Neural Networks 105 (2018) 304–315

Contents lists available at ScienceDirect

Neural Networks

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

# Nonlinear analysis and synthesis of video images using deep dynamic bottleneck neural networks for face recognition

Saeed Montazeri Moghadam, Seyyed Ali Seyyedsalehi *

*Department of Biomedical Engineering, Amirkabir University of Technology, 424Hafez Ave, Tehran, Iran*

ARTICLE INFO

*Article history:*
Received 15 December 2017
Received in revised form 17 May 2018
Accepted 23 May 2018
Available online 31 May 2018

*Keywords:*
Nonlinear video analysis
Video synthesis
Expression intensity
Deep neural networks
Facial expression

ABSTRACT

Nonlinear components extracted from deep structures of bottleneck neural networks exhibit a great ability to express input space in a low-dimensional manifold. Sharing and combining the components boost the capability of the neural networks to synthesize and interpolate new and imaginary data. This synthesis is possibly a simple model of imaginations in human brain where the components are expressed in a nonlinear low dimensional manifold. The current paper introduces a novel *Dynamic Deep Bottleneck Neural Network* to analyze and extract three main features of videos regarding the expression of emotions on the face. These main features are identity, emotion and expression intensity that are laid in three different sub-manifolds of one nonlinear general manifold. The proposed model enjoying the advantages of recurrent networks was used to analyze the sequence and dynamics of information in videos. It is noteworthy to mention that this model also has also the potential to synthesize new videos showing variations of one specific emotion on the face of unknown subjects. Experiments on discrimination and recognition ability of extracted components showed that the proposed model has an average of 97.77% accuracy in recognition of six prominent emotions (Fear, Surprise, Sadness, Anger, Disgust, and Happiness), and 78.17% accuracy in the recognition of intensity. The produced videos revealed variations from neutral to the apex of an emotion on the face of the unfamiliar test subject which is on average 0.8 similar to reference videos in the scale of the SSIM method.



## 1. Introduction

Analyzing sequential images, which are continuously acquired by the human visual system, yields a clear recognition and understanding of the environmental events. This recognition is the result of processing a combination of various kinds of information such as online input image, variations between sequential images through time, previous experiences from similar situations and other informative sources (Ullman, 1996). It is documented that a human extracts these important information from his/her visual input as low-dimensional high-level concepts, i.e. principal components (Nejadgholi & Seyyedsalehi, 2012). Instead of the whole images, these components are stored in mind (Nejadgholi & Seyyedsalehi, 2012). Subsequently, this component, which lies in a nonlinear manifold space in the brain, can be combined with other components in order to synthesize an imagination or vision and reconstruct a memory.

The recognition of facial emotions and image understanding are two tasks that human beings perform continuously and effortlessly (Jain & Li, 2011), but these tasks are not yet accomplished perfectly by computers. Nowadays, the recognition of identity and emotions has a large variety of applications in animation, biometry, security, neuromarketing, interactive games, and sociable robotics (Lopes, Aguiar, De Souza, & Oliveira-Santos, 2017).

Some prominent works concentrate on using methods of image analysis and techniques of machine learning (Kacem, Daoudi, Amor, & Alvarez-Paiva, 2017). Chen, Chen, and Fu (2016) proposed a new feature descriptor called Histogram of Oriented Gradients from Three Orthogonal Planes (HOG-TOP) to extract dynamic textures from video sequences to characterize changes in facial appearance. Additionally, a new geometric feature derived from the warp transformation of facial landmarks has been proposed to capture changes in facial configuration. Their experiments on the extended Cohn–Kanade (CK+) database demonstrated 95.7% accuracy. Agrawal, Cosgriff, and Mudur (2009) showed that some particular salient regions of a facial image carry major expression-related information compared with other face regions. Therefore, only a few features from each salient face region are enough for expression representation. The recognition accuracy of this method on the CK+ was 95.8%.

Ding, Zhou, and Chellappa (2017), motivated by Zhao et al. (2016), proposed a *Convolutional Neural Network* (CNN) called *expression net* based on the *face net*. They used the trained face

* Corresponding author.
*E-mail address:* ssalehi@aut.ac.ir (S.A. Seyyedsalehi).

https://doi.org/10.1016/j.neunet.2018.05.016
0893-6080/

net to regularize the training of the convolutional layers of the expression net. Their experiments on CK+ showed 98.6% accuracy. More recently, Ranjan, Patel, and Chellappa (2017) and Wang et al. (2017) also utilized CNNs with regard to the problem of face understanding. More recently, Jung et al. (2015) trained a deep temporal geometry network and a deep temporal appearance network with a joint fine-tuning method. In Mollahosseini, Chan, and Mahoor (2016), Mollahosseini et al. showed that the architecture of inception network works very well for an expression recognition task.

Recently, several experiments focusing on manifold extraction and manifold learning are reported (Gu, Xiang, Venkatesh, Huang, & Lin, 2012; Hamedani, Seyyedsalehi, & Ahamdi, 2016; Khandait, Thool, & Khandait, 2012; Li, Imai, & Kaneko, 2010; Rao & Thiagarajan, 2010; Seyyedsalehi & Seyyedsalehi, 2014; Wan, Tian, & Liu, 2012; Zhang & Zhou, 2014; Zhao & Pietikainen, 2007). Manifold extraction is defined as extracting high-level and conceptual information by compressing the input data in a very low dimensional space (Hinton & Salakhutdinov, 2006; Nejadgholi & Seyyedsalehi, 2012). For instance, in the field of understanding the facial images, the manifold can be made up of two sub-manifolds of identity and emotion. Based on the number of subjects and emotions of study, each sub-manifold consists of different components. With respect to the problem of identity and recognition of emotions, manifold learning has demonstrated better results compared to other methods.

Chang, Hu, Feris, and Turk (2006) used a mixed model to map facial contour features into a low dimensional manifold space. Mohammadian, Aghaeinia, and Towhidkhah (2016) presented a *Manifold Based Parametric Model* (MBPM) to analyze and produce a virtual image of different emotions on the face of different identities based on a number of dynamic parameters. Both researchers achieved image synthesizing by manipulating manifold features.

It is illustrated that auto-associative neural networks have a great ability in nonlinear feature and component extraction and manifold formation (Hinton & Salakhutdinov, 2006; Mase, 1991). *Bottleneck neural networks* (BNN) are the typical and more common types of auto-associative neural networks which can do dimension reduction along with feature extraction (Seyyedsalehi & Seyyedsalehi, 2015). In each layer of BNNs, the dimension of input data is reduced by decreasing the number of neurons. More recent research studies on dimension reduction and feature extraction using deep BNN (DBNN) demonstrated fascinating results in discriminatory feature extraction and more accuracy in recognition (Huang, Wang, Wang, & Tan, 2014; Schmidhuber, 2015). By increasing the number of processing layers in BNNs, the network and also the processing become deeper. However, the possibility of trapping in *local minima* increases. Therefore, using pre-training methods that prepare proper weights (instead of using random weights) for the fine-tuned training stage is demanded. Seyyedsalehi (Kacem et al., 2017) achieved 92.86% accuracy in the recognition of emotions on CK+ faces using layer-by-layer pre-training algorithm (Seyyedsalehi & Seyyedsalehi, 2015).

Moreover, manifolds of sequential data can represent the data characteristics better if and only if the variations of information lying in the sequences are included in their formation. The *Hidden Markov Models* (HMMs) and *Recurrent Neural Networks* (RNNs) are the two top common methods in dynamic modeling and sequence analyses. Hence, some of the investigations are done in video analyses especially in the area of facial understanding using these two methods (Aleksic & Katsaggelos, 2006; Chao, Tao, Yang, Li, & Wen, 2015; Rahimi, Darrell, & Recht, 2005; Yan, Chang, Shan, & Chen, 2014). Lien, Kanade, Cohn, and Li (1998) used HMM and *Facial Action Coding System* (FACS) for discrimination of information in facial expression and classification of the expressions, respectively.

In the field of facial understanding, some studies are done on the detection of expression intensity from the face especially for clinical and psychological purposes (Calvo & Marrero, 2009; Suslow, Junghanns, & Arolt, 2001; Lijun, Wei, Sun, Wang, & Rosato, 2006). Expression intensity is defined as the intensity of an emotion in the face, which is expressed by a person. The intensity is usually quantized at some levels based on the experts' definition. Liu et al. (2014) clustered the intensities of different emotions and then extracted partial expression transitions between each cluster and saved them in a dictionary. Then, he matched the extracted expression transition(which was sparse and partial) of an input video with the dynamic dictionary to recognize the emotion from the video. He achieved 92.08% accuracy in the recognition of emotion for the CK+ dataset. This paper presents a model analyzing a video of a person showing the variation of an emotion in his/her face and then synthesizes these variations on the face of an unknown person. Thus, it seems that analyzing and detecting the intensity of expression plays an important role in this model. Lee and Xu (2003) has developed a system which can automatically estimate the intensity of facial expressions in real-time. In his model, the intensity of expression is extracted from training facial transition sequences based on isometric feature mapping. Then, he used *cascade neural networks* and *Support Vector Machine* (SVM) to model the relationship between the trajectories of facial feature points and the expression intensity level.

In our previous work (Moghadam & Seyyedsalehi, 2017), we presented emotion recognition DBNN. In this paper, we now propose a dynamic DBNN (DDBNN) model to extract low-dimensional high-level components from the input video and represent them in a manifold with three sub-manifolds. These sub-manifolds are identity, emotion and expression intensity. The model should synthesize variations of emotions on the face of a test subject in a video stream. This aim will be achieved by combining the information of components from two sub-manifolds.

The remainder is organized into 4 sections: Section 2 describes the presented novel dynamic DBNN model and the special learning algorithm for the proper extraction of components in the bottleneck part of the model. In Section 3, the results of simulations and analyses on the ability of the model in recognition and synthesis of different variations of an emotion are presented. Further, the ability of the model in interpolating a new intensity of emotion between two labeled intensities is assessed. Section 4 summarizes the results of this work and draws conclusions.

## 2. Materials and methods

The purpose of this paper is to extract a unified manifold consisting of principal components in three different sub-manifolds (identity, emotion and expression intensity). This manifold is extracted from videos using the encoder part of the model. Furthermore, this model should synthesize sequential images of variations in emotion intensity using its dynamic structure. We will discuss these two tasks of the model in the following two subsections, separately. Then in the last part of this section, these two parts are combined and the overall model is presented.

### 2.1. Dataset

The extended Cohn–Kanade (CK+) database is utilized to train and test the ability of the model in analyzing video sequences and manifold learning (Kanade, Cohn, & Tian, 2000; Lucey et al., 2010). This dataset has 593 video sequences from 123 university students within the age range of 18 to 30. In every experiment, each subject was asked to show some emotional facial expressions in sequential mode from neutral to the apex. Furthermore, in order to omit any in-plane head rotation in each frame, all the

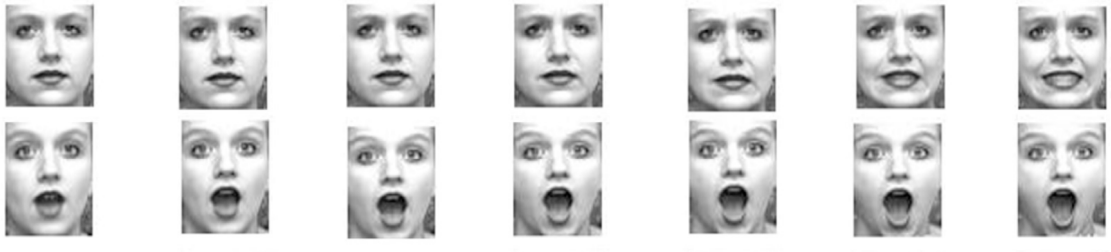

**Fig. 1.** Some samples extracted from preprocessed CK+ videos showing variations from neutral to the apex of "fear" (up) and "surprise" (down) emotions.

pixels were automatically rotated such that the line connecting two landmarks located in the inner corner of each eye was aligned horizontally. The landmarks were extracted from text files, which the CK+ database provides for every frame to facilitate the face detection. After this pre-processing, every frame was resized to a 50*50 gray-scale image and normalized to be ready to be fed to the network. Fig. 1 shows two sets of frame sequences in "fear" and "surprise" emotions from the video of one training subject after pre-processing.

The database is used in two different sets. One bigger set is used to train and test the proposed final model. This set has 3528 images consisted of 7 sequential frames (7 different intensities), from the start of showing expression to apex in 6 different emotions (fear, surprise, sadness, anger, disgust and happiness) from 95 subjects. The second smaller set is considered to train and test each one of the four proposed multi-component separator models separately in order to choose the one with higher performance. It contains 7 intensities in 6 mentioned emotions from 20 subjects. The smaller set provides better time and computation efficiency in training. So it is first used in order to choose the best multi-component separator model and the bigger set is then used to train the final model with the winner separator model in the Principal Components Extractor part of Fig. 2.

### 2.2. Basic model

Introducing the basic model, a nine-layer static DBNN model is considered and demonstrated in Fig. 2. The presented DBNN consists of one encoder and one decoder part. The encoder part is responsible for feature extraction and dimension reduction from input to bottle-neck, in contrast to the decoder part which is responsible for feature combination from bottleneck to output. This model is a static neural network, which is suitable for dealing with static images. The number of neurons in each layer has been chosen based on extensive experiments that have been done on 10,000 sample images from the dataset. However, the pioneer papers such as Abdolali and Seyyedsalehi (2011), Hamedani et al. (2016) and Seyyedsalehi and Seyyedsalehi (2014) are also considered as references to choose the appropriate number of neurons in each layer. The size of input is 2500 neurons with one bias which is always a constant value of one. So the size of input picture was 50*50 pixels showing the faces of subjects. The number of neurons in the multi-component separator layer of the baseline model, which works with the smaller dataset, is set to 3 for expression intensity and emotion sub-manifolds, and 5 for identity sub-manifolds. Table 1 presents other characteristics of the baseline model. All the weights of this network are pre-trained with the layer-wise pre-training algorithm presented by Seyyedsalehi and Seyyedsalehi (2015). The most important part of the presented model is the middle part (bottleneck) of the network. In the bottleneck layer, all the understanding of the network from the input is summarized in high-level low-dimensional components (Seyyedsalehi & Seyyedsalehi, 2014). Thus, the structure and

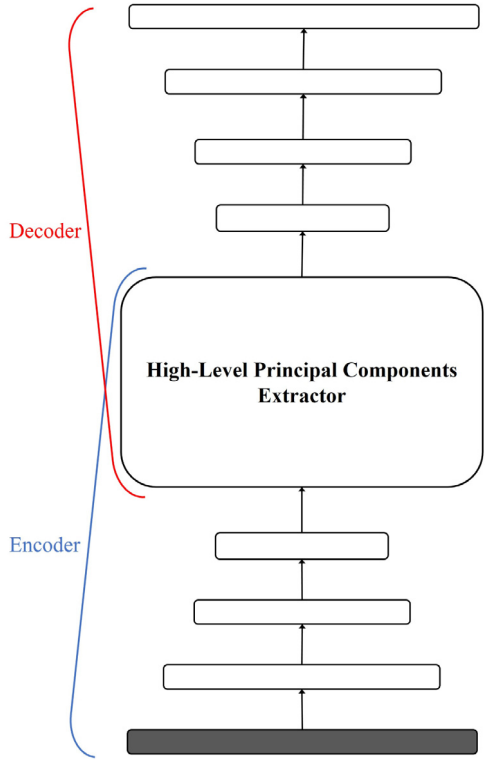


**Fig. 2.** Basic model of a nine-layer DBNN.

**Table 1**
Characteristics of basic model.

| Properties name | Specification |
|---|---|
| Number of neurons in each layer (input → output) | 2500-500-200-100-100-200-500-2500 |
| Activation functions (input → output-1) | Logsig |
| Activation function (output) | Linear |
| Learning algorithm | Back propagation |

training procedure of this part play an important role in the overall performance of the model. The following subsection presents four different approaches for the training of this part in the framework of supervised/unsupervised nonlinear manifold extraction.

### 2.3. Bottleneck and principal component extractor

In order to extract components of three different subspaces, the encoder part of a baseline DBNN is studied. These three different sub-manifolds are identity, emotion and expression intensity which should be extracted from the input videos. The

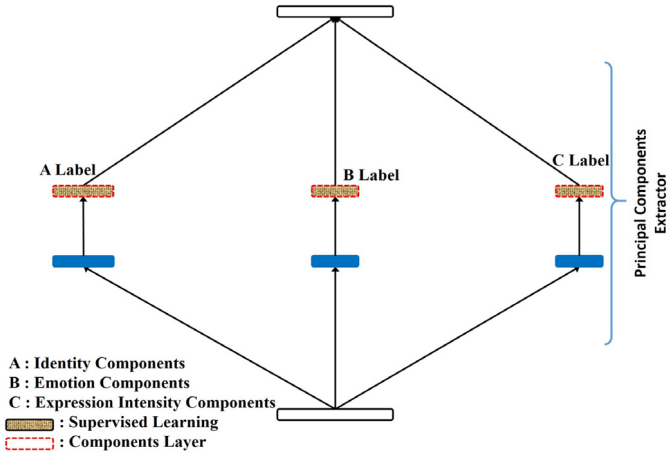


**Fig. 3.** Proposed structure for manifold extraction in the first and second method. This model follows a supervised training, without any clustering forces in the first method and with clustering in the second method. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

videos were taken from the face of subjects showing different emotions and variations of each emotion through time. The encoder part of the model should analyze the input video and extract principal components. It is expected from these principal components to fully discriminate the three above-mentioned sub-manifolds of the unified manifold. As the first step, the model should work with static images (i.e. each frame of the video separately). Four different approaches for better training of the bottleneck layer in the static encoder part of DBNN are presented as follows.

*2.3.1. Supervised and without clustering structure of neural network for nonlinear component (manifold) analysis*

Considering the static structure of DBNN, all the weights of this network except the middle part (bottleneck) are pre-trained using the layer-wise algorithm presented by Seyyedsalehi and Seyyedsalehi (2015). In this subsection, the focus is on presenting the best structure and training algorithm for the middle part, which is responsible for high-level input analysis and feature extraction illustrated in Fig. 2.

A layer is inserted before the "*components layer*" (bottleneck layer) with the same connections as it is due to the complexity of data (Blue Boxes in Figs. 3 and 4). We named this layer the "*feature separator*". The number of neurons for each part of this layer has been set to 20 based on the dimension reduction theory, i.e. the number of neurons in a layer should be less than the previous layer and greater than the next layer. The duty of each part is to extract the relevant information about the corresponding sub-manifold. For example, the first part connected to the identity component layer is meant to extract the identity features from mixed features of the lower level. The first presented model follows a supervised learning algorithm. For all the three sub-manifolds (identity, emotion, and expression intensity) a binary, unique and sparse labeling code is prepared and labels are directly fed to the component layers. We called this procedure as supervised learning which is followed by both the first and second models. So, in these models, the component layer must identify the sparse labels as the components. Fig. 3 illustrates the proposed structure for the first and second models. The difference between model 1 and model 2 lies in the existence of clustering rules in model 2 which are absent in model 1. In this paper, clustering rules are defined as some cost functions to be minimized with the output reconstruction error (Eq. (1)) and label reconstruction error (Eq. (2)). These rules force the component layer to allocate one specific feature to each class of sub-spaces separately (e.g. one specific feature to happiness in emotion sub-space). Without clustering rules, for example, the network can make different sets of components for one identity that shows different emotions on his/her face. However, it is expected to have a unique set of components for one identity showing different variations on his/her face. This procedure is exactly the same for the other two sub-manifolds.

The component layer should identify the labels as its components and simultaneously reconstruct the input data in the output layer. Considering $X$ as input vector and $\hat{X}$ as reconstructed $X$ in the output layer, the reconstruction error is defined as $X - \hat{X}$. This error is distributed back through the network layers, known as back-propagation, to the components layer. Eq. (1) is the back-propagated reconstruction error. In this equation, $\delta^{O}_{c(1:n_A)}$ is the back propagated reconstruction error from the output layer ($O$) for the component layer ($c$) of identity (A part) which has $n_A$ number of neurons as we assumed that neurons number 1 to $n_A$ are considered for identity part. Likewise, $\delta^{O}_{c(n_A+1:n_A+n_B)}$ is for emotion (B part) and $\delta^{O}_{c(n_A+n_B+1):(n_A+n_B+n_C)}$ is for intensity (C part). Also, $\delta^{O}_{c+1}$ is the reconstruction error back-propagated to the upper layer $(c+1)$, and $W'_{c(c+1)}$ is the transposed weight matrix which connects each part of the component layer ($cA$, $cB$ and $cC$) to the upper layer

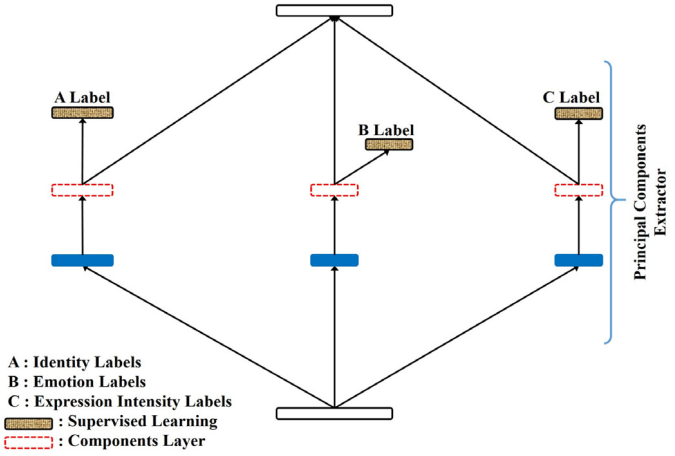


**Fig. 4.** The second proposed structure of bottleneck for manifold extraction. This model follows an unsupervised training without any clustering forces. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

$(c+1)$.

$$\begin{aligned}
\delta^O_{c(1:n_A)} &= (\frac{1}{3})(\delta^O_{c+1} * W'_{cA(c+1)}) \\
\delta^O_{c(n_A+1:n_A+n_B)} &= (\frac{1}{3})(\delta^O_{c+1} * W'_{cB(c+1)}) \\
\delta^O_{c(n_A+n_B+1:n_A+n_B+n_C)} &= (\frac{1}{3})(\delta^O_{c+1} * W'_{cC(c+1)})
\end{aligned} \tag{1}$$

Eq. (2) represents the label reconstruction error ($\delta^L_c$) for each part separately in which LA, LB and LC are the labels of identity part ($A = 1 : n_A$), emotion part ($B = n_A + 1 : n_A + n_B$) and emotion intensity part ($C = n_A + n_B + 1 : n_A + n_B + n_C$), respectively. $Y$ is also the output of the component layer for each part. The reconstruction error which is back propagated to the component layer, accumulates with label reconstruction error (Eq. (2)), which has been calculated for each part separately.

$$\begin{aligned}
\delta^L_{cA} &= (LA - Y_{cA}) \\
\delta^L_{cB} &= (LB - Y_{cB}) \\
\delta^L_{cC} &= (LC - Y_{cC})
\end{aligned} \tag{2}$$

The gradient of error in component layer based on these two errors is presented in Eq. (3). This error is used to update the weights and it is also back propagated in the underneath layers to finally reach the first layer.

$$\begin{aligned}
\delta_{cA} &= Y_{cA}(1 - Y_{cA})(\delta^O_{cA} + \delta^L_{cA}) \\
\delta_{cB} &= Y_{cB}(1 - Y_{cB})(\delta^O_{cB} + \delta^L_{cB}) \\
\delta_{cC} &= Y_{cC}(1 - Y_{cC})(\delta^O_{cC} + \delta^L_{cC})
\end{aligned} \tag{3}$$

Finally, Eq. (4) shows how the weights of component layer are updated. In this equation, $Y'_{(c-1)}$ is the transposed output of the previous layer and $\Delta W$ is the delta of weights in each iteration ($\Delta W = W_t - W_{t-1}$).

$$\begin{aligned}
\Delta W_{(c-1)c}(:, 1 : n_A) &= -\eta Y'_{c-1}\delta_{cA} \\
\Delta W_{(c-1)c}(:, n_A + 1 : n_A + n_B) &= -\eta Y'_{c-1}\delta_{cB} \\
\Delta W_{(c-1)c}(:, n_A + n_B + 1 : n_A + n_B + n_C) &= -\eta Y'_{c-1}\delta_{cC}
\end{aligned} \tag{4}$$

The back-propagated error from feature extractor layer toward the input layer is made up of the reconstruction error and the label reconstruction error and follows the *Back-Propagation* (BP) algorithm.

### 2.3.2. Supervised and clustered structure of neural network for nonlinear component (manifold) analysis

In this model, clustering conditions are also involved. Any single sample frame in the input is analyzed through the encoder to be recognized and clustered in the sub-manifolds identity, emotion and expression intensity. The Euclidean distance between the extracted components and the center of the corresponding cluster should be minimized. This error is aggregated with the errors discussed in the previous chapter. $MA_i$ is defined as the center of the $i$th cluster related to the $i$th identity in the identity sub-manifold, $MB_j$ as the center of the $j$th cluster related to the $j$th emotion in emotion sub-manifold and $MC_k$ as the center of the $k$th cluster related to the $k$th expression intensity. The centers are estimated and updated in the final stage of each epoch by Eq. (5) and they are constant over all the next iteration calculations.

$$\begin{aligned}
MA_i(t) &= \gamma MA_i(t-1) + (1-\gamma)Y^i_{cA} \\
MB_j(t) &= \gamma MB_j(t-1) + (1-\gamma)Y^j_{cB} \\
MC_k(t) &= \gamma MC_k(t-1) + (1-\gamma)Y^k_{cC}
\end{aligned} \tag{5}$$

$\gamma$ is a normalization factor to make a trade-off between the previous centers and the newly calculated component output.

Eq. (6) represents the distance between the extracted components and the center of the corresponding cluster defined as a clustering error for each subspace separately.

$$\begin{aligned}
\delta^M_{cA} &= MA - Y_{cA} \\
\delta^M_{cB} &= MB - Y_{cB} \\
\delta^M_{cC} &= MC - Y_{cC}
\end{aligned} \tag{6}$$

The gradient of error in component layer based on the two previous errors and clustering error is presented in Eq. (7). This error is used to update the same weights as in Eq. (4) and it is also back propagated in the underneath layers to finally reach the first layer.

$$\begin{aligned}
\delta_{cA} &= Y_{cA}(1 - Y_{cA})(\delta^O_{cA} + \delta^L_{cA} + \delta^M_{cA}) \\
\delta_{cB} &= Y_{cB}(1 - Y_{cB})(\delta^O_{cB} + \delta^L_{cB} + \delta^M_{cB}) \\
\delta_{cC} &= Y_{cC}(1 - Y_{cC})(\delta^O_{cC} + \delta^L_{cC} + \delta^M_{cC})
\end{aligned} \tag{7}$$

The back-propagated error from the feature extractor layer toward the input layer is made up of reconstruction error, label reconstruction error and clustering error, which follows the BP algorithm.

### 2.3.3. Unsupervised and non-clustering structure of neural network for nonlinear component (manifold) analysis

The third model is presented to follow unsupervised training. In other words, the labels did not directly feed to the component layer and the encoder extracts the components in an unsupervised way. Besides, a new layer, which is broken into three parts, is added to the outside of the main path of the network, each part of which is connected to one sub-manifold component layer. The labels are directly assigned to the neurons in this layer, making it a supervised training for this layer. Consequently, in this model, the component layer has two tasks to do, the first one is to extract features which help the network to reproduce the input sample in the output and reduce the reconstruction error. The second task is to help the new added outside layer to reconstruct the labels and recognize the identity, emotion and intensity of the input image. Subsequently, as demonstrated in Fig. 4, the component layer is free to choose its components but they should have the ability to make the labels for the next layer and simultaneously reconstruct the input data in high quality.

Eq. (8) shows the label reconstruction error "*recognition error*" calculated for the extra layer, in which $Y_i(i = A, BandC)$ are the outputs of this layer and $LA, LBandLC$) are the labels.

$$\begin{aligned}
\delta^{LA}_A &= Y_A(1 - Y_A)(LA - Y_A) \\
\delta^{LB}_B &= Y_B(1 - Y_B)(LB - Y_B) \\
\delta^{LC}_C &= Y_C(1 - Y_C)(LC - Y_C)
\end{aligned} \tag{8}$$

This error back-propagates (Eq. (9)) and accumulates with back-propagated output reconstruction error to the component layer and results in the gradient of error Eq. (10). Similar to the previous models, this error is used to update the same weights as in Eq. (4) and it is also back propagated in the underneath layers to finally reach the first layer.

$$\begin{aligned}
\delta^L_{cA} &= \delta^{LA}_A * W'_{cA(c+1)A} \\
\delta^L_{cB} &= \delta^{LB}_B * W'_{cB(c+1)B} \\
\delta^L_{cC} &= \delta^{LC}_C * W'_{cC(c+1)C}
\end{aligned} \tag{9}$$

$$\begin{aligned}
\delta_{cA} &= Y_{cA}(1 - Y_{cA})(\delta^O_{cA} + \delta^L_{cA}) \\
\delta_{cB} &= Y_{cB}(1 - Y_{cB})(\delta^O_{cB} + \delta^L_{cB}) \\
\delta_{cC} &= Y_{cC}(1 - Y_{cC})(\delta^O_{cC} + \delta^L_{cC})
\end{aligned} \tag{10}$$

*2.3.4. Unsupervised and clustered structure of neural network for nonlinear component (manifold) analysis*

The fourth model, which is the last proposed possible model, has the same structure as the third model, but like the second one, the clustering rules are also involved. In this model, the unsupervised learning of the component layer makes the clustering rules more important; despite the existence of the supervised learning, components in this model are not directly assigned to the labels. The error gradient equation and the weight modification equations are the same as those in Eqs. (8) and (7).

These four presented manifold separator models are used in four exactly similar base models and trained on the smaller preprocessed dataset. Results of extensive evaluation and validations on these four models are documented in the first part of the next section. It is concluded that the fourth model has a better performance in terms of component extraction and manifold representation against the other models. So, the fourth unsupervised and clustered multi-component separator model is the winner and will be used in the final novel video analyzer.

### *2.4. Final structure of proposed DDBNN*

One of the novelties of this paper pertains to analyzing videos in order to recognize the identity, emotion and expression intensity. Thus, the presented model should be able to deal with dynamics and sequences. Considering this aim, the ability of the recurrent networks in dynamic modeling and sequential data analysis is used in the presented model. In the CK+, each video shows a person expressing variation in expression intensities from neutral to the apex of a specific emotion. Thus, along with this video, the intensity of emotions is varying through the time when the identity and emotion are persistent. Distinguishing between different emotions from the initial frames of the video is difficult due to their similarities. However, by following the nonlinear variations appearing in the subject's face, it is possible to estimate what emotion the set of variations belong to. Therefore, in the final frames with high expression intensities, the recognition rate is high. Hence, a recurrent link with one step delay is added to the Emotion part (B) of the feature separator (Fig. 5). This link connects the information of variations between two frames to help the encoder in recognizing an emotion based on the previous frames.

Another recurrent network is connected to the component layer (C) of expression intensity to follow video synthesis purposes. The next subsection elaborates further on the proposed network.

### *2.5. Video synthesis*

In this subsection, we are going to find the best decoder model in order to synthesize a video showing dynamics of emotions. These dynamics will change the intensity of expressions on the face using the extracted manifold. The decoder part consists of 4 layers of neurons (3 nonlinear sigmoidal functions and 1 linear output function), which are responsible for decomposing one static picture in the output at a time. The reconstructed static pictures for a test subject are nonlinear combinations of different identities from the training dataset with more similarity to the test subject. These reconstructed pictures show an average of emotion and expression intensities from what the network understands. This synthesis is an imagination and interpolation of the network for an unknown subject's face. By adding a recurrent network connected to the component layer (C) of expression intensity, the model has the ability to synthesize sequential pictures showing variations in intensities from low to high and vice versa. The recurrent network predicts and produces the next intensity components from the present components. Components represent a specific picture with a specific intensity for an emotion, so by changing the intensity through the sequence, the video is synthesized. The number of nonlinear neurons for the recurrent network has been chosen as 20 considering data complexity and discriminability of components (Purple Box in Fig. 5).

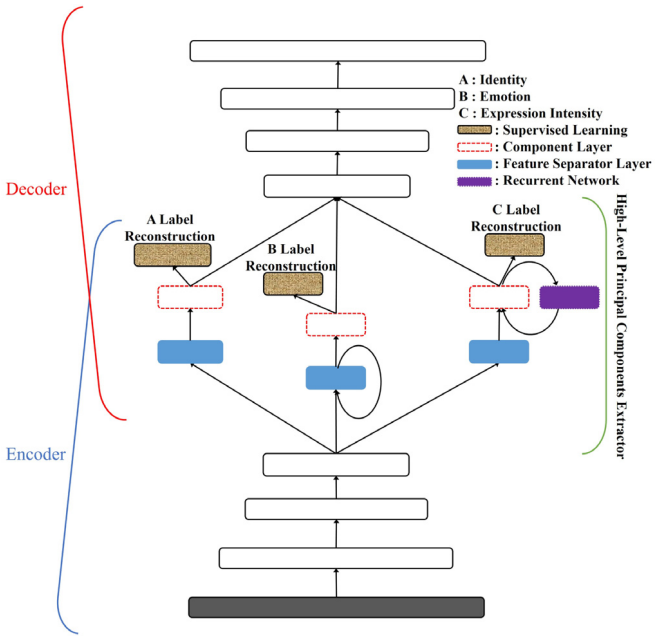


**Fig. 5.** The final structure of proposed DDBNN. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

To sum up, one decoder and four encoder models are introduced. Next, the fourth model presented in Section 2.3.4 outperforms the other three in terms of multi-component manifold extraction and is chosen as the winner model and it is utilized during the rest of the simulations. Finally, the encoder and decoder parts are assembled and the final structure of the proposed DDBNN is constructed and illustrated in Fig. 5.

## 3. Results and discussion

In the present section, the results of our evaluations on the performance of the four presented multi-component extractor models are shown. Based on these evaluations the best extractor model to be placed in the final structure of DDBNN is chosen. After that, the presented DDBNN has been fully studied on the complete dataset.

### *3.1. Choosing the best multi-component (manifold) extractor*

All of the four presented models in the previous section are trained by the smaller dataset showing six emotions in seven expression intensities. Those four models were trained under equal and constant conditions and the same parameters and then all the components in three sub-spaces were extracted after 50,000 iterations of training. Findings of the complexity and time efficiency in identity, emotion and expression intensity sub-manifolds extraction showed that the identity components are the first and easiest sub-manifolds to be extracted. The emotion components are in second and after that, the expression intensity components are the last and most difficult sub-manifold to be trained and extracted. Thus, we designed an evaluation method in three stages to choose the best manifold extractor. As the first stage of evaluation, all the four models were tested in the identity recognition. The quality and discrimination of the recognized identities between all the four models were evaluated. It is obvious that each model that fails to

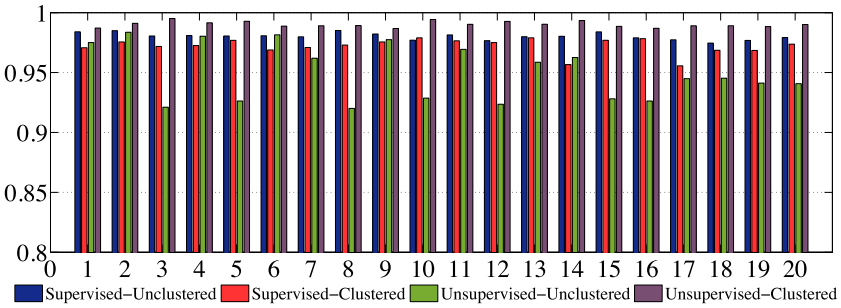


**Fig. 6.** Value of reconstructed labels from proposed four models for 20 subjects. Fully-recognized identity is accomplished when the reconstructed value is one. The unsupervised model with clustering considerations works better than the other models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

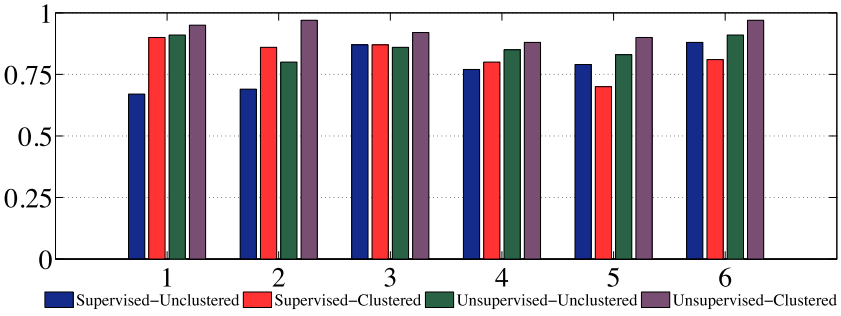


**Fig. 7.** Value of reconstructed labels from proposed four models for 6 emotions (1: Fear, 2: Surprise 3: Sadness, 4: Anger, 5: Disgust, 6: Happiness). Fully-recognized emotion is accomplished when the reconstructed value is one. The unsupervised model with clustering considerations works better than the other models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

fully discriminate the identities (the easiest sub-manifold), obviously cannot work properly in the more difficult sub-manifolds. The same evaluation process in terms of emotion and intensity recognition was conducted, respectively. Similar to identity, the models with low discrimination in emotion components were not acceptable. The final stage was evaluating the recognition of the expression intensity.

All the four models were trained and the mean of outputs of identity label neurons for all the training samples were extracted and illustrated in Fig. 6.

The mean of the outputs for each subject represents the performance of the presented models in recognition and discrimination of different input samples. The highest value of each neuron with logistic function is one, so if the mean of the outputs for one subject is near to one, the model has recognized him/her and if it is near zero, the model has not recognized the subject's identity yet. Fig. 6 shows that all the four models have the potential to analyze and recognize all the 20 different input identities with no mismatch or confusion. So, all the four models have acceptable performance in identity discrimination and recognition.

Fig. 7 shows the mean of reconstructed emotion labels in all models. As this figure shows, the reconstructed labels for the first two models with the supervised learning algorithm do not show acceptable performance. We have investigated the output of these models on every subject of the testing dataset. The results showed that the main failure of these models was in distinguishing between "surprise" and "fear" emotions among some input samples. However, these results are not general and they depend on the training procedure.

The third model follows an unsupervised learning algorithm. Performance of this model in recognition and discrimination of all the emotions, except for surprise, is much better than the supervised models. Nevertheless, some misclassification and unclassified emotions were observed. In contrast with the above-mentioned models, the fourth model with clustering rules fully recognizes and clusters all the emotions correctly.

Fig. 8 demonstrates the mean of recognition for expression intensity. The quality of classification and clustering for the components of the hardest sub-manifold is lower in comparison to the other sub-manifolds for all the models. However, the fourth model has correctly clustered almost all of the intensities.

Considering the results of these three stages, the fourth model, which follows unsupervised with clustering rules in the learning algorithm, performs better than the other models in recognition and clustering the component. The quality of the extracted component in each sub-manifold straightly affects the representation of the nonlinear manifold. This model is chosen as the high-level principal component extractor to construct the final model.

### 3.2. Training and simulations of DDBNN

The final presented model demonstrated in Fig. 5 consists of eight nonlinear layers and one linear output layer. All the layers were pre-trained and then fine-tuned by *Real-time Recurrent Learning* (RTRL) (Ronald & Zipser, 1989) on the bigger dataset of CK+ and the root mean square reconstruction error decreased to 0.015 after over 63,000 iterations.

The reconstructed labels for three sub-manifolds in the label reconstruction layer are shown in Fig. 9a–c Fig. 9a shows the recognized and discriminated labels for 95 identities. In this plot, when the recognition rate of each identity number is near to one, this means that corresponding identity is fully recognized. It shows that all the 95 subjects are fully recognized and no identity remains undecided between zero and one. Taking a glance at the

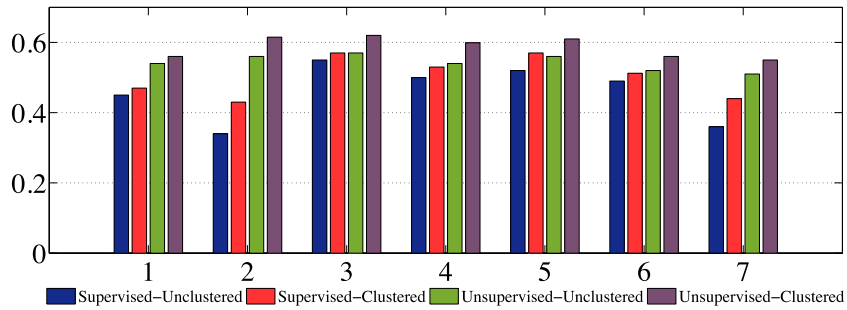


**Fig. 8.** Value of reconstructed labels from proposed four models for 7 intensities. Fully-recognized intensity is accomplished when the reconstructed value is one. The unsupervised model with clustering considerations works better than the other models. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

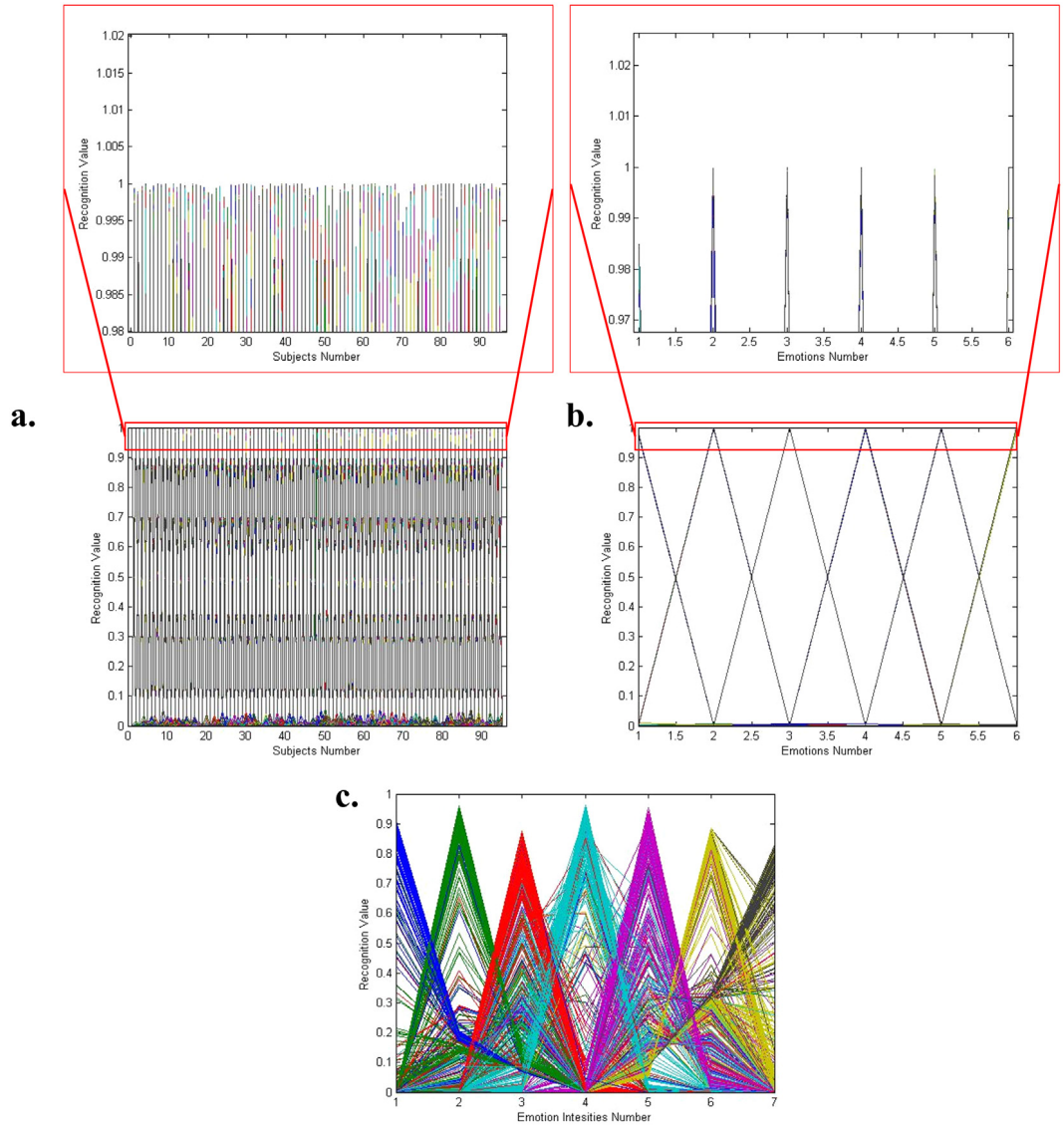


**Fig. 9.** The reconstructed labels for three sub-manifolds. a: reconstructed labels for identity of 95 training subjects. When the reconstructed value is near to one (magnified box), that identity is fully-recognized. b: reconstructed labels for six emotions (1: fear, 2: surprise, 3: sadness, 4: anger, 5: disgust, 6:happiness). When the reconstructed value is near to one (magnified box), that emotion is fully-recognized. c: reconstructed labels for emotions intensities in 7 different levels (1[Blue]: low level (neutral), 2[Green], 3[Red], 4[Light Blue], 5[Purple], 6[Yellow] and 7[Gray]: high level (apex)). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

unsupervised extracted identity component presented in Fig. 10, we realized that the model used all the seven neurons (seven-dimensional component space) to represent and discriminate all the input identities. Our aim of showing this figure is to illustrate the differences of components in the identity sub-manifold.

Fig. 9b shows the recognized and discriminated labels for 6 emotions. This plot shows that all of the six emotions are recognized with a value of one and no input sample remained undecided or misclassified. The unsupervised extracted component for emotions is presented in Fig. 11 and three neurons for representing

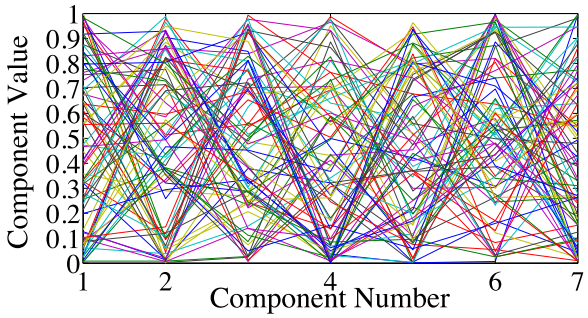


**Fig. 10.** Unsupervised extracted identity component in a seven-dimensional space using seven neurons.

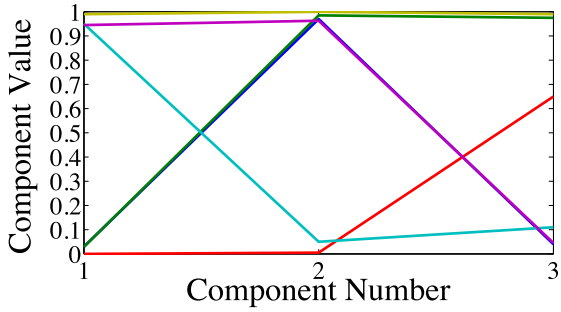


**Fig. 11.** Unsupervised extracted Emotion component in a three-dimensional space using three neurons.

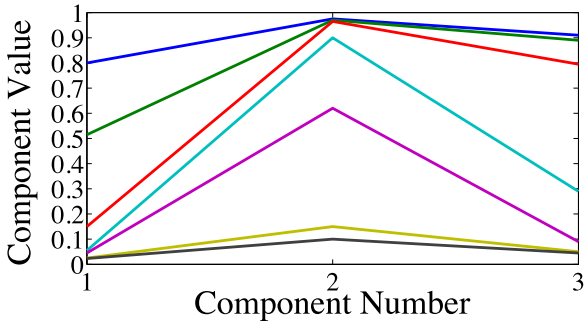


**Fig. 12.** Unsupervised extracted Intensity component in a three dimensional space using three neurons.

six emotions in the component layer are used. Table 2 obtained from the components in Fig. 11 shows that these three neurons have the ability to fully recognize and discriminate all the six emotions.

Moreover, Fig. 9c shows the recognized and discriminated labels for 7 different expression intensities. It shows that almost all of the intensities are recognized, however, some input samples remained undecided or wrongly recognized. Fig. 12 also shows that the unsupervised extracted components, which are represented by three neurons, can recognize and discriminate all the seven different intensities.

In order to analyze the ability of the presented model in video understanding and synthesis, we have focused on these topics separately in the following subsections.

### 3.2.1. Video understanding

The performance of the presented model in emotion understanding has been evaluated in this subsection. The confusion matrix in emotion recognition is calculated by applying a mathematical mean function to the results of five-fold cross-validation and is shown in Table 3. Each row of this table represents the emotion of input sequences and each column represents the results of recognition. This table shows that the accuracy of recognition for sadness is 100%. However, for other emotions such as anger and disgust, the recognition rate is trivially less than 100% because of similarity. The confusion matrix of intensity recognition is also extracted (840 frames) and presented in Table 4. Each row of this table represents the labels of seven input intensities and each column represents the results of intensity recognition. The presented model achieves 97.77% and 78.17% accuracy in recognition of emotion and intensity, respectively. The reported results are simultaneously obtained from the model. Our experimental results are also compared with some prominent emotion recognition approaches on the CK+ and presented in Table 5. This table shows the comparison of the referenced methods in different benchmarks partitioned in 6 columns. Nevertheless, a fair comparison among the approaches is very difficult due to the lack of a common benchmark and a unified testing protocol. Hence, we are satisfied with the claim that the recognition rate of the proposed DDBNN model is competitive with almost all of the state-of-the-art.

**Table 2**
Extracted binary codes from components for each emotion.

| Emotion | Extracted code |
|---|---|
| Fear | [0 1 0] |
| Surprise | [0 1 1] |
| Sadness | [0 0 1] |
| Anger | [1 0 0] |
| Disgust | [1 1 0] |
| Happiness | [1 1 1] |

### 3.2.2. Video synthesizing

The presence of the recurrent network connected to the component layer of intensity equips the model with the ability to synthesize videos showing the variation of different intensities from low to high and vice versa. When a single image of an unfamiliar identity is fed to the input of the network, the model enables us to manipulate all the components of the extracted manifold. For example, the extracted identity components can be preserved while the desired emotion is applied to an image via emotion codes in Table 2. When the components are determined, the model reconstructs the first frame of the video in the output. Afterwards, the recurrent network changes the intensity and reconstructs the following frames of the video to reach the extreme intensity. The model is also able to learn to synthesize the video in a reverse direction from high intensity to neutral. The following flowchart (Fig. 13) shows the algorithm of video synthesizing for a single input image.

In Fig. 14, the synthesized frames of the presented DDBNN model in three different emotions for a test subject are shown. In order to evaluate the quality of the produced sequences and their similarity to the original test, we applied the *Structural Similarity Index Method* (SSIM). This index shows a specific number between −1 to 1 for each of the two-paired images. −1 represents the minimum similarity while 1 shows a duplicate image (Dosselmann & Yang, 2011; Wang, Bovik, Sheikh, & Simoncelli, 2004). The results are shown in Table 6 with 0.8 similarities in average between the produced and original images. The worst result corresponds to a pair of images with 0.1 SSIM.

Finally, to evaluate the applicability and generality of the proposed model, we tested static pictures of three different people in order to synthesize videos of expressions related to these samples, which are illustrated by Fig. 15. The pictures are taken by an *Apple iPhone 6* mobile phone and subjects were asked to show

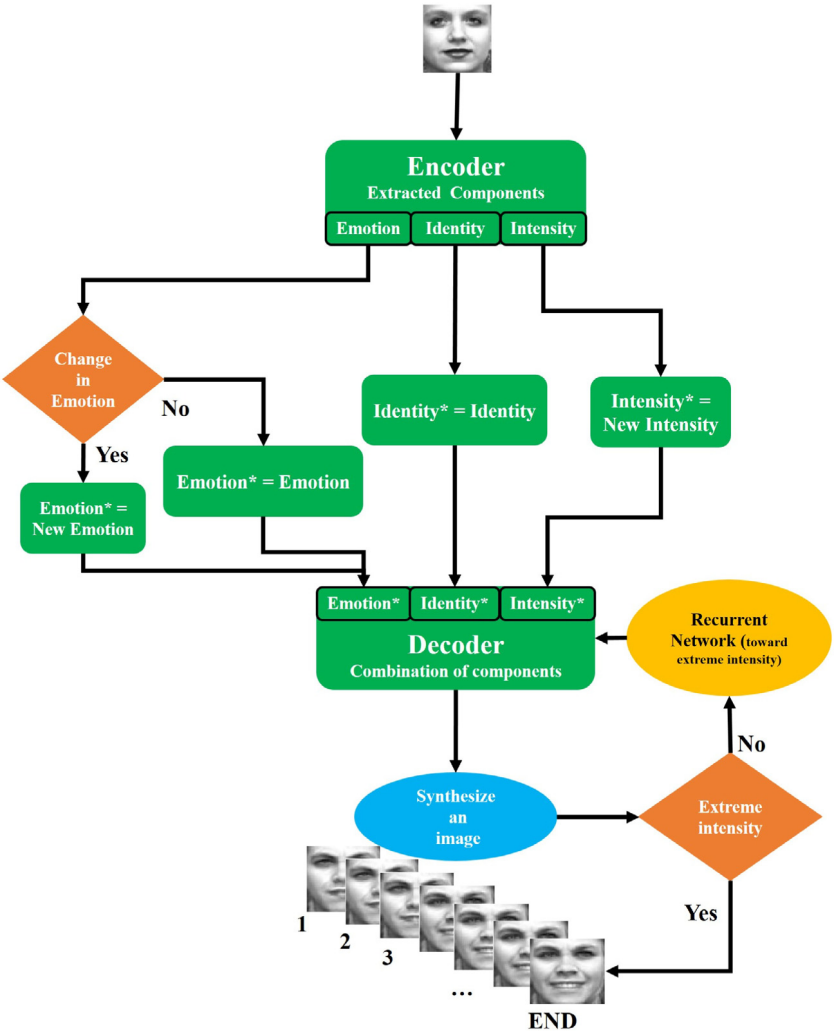


**Fig. 13.** The algorithm of video synthesizing for a single input image.

**Table 3**
Confusion matrix of CK+for the six emotions.

| | Fear | Surprise | Sadness | Anger | Disgust | Happiness | Unrecognized |
|---|---|---|---|---|---|---|---|
| Fear | **0.98** | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| Surprise | 0.00 | **0.94** | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 |
| Sadness | 0.00 | 0.00 | **1.00** | 0.00 | 0.00 | 0.00 | 0.00 |
| Anger | 0.00 | 0.00 | 0.00 | **0.98** | 0.02 | 0.00 | 0.00 |
| Disgust | 0.00 | 0.00 | 0.00 | 0.00 | **0.98** | 0.00 | 0.02 |
| Happiness | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | **0.98** | 0.00 |

**Table 4**
Confusion matrix of emotion intensity for CK+database. Each column shows the mean of recognized Intensity Level out of five-fold cross validation.

| Frame number | 1st level | 2nd level | 3rd level | 4th level | 5th level | 6th level | 7th level | Unrecognized |
|---|---|---|---|---|---|---|---|---|
| 1st | **0.79** | 0.10 | 0.10 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2nd | 0.10 | **0.79** | 0.09 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3rd | 0.05 | 0.01 | **0.79** | 0.07 | 0.00 | 0.00 | 0.02 | 0.06 |
| 4th | 0.02 | 0.00 | 0.07 | **0.76** | 0.07 | 0.02 | 0.05 | 0.01 |
| 5th | 0.00 | 0.02 | 0.03 | 0.07 | **0.74** | 0.07 | 0.05 | 0.02 |
| 6th | 0.00 | 0.00 | 0.04 | 0.05 | 0.07 | **0.71** | 0.07 | 0.06 |
| 7th | 0.00 | 0.00 | 0.02 | 0.00 | 0.03 | 0.05 | **0.89** | 0.01 |

an emotion on their face. After converting the obtained images to gray-scale format and cropping their no-face areas, the pictures were fed to the networks as input. Fig. 15 shows that the model makes an attempt to preserve the face characteristics of the subject (such as beard) and synthesize variations based on the similarity of those subjects to its learning dataset.

## 4. Conclusion

DBNNs show great ability in simultaneous multi-subspace manifold extraction. Based on this, we propose a novel Dynamic DBNN (DDBNN), which considers the sequential information underlying the video and extracts the intensity of emotion as well as

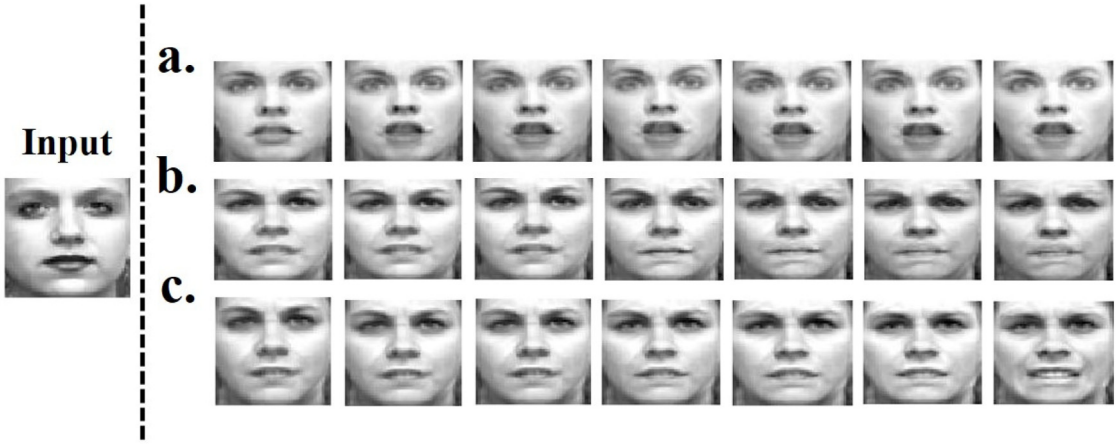


**Fig. 14.** Synthesized frames of the presented DDBNN model for a test (Input) subject in, a: Surprise, b: Anger and c: Fear.

**Fig. 15.** Synthesized frames of presented DDBNN model on the provided pictures of three participants. The intensity constituently increases from right to left.

**Table 5**
Comparison of proposed method with other state-of-the-art (Non-CNN) approaches on the CK+database.

| Dynamic | Emotion classes | Subjects | Measure | Method | Accuracy (%) | Reference |
|---|---|---|---|---|---|---|
| No | 7 | 96 | %ERR | NMSNN | 92.86 | Seyyedsalehi and Seyyedsalehi (2014) |
| No | 7 | 94 | Ten-fold | Radial encoding | 91.51 | Gu et al. (2012) |
| No | 7 | 118 | Ten-fold | Trajectory representation | 96.87 | Kacem et al. (2017) |
| Yes | 6 | 97 | – | HMMs | 93.66 | Aleksic and Katsaggelos (2006) |
| Yes | 6 | 90 | LOSO | PHOG | 96.33 | Li et al. (2010) |
| Yes | 6 | 97 | Ten-fold | D-LBP | 96.26 | Zhao and Pietikainen (2007) |
| Yes | 6 | 95 | Five-fold | DDBNN | 97.77 | This work |

sub-manifolds of identity and emotion. The ability of different approaches in manifold extractor models in terms of supervised/unsupervised and clustering/non-clustering methods is studied. The results of preliminary experiments showed that the unsupervised-clustered model is the best extractor model on the CK+ dataset.

The final model has a total of nine layers with one recurrent link connected to the emotion feature extractor layer and a recurrent network connected to the intensity component layer. The layer by layer pre-training and Real-time recurrent learning (RTRL) (Ronald & Zipser, 1989) joint training were used to train the proposed DDBNN network. After the training, the extracted principal component and the reconstructed labels were analyzed and the results demonstrated distinct and unique components. Results highlight that the presented model has remarkable 97.77% and 78.17% accuracy rates in emotion and intensity recognition, respectively.

**Table 6**
Calculated SSIM for synthesized images.

| | SSIM value |
|---|---|
| Fear | 0.81 |
| Surprise | 0.82 |
| Sadness | 0.79 |
| Anger | 0.75 |
| Disgust | 0.76 |
| Happiness | 0.86 |

Experimental results are also compared with some prominent emotion recognition approaches in Table 5. As the results suggest, the recognition rate of the proposed model is competitive with almost all state-of-the-art approaches.

The great ability of the model in video synthesizing creates wonderful videos based on the networks' imagination which has 0.8 similarities with the original videos on SSIM scale.

## Acknowledgments

The authors would like to thank Mr. Nima Amini who reviewed the manuscript and also Mr. Ali Niaty, Mr. Soroush Shahidi and Mr. Arash Foroudi for participation and their permission to publish their pictures in this study.

## References

Abdolali, Fatemeh, & Seyyedsalehi, SeyyedAli (2011). Improving pose manifold and virtual images using bidirectional neural networks in face recognition using single image per person. In *International symposium on artificial intelligence and signal processing, AISP, 2011*. IEEE.

Agrawal, Neeraj, Cosgriff, Rob, & Mudur, Ritvik (2009). *Mood detection: Implementing a facial expression recognition system*. CS229 project.

Aleksic, Petar S., & Katsaggelos, Aggelos K. (2006). Automatic facial expression recognition using facial animation parameters and multistream HMMs. *IEEE Transactions on Information Forensics and Security*, *1*(1), 3–11.

Calvo, Manuel G., & Marrero, Hipólito. (2009). Visual search of emotional faces: The role of affective content and featural distinctiveness. *Cognition and Emotion*, *23*(4), 782–806.

Chang, Ya, Hu, Changbo, Feris, Rogerio, & Turk, Matthew (2006). Manifold based analysis of facial expression. *Image and Vision Computing*, *24*(6), 605–614.

Chao, Linlin, Tao, Jianhua, Yang, Minghao, Li, Ya, & Wen, Zhengqi (2015). Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th international workshop on audio/visual emotion challenge* (pp. 65–72). ACM.

Chen, Junkai, Chen, Zenghai, Chi, Zheru, & Fu, Hong (2016). Facial expression recognition in video with multiple feature fusion. *IEEE Transations on Affective Computing*.

Ding, Hui, Zhou, Shaohua Kevin, & Chellappa, Rama (2017). Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In *12th IEEE international conference on automatic face & gesture recognition, FG 2017, 2017* (pp. 118–126). IEEE.

Dosselmann, Richard, & Yang, Xue Dong (2011). A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, *5*(1), 81–91.

Gu, Wenfei, Xiang, Cheng, Venkatesh, Y. V., Huang, Dong, & Lin, Hai (2012). Facial expression recognition using radial encoding of local Gabor features and classifier synthesis. *Pattern Recognition*, *45*(1), 80–91.

Hamedani, Kian, Seyyedsalehi, Seyyed Ali, & Ahamdi, Reza (2016). Video-based face recognition and image synthesis from rotating head frames using nonlinear manifold learning by neural networks. *Neural Computing and Applications*, *27*(6), 1761–1769.

Hinton, Geoffrey E., & Salakhutdinov, Ruslan R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Huang, Yan, Wang, Wei, Wang, Liang, & Tan, Tieniu (2014). A general nonlinear embedding framework based on deep neural network. In *Pattern recognition, ICPR, 2014 22nd international conference on* (pp. 732–737). IEEE.

Jain, Anil K., & Li, Stan Z. (2011). *Handbook of face recognition*. New York: springer.

Jung, H., Lee, S., Park, S., Lee, I., Ahn, C., & Kim, J. (2015). Deep temporal appearance-geometry network for facial expression recognition. In *ICCV*.

Kacem, Anis, Daoudi, Mohamed, Amor, Boulbaba Ben, & Alvarez-Paiva, Juan Carlos (2017). A novel space-time representation on the positive semidefinite cone for facial expression recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3180–3189).

Kanade, Takeo, Cohn, Jeffrey F., & Tian, Yingli (2000). Comprehensive database for facial expression analysis. In *Fourth IEEE international conference on automatic face and gesture recognition, 2000 proceedings*. IEEE.

Khandait, S. P., Thool, Ravindra C., & Khandait, P. D. (2012). Automatic facial feature extraction and expression recognition based on neural network. arXiv preprint arXiv:1204.2073.

Lee, Ka Keung, & Xu, Yangsheng. (2003). Real-time estimation of facial expression intensity. In *IEEE international conference on robotics and automation, 2003 proceedings. ICRA'03. Vol. 2*. IEEE.

Li, Zisheng, Imai, Jun-ichi, & Kaneko, Masahide. (2010). Facial expression recognition using facial-component-based bag of words and PHOG descriptors. *The Journal of the Institute of Image Information and Television Engineers*, *64*(2), 230–236.

Lien, JJ-J., Kanade, Takeo, Cohn, Jeffrey F., & Li, Ching-Chung (1998). Subtly different facial expression recognition and expression intensity estimation. In *IEEE computer society conference on computer vision and pattern recognition, 1998 proceedings. 1998* (pp. 853–859). IEEE.

Liu, Yunfan, Hou, Xueshi, Chen, Jiansheng, Yang, Chang, Su, Guangda, & Dou, Weibei (2014). Facial expression recognition and generation using sparse autoencoder. In *International conference on smart computing, SMARTCOMP, 2014* (pp. 125–130). IEEE.

Lopes, André Teixeira, Aguiar, Edilson de, De Souza, Alberto F., & Oliveira-Santos, Thiago (2017). Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order. *Pattern Recognition*, *61*, 610–628.

Lucey, Patrick, Cohn, Jeffrey F., Kanade, Takeo, Saragih, Jason, Ambadar, Zara, & Matthews, Iain (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *IEEE computer society conference on computer vision and pattern recognition workshops, CVPRW, 2010* (pp. 94–101). IEEE.

Mase, Kenji (1991). Recognition of facial expression from optical flow. *IEICE Transactions (E)*, *74*, 3474–3483.

Moghadam, Saeed Montazeri, & Seyyedsalehi, Seyyed Ali (2017). Nonlinear analysis of video images using deep recurrent auto-associative neural networks for facial understanding. In *3rd international conference on pattern recognition and image analysis, IPRIA, 2017* (pp. 20–25). IEEE.

Mohammadian, Amin, Aghaeinia, Hassan, & Towhidkhah, Farzad (2016). Diverse videos synthesis using manifold-based parametric motion model for facial understanding. *IET Image Processing*, *10*(4), 253–260.

Mollahosseini, A., Chan, D., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. In *2016 IEEE winter conference on applications of computer vision, WACV* (pp. 1–10). IEEE.

Nejadgholi, Isar, & Seyyedsalehi, Seyyed Ali (2012). A new brain-inspired robust face recognition through elimination of variation features. *Procedia-Social and Behavioral Sciences*, *32*, 204–212.

Rahimi, Ali, Darrell, T., & Recht, B. (2005). Learning appearance manifolds from video. In *IEEE computer society conference on computer vision and pattern recognition, 2005 CVPR 2005. Vol. 1*. IEEE.

Ranjan, Rajeev., Patel, Vishal M., & Chellappa, Rama (2017). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Rao, Adithya, & Thiagarajan, Narendran (2010). Recognizing facial expressions from videos using Deep Belief Networks. In *Stanford CS 229 machine learning final projects*, Technical report.

Schmidhuber, Jürgen (2015). Deep learning in neural networks: An overvie. *Neural networks*, *61*, 85–117.

Seyyedsalehi, Seyyede Zohreh, & Seyyedsalehi, Seyyed Ali (2014). Simultaneous learning of nonlinear manifolds based on the bottleneck neural network. *Neural Processing Letters*, *40*(2), 191–209.

Seyyedsalehi, Seyyede Zohreh, & Seyyedsalehi, Seyyed Ali (2015). A fast and efficient pre-training method based on layer-by-layer maximum discrimination for deep neural network. *Neurocomputing*, *168*, 669–680.

Suslow, Thomas, Junghanns, Klaus, & Arolt, Volker (2001). Detection of facial expressions of emotions in depression. *Perceptual and Motor Skills*, *92*(3), 857–868.

Ullman, Shimon (1996). *High-level vision: Object recognition and visual cognition. Vol. 2*. Cambridge, MA: MIT press.

Wan, Chuan, Tian, Yantao, & Liu, Shuaishi (2012). Facial expression recognition in video sequences. In *10th world congress on intelligent control and automation, WCICA, 2012*. IEEE.

Wang, Zhou, Bovik, Alan C., Sheikh, Hamid R., & Simoncelli, Eero P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612.

Wang, Feng, Xiang, Xiang, Liu, Chang., Tran, Trac D, Reiter, Austin, & Hager, Gregory D, et al. (2017). Regularizing face verification nets for pain intensity regression. In *Proceedings of the IEEE conference on image processing*.

Williams, Ronald J., & Zipser, David (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, *1*(2), 270–280.

Yan, Xing, Chang, Hong, Shan, Shiguang, & Chen, Xilin (2014). Modeling video dynamics with deep dynencoder. In *European conference on computer vision* (pp. 215–230). Cham: Springer.

Yin, Lijun, Wei, Xiaozhou, Sun, Yi, Wang, Jun, & Rosato, Matthew J. (2006). A 3D facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition, 2006 FGR 2006* (pp. 211–216). IEEE.

Zhang, Bailing, & Zhou, Juntao (2014). Video-based face recognition by Auto-Associative Elman Neural network. In *7th international congress on image and signal processing, CISP, 2014*. IEEE.

Zhao, X., Liang, X., Liu, L., Li, T., Vasconcelos, N., & Yan, S. (2016). Peakpiloted deep network for facial expression recognition, arXiv preprint arXiv:1607.06997.

Zhao, Guoying, & Pietikainen, Matti. (2007). Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 915–928.