems are not yet feasible/affordable to be employed in real world applications.

To overcome the additional complexity posed in DCN, a decent but feasible DL approach (based on auto-encoders) focused only on pose variation problem was proposed in [15]. They argue that the pose variations change non-linearly and

978-1-4673-6795-0/15/$31.00 ©2015 IEEE

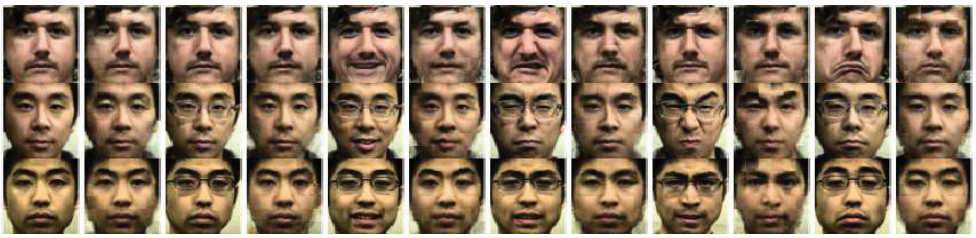

Fig. 1. Recovered neutral expressions faces against the noisy version of it. The faces with expressions (1, 3, 5, 7, 9 and 11) are neutralized into (2, 4, 6, 8, 10 and 12) neutral expressions correspondingly. Glass removal examples are also shown in this figure thus the glasses can be considered as noise as well.

smoothly along the manifold thus a stacked progressive auto-encoder model is designed where shallow progressive auto encoders are used to map a face image at a larger pose[1] to a virtual view at smaller poses[2] while keeping those images at smaller poses unchanged. This method works well for pose variation problems. However since an expression cannot be learnt progressively or subdivided into different stages of posing (non-sequenced face images), it is not possible to follow this approach directly for expression-robust FR. To the best of our knowledge, no comparative research work has been done on the auto encoder (AE) framework to determine suitable methods for expression invariant face recognition.

Deep Auto Encoder (DAE) models [16] [17] allow effective feature learning through hierarchical non-linear mappings via the model's hidden layers. The simple design and ease of training (back-propagation of error gradient) of a DAE model reduce the central complexity comparatively to DCN. Inspired by [15], we propose in this paper a thoroughly designed DAE model which also provides optimal solutions for problems of highly non-linear nature due to its impressive learning capacity. Our contributions are as follows:

- Problem formulation to yield a better feature space to prominently improve the recognition accuracies of FR under varying expressions.
- A novel Multiple-Encoder Single-Decoder Colour Fusion Model and other related techniques to breakdown the ultimate objective into smaller but tractable goals while generalizing the model to a decent level.

The rest of the paper is organized as follows. Section II details the Stacked Face De-noising Auto Encoder (SFDAE) model and the implementation steps; Section III evaluates the performances of applying the proposed model on the AR [18] and Curtin [4] databases against various expression variations followed by the conclusion in the last section.

II. Stacked Face De-Noising Auto Encoders

This section is divided into two sub-sections to discuss the evolution of the proposed SFDAE model from the typical deep auto encoder networks and the step-wise approach of utilizing the SFDAE model in face recognition under varying expressions.

[1]Poses that are larger than 30 degrees.
[2]Poses that are less than or equal to 15 degrees.

A. Evolution of the SFDAE model

In this section, we first look at the typical auto encoder DL network, followed by the motivation behind the proposed approach and the formulation of it.

1) Auto Encoder (AE): A traditional unsupervised auto encoder DL network [19] consists of two core segments: encoder and decoder [16] - with a single hidden layer.

Encoder: The deterministic mapping $f(\mathbf{x})$ that transforms an d-dimensional input vector $\mathbf{x} \in \mathbb{R}^d$ into a r-dimensional hidden representation $\mathbf{h} \in \mathbb{R}^r$ is called an encoder. Its typical form is an affine mapping followed by a non-linearity as follows:

$$\mathbf{h} = f(\mathbf{x}) = \Phi(W\mathbf{x} + \mathbf{b}) \tag{1}$$

where $W \in \mathbb{R}^{rxd}$ denotes the affine mapping, $\mathbf{b} \in \mathbb{R}^r$ is the bias and $\Phi(\mathbf{x}) = sigmoid(\mathbf{x}) = \frac{1}{1+e^{-x}}$ is the activation function of each element which is usually a squashing non-linear function.

Decoder: The mapping $g(\mathbf{h})$ that transforms the hidden representation $\mathbf{h}$ (observed in the above step) back into a reconstructed d-dimensional vector $\mathbf{z}$ in the input space is called a decoder. The typical form of a decoder also has an affine mapping optionally followed by a squashing non-linearity.

$$\mathbf{z} = g(\mathbf{h}) = \Phi(\widehat{W}\mathbf{h} + \widehat{\mathbf{b}}) \tag{2}$$

where $\widehat{W} \in \mathbb{R}^{dxr}$ is the affine mapping, $\widehat{\mathbf{b}} \in \mathbb{R}^d$ is the bias and $\Phi(.)$ is the activation function which is described above. $\mathbf{z}$ can be interpreted as the exact reconstruction of $\mathbf{x}$ as well as in probabilistic terms as the mean of a distribution $p(\mathbf{X}{=}\mathbf{x}|\mathbf{Z}{=}\mathbf{z})$ that may generate $\mathbf{x}$ with high probability [16] where $loss(x, z) \propto -log(p(\mathbf{x}|\mathbf{z}))$. For a real value x, the squared error can be derived as the loss incurred if a Gaussian is chosen as $p(\mathbf{x}|\mathbf{z})$. Hence training an auto encoder to minimize the reconstruction error amounts to maximizing the lower bound on the mutual information between input $\mathbf{x}$ and the representation $\mathbf{h}$.

2) Motivation of the proposed SFDAE: The typical auto-encoder learns the representation $\mathbf{h}$ that retrains information about $\mathbf{x}$ which is actually not useful by itself for efficient feature learning [16]. One can easily maximize this mutual information by choosing the hidden layer dimension r to be the same or larger than the input dimension d. Hence this could yield a perfect reconstruction simply by learning an identity mapping but at the same time it could fail in learning a representation more useful than the input itself. To avoid this problem, the traditional auto encoders promote an under-complete representation where $r < d$ thus $\mathbf{h}$ can be seen

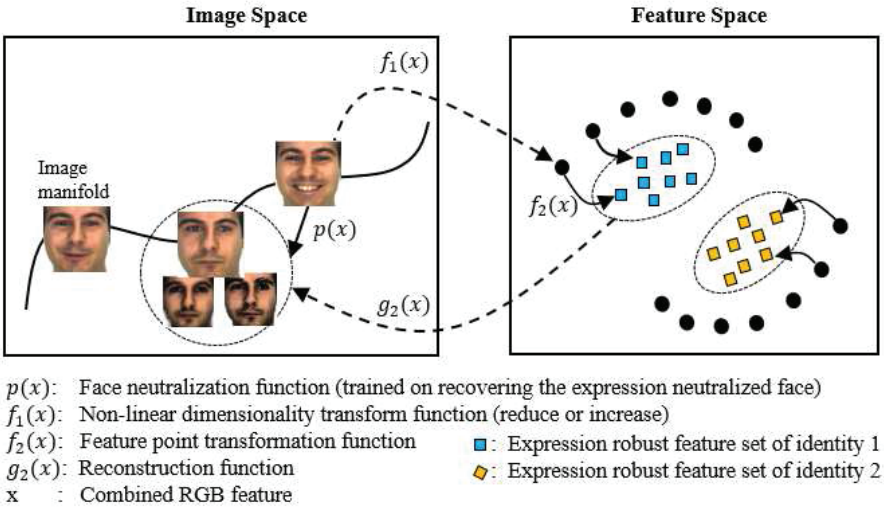

Fig. 2. Transformation functions that are learnt by the proposed SFDAE model via its hidden layers and the formation of the low dimensional feature space.

as a lossy compression of **x**. This can be viewed as a type of constrained learning. However, an AE could fail to learn effective features in reconstructing the original input from a noisy input due to the typical image reconstruction squared error function that is used to train the model. In addition, when a stacked auto encoder (deep auto encoder) model is used on an objective of highly non-linear nature as depicted in [15], it is prone to be trapped in local minima that could be far from the true global minima if there is only a relatively small number of training samples.

In our approach we try to exploit the fact that expressions generally change smoothly along the manifold. Hence recovering the neutral expression from a simple expression is more accurate than recovering it from an extreme one. We consider a face with an expression as a neutral face exposed to noise. Our model is trained to de-noise this face noise while learning a better low-dimensional feature space to perform the recognition (Fig. 1). The proposed approach performs non-linear dimension reduction and de-noising of images under strong regularization to yield better features of the learnt feature space. This is achieved by our novel single-decoder-multiple-encoder AE model that can de-noise the noisy representation learnt from a regular AE. We ensure that de-noising is supported by a strong supervisory signal which is indeed the neutral face. Hence the under-complete representation that is learnt in the de-noising layer will maximize the mutual information between the neutral face and an expression face (Fig. 2). In this way, a typical auto-encoder learnt representation h, which is affected by noise, can be de-noised thus resulting in useful features at the right-most hidden layer for a better representation that is invariant under varying noise. "A good representation is one that can be obtained robustly from a corrupted input and that will be useful for recovering the corresponding clean input" [16]. In a nutshell:

- The proposed SFDAE model which is supported by the deep architecture subdivides the global non-linear transformation into series of sub objectives where each hidden layer of SFDAE is trained on each sub objective to support the face neutralization process - Section II (A.3)
- SFDAE learns better weights for each colour component at pixel level in observing low dimensional expression-robust feature space -Section II (B.2)
- SFDAE adds invariability to expression-induced spatial deformities undergone by a face at the patch level due to patch-based training thus promoting further division of the global learning objective into limited but tractable learning goals -Section II (B.1)

3) Formulation of SFDAE: We formulate the SFDAE by utilizing a deep architecture where the 1st layer performs the colour fusion based on non-linear dimensionality reduction; the 2^{nd} layer sanitizes (de-noising) the low dimensional noisy feature observed in the first hidden layer; and the 3rd layer guides the entire learning process with a strong supervisory signal (acting as a strong regularizer). The low dimensional feature space observed at the 2^{nd} layer will be more robust to noisy features such as facial expressions. Formally, the first layer of SFDAE employs the following cost function to find the optimal parameters for patch j:

$$\left[W^*_{l=1}, b^*_{l=1}, \widehat{W}^*_{l=1}, \widehat{b}^*_{l=1}\right]_j = argmin_{W,b,\widehat{W},\widehat{b}} \sum_{i=1}^{S}\sum_{r=1}^{N_i} \left\| c^j_{ir} - g_1(f_1(c^j_{ir})) \right\|_2^2 \quad (3)$$

where S is the total number of identities, N_i is the number of images that belongs to the i^{th} class, $c^j_{ir} \in \mathbb{R}^{108}$ is the combined feature (Section II, B.2) for the j^{th} patch of the r^{th} image that belongs to the i^{th} class and $g_1(.), f_1(.)$ are the decoder and the encoder functions respectively (Fig. 3). The 2^{nd} layer of SFDAE de-noises the low dimensional features obtained from the 1^{st} layer via the following cost function:

$$\left[W^*_{l=2}, b^*_{l=2}, \widehat{W}^*_{l=2}, \widehat{b}^*_{l=2}\right]_j = argmin_{W,b,\widehat{W},\widehat{b}} \sum_{i=1}^{S}\sum_{r=1}^{N_i} \left\| \left\{ c^j_{ir} \right\}_F - g_2(f_2(h^j_{ir})) \right\|_2^2 \quad (4)$$

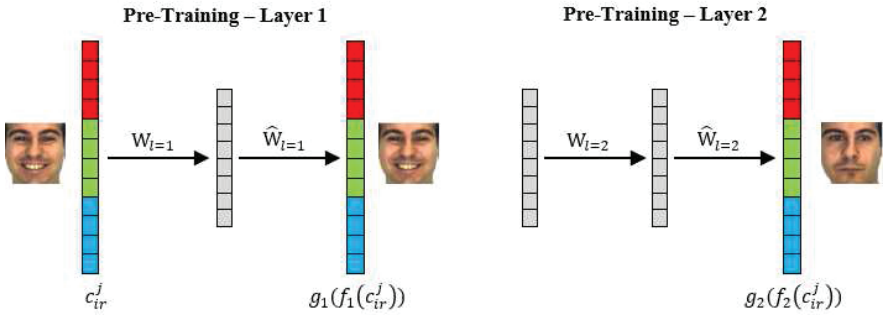

Fig. 3. Progressive pre-training of respected layers to achieve better initial weights prior to the training phase. Left figure denotes the non-linear dimension reduction layer pre-training while de-noising layer pre-training is shown in the right figure.

where $h^j_{ir} \in \mathbb{R}^{50}$ 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 $\{c^j_{ir} \in \mathbb{R}^{108}\}_F$ is the corresponding combined feature of the frontal face of the corresponding $c^j_{ir} \in \mathbb{R}^{108}$ described above (Fig. 3). All other parameters are the same as above. Furthermore, each layer is optimized to obtain better parameters via the full batch gradient descent algorithm on the cost functions when necessary. Once the optimal parameters are observed, the whole network is tuned again to optimize all layers (L) jointly as below:

$$\left[W^*_l|^L_{l=1}, b^*_l|^L_{l=1}, \widehat{W}^*_L, \widehat{b}^*_L\right] = argmin_{W_l|^L_{l=1}, b_l|^L_{l=1}, \widehat{W}_L, \widehat{b}_L} \sum_{i=1}^{S}\sum_{r=1}^{N_i} \left\| \left\{c^j_{ir}\right\}_F - p(c^r_{ij}))\right\|^2_2 \quad (5)$$

where $p(x_i) = g_L(f_L(f_{L-1}(x_i)))$ with $L = 2$ and $W_l|^L_{l=1}$ denotes the encoder's weights while W_L denotes the decoder's weights as shown in Fig. 4. All other parameters are described above. By jointly optimizing the objective function (Eq. 5) towards neutralizing facial expressions, the model is expected to learn a better feature space which is invariant to face noise.

B. Utilizing SFDAE model in the FR context

In this section, we describe the steps involved in utilizing the proposed model in the face recognition (FR) context in detail. The steps are mentioned in order.

1) Patch-based division of the face images: Patch-based division divides the global non-linear objective into small but tractable goals for SFDAE by limiting the number of parameters (complexity) that the model should learn. Processing a low resolution patch itself requires less computing resources (in terms of CPU and memory) in training a SFDAE due to its reduced complexity. This step enables us to train a SFDAE model for each individual patch in a parallel cluster environment. Multiple resolution images can also be handled efficiently. Another important fact is the consideration of independency in training all patches from one another. Due to the cost function that is occupied (Eq.5) at the output layer, it is conceivable that the image patches should be treated as a whole to calculate the cost, thus it preserves the correlation between each patch at the output layer and promotes learning the relationships between the patches. The choice of good image patch size is a vital factor in this model because of the misalignment of faces in the training databases used. Also it yields the opportunity of capturing more dependencies [16] between dimensions by producing a high dimensional input. After a few properly designed test runs, it was decided that the best image patch size is $6x6$. We denote a patch by the notation p^j_{ir}, where $p^j_{ir} \in \mathbb{R}^{36}$ is the j^{th} image patch of the r^{th} image that belongs to class i.

2) Combined Feature Formation and Pre-Processing: Since every colour component of a colour image has its own strengths and weaknesses in the recognition process, we stack all 3 channels (RGB) of an image into one big vector so that the SFDAE model can learn the optimal combination of each colour component in producing expression-robust features. Each pixel's colour component will be assigned a weight after the training phase and those weights are decided on the basis of how much meaningful information each pixel/dimension carries for its neutral expression face image reconstruction. Moreover, to eliminate the contrast effect of images, histogram equalization is performed on the V-channel after converting the RGB image to the HSV colour space. Once the contrast normalization is done, the image is converted back to its original colour space (RGB) and all image pixel values are normalized from 0-255 to the range of 0-1. This is due to the operating range of the non-linear squashing function that is used at each hidden unit of every layer. Formally, a combined feature can be denoted as:

$$c^j_{ir} = [p^{jR}_{ir}\ p^{jG}_{ir}\ p^{jB}_{ir}]^T \quad (6)$$

where $p^{jR}_{ir}, p^{jG}_{ir}, p^{jB}_{ir} \in \mathbb{R}^{36}$ are the corresponding colour channels RGB for the j^{th} patch of the r^{th} image that belongs to the i^{th} class. Hence $c^j_{ir} \in \mathbb{R}^{108}$ is the combined patch feature of the channels R, G, B of the r^{th} image in the i^{th} class. This concatenated high dimensional feature is used as the input of the SFDAE model in the pre-training and training stages.

3) Pre-training, Training and Validation: It is vital that the SFDAE model is initialized to have better initial weights prior to the training phase. Better initial weights ensure that a DL model has a larger probability to avoid some bad local minima [20]. It is also claimed that a DL model yields a better classification performance with a well-engineered pre-training process compared with having randomly initialized weights [16] [20]. A typical pre-training of the SFDAE model can be done via any face database to learn better starting weights which can help to tackle the global non-linearity progressively in each hidden layer. In the pre-training

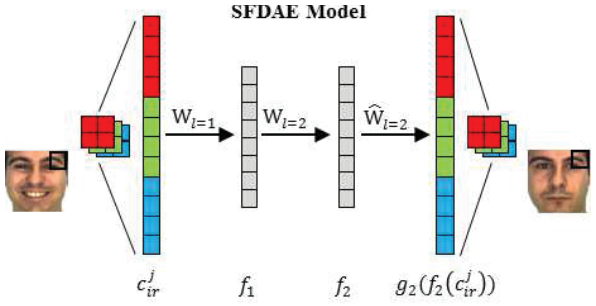

Fig. 4. The proposed SFDAE model where $f_1 \in \mathbb{R}^{50}$ denotes low dimensional noisy feature learnt at layer 1, while $f_2 \in \mathbb{R}^{50}$ denotes the noiseless feature learnt at layer 2 in the observed low dimensional space. We halves the image space by 50% to constraint the model to learn an effective low dimensional feature.

phase, the cost functions denoted by Eq. 3 and Eq. 4 will be optimized respectively. Once the optimal initial weights $[W^*_{l=1}, b^*_{l=1}, W^*_{l=2}, b^*_{l=2}, \widehat{W}^*_{l=2}, \widehat{b}^*_{l=2}]$ are observed in the pre-training stage, the training will be performed on the database of interest by stacking the pre-trained layers one after another (Fig.3) to form a deep architecture. The whole network is then trained to optimize all layers jointly (Eq. 5) to yield better expression-robust features in the low dimensional feature space learnt at layer 2. Training over a strong supervised signal (neutral face) will pose a difficult problem to be solved thus the SFDAE model needs to learn a weighted combination of each colour channel for each pixel rather than classifying each image to one of the pre-defined categories. This challenging task will utilize the model's super learning capacity while acting as a strong regularizer in the learning process. In order to regularize the SFDAE model further, we use a validation dataset to further constraint the learning process using the validation error. One of the sample datasets can be used as a validation set which can improve the model's generalizability while befitting from the model's super capacity of learning the task.

4) Expression Invariant Feature Acquisition: Since the right-most hidden layer f_2 performs the de-noising of the input in the observed low dimensional feature space, the representation at the f_2 layer should almost have no expression variations (Fig.4). This layer will ensure that the learnt low dimensional features are highly favourable in reconstructing the neutral expression face rather than reconstructing the input image itself. As opposed to the typical stacked auto-encoder framework which is used in non-linear dimension reduction, our framework will ensure that the essential feature components are learnt in reconstructing the neutral expression face. Hence a trained model can observe expression-robust features for the faces that may carry various expressions such as closed eyes and etc. As stated in [21], most recognition methods are robust to small expression variations. The output of the right-most hidden layer f_2 in our system can hence be used as expression-robust features for most face recognition methods.

The learnt feature f_2 from the proposed SFDAE model is unsupervised and therefore cannot be expected to be discriminative. As a remedy, once the expression-robust features are obtained for both the gallery images and the test image via the model, we apply LDA for supervised dimensionality reduction and the nearest neighbour classifier for recognition. Furthermore we evaluate the LDA performance against PCA performance in order to better understand the structure of the obtained low dimensional space.

Algorithm 1 Training Algorithm:

$X^j = [c^j_{11}...c^j_{1N_1}c^j_{21}...c^j_{2N_2}...c^j_{S1}...c^j_{SN_S}]$;
$\{X^j\}_F = corresponding\ frontal face\ patches\ of X^j$;
$\{V^j\}_F = repective\ frontal\ face\ patches\ of\ validation\ patch\ set$;
$V^j = validation\ patch\ set$;
$W^j_0 = weights\ obtained\ after\ pretraining\ stage$;
$ve^j_0 = +\infty,\ m = 0;\ \ //temporary\ variables$;
$count_{epoch} = 0$;
$error_{validation} = +\infty$;
while $(error_{validation} > 0) AND (count_{epoch} < t)$ **do**
  $F^j = SPDAE(X^j, W^j_c)$
  $m = (\frac{1}{S}) * \sum_{i=1}^{S} N_i$
  $e^j = (\frac{1}{2m}) * \left|\left|\{X^j\}_F - F^j\right|\right|_2^2$
  $\nabla F^J = \frac{\partial e^j}{\partial F^j}$
  $\nabla W^j_c = \nabla F^j * \frac{\partial SPDAE(X^j, W^j_c)}{\partial W^j_c}$
  $\nabla W^j_{c+1} = W^j_c - \lambda(\nabla W^j_c)$
  $VF^j = SPDAE(V^j, W^j_{c+1})$
  $ve^j_{c+1} = (\frac{1}{2m}) * \left|\left|\{V^j\}_F - VF^j\right|\right|_2^2$
  $error_{validation} = ve^j_c - ve^j_{c+1}$
  $count_{epoch} = count_{epoch} + 1$
end while
{where c is the iteration index, λ denotes the learning rate and t is a large positive integer to denote the maximum epoch count. In training, the breaking condition that will mostly occur is the validation error criteria.}

III. Experiments

In this section, the proposed model is assessed against the state-of-the-art methods on two publicly available colour image databases (AR [18] and Curtin [4]) where each of the setups are based on frontal faces with different facial expressions in uniform illumination. All images were cropped, aligned and resized into the resolution 66x66. The main objective is to find the identity of a person irrespective of the facial expressions that a face can pose. The evaluation of the model was done through three different experiments, each of them consisting of six (6) test cases to evaluate the model's invariability on six (6) different expressions as shown in Fig. 5. These expressions include: open mouth (smiling), closed eyes and etc. Neutral face is included in these expressions to evaluate the model's performance over typical FR scenarios as well. Furthermore, the performance graphs compare the results for:

- SRC [12], LDA, PCA
- Deep learnt feature (at the expression-robust hidden layer) on SRC(SFDAE-SRC), LDA(SFDAE-LDA),

PCA(SFDAE-PCA)

- Reconstructed neutral face (at the output layer) on SRC(Reconst), LDA(Reconst), PCA(Reconst), Nearest Neighbor with V channel of the HSV color space (HSV-V)

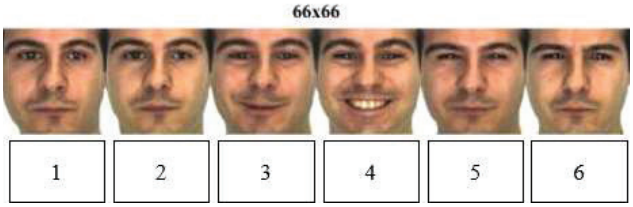

Fig. 5. Images with different expressions and their corresponding indices.

A. *Isolated Database Experiment*

In the first experiment, training and testing were performed on the same database. AR database was used primarily in this experiment.

Same Identity Testing. In the first setup, the training set included 5 images from each subject whereas the testing set included only 1 image that was left out from the same subject. This is to ensure that the model is trained with the maximum number of expression variations in its training phase. Such settings generally demonstrate better performance than using one image as the gallery [15]. The image selections were done in a round robin fashion for each test case and details of the six (6) test cases are given below:

1^{st} Test Case: Images $2 \sim 6$ of each identity were taken for training and image 1 of each identity was used for testing.

For each Test Case i: Test on i^{th} image of each identity and train on remaining 5 images of each identity.

Fig.6 demonstrates the distribution of training dataset, validation dataset and testing dataset of 100 identities from the AR database for Test Case 3. The recognition accuracies for the first experiment are shown in Fig.7.

Fig. 6. Data splitting in Test Case 3. Images 1, 2, 4-6 from 75 identities were taken for training. Images of the same indices (1, 2, 4-6) from the remaining 25 identities were used for validation. Images 3 of the respective 75 identities were used for testing.

Discussion: As denoted by the red solid line in Fig.7, SFDAE model performs consistently high and stable compared to other methods in different types of expression noise. The differences in SFDAE-LDA vs. LDA and SFDAE-PCA vs. PCA clearly

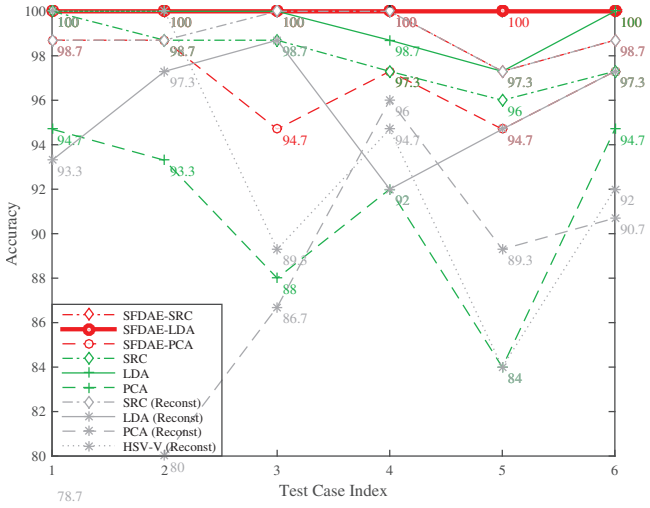

Fig. 7. Results of the tests performed with 75 identities in AR database.

show the improvement of SFDAE over the two popularly used methods. The proposed model which can be seen as a non-linear dimensionality reduction technique clearly outperforms the other linear dimensionality reduction techniques. In addition, it performs comparatively better than the popular SRC classification algorithm.

Cross Identity Testing. In this sub arrangement, training was performed with images that belong to a set of identities whereas testing was performed with images that belong to another set of identities in the same database, with no overlap between the identities used for training and testing. This setting intends to evaluate the model's generalization ability on mutually exclusive datasets that were built under the same environmental conditions such as lighting, reflection, camera alignment etc. Once the model is trained, we compare and contrast the recognition accuracies under each expression as shown below. Fig.8 shows the data split from AR database for Test Case 3.

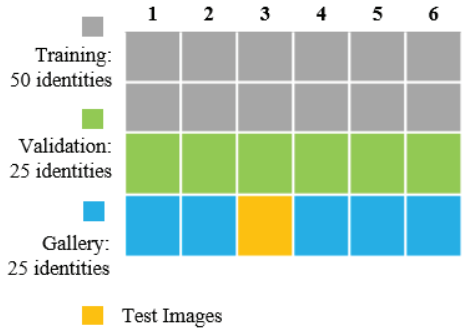

Fig. 8. The data splitting of a cross identity test case. All images (1-6) from 50 identities were taken for training whereas the other 50 identities were split (25 identities each) for validation and testing. Each test case concerns one expression for recognition accuracies. Shown in the figure is when Images 3 were used for testing.

Discussion: As shown in Fig.9, the results show that the proposed model is highly generalized over cross identity datasets that were built under the same environmental conditions. Since all expression variations of one dataset were used in training,

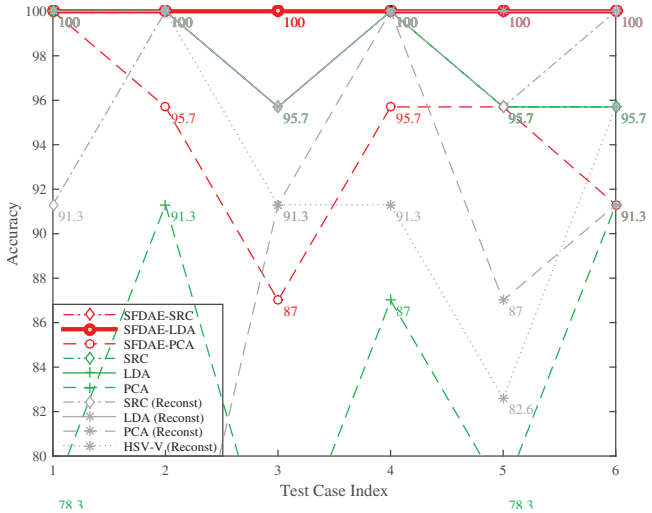

Fig. 9. Results of the experiments performed on cross subject arrangements.

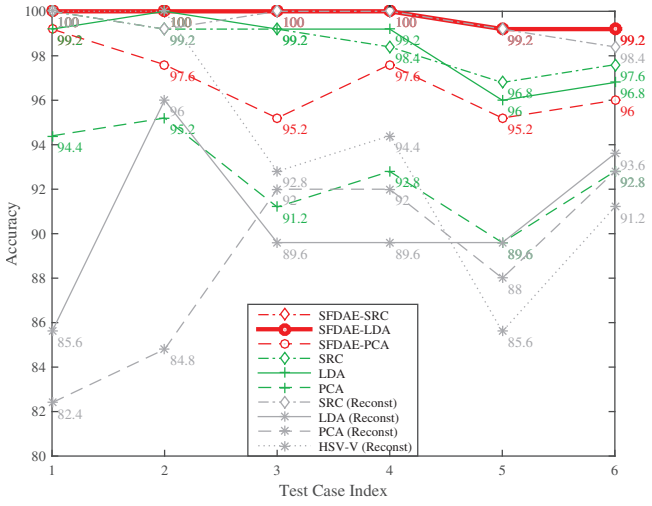

Fig. 11. Results of the experiments performed on combined database.

the model is highly capable of observing an expression-robust feature space for classification. Since cross training cannot be performed on a classification algorithm (SRC), the SRC results were excluded.

B. Combined Database Experiment

In the second experiment, training and testing were performed on a combined database built by merging 100 identities from AR database and 50 identities from Curtin database. This setting intends to evaluate the model's flexibility of handling a large number of subjects despite the fact that the images of those subjects were obtained under different environmental conditions such as illumination, lighting etc. The six (6) Test Cases were formed in the same fashion as described in Section III(A). Once the model is trained, we compare and contrast the recognition accuracies under each expression as shown in Fig.11. The distribution of training dataset, validation dataset and testing dataset of 150 identities from the combined database for Test Case 3 is shown in Fig.10. All of the splits consist of a random mix from the two databases.

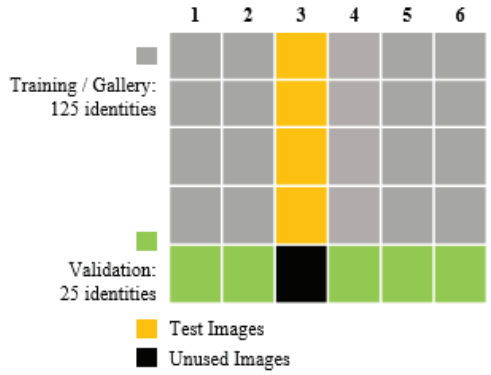

Fig. 10. Data splitting in Test Case 3. Images 1, 2, 4-6 from 125 identities were taken for training. Images of the same indices (1, 2, 4-6) from the remaining 25 identities were used for validation. Images 3 of the respective 125 identities were used for testing.

Discussion: As shown in Fig.11, SFDAE still performs consistently higher than the other methods including the SRC classification algorithm. Hence the proposed model possesses the ability to perform learning on subjects that were taken in different environmental conditions which demonstrates the generalization ability of the model.

C. Cross Database Experiment

In the 3rd experiment, training was performed on the 100 identities in the AR database and testing was performed entirely on another database (Curtin). Due to the difference in the environments where the two databases are captured, this setup is more challenging and is used to demonstrate the immense generalization ability of the proposed SFDAE model. The six (6) test cases were formed in the same manner as described in Section III(A). In the AR database, the training set consisted of 75 subjects and the remaining 25 subjects were taken as the validation set. The 50 subjects from Curtin database were used for testing. The recognition accuracies for this experiment are shown in Fig.12.

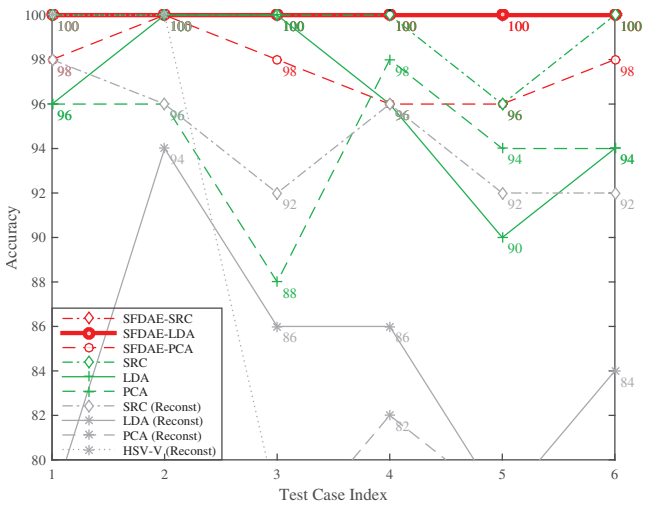

Fig. 12. Results of the experiments performed on cross database arrangement.

Discussion: As shown in Fig.12, the proposed model depicts

a very strong generalization ability towards the unseen subjects and expressions in training, even if the training and testing images are taken in different environmental settings. It consistently outperforms other methods, showing the true invariability towards the facial expressions.

IV. Conclusion

We proposed a stacked face de-noising auto encoder framework to map a noisy face (face with an expression) to a neutral face hence learning a complex non-linear face neutralizing function. The framework is able to yield a better reduced dimension feature space for classification while fusing different colour components at the input layer. The global complicated non-linearity involved in neutralizing the facial expression is divided into pieces of more tractable objectives which are modelled by multiples shallow auto encoders focused on specific simpler goals while following a patch-based approach. As shown in the experiments, SFDAE can effectively learn expression variations, and improve the performance of face recognition. In the future, we wish to extend the SFDAE framework to deal with illumination and occlusions while incorporating discriminative information into the design of the framework. Possibility of online learning with a video sequence of face deformations during an expression will also be exploited to learn temporal relationships between different frames for much better learning.

References

[1] 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*, vol. 19, no. 7, pp. 711–720, 1997.

[2] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama, "3d face recognition under expressions, occlusions, and pose variations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 9, pp. 2270–2283, 2013.

[3] X. Wang, Q. Ruan, Y. Jin, and G. An, "Three-dimensional face recognition under expression variation," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, pp. 1–11, 2014.

[4] 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*. IEEE, 2013, pp. 186–192.

[5] M. Murtaza, M. Sharif, M. Raza, and J. H. Shah, "Analysis of face recognition under varying facial expression: A survey." *International Arab Journal of Information Technology (IAJIT)*, vol. 10, no. 4, 2013.

[6] P. Tsai and T. Jan, "Expression-invariant face recognition system using subspace model analysis," in *Systems, Man and Cybernetics, 2005 IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1712–1717.

[7] G. Y. An and Q. Ruan, "Novel mathematical model for enhanced fisher's linear discriminant and its application to face recognition," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 2. IEEE, 2006, pp. 524–527.

[8] T. Zhang, B. Fang, Y. Y. Tang, Z. Shang, and B. Xu, "Generalized discriminant analysis: A matrix exponential approach," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 40, no. 1, pp. 186–197, 2010.

[9] X. Chai, S. Shan, X. Chen, and W. Gao, "Local linear regression (llr) for pose invariant face recognition," in *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*. IEEE, 2006, pp. 631–636.

[10] I. Naseem, R. Togneri, and M. Bennamoun, "Linear regression for face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 11, pp. 2106–2112, 2010.

[11] N. Kumar, R. Agrawal, and A. Jaiswal, "A comparative study of linear discriminant and linear regression based methods for expression invariant face recognition," in *Advances in Signal Processing and Intelligent Recognition Systems*. Springer, 2014, pp. 23–32.

[12] 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*, vol. 31, no. 2, pp. 210–227, 2009.

[13] 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.

[14] ——, "Deep learning identity-preserving face space," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 113–120.

[15] 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*. IEEE, 2014, pp. 1883–1890.

[16] 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," *The Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[17] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[18] A. Martínez and R. Benavente, "The ar face database," Computer Vision Center, Bellatera, Tech. Rep. 24, Jun 1998, cites in Scholar Google: http://scholar.google.com/scholar?hl=en&lr=&client=firefox-a&cites=1504264687621469812. [Online]. Available: http://www.cat.uab.cat/Public/Publications/1998/MaB1998

[19] Y. Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[20] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[21] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognition*, vol. 42, no. 11, pp. 2876–2896, 2009.