Neurocomputing 367 (2019) 188–197

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Semi-supervised facial expression recognition using reduced spatial features and Deep Belief Networks

Aswathy Rajendra Kurup*, Meenu Ajith, Manel Martínez Ramón

Department of Electrical and Computer Engineering, The University of New Mexico, New Mexico, USA

ARTICLE INFO

Article history:
Received 20 February 2019
Revised 31 July 2019
Accepted 8 August 2019
Available online 13 August 2019

Communicated by Dr. Zhiyong Wang

Keywords:
Emotion recognition
Semi-supervised learning
Dimensionality reduction
Contrastive divergence
Backpropagation
K-Fold cross-validation

ABSTRACT

A semi-supervised emotion recognition algorithm using reduced features as well as a novel feature selection approach is proposed. The proposed algorithm consists of a cascaded structure where first a feature extraction is applied to the facial images, followed by a feature reduction. A semi-supervised training with all the available labeled and unlabeled data is applied to a Deep Belief Network (DBN). Feature selection is performed to eliminate those features that do not provide information, using a reconstruction error-based ranking. Results show that HOG features of mouth provide the best performance. The performance evaluation has been done between the semi-supervised approach using DBN and other supervised strategies such as Support Vector Machine (SVM) and Convolutional Neural Network (CNN). The results show that the semi-supervised approach has improved efficiency using the information contained in both labeled and unlabeled data. Different databases were used to validate the experiments and the application of Linear Discriminant Analysis (LDA) on the HOG features of mouth gave the highest recognition rate.

1. Introduction

Amongst the various modes of emotion recognition (ER), the facial expression is one of the conveying forms used for the display of emotions. ER can be applied in various fields like medicine, marketing, entertainment. For example, a medical robot can be designed to continuously monitoring their emotional state [1,2], or a diagnostic suggestion system for therapists [3]. In Human-Computer Interaction, a system endowed with emotional intelligence can be used to create effective communication with users [4]. In emergency situations, as part of the corresponding situational awareness, real-time decisions can be made from the behavioral patterns of the subjects.

The development of a facial ER system is challenging since the images of the same person with the same facial expression can vary with the lighting conditions, background, and occlusions [5], which precludes homogeneity. Certain emotions have only subtle distinctions which make them harder to analyze and describe. The state-of-the-art approaches in facial ER used feature-based methods [6,7] and template-based methods [8]. The first ones focus appearance and geometric modelled feature extraction. Template-based methods were less reliable because they are limited to only frontal faces and the accuracy rate changed with variations in pose, scale, and shape. Feature extraction is mostly based on Histogram of Oriented Gradients (HOG) [9] and Local Binary Patterns (LBP) [10]. HOG descriptors were used to encode facial components since it projected the appearance of gradient orientation in an image. Other works use Discrete Wavelet Transform (DWT) for feature extraction and Neural Networks for classification [11]. Dimensionality reduction (DR) techniques in ER include principal components analysis (PCA) [12] and linear discriminant analysis (LDA) [13]. Recently, PCA based facial feature projection has also been used for age progression application [14]. These methods cannot be used to find the nonlinear structure of the data. To overcome this limitation various nonlinear DR algorithms such as kernel PCA [15], locally linear embedding (LLE) [16], isometric feature mapping (Isomap) [17] and T-distributed Stochastic Neighbor Embedding (t-SNE) [18] have been proposed. The sparse representation-based methods for classification (SRC) are also widely used since 2009 [19]. SRC is most effective when there is high separability between the subspaces [20–22]. But its main disadvantage over classical subspace learning algorithms is that the classification criterion of SRC fails and leads to misclassification when the samples are highly correlated. Deep Neural Networks [23] have gained popularity in the recent years as a choice for supervised learning. The major drawback of supervised learning comes from the fact that most of the data available in general is unlabeled., something particularly evident in the case of human face images. In 2004,

* Corresponding author.
E-mail address: aswathyrk@unm.edu (A. Rajendra Kurup).

https://doi.org/10.1016/j.neucom.2019.08.029
0925-2312/© 2019 Elsevier B.V. All rights reserved.

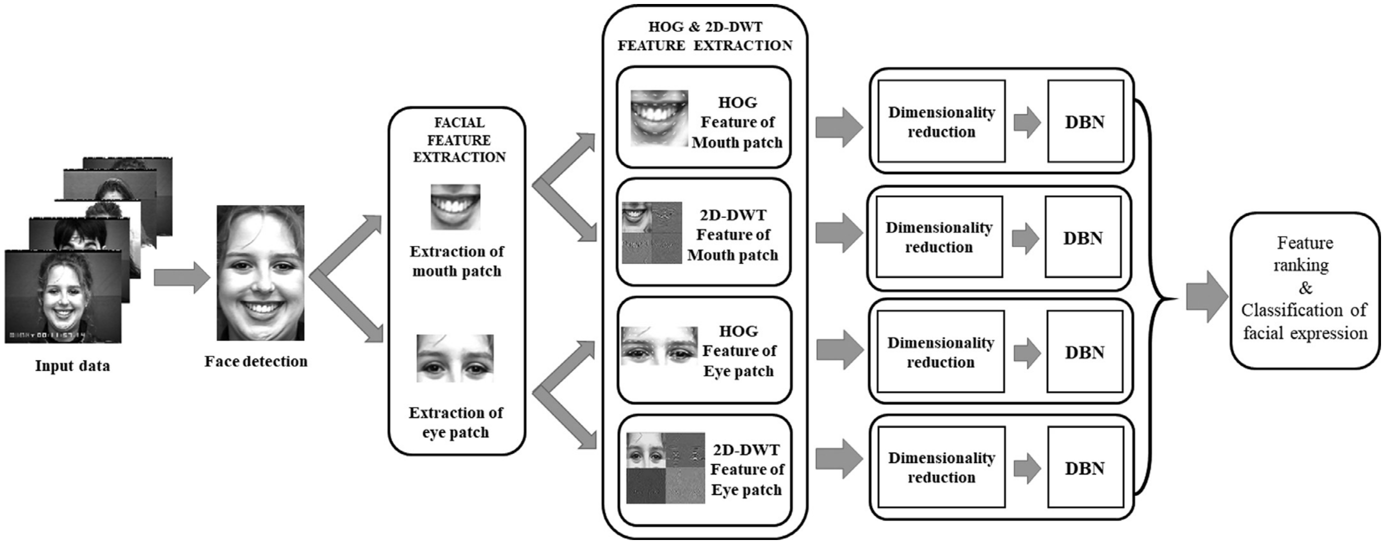
Input data
Face detection
FACIAL FEATURE EXTRACTION
Extraction of mouth patch
Extraction of eye patch
HOG & 2D-DWT FEATURE EXTRACTION
HOG Feature of Mouth patch
2D-DWT Feature of Mouth patch
HOG Feature of Eye patch
2D-DWT Feature of Eye patch
Dimensionality reduction
DBN
Dimensionality reduction
DBN
Dimensionality reduction
DBN
Dimensionality reduction
DBN
Feature ranking & Classification of facial expression

Fig. 1. Block diagram of the proposed method.

Hinton and co-workers Hinton et al. [24] proposed the idea of the Restricted Boltzmann Machines (RBM) and its generalization to Deep Belief Networks (DBN), where unsupervised techniques can model the probabilistic distribution of the data and cluster it [25].

In this paper, a semi-supervised DBN is used to include unlabeled and labeled data to improve the accuracy of the classifier. Semi-supervised learning is comparable to human learning, which involves a small amount of labeled data along with greater amounts of unlabelled observations [26]. To make use of unlabeled data, DBNs are applied to learn the model [27] and the obtained discriminative model is fitted to a labeled dataset by performing Backpropagation (BP).

The major contributions in this paper are two. First, we propose to use semi-supervised learning during feature selection to determine the features that are more explanatory of the human emotions in the available data. The proposed DBN has an input layer taking in the dimensionality-reduced feature vectors corresponding to Histogram Oriented Gradients (HOG) of mouth, HOG of the eye, Wavelet Transform of mouth and Wavelet transform of the eye. Reconstruction error and validation accuracy were used to find the most significant feature vector. Second, a semisupervised learning process is proposed that uses non labelled data to train a DBN. After convergence, we proceed to fine tune the structure with BP and the available labelled data. The data is previously processed by a dimensionality reduction method. The most efficient linear method was LDA and amongst the nonlinear approaches, the best one was Isomap.The proposed semi-supervised framework was evaluated on the CK+, MMI and RAFD databases. The results show that the presented approach have a performance similar or better than those of the state of the art (SVM and CNN) with the additional benefits of using significantly less labelled data and a dramatically reduced training and test computational burdens.

2. Proposed approach

The introduced semi-supervised deep belief network for facial ER is shown in Fig. 1. The proposed method incorporates different feature extraction methods and dimensionality reduction techniques prior to passing the data into the DBN. Based on the characteristics of the different facial expressions, the mouth and eye patches are extracted from the facial data. Then two feature extraction methods namely HOG and 2D-DWT were used to compute the significant spatial components from the mouth and eye data. Histogram of oriented gradients (HOG) is a feature descriptor which extracts the information regarding the image gradients to formulate the shape of structures in an image [28,29]. HOG was first used for pedestrian detection in 2005 and it was implemented by dividing the images into overlapped cells of $n \times n$ pixels. The cells are organized as overlapping blocks. Inside each cell, the pixel gradients are computed by using vertical and horizontal kernels $[-1, 0, 1]$ and $[-1, 0, 1]^\top$. Varying the cell size helps in capturing information at different scale. The number of orientation histogram bins helps in recording the details of orientation. Increasing the number of bins helps in capturing finer orientation details. The gradient vectors are arranged to form a histogram to compress the feature descriptor as well as to generalize the information contained in the cells. By normalizing the gradient vectors, HOG also becomes invariant to geometric and photometric transformations. HOG characterizes the local shape by capturing the edge or gradient structure and the features being a lot smaller compared to the local spatial or orientation bin size makes it invariant to translations and rotations. Whereas, in 2D-DWT [30–32] the image is decomposed into a set of basis functions called wavelets which provides both frequency and time information. The 2D wavelet decomposition of an image is implemented as a set of filter banks in which 1D-DWT is at first applied along the rows and then along the columns. The filter banks comprise of a cascaded design of high pass and low pass filters and the decomposition results in four sub-band images namely low - low (LL), low - high (LH), high - low (HL), and high - high (HH).

In this paper, four different feature vectors, which are HOG of mouth, HOG of eye, 2D-DWT of mouth and 2D-DWT of eye are evaluated to obtain the most suitable feature extraction method for this application. However, the dimensionality of these feature vectors are further decreased to accelerate the training of the DBN. These reduced features are given as input to the semi-supervised DBN and the recognition rate for different feature vectors are computed. Based on this accuracy, a ranking is assigned to the reduced HOG and DWT features of mouth and eye. Therefore this proposed approach, as shown in Fig. 1 can be used to predict the most relevant facial features using lesser computations.

2.1. Dimensionality reduction

Dimensionality reduction can be used for both feature extraction as well as feature selection. This work focuses on feature extraction using linear methods such as PCA and LDA and nonlinear techniques such as Kernel PCA, Isomap and t-SNE. PCA [33] is an extensively used technique which outputs a linear approximation of dimension d which is lesser than the input dimension n. The input data is projected such that the variance of each principal component is maximal [34]. Though PCA is not computationally demanding, it fails to model the nonlinear variabilities in high dimensional data. This problem was addressed by introducing Kernel PCA [35] which uses different kernels to project the input to a nonlinear feature space. The most popular kernels used are gaussian, polynomial and hyperbolic tangent [36]. However, for Kernel PCA as the number of data points increases, the kernel matrix grows quadratically and hence the eigenvalue decomposition of this matrix becomes computationally expensive. Another supervised but linear dimensionality reduction algorithm is LDA [37], which can be defined as an optimization problem to compute linear combinations with coefficients **w**, which tries to maximize the ratio of between the class variance to within the class variance. The objective function used is as follows:

$$J(W) = \frac{\mathbf{w}^\top S_\mathbf{B} \mathbf{w}}{\mathbf{w}^\top S_\mathbf{W} \mathbf{w}} \tag{1}$$

Where $S_\mathbf{B}$ is between the class variance and and $S_\mathbf{W}$ is within the class variance. Despite being a popular dimensionality reduction method, LDA has its limitations. For classification purposes, when the distribution of the data is non-gaussian, LDA won't be able to preserve the complex structure of the data [38].

Isomap uses the geodesic distance to create a lower dimensional embedding to preserve the manifold structure. Though it provides an estimate of the underlying geometry of the data [39], the main disadvantage of isomap is that it generally fails for manifolds which have holes. Thus in contrast to isomap, LLE [40] computes neighborhood-preserving embeddings of the high-dimensional data. LLE represents each data point as the weighted sum of the k nearest neighbors. This linear mapping to lower dimension helps to retain the weights learned in the higher dimension. Further, t-SNE is a nonlinear algorithm developed by Laurens van der Maaten and Geoffrey Hinton [41] for embedding high-dimensional data for visualization in a low-dimensional space. It models the affinities of data points to probabilities and hence identical objects are characterized by nearby points and non-identical are modeled by distant points.

2.2. Deep belief networks

DBNs are a representative deep learning model built using stacked RBMs. It is an unsupervised learning algorithm in contrast to perceptron [42] and BP NNs [43]. The DBN training process consists of a pretraining phase and fine-tuning phase [44,45]. Each RBM is pre-trained in an unsupervised manner, the output of one layer being the input of the next one. Fine tuning is done in a supervised manner using labeled data and the BP algorithm.

An RBM is restricted since there are only connection between hidden and visible units. Structure of an RBM is shown in Fig. 2. The energy function associated with a joint distribution of two layers of a binary RBM is given by

$$E(\mathbf{v},\mathbf{h}) = -\sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij} v_i h_j - \sum_{i=1}^{V} b_i v_i - \sum_{j=1}^{H} a_j h_j \tag{2}$$

where a_i and b_j are the bias terms, V and H are the number of visible and hidden units, w_{ij} is the weight of the connection between visible unit i and hidden unit j, v_i, h_j are the binary states of the units, and **v, h** are column vectors containing them. A joint probability of **h** and **v** is defined as:

$$p(\mathbf{v},\mathbf{h}) = \frac{1}{Z} e^{-E(\mathbf{v},\mathbf{h})} \tag{3}$$

where Z is the partition function

$$Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \tag{4}$$

its marginalization over **h** is:

$$p(\mathbf{v}) = \frac{1}{Z} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})} \tag{5}$$

The posterior probability of a hidden node given **v** is modelled as

$$p(h_j = 1|\mathbf{v}) = \sigma(b_j + \sum_i v_i w_{ij}) \tag{6}$$

where the sigmoid function $\sigma(x) = (1 + e^{-x})^{-1}$. Similarly, the posterior of a visible node given **h** is modelled as

$$p(v_i = 1|\mathbf{h}) = \sigma(a_i + \sum_j h_j w_{ij}) \tag{7}$$

An efficient training procedure called contrastive divergence (CD) is introduced in [46], where the incremental learning rule is

$$\Delta w_{ij} \propto \langle h_i v_j \rangle_{data} - \langle h_i v_j \rangle_{rec} \tag{8}$$

where $\langle h_i v_j \rangle_{data}$ denotes the measured correlation between h_i and v_j when the states of the hidden units are determined by Eq. (6) and the visible vectors are samples from the training set. The term $\langle h_i v_j \rangle_{rec}$ is the correlation where h_i and v_j are

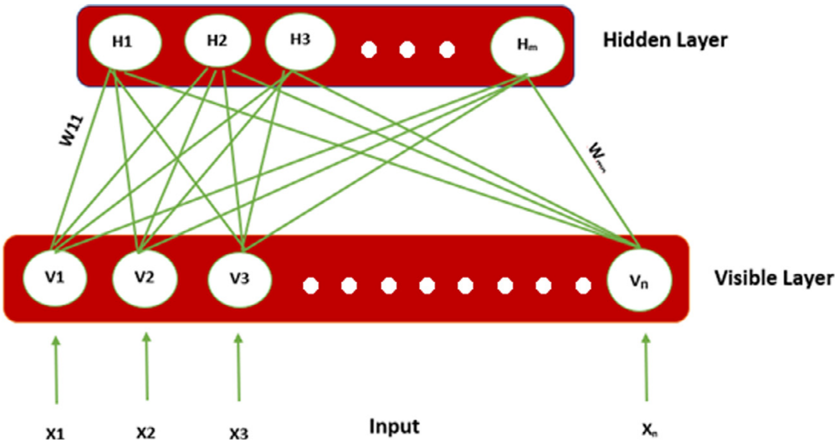

Fig. 2. Structure of RBM. An RBM is a bipartite graph in which observations are represented using visible units in the visible layer, which are connected to the hidden units in the hidden layer through undirected weighted connections.

both *reconstructions* of the data vectors and the states of the hidden units are chosen at random from distributions (6). The above pre-training stage aims at reconstructing the input and it is unsupervised.

A Deep Boltzmann Machine (DBN) can be constructed by stacking L layers with parameters $\mathbf{W}^{(l)}$, $\mathbf{a}^{(l)}$ and $\mathbf{b}^{(l)}$, $1 \leq l \leq L$. This machine can be trained layerwise, as each layer of nodes is conditionally independent of each other given the knowledge of the previous or following layer.

2.3. Backpropagation

Backpropagation [47] is a very well known technique to optimize neural networks (see, e. g. [25] or [48]). Given a training set $(\mathbf{x}_i, \mathbf{y}_i)$, $1 \leq i \leq m$ where targets $\mathbf{y}_k \in \mathbb{R}^D$. A common approach is to minimize a cost function that takes into account the averaged negative log-likelihood of the training targets given the input data and a prior over the parameters as

$$\begin{aligned} J(\mathcal{W}) &= -\log\left(\prod_{n,k} p(y_{n,k}|\mathbf{X}, \mathcal{W})p(\mathcal{W})\right) \\ &= -\sum_{n,k} \log p(y_{n,k}|\mathbf{X}, \mathcal{W}) - \log p(\mathcal{W}) \end{aligned} \tag{9}$$

where matrix $\mathbf{X}$ contains all data $\mathbf{x}_n$ and $\mathcal{W}$ contains all the linear parameters of the DBN. The posterior probability is proportional to the prior times the data likelihood. thus, this setup is equivalent to maximize the parameter log-posterior, given the training data. Usually, the prior for the parameters is a zero mean Gaussian function with covariance $\mathbf{\Sigma}$.

If the likelihood model for the data is a Gaussian distribution, then the function to optimize is simply

$$J(\mathcal{W}) = \frac{1}{2\sigma^2} \sum_{n,k} \|y_{n,k} - o_k(x_n)\|^2 + \boldsymbol{w}^\top \mathbf{\Sigma}_p^{-1} \boldsymbol{w} \tag{10}$$

and if the distribution is a multinoulli, then

$$J(\mathcal{W}) = \frac{1}{2\sigma^2} \sum_{n,k} y_{n,k} \log o_k(x_n) + \boldsymbol{w}^\top \mathbf{\Sigma}_p^{-1} \boldsymbol{w} \tag{11}$$

where $\boldsymbol{w}$ is a vector containing all parameters and $o_k(\mathbf{x}_n)$ is the response of the $k_t h$ output of the DBN to input $\mathbf{x}_n$. Usually, the simplification $\mathbf{\Sigma}_p = \mathbf{I}$ is taken, so the regularization term of the cost function is simply the norm of the vectorized parameters, or $\boldsymbol{w}^\top \mathbf{\Sigma}_p^{-1} \boldsymbol{w} = \sum_{l=1}^{n_l-1} \sum_{i=1}^{s_l} \sum_{j=1}^{s_l+1} (w_{ji}^{(l)})^2$, assuming that $w_{ji}^{(l)}$ is a weight connecting node i of layer $l-1$ and node j of layer l.

2.4. Architecture and training of the network

The network has five layers (Fig. 3). The input layer has two nodes. The classification is designed using a 4 bit code for the classes. Since the classification problem at hand consists of 7 classes, 4 bits are necessary for representing them. Then, the number of nodes at the output layer is 4. The rest of the layers have 3, 3, and 4 nodes, all of them with sigmoid activation. The output nodes are normalized using a softmax activation. The network is first trained in an unsupervised way using CD, and then a softmax activation of the output layer is used to apply a BP with the labelled data. The various steps involved in semi-supervised training are as follows.

The data is partitioned into a training set and test set. The training set consists of $N_l^{(tr)}$ non labelled data and $N_n^{(tr)}$ labelled data. The test set consists of N^{tst} labelled data. During the training of DBN, RBMs are trained layer after layer using the process explained above. For each epoch, all the RBMs were trained individually 5 times. In this process, and using Eqs. (6) and (7), the outputs $\mathbf{v}$ and $\mathbf{h}$ of each RBM are computed using the $N_l^{tr} + N_n^{tr}$ data. After that, the learning rule is used to find the weight update and Eq. (8) is used to compute the new weights and biases. For training the network, different learning rates were used. The

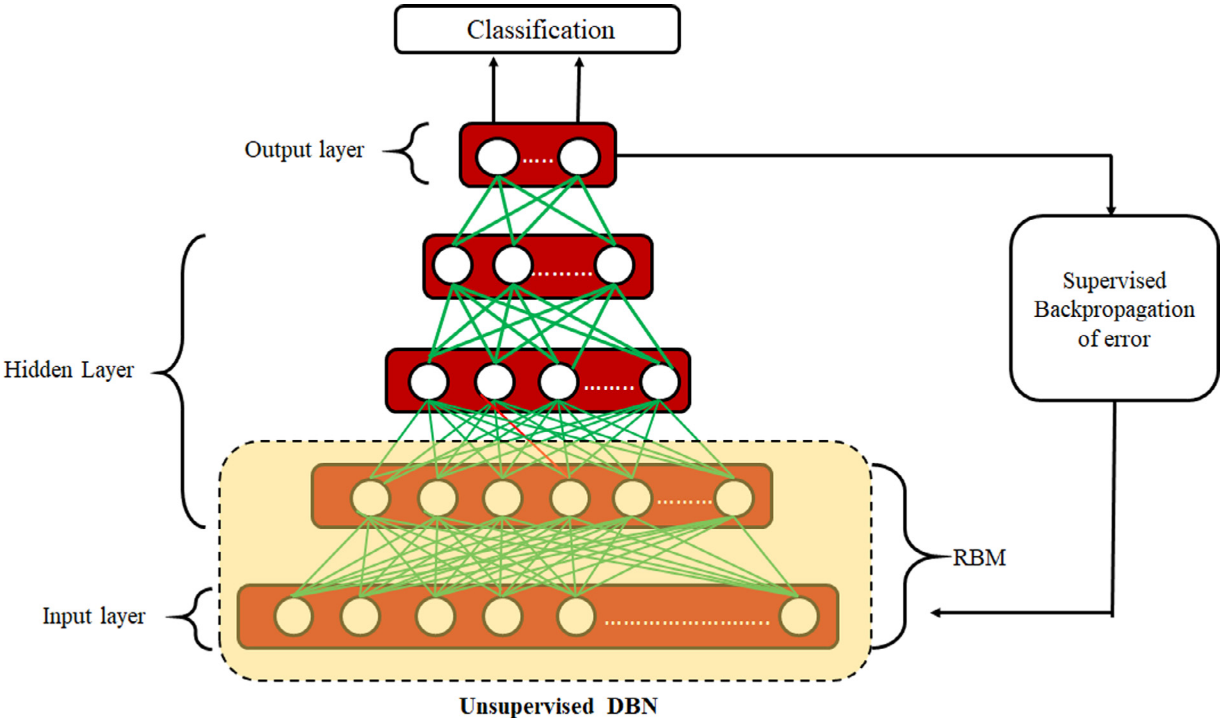

Fig. 3. Architecture of the semi-supervised DBN. The structure consists of five layers. The number of nodes, from input to output layers, is 2, 3, 3, 4 nodes. The number of staked RBMs is 4. The dashed area highlights the structure of the first stacked RBMs. The network is first trained with unlabelled data using CD until convergence, and then a softmax activation of the output layer is used to apply a BP with the labelled data.

hidden output of the previous RBM is passed onto the next RBM and the output of the final layer is further fine-tuned using BP with the N_l^{tr} labelled data. Hence the entire training procedure uses both labeled and non-labeled data to classify facial expression. A portion of the training data is also used for validation to obtain the optimized training parameters and model.

3. Experiments and results

3.1. Databases

The extended Cohn-Kanade database (CK+) [49], the Radboud Faces database (RaFD) [50] and MMI database [51] were used to test the proposed method for facial ER. The CK+ and MMI databases were captured from a lab-based environment whereas the RaFD database contained facial images with varying poses and gaze directions. Firstly, in the case of CK+ database, there are 327 image sequences with 7 expression labels namely anger, neutral, disgust, fear, happy, sad, and surprise. The last frame of each sequence consists of the peak expression with an image size of 640 × 490 or 640 × 480 pixels. A total of 1400 images i.e. 200 images from each of the seven emotion classes were used for all the experimentation. Secondly, the MMI database consist of 205 image sequence and the peak expression can be found in the middle of this sequence. This database consists of 6 facial expression classes and with an image resolution of 720 × 576 pixels. Finally, the RaFD database used for this experimentation consists of images from 67 subjects displaying 8 different expressions. The image data belonging to the category of contempt has been eliminated for this experiment. Further there exist samples belonging to 5 different camera angles and amongst them, only three angles 135°, 90° and 45° are taken into consideration. The frontal face images in this database have a size of 125 × 160 pixels whereas the images including the distinct camera poses have a size of 284 × 284 pixels.

3.2. Overview of the experiments

The Viola-Jones face detector [52] was used to detect the face from the given image and after which it detects possible mouth and eye regions. From these image regions, the HOG and wavelet features were extracted to form different feature vectors. Further different dimensionality reduction methods such as PCA, LDA, Kernel PCA, Isomap, t-SNE, and LLE were compared and evaluated to obtain the reduced features for easier computation during training [53]. The selection of the most significant features was done by cross-validating the model using the different feature vectors. It was found that the HOG features of the mouth provided more accuracy compared to the rest of the features. Finally, these HOG features were passed into a 5 layered DBN for pre-training [54]. A fine-tuning procedure using BP was used at the end of the DBN in order to reduce the reconstruction error of the entire network. Different experimentation was done to test the semi-supervised, supervised and unsupervised models. The proposed semi-supervised architecture was obtained by cascading the unsupervised DBN along with the supervised BP. The main test parameters used were the learning rate, the number of epochs, reconstruction error, and accuracy. K-fold cross-validation method was used here for testing the semi-supervised DBN models. The training data is first shuffled and split into 5 groups. In each case, a unique fold is chosen as the validation data set and the remaining 4 folds are iteratively used to train the model. The evaluation accuracy is retained, and the model is discarded. Finally, the recognition rate over the 5 folds is averaged to estimate the accuracy and the results were compared with a multiclass SVM classifier [55]. An NVIDIA GeForce GTX 1060 GDDR5 6.0 GB GPU and supercomputers from The Center for Advanced Research Computing (CARC) of the University of New Mexico were used to carry out all the experiments.

Table 1
Recognition rate of the CK+ database using different dimensionality reduction algorithms and the proposed semisupervised classification algorithm.

Method	Feature	Recognition rate %	
		2 dimension	5 dimension
PCA+DBN	HOG eye	55.71	58.1
	HOG mouth	63.81	65.71
	2D-DWT eye	41.90	43.81
	2D-DWT mouth	55.24	60.0
LDA+DBN	HOG eye	62.86	61.31
	HOG mouth	**98.57**	94.68
	2D-DWT eye	80.0	76.57
	2D-DWT mouth	90.0	86.78
Kernel PCA+DBN	HOG eye	27.12	27.62
	HOG mouth	36.19	37.14
	2D-DWT eye	28.10	29.52
	2D-DWT mouth	25.24	31.43
LLE+DBN	HOG eye	40.95	55.71
	HOG mouth	70.00	72.86
	2D-DWT eye	42.38	46.67
	2D-DWT mouth	50.95	69.05
Isomap+DBN	HOG eye	55.71	66.19
	HOG mouth	69.52	74.29
	2D-DWT eye	40.95	56.19
	2D-DWT mouth	63.81	75.71
t-SNE+DBN	HOG eye	27.12	27.62
	HOG mouth	36.19	37.14
	2D-DWT eye	28.10	29.52
	2D-DWT mouth	25.24	31.43

3.3. Comparison of recognition rate

Table 1 shows the comparative analysis of various linear and nonlinear dimensionality reduction algorithms. The feature vectors HOG eye, HOG mouth, 2D-DWT eye, and 2D-DWT mouth are projected to a lower dimensional space using the dimensionality reduction methods. Here, 2 dimension and 5 dimension refers to the number of dimensions after performing the dimensionality reduction. The recognition rate is similar for 2 dimensions and 5 dimensions when the semi-supervised DBN is used along with PCA, Kernel PCA, LLE, Isomap, and t-SNE. However, LDA and semi-supervised DBN with a 2-dimensional feature vector give a reasonable increase in recognition rate compared to using 5 dimensions. The dimensionality-reduced HOG mouth features show the best performance by giving a recognition rate of 98.57%. Similarly, the 2D-DWT mouth features also gave a comparable accuracy of 90%. The dimensionality-reduced mouth features were able to classify the emotions more effectively than the others. Further, it was observed that LDA based dimensionality reduction gave a huge increase in recognition rate compared to the rest of the methods.

3.4. Confusion matrices for different feature vectors

The test confusion matrices comparing the four different feature vectors are shown in Fig. 4 for CK+ database. Here the dimensionality reduction of the feature vectors is done using LDA and these features are passed on to the semi-supervised DBN using 60% of labeled data and the rest non-labeled data. in order to evaluate its performance. In the case of HOG eye features, it is unable to classify sadness, disgust and neutral emotions whereas it shows better accuracy for the rest of the cases. The 2D-DWT features of eye show confusion only in case of surprise and disgust whereas the 2D-DWT features of mouth show good performance in all cases except in case of sadness.

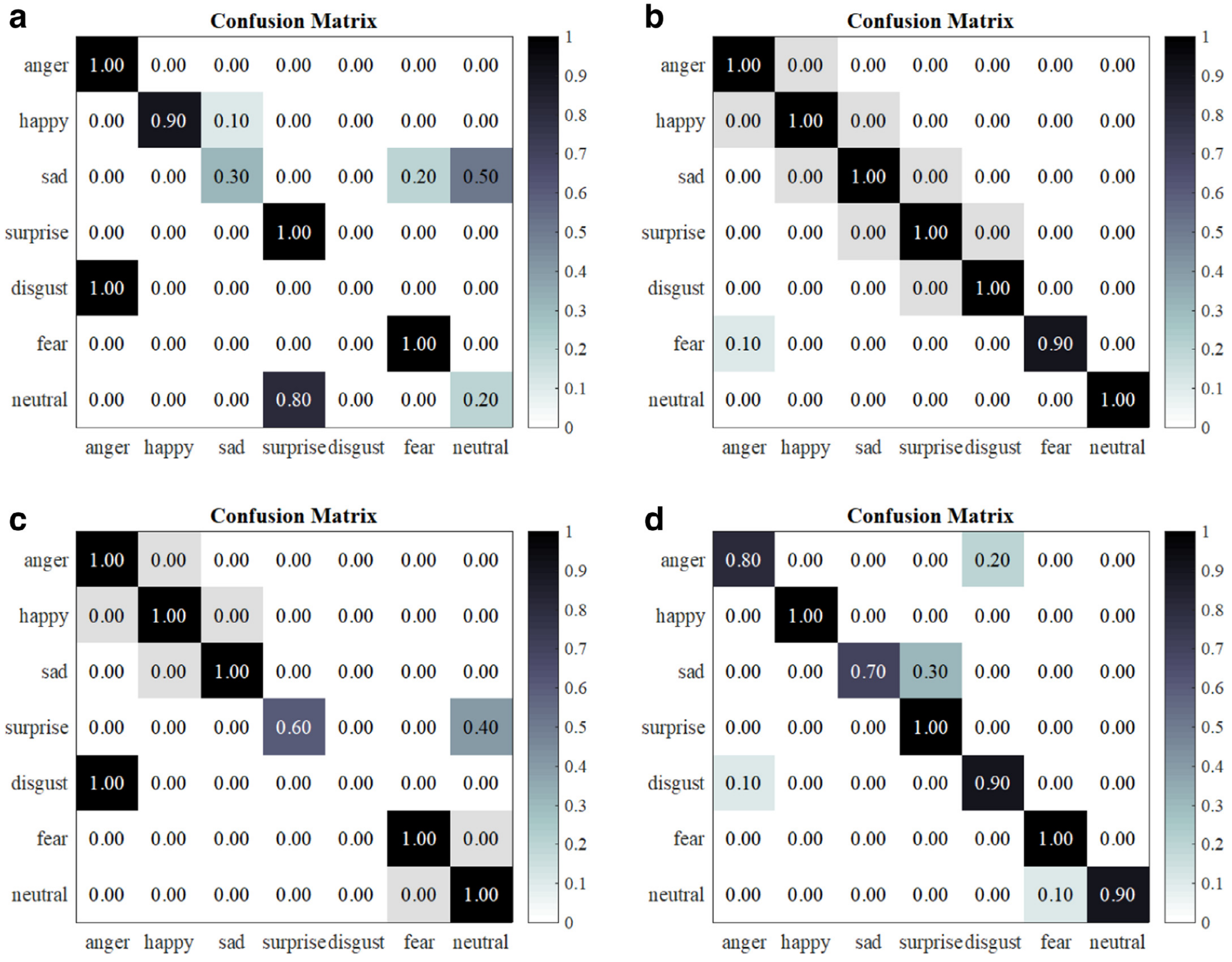

Fig. 4. Test confusion matrix of CK+ database for semi-supervised DBN with LDA, using (a) HOG eye (b) HOG mouth (c) 2D-DWT eye (d) 2D-DWT mouth features.

However, the overall recognition rate of 2D-DWT of mouth is higher than that of HOG eye and 2D-DWT eye features. Amongst the four features used, the best performance was given by HOG mouth features since it gave a recognition rate of 90% for fear and 100% for the rest of the emotions. In general, the mouth features after dimensionality reduction were found to be more suitable for the application of facial ER.

3.5. Test accuracy comparing different training criteria

Fig. 5 compares the test accuracies obtained for supervised, semi-supervised and unsupervised training of the DBN. It can be observed from the bar graph that the best performance in terms of the most significant features is shown by HOG features of mouth. Semi-supervised training is able to give better test accuracy for each of the feature sets, compared to supervised and unsupervised training methods. The experimental data undergoes dimensionality reduction using LDA before inputting into the DBN system. The semi-supervised DBN network using HOG features of mouth gives the highest test accuracy of 98.57% with CK+ database.

Semi-supervised training uses both labeled and unlabeled data for training and hence it is able to provide the network with more information. This training technique uses 60% as labeled data and the label information of the rest of the data is discarded before the experiment. The supervised training uses only the labeled data for recognizing the facial expressions whereas the unsupervised DBN uses the entire data as unlabeled and the training of the network does not include the minimization of BP error.

3.6. Performance evaluation based on dimensionality

Fig 6 shows the comparison between the performance of semi-supervised DBN using high dimensional and lower dimensional features. The relevance of dimensionality reduction is well-explained through the bar graph. Using the reduced features the test accuracies have improved significantly. The test accuracy improved from 75.24% to 98.57% in case of HOG features of mouth after using low dimensional features.

It is very clear that dimensionality reduction, in particular using LDA, is able to classify the facial expressions better compared to the higher dimensional data. For example, using LDA the dimensionality was reduced from 6048 to 2 dimension for HOG features of mouth. This huge reduction of dimension guarantees that the relevant information for classification of facial expression is preserved, hence, it shows improvement in the test accuracy. Further, it can be observed that for all the feature vectors the low dimensional features show much better performance compared to the high dimensional features.

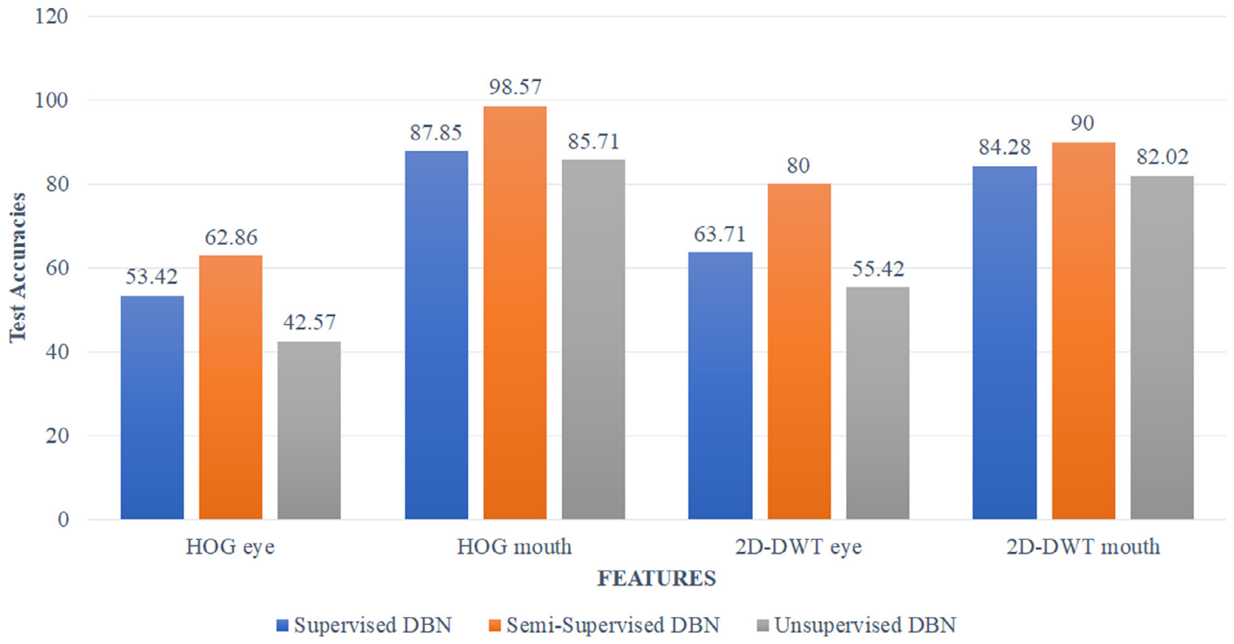

Fig. 5. Test accuracies of supervised, semi-supervised and unsupervised DBN for dimensionality reduced features using LDA on CK+ database.

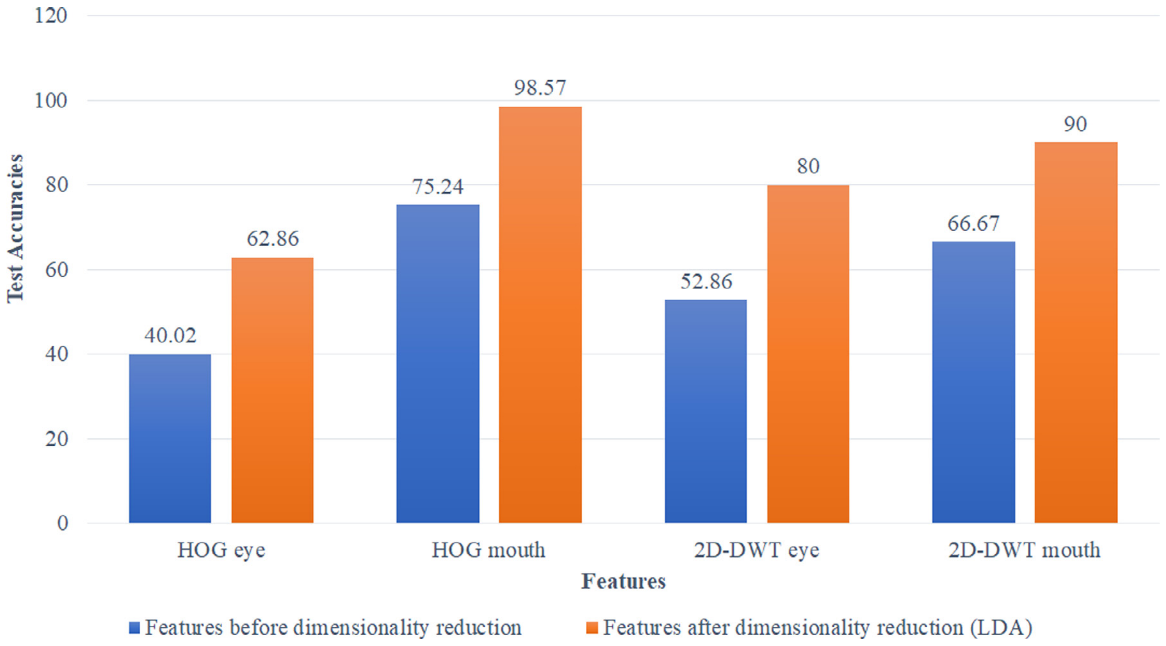

Fig. 6. Test accuracies of semi-supervised DBN model using high dimensional and low dimensional features.

3.7. Feature analysis using reconstruction error

Minimization of the quadratic reconstruction error was one of the methods used for the analysis of different feature vectors. Reconstruction of the input is a mapping from the output to input. The mean square error of the input and the reconstructed input is computed for each epoch during training for different feature matrices as inputs. Fig. 7 shows the plot of reconstruction error vs epoch for validation.

It is observed that mouth HOG features have the lowest reconstruction error, hence the most informative among all the feature sets. The reconstruction error attains a steady value around 1000 epochs after which there is not much change recorded. It can be seen that the mouth features perform better compared to that of the features from the eyes. Further, the HOG features of mouth perform the best among all the feature sets with the highest test accuracy of 98.57%.

3.8. Computational analysis based on dimensionality of data

Table 2 shows the run-time for different dimensions of the input feature vector. The feature vectors were input to the semi-supervised DBN and the run-time was recorded for testing and over individual epoch for training. The experiment was performed for different feature sets. The feature vector with the exact dimensions before dimensionality reduction and after dimensionality reduction was used to compute the run-time for training and testing. For example, the HOG features detected from the eye patch

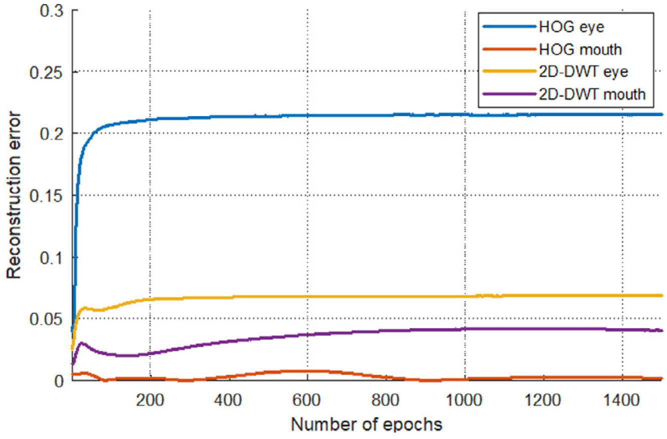

Fig. 7. Graph showing the reconstruction error versus epoch during semi-supervised training of DBN for different features.

Table 2
Performance based on the run-time for different dimensions of input feature vector.

Feature	Dimensionality	Run time (in seconds)	
		Training (per epoch)	Testing
HOG eye	9504	1.832	0.1534
	5	0.0635	0.0283
	2	**.0591**	**0.0228**
HOG mouth	6048	1.082	0.0891
	5	0.0628	0.0263
	2	**0.0578**	**.0231**
2D-DWT eye	5832	1.038	0.0854
	5	0.0621	0.0256
	2	**0.0569**	**0.0221**
2D-DWT mouth	2916	0.501	0.052
	5	0.0645	0.0272
	2	**0.0595**	**0.023**

was of dimension 9504 × 1 for each sample data and the dimensionality of the feature vector was reduced to 5 dimensions and 2-dimension to train using semi-supervised DBN.

From Table 2, it can be observed that dimensionality reduction results in fast training. The run-time recorded for reduced features is half of that of the whole feature set. The dimensionality reduction technique used here is LDA. It was able to capture the most relevant features required for the classification of facial expression as it was able to give a good performance in terms of accuracy. The significant reduction in the run-time shows the relevance of dimensionality reduction as it saves computational time and also memory usage.

3.9. Performance comparison with multi-class SVM and CNN

Fig. 8 shows the test error comparing the semi-supervised DBN and SVM. The test error graphs of HOG mouth features were plotted against different percentages of labeled data in order to evaluate the performance of the classifiers.

In the case of the semi-supervised DBN approach, it was found that each of the feature vectors exhibited the lowest test error when there was just 60% of labeled data and the rest unlabeled data. The test error of HOG features of mouth oscillates around zero, thereby showing the significance of these features in improving the classifier. Moreover, the multiclass SVM does not show a comparable performance even while utilizing the entire amount of labeled data. It has a high and fluctuating test error irrespective of the feature vector used. Thus the proposed semi-supervised approach was able to attain a superior performance with a lesser amount of labeled data compared to the traditional SVM. Table 3 shows the accuracy for different approaches. Our proposed approach using Semi-supervised DBN was able to obtain a comparable accuracy with CNN based approaches. Semi-supervised DBN required label information from just 60% of the total data to train the system whereas CNNs undergo supervised training with label information from the entire data. Further, the training time and memory usage of CNNs [56–58] were very high compared to our approach. The combined training and test time for our approach was 197.5 seconds whereas the CNN based approach [57] takes about 12,444 s.

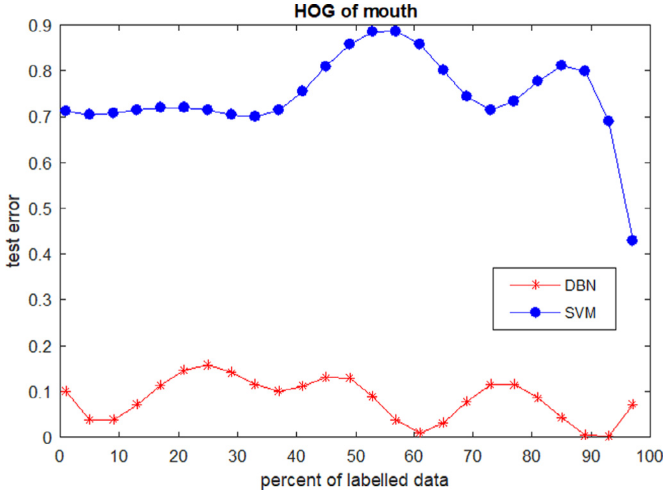

Fig. 8. Test error of semi-supervised DBN with different ratios of labeled and unlabeled data using HOG mouth features.

Table 3
Accuracy comparison with different databases and classifiers. For Semi-supervised training 60% of the data had the label information and the rest 40% were used as unlabeled data. But all the supervised methods used the entire labeled data for training.

Method	Training	Feature	Database	Accuracy
DBN(Ours)	Semi-supervised	HOG mouth	CK+	98.57%
			RaFD 135°	91.95%
			RaFD 90°	94.50%
			RaFD 45°	92.75%
			MMI	98.75%
SVM	Supervised	HOG mouth	CK+	28.57%
			RaFD 135°	84.00%
			RaFD 90°	88.20%
			RaFD 45°	86.28%
			MMI	84.57%
STC-NLSTM [56]	Supervised	Convolutional features	CK+	99.80%
			MMI	84.53%
CNN [57]	Supervised	Convolutional features	RaFD 90°	93.41%
CNN with visual attention [58]	Supervised	Convolutional features	RaFD 135° + RaFD 45°	93.10%
			RaFD 90°	95.20%

4. Conclusion

A semi-supervised approach for facial ER utilizing reduced facial features with most of the data being unlabeled is introduced with a four-layered neural network. They are convenient to use due to their easy training. Since we use CD and BP, training can be done sequentially. Semi-supervised learning was achieved by combining CD and BP, as CD is unsupervised, and BP is supervised. The facial features used were mouth and eye HOG, 2D-DWT of eyes and 2D-DWT of mouth. Further, the analysis was done with different dimensionality reduction algorithm on each of the feature sets. The test accuracies were compared for the semi-supervised training using DBN and supervised training using SVM and CNN. It was observed that the semi-supervised training showed the best performance with a test accuracy of 98.57% and outperformed SVM in terms of accuracy and CNN in terms of computational complexity. DBN used the information in unlabeled data to give better performance and the most accurate model required 60% of labeled data and 40% of unlabeled data. The semi-supervised training with HOG features of mouth also showed a consistent performance even when there was just 10% labeled data and rest unlabeled data. Furthermore, the DBN was trained in a supervised, semi-supervised and unsupervised manner using the reduced features to examine the difference in performance and again semi-supervised training was able to give better accuracy compared to the supervised and unsupervised training. Feature analysis with the reduced-dimensional features (LDA) was performed using the reconstruction error technique. Based on the experiment using reconstruction error it was found that reduced HOG features of mouth contained the most relevant information to classify facial expression. Analysis based on training run-time and test accuracies for features of different dimension was also carried out. It was observed that the best performance, that is minimum run-time and high accuracy was given by the dimensionally reduced features (LDA) with 2-dimensions. The test accuracies improved significantly after using dimensionality reduction technique. Future work aims at the use of ER technology in videos, in particular in emergency response situational awareness systems with thermal imaging, to detect emotions in civilians in the emergency scenario.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by National Science Foundation S&CC EAGER grant 1637092. Authors would like to thank UNM Center for Advanced Research Computing, supported in part by the NSF, for providing high performance computing, large-scale storage and visualization resources.

References

[1] J. Park, J. Kim, Y. Oh, Feature vector classification based speech emotion recognition for service robots, IEEE Trans. Consum. Electr. 55 (3) (2009) 1590–1596, doi:10.1109/TCE.2009.5278031.

[2] D.J. France, R.G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk, IEEE Trans. Biomed. Eng. 47 (7) (2000) 829–837, doi:10.1109/10.846676.

[3] P. Liu, S. Han, Z. Meng, Y. Tong, Facial expression recognition via a boosted deep belief network, in: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1805–1812, doi:10.1109/CVPR.2014.233.

[4] M.F. Valstar, M. Pantic, Combined support vector machines and hidden markov models for modeling facial action temporal dynamics, in: Proceedings of the 2007 IEEE International Conference on Human-computer Interaction, HCI'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 118–127.

[5] C. Soladi, N. Stoiber, R. Sguier, A new invariant representation of facial expressions: definition and application to blended expression recognition, in: Proceedings of the 2012 19th IEEE International Conference on Image Processing, 2012, pp. 2617–2620, doi:10.1109/ICIP.2012.6467435.

[6] M. Pantic, I. Patras, Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences, IEEE Trans. Syst. Man Cybern. Part B (Cybern.) 36 (2) (2006) 433–449, doi:10.1109/TSMCB.2005.859075.

[7] Y.-l. Tian, T. Kanade, J.F. Cohn, Recognizing action units for facial expression analysis, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2) (2001) 97–115, doi:10.1109/34.908962.

[8] M. Pantic, L.J.M. Rothkrantz, Automatic analysis of facial expressions: the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) (2000) 1424–1445, doi:10.1109/34.895976.

[9] Y. Hu, Z. Zeng, L. Yin, X. Wei, X. Zhou, T.S. Huang, Multi-view facial expression recognition, in: Proceedings of the 2008 8th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2008, 2008, doi:10.1109/AFGR.2008.4813445.

[10] C. Shan, S. Gong, P.W. McOwan, Facial expression recognition based on local binary patterns: a comprehensive study, Image Vis. Comput. 27 (6) (2009) 803–816, doi:10.1016/j.imavis.2008.08.005.

[11] S.B. Kazmi, Q. Ain, M.A. Jaffar, Wavelets based facial expression recognition using a bank of neural networks, in: Proceedings of the 2010 5th International Conference on Future Information Technology, 2010, pp. 1–6, doi:10.1109/FUTURETECH.2010.5482717.

[12] M. Turk, A. Pentland, Eigenfaces for recognition, J. Cognit. Neurosci. 3 (1) (1991) 71–86, doi:10.1162/jocn.1991.3.1.71. PMID: 23964806

[13] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720, doi:10.1109/34.598228.

[14] x. shu, J. Tang, Z. Li, H. Lai, L. Zhang, S. Yan, Personalized age progression with bi-level aging dictionary learning, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2018) 905–917, doi:10.1109/TPAMI.2017.2705122.

[15] T. Balachander, R. Kothari, Kernel based subspace pattern classification, in: Proceedings of the International Joint Conference on Neural Networks. Proceedings (Cat. No.99CH36339), IJCNN'99., volume 5, 1999, pp. 3119–3122 vol.5, doi:10.1109/IJCNN.1999.836149.

[16] L.K. Saul, S.T. Roweis, Think globally, fit locally: unsupervised learning of low dimensional manifolds, J. Mach. Learn. Res. 4 (2003) 119–155, doi:10.1162/153244304322972667.

[17] J.B. Tenenbaum, V.d. Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323, doi:10.1126/science.290.5500.2319.

[18] L. van der Maaten, Accelerating t-SNE using tree-based algorithms, J. Mach. Learn. Res. 15 (2014) 3221–3245.

[19] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227, doi:10.1109/TPAMI.2008.79.

[20] X. Shu, J. Tang, G. Qi, Z. Li, Y. Jiang, S. Yan, Image classification with tailored fine-grained dictionaries, IEEE Trans. Circuits Syst. Video Technol. 28 (2) (2018) 454–467, doi:10.1109/TCSVT.2016.2607345.

[21] X. Lan, S. Zhang, P.C. Yuen, R. Chellappa, Learning common and feature-specific patterns: a novel multiple-sparse-representation-based tracker, IEEE Trans. Image Process. 27 (4) (2018) 2022–2037, doi:10.1109/TIP.2017.2777183.

[22] J. Gu, L. Jiao, F. Liu, S. Yang, R. Wang, P. Chen, Y. Cui, J. Xie, Y. Zhang, Random subspace based ensemble sparse representation, Pattern Recognit. 74 (2017), doi:10.1016/j.patcog.2017.09.016.

[23] S. Hochreiter, Y. Bengio, P. Frasconi, Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, in: J. Kolen, S. Kremer (Eds.), Field Guide to Dynamical Recurrent Networks, IEEE Press, 2001.

[24] G.E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (7) (2006) 1527–1554, doi:10.1162/neco.2006.18.7.1527.

[25] K.P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press, 2012.

[26] A. Rasmus, H. Valpola, M. Honkala, M. Berglund, T. Raiko, Semi-supervised learning with ladder networks, in: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15, MIT Press, Cambridge, MA, USA, 2015, pp. 3546–3554.

[27] H. Xu, K.N. Plataniotis, Affective states classification using EEG and semi-supervised deep learning approaches, in: Proceedings of the 2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP), 2016, pp. 1–6, doi:10.1109/MMSP.2016.7813351.

[28] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, 2005, pp. 886–893 vol. 1, doi:10.1109/CVPR.2005.177.

[29] Y. Ouyang, N. Sang, R. Huang, Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers, Neurocomputing 149 (2015) 71–78, doi:10.1016/j.neucom.2014.03.073. Advances in neural networks Advances in Extreme Learning Machines

[30] H. Ali, V. Sritharan, M. Hariharan, S.K. Zaaba, M. Elshaikh, Feature extraction using radon transform and discrete wavelet transform for facial emotion recognition, in: Proceedings of the 2nd IEEE International Symposium on Robotics and Manufacturing Automation (ROMA), 2016, pp. 1–5, doi:10.1109/ROMA.2016.7847840.

[31] F. Dornaika, F. Davoine, Simultaneous facial action tracking and expression recognition in the presence of head motion, Int. J. Comput. Vis. 76 (3) (2008) 257–281.

[32] F.Y. Shih, C.-F. Chuang, P.S.P. Wang, Performance comparisons of facial expression recognition in JAFFE database, Int. J. Pattern Recognit. Artif. Intell. 22 (03) (2008) 445–459.

[33] H. Hotelling, Analysis of a complex of statistical variables with principal components, J. Educ. Psychol. 24 (1933) 417–441.

[34] I. Jolliffe, Principal Component Analysis, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011, pp. 1094–1096, doi:10.1007/978-3-642-04898-2_455.

[35] S. Mika, B. Schölkopf, A.J. Smola, K.-R. Müller, M. Scholz, G. Rätsch, Kernel PCA and de-noising in feature spaces, in: M.J. Kearns, S.A. Solla, D.A. Cohn (Eds.), Advances in Neural Information Processing Systems 11, MIT Press, 1999, pp. 536–542.

[36] B. Scholkopf, A.J. Smola, Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, Cambridge, MA, USA, 2001.

[37] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, Newark, NJ, 2005.

[38] B.A. Draper, K. Baek, M.S. Bartlett, J.R. Beveridge, Recognizing faces with PCA and ICA, Comput. Vis. Image Underst. 91 (1–2) (2003) 115–137, doi:10.1016/S1077-3142(03)00077-8.

[39] J.B. Tenenbaum, Mapping a manifold of perceptual observations, in: M.I. Jordan, M.J. Kearns, S.A. Solla (Eds.), Advances in Neural Information Processing Systems 10, MIT Press, 1998, pp. 682–688.

[40] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326, doi:10.1126/science.290.5500.2323.

[41] G. Hinton, Y. Bengio, Visualizing data using t-SNE, Cost-sensitive Machine Learning for Information Retrieval 33, 2008.

[42] F. Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, Psychol. Rev. 65 6 (1958) 386–408.

[43] Y. LeCun, L. Bottou, G. Orr, K. Müller, Efficient backprop, in: G. Orr, K. Muller (Eds.), Neural Networks: Tricks of the trade, Springer, 1998.

[44] R. Salakhutdinov, G. Hinton, Deep Boltzmann machines, in: D. van Dyk, M. Welling (Eds.), Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, volume 5, PMLR, 2009, pp. 448–455.

[45] R. Salakhutdinov, H. Larochelle, Efficient learning of deep boltzmann machines, in: Y.W. Teh, M. Titterington (Eds.), Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, volume 9, PMLR, Chia Laguna Resort, Sardinia, Italy, 2010, pp. 693–700.

[46] G.E. Hinton, Training products of experts by minimizing contrastive divergence, Neural Comput. 14 (8) (2002) 1771–1800, doi:10.1162/089976602760128018.

[47] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning representations by back--propagating errors, Nature 323 (1986). 533–

[48] C.M. Bishop, Pattern Recognition and Machine Learning (Information Science and Statistics), Springer-Verlag, Berlin, Heidelberg, 2006.

[49] P. Lucey, J.F. Cohn, T. Kanade, J.M. Saragih, Z. Ambadar, I.A. Matthews, The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression, in: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2010, pp. 94–101.

[50] O. Langner, R. Dotsch, G. Bijlstra, D.H.J. Wigboldus, S.T. Hawk, A. van Knippenberg, Presentation and validation of the radboud faces database, Cognit. Emot. 24 (8) (2010) 1377–1388, doi:10.1080/02699930903485076.

[51] M. Valstar, M. Pantic, Induced disgust, happiness and surprise: an addition to the MMI facial expression database, in: Proceedings of the International Conference Language Resources and Evaluation, Workshop EMOTION, 2010, pp. 65–70.

[52] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001., volume 1, 2001, doi:10.1109/CVPR.2001.990517.

[53] H. Yan, Biased subspace learning for misalignment-robust facial expression recognition, Neurocomputing 208 (2016) 202–209, doi:10.1016/j.neucom.2015.11.115. SI: BridgingSemantic

[54] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, A.M. Dobaie, Facial expression recognition via learning deep sparse autoencoders, Neurocomputing 273 (2018) 643–649, doi:10.1016/j.neucom.2017.08.043.

[55] B. Heisele, P. Ho, T. Poggio, Face recognition with support vector machines: global versus component-based approach, in: Proceedings of the Eighth IEEE International Conference on Computer Vision. ICCV 2001, volume 2, 2001, pp. 688–694 vol.2, doi:10.1109/ICCV.2001.937693.

[56] Z. Yu, G. Liu, Q. Liu, J. Deng, Spatio-temporal convolutional features with nested LSTM for facial expression recognition, Neurocomputing 317 (2018) 50–57, doi:10.1016/j.neucom.2018.07.028.

[57] W. Sun, H. Zhao, Z. Jin, A complementary facial representation extracting method based on deep learning, Neurocomputing 306 (2018) 246–259, doi:10.1016/j.neucom.2018.04.063.

[58] W. Sun, H. Zhao, Z. Jin, A visual attention based ROI detection method for facial expression recognition, Neurocomputing 296 (2018) 12–22, doi:10.1016/j.neucom.2018.03.034.

Aswathy Rajendra Kurup received the bachelor's degree in Electronics and Communication Engineering from Amrita school of Engineering in 2015 and the master's degree in Electrical Engineering from The University of New Mexico in 2017. She is currently working towards her Ph.D. degree in Electrical Engineering from The University of New Mexico. Her research interests are Image Processing, Signal Processing and Machine Learning.

Meenu Ajith received the bachelor's degree in Electronics and Communication Engineering from Amrita school of Engineering in 2015 and the master's degree in 2017 in Electrical Engineering from The University of New Mexico in 2017. She is currently working towards her Ph.D. degree in Electrical Engineering from The University of New Mexico. Her research interests are Machine Learning, Computer Vision, Pattern Recognition and Image Processing.

Manel Martínez Ramón is a professor with the ECE department of The University of New Mexico. He holds the King Felipe VI Endowed Chair of the University of New Mexico, a chair sponsored by the Household of the King of Spain. He is a Telecommunications Engineer (Universitt Politecnica de Catalunya, Spain, 1996) and Ph.D. in Communications Technologies (Universidad Carlos III de Madrid, Spain, 1999). His research interests are in Machine Learning applications to smart antennas, neuroimage, first responders and other cyber-human systems, smart grid and others. His last work is the monographic book "Signal Processing with Kernel Methods", Wiley, 2018.