A Deep Learning Approach for Subject Independent Emotion Recognition from Facial Expressions

VICTOR-EMIL NEAGOE[*], ANDREI-PETRU BĂRAR[*], NICU SEBE[**], PAUL ROBITU[*]
[*]Faculty of Electronics, Telecommunications & Information Technology
Polytechnic University of Bucharest
Splaiul Independentei No. 313, Sector 6, Bucharest,
ROMANIA
victoremil@gmail.com, andreibarar@gmail.com, robitupaul@gmail.com
[**]Department of Information Engineering and Computer Science
University of Trento
ITALY
sebe@disi.unitn.it

Abstract: - This paper proposes Deep Learning (DL) models for emotion recognition from facial expressions. We have focused on two "deep" neural models: Convolutional Neural Networks (CNN) and Deep Belief Networks (DBN). For each of these DL neural models, we have chosen several architectures. We have considered both the case of subject independent emotion recognition and also that of subject dependent emotion recognition. One has selected the Support Vector Machine (SVM) as a benchmark algorithm. We have chosen the JAFFE database to evaluate the above proposed models for person independent/dependent facial expression recognition. Using DL approach, we have obtained a subject-independent emotion recognition score of 65.22%, corresponding to an increase of 6% over the best score given by the considered benchmark methods. For person dependent emotion recognition, the DL model leads to the recognition score of 95.71%, representing an increase of 3% over the best of chosen benchmark methods.

Key-Words: - Facial expression, Deep Learning (DL), Convolutional Neural Networks (CNN), Deep Belief Networks (DBN), subject independent/dependent emotion recognition

1 Introduction

In recent years there has been a growing interest in improving the aspects of interaction between humans and computers. *Human-computer intelligent interaction (HCII)* implies a natural interaction with the user, similar with human-human interaction [2], [3], [16].

Humans interact mainly through speech, but also use body gestures to emphasize the speech and display emotions.

One of the important ways to display emotions is through facial expressions; it contains important information regarding the user current emotion.

Chibelushi has stated that facial expressions have a considerable impact on a listening interlocutor [2]; he established that the facial expression of a speaker are responsible for about 55 percent of the impact, 38 percent by voice intonation and 7 percent by the actual spoken words [2], [16].

Ekman and his colleagues in their studies of human facial expressions identified six states as "universal facial expressions" representing *happiness, sadness, anger, fear, surprise* and *disgust* [7].

The human visual system efficiently recognizes those emotions within difficult scenes (variations in illumination or perspective). Computer algorithms designed to automatically recognize different emotions in facial expressions (for HCII) have to solve a difficult task [1], [8], [19].

Making a computer able to detect subtle changes in user's affective behaviour and to initiate interactions based on this is the required step for future human-centred designs and interfaces [21].

Even though there are a lot of facial expression recognition approaches [2], [3], [6], [14], [15], [17], [21], there are only a few researches dealing with *subject independent* facial expression recognition [4], [16], [20] proving that recognizing facial expression over different persons is still a difficult task.

With the recent rise of Deep Learning (DL) neural models, due to their increase in computational power availability and affordability, there becomes of interest and value to approach emotion recognition from the new perspective of DL [11].

Deep neural models mimic the nature of visual cortex and involve learning a hierarchy of internal representations. DL techniques have been shown to perform better than other shallow techniques for image classification tasks [10], [11], [12], [18].

Within this paper we extend our previously study on DL neural models used for image classification for the challenging task of subject independent facial expression recognition. The paper is structured as follows.

Second section presents the proposed neural models with several configurations, first using Convolutional Neural Networks (CNN) and the second one using Deep Belief Networks (DBN).

Third section contains the experimental results of the proposed DL neural models for the Japanese Female Facial Expression (JAFFE) dataset [13]; we have considered the CNN and DBN variants of DL. For the comparative evaluation of the DL models, the following benchmark algorithms have been selected: Nearest Neighbor (NN), Support Vector Machine (SVM) with RBF kernel and SVM with linear kernel.

In the fourth section we present the concluding remarks.

2 Model description

In this paper we propose two *Deep Learning* architectures, for subject dependent / independent facial expressions emotion recognition on JAFFE database in the following variants:

- **CNN** – Convolutional Neural Network;
- **DBN** – Deep Belief Network.

2.1 Convolutional Neural Networks (CNN)

CNNs are variants of MLPs inspired from biology [9], [10]; they represent hierarchical neural networks with alternating convolutional and max-pooling/subsampling layers (see Fig.1).

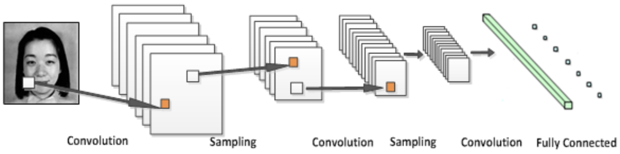


Fig.1 CNN structure: input, convolutional, sub-sampling / max-pooling and classification layers

CNNs use spatially local correlation with a local connectivity pattern between neurons of adjacent layers. The hidden units (inputs) in the *m*-th layer are connected to a local subset of units in the *(m-1)*-th layer (see Fig.2).

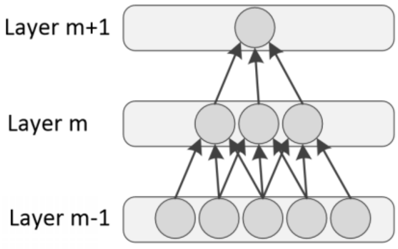


Fig.2 CNN structure: connectivity pattern between layers

2.1.1 Convolutional layer

A convolutional layer is described by the number of feature maps and kernel sizes. Each layer has *M* maps with the same size (M_x, M_y). A kernel with (K_x, K_y) size will be shifted over the input map.

Output map size:

$$M_x^k = \frac{M_x^{k-1} - K_x^k}{S_x^k + 1} + 1 \,, \qquad (1)$$

$$M_y^k = \frac{M_y^{k-1} - K_y^k}{S_y^k + 1} + 1 \qquad (2)$$

where index *k* indicates the layer and S_x, S_y are the skipping factors – number of pixels skipped by the kernel between convolutions.

A feature map is obtained by convolution between the input map and a linear filter, adding a bias term and applying a non-linear function.

For the *k*-th feature map M^k using the weights W^k and bias b_k and applying the *tanh* function, we have

$$M_{i,j}^k = \tanh\left(\left(W^k * x\right)_{ij} + b_k\right) \qquad (3)$$

The definition of convolution for one dimensional signal leads to

$$o[n] = f[n] * g[n] = \\ \sum_{u=-\infty}^{\infty} f[u]g[u-n] = \sum_{u=-\infty}^{\infty} f[n-u]g[u] \quad (4)$$

CNN's use two dimensional signal convolution (an extension of equation (4)) as follows

$$o[m,n] = f[m,n] * g[m,n] = \\ \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} f[u,v]g[u-m,v-n] \qquad (5)$$

Neurons from a specified map share their weights; this increases the efficiency and reduces the number of parameters to be learned. Using this approach, CNNs achieve better generalization.

2.1.2 Max-pooling layer

Max-pooling is preferred over sub-sampling. This improves convergence speed (by reducing computational effort) and it also increases generalization due to position invariance over larger regions.

The output of this layer is given by the maximum activation over the kernel regions, providing a down sample effect of the input maps by a factor equal with the kernel size along each axis.

2.1.3 Classification layer

A fully connected layer combines the outputs of the last convolutional layer into a 1D feature vector with one output neuron per class label.

2.2 Deep Belief Network (DBN)

DBNs are graphical models with the ability to learn a hierarchical representation of the training data [11], [12], [18].

DBNs model the joint distribution between observed vector x and the ℓ hidden layers h^k as follows:

$$P(x, h^1, \ldots, h^\ell) = \left(\prod_{k=0}^{\ell-2} P(h^k|h^{k+1}) \right) P(h^{\ell-1}, h^\ell) \quad (6)$$

where $x = h^0$, $P(h^{k-1}|h^k)$ is a conditional distribution for the visible units conditioned on the hidden units of the Restricted Boltzmann Machine (RBM) at level k, and $P(h^{\ell-1}, h^\ell)$ is the visible-hidden joint distribution in the top-level RBM (See Fig.3).

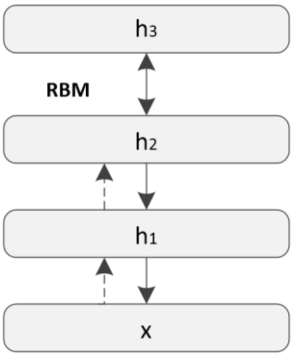


Fig.3 DBN structure

DBNs with RBMs layers use greedy layer-wise unsupervised training [9]. The training process is as follows:

1. Initially train the first layer considering the input $x = h^{(0)}$ as the visible layer of the first RBM in the stack;

2. The trained first layer is used to obtain the input data for the second layer using the mean activations *p* $(h^{(1)} = 1 \mid h^{(0)})$;

3. Train the second layer also as an RBM, using the data obtained in step 2 as training data;

4. Iterate through steps 2 and 3 for the chosen number of layers, each time propagating upward the mean values;

5. Fine tune the parameter via supervised gradient descent.

We use the logistic regression classifier to classify the input *x* based on $h^{(l)}$ (last hidden layer of DBN model).

This is equivalent to an initialization of a Multilayer Perceptron (MLP) network using the weights and hidden layer biases obtained during the unsupervised training.

3 Experimental Results

3.1 JAFFE Dataset

We have chosen *Japanese Female Facial Expression (JAFFE)* database to evaluate the proposed neural models. JAFFE database was planned and assembled by Michael Lyons, Miyuki Kamachi and Jiro Gyoba with photos taken at the Psychology Department, Kyushu University [13].

JAFFE dataset contains 213 images of 7 facial expressions - six basic facial expressions: happiness, sadness, surprise, anger, disgust, fear and neutral face and one neutral (see Fig. 4) - posed by 10 Japanese female models. Each image has been rated on 7 emotion adjectives (including the neutral one) by 60 Japanese subjects.

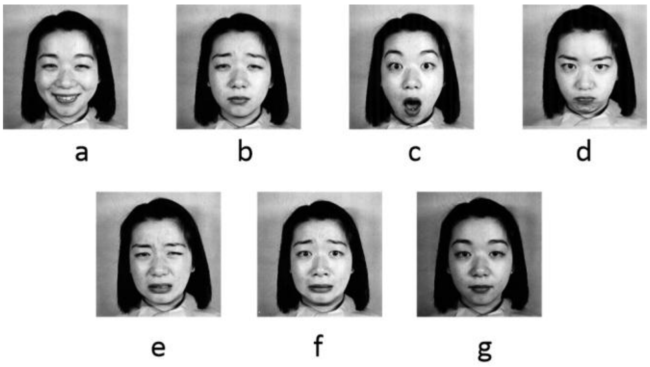


Fig. 4. Emotion example for a subject from *JAFFE* dataset: (a) happiness; (b) sadness; (c) surprise; (d) anger (e) disgust (f) fear (g) neutral

3.2 Preprocessing

We have used 210 images from the 213 initial JAFFE dataset [13], [14].

•In order to remove background influences, we have cropped each image from its original size of 256 x 256 pixels into an 128 x 128 pixel segment.

•To compensate the various illumination conditions, we have applied a histogram equalization.

3.3 Test Procedure

The preprocessed images of 128 x 128 pixels of the JAFFE database are represented as 16384-dimensional vectors, and they are used to build several datasets.

We have considered J = 10 subjects, M = 7 emotions and k = 3 images/subject for each emotion class.

3.3.1 Subject independent emotion recognition

Subject independent emotion recognition implies building datasets for each of the J=10 subjects (for JAFFE database).

We have 10 datasets DS$_i$, where i = (1,…,J); each one is split into a learning dataset LDS$_i$ and a testing dataset TDS$_i$:

- LDS$_i$ – the learning dataset for the subject “i”, is created using (J-1) pictures of the other persons from the dataset, and has a size of k x (J-1) x M pictures; k is the number of images for each of the M facial expressions (k=3 for JAFFE database);

- TDS$_i$ – the test dataset for the subject “i”, consists of k x M pictures of the person “i”.

3.3.2 Subject dependent emotion recognition

Subject dependent emotion recognition uses a single dataset containing all J = 10 subjects (for JAFFE database).

This dataset consists of two data subsets:

- LDS – the learning dataset built by using images of all J persons and has a size of n x J x M; n is the number of images for each emotion of each person used for learning (n<k; we have chosen n=2);

- TDS – the test dataset consisting of (k-n) x J x M pictures of all J persons.

3.4 DL Model Specifications

3.4.1 CNN model

Network parameters:

- Batch size = 20;
- Learning rate = 0.1

CNN layer types:

- **I** – input layer;
- **C** – convolutional layer;
- **MP** – max pooling layer;
- **FC** – fully connected layer.

The specifications of the experimented CNN1 and CNN2 models are shown in Tables 1 and 2.

Table 1. *CNN1*: version 1 of *CNN* proposed architecture

Layer	Type	Out Maps [neurons]	Kernel Size	Pooling Size
0	I	1 map 128x128	-	-
1	C	128 maps 100x100	29x29	-
2	MP	128 maps 25x25		4x4
3	C	256 maps 10x10	16x16	-
4	MP	256 maps 5x5	-	2x2
5	FC	25 neurons	-	-
6	FC	8 neurons	-	-

Table 2. *CNN2*: version 2 of *CNN* proposed architecture

Layer	Type	Out Maps [neurons]	Kernel Size	Pooling Size
0	I	1 map 128x128	-	-
1	C	128 maps 120x120	9x9	-
2	MP	128 maps 60x60	-	2x2
3	C	256 maps 50x50	11x11	-
4	MP	256 maps 25x25	-	2x2
5	C	512 maps 20x20	6x6	-
6	MP	512 maps 10x10	-	2x2
9	FC	64 neurons	-	-
10	FC	8 neurons	-	-

3.4.2 DBN model

Network parameters:

- Pre-training epochs = 100;
- Training epochs = 200;
- Batch size = 20;
- Learning rate = 0.1;
- Fine tuning learning rate = 0.01

DBN layer types:

- **I** – input layer;
- **RBM** – Restricted Boltzmann Machine layer;
- **FC** – fully connected layer.

The specifications of the experimented DBN model are given in Table 3.

Table 3. *DBN* proposed architecture

Layer	Type	Out Maps [neurons]
0	I	1 map 128x128
1	RBM	128
2	RBM	256
6	FC	8

3.5 Performance Evaluation

• Subject independent emotion recognition

$\mathbf{RS_i}$ – the emotion recognition score for the subject "i" for the corresponding dataset TDS_i;

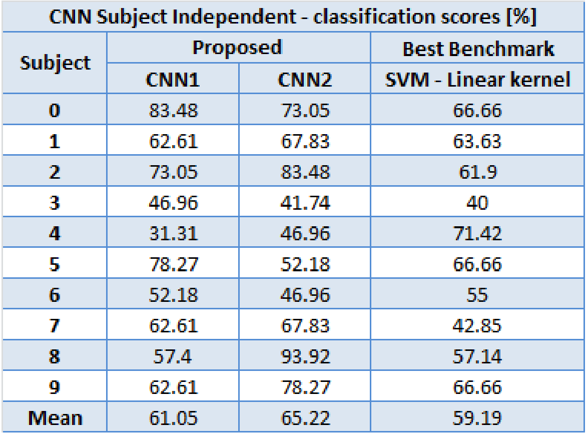
Mean RS – average recognition score on all M emotions.

• Subject dependent emotion recognition

RS – average recognition score on all M emotions (M=7) using the TDS (test dataset).

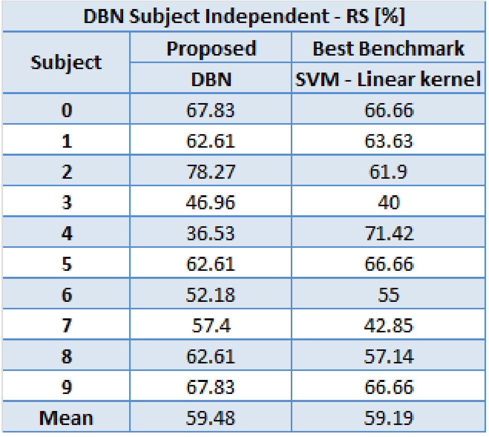
The experimental performances of the DL models versus those of the best benchmark methods are given in Tables 4, 5 and 6.

Table 4. Recognition score (RS) for *CNN* subject independent emotion recognition



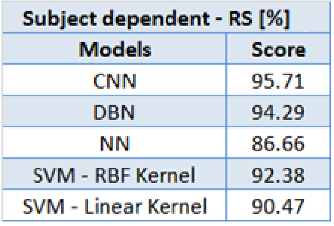
CNN Subject Independent - classification scores [%]			
Subject	**Proposed**		**Best Benchmark**
	CNN1	**CNN2**	**SVM - Linear kernel**
0	83.48	73.05	66.66
1	62.61	67.83	63.63
2	73.05	83.48	61.9
3	46.96	41.74	40
4	31.31	46.96	71.42
5	78.27	52.18	66.66
6	52.18	46.96	55
7	62.61	67.83	42.85
8	57.4	93.92	57.14
9	62.61	78.27	66.66
Mean	61.05	65.22	59.19

Table 5. Recognition score (RS) for *DBN* subject independent emotion identification



DBN Subject Independent - RS [%]		
Subject	**Proposed**	**Best Benchmark**
	DBN	**SVM - Linear kernel**
0	67.83	66.66
1	62.61	63.63
2	78.27	61.9
3	46.96	40
4	36.53	71.42
5	62.61	66.66
6	52.18	55
7	57.4	42.85
8	62.61	57.14
9	67.83	66.66
Mean	59.48	59.19

Table 6. Recognition score (RS) for subject dependent emotion identification



Subject dependent - RS [%]	
Models	**Score**
CNN	95.71
DBN	94.29
NN	86.66
SVM - RBF Kernel	92.38
SVM - Linear Kernel	90.47

4 Concluding Remarks

1) In this paper, we apply Deep Learning architectures – Convolutional Neural Network and Deep Belief Network – for subject independent / dependent emotion recognition from facial expressions. These architectures are evaluated on JAFFE database (210 images, 10 subjects, 7 emotions – *happiness*, *sadness*, *surprise*, *anger*, *disgust*, *fear* and *neutral*).

2) We have proposed two configuration versions for CNN (CNN1, CNN2 according to Tables 1-2) and one for DBN (Table 3) and we have compared the above DL models with several benchmark algorithms (NN, SVM with RBF kernel and SVM with linear kernel).

3) Table 4 shows the subject-independent emotion recognition scores using CNN model. The best average recognition score of 65.22% corresponds to CNN2 architecture variant. The above CNN performance represents an increase with 6% over the best considered benchmark score (SVM with linear kernel).

4) In Table 5 we present the subject-independent emotion recognition scores using DBN. The average DBN recognition score of 59.48% is better than the benchmark score (SVM with linear kernel).

5) From Table 6 we can see that the best subject dependent emotion recognition score of 95.71% is obtained using CNN2 architecture; the above performance is about 3% higher than the best benchmark result (SVM with RBF kernel).

References:

[1] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.

[2] C. Chibelushi, F. Bourel, Facial Expression Recognition: A Brief Tutorial Overview, in *CVonline: On-Line Compendium of Computer Vision*, editor Robert Fisher, January 2003.

[3] Cohen, A.Garg, T.S. Huang, Emotion Recognition from Facial Expressions using Multilevel HMM, *Science and Technology*, 2000.

[4] Cohen, N. Sebe, A. Garg, L.S. Chen, T.S. Huang, Facial Expression Recognition From Video Sequences: Temporal and Static Modeling, *Computer Vision and Image Understanding (CVIU)*, Vol. 91, No. 1-2, 2003, pp. 160–187.

[5] J. Daugman, Face Recognition by Feature Demodulation, *Proc. Int'l Workshop Automatic*

Face and Gesture-Recognition, Zurich, 1995, pp. 350-355.

[6] G. Donato, M.S. Bartlett, J.C. Hager, P. Ekman, T.J. Sejnowski, Classifying Facial Actions, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 21, No. 10, Oct. 1999, pp. 974-989.

[7] P. Ekman, W.V. Friesen, *Facial Action Coding System: Investigators Guide*, Consulting Psychologists Press, Palo Alto, CA, 1978.

[8] D.M. Gavrila, J. Giebel, Virtual Sample Generation for Template-Based Shape Matching, in *Proc. Computer Vision and Pattern Recognition Conf. (CVPR'01)*, Kauai, Hawaii, Vol. 1, Dec. 8-14, 2001, pp. 676-681.

[9] G.E. Hinton, R.R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, *Science*, 28 July 2006, Vol. 313, No. 5786, pp. 504-507.

[10] Krizhevsky, I. Sutskever, G. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, *Advances in Neural Information Processing Systems 25 (NIPS'11)*, 2012, pp. 1106-1114.

[11] H. Lee, R. Grosse, R. Ranganath, A. Y. Ng, Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations, *Proc. 26th Annual Int'l Conf. Machine Learning*, June 14-18, 2009, Montreal, Quebec, Canada, pp. 609-616.

[12] H. Lee, C. Ekanadham, A. Y. Ng, Sparse Deep Belief Net Model for Visual Area V2, in *Advances in Neural Information Processing Systems 20 (NIPS'07)*, (J. Platt, D. Koller, Y. Singer, and S. P. Roweis, eds.), Cambridge, MA: MIT Press, 2008, pp. 873-880.

[13] M. Lyons, J. Budynek, S. Akamatsu, Automatic Classification of Single Facial Images, *IEEE Trans. Pattern Analysis and Machine Intelligence.*, Vol. 13, March 1991, pp. 252-263.

[14] M. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba, Coding Facial Expressions with Gabor Wavelets, in *Third IEEE Int'l Conf. Automatic Face and Gesture Recognition*, Nara, Japan, April 14-16 1998, pp. 200-205.

[15] M. Matsugu, K. Mori, Y. Mitari, Y. Kaneda, Subject Independent Facial Expression Recognition with Robust Face Detection Using a Convolutional Neural Network, *Neural Networks,* Vol. 16, 2003, pp. 555-559.

[16] V. E. Neagoe, A. D. Ciotec, Subject-Independent Emotion Recognition from Facial Expressions using a Gabor Feature RBF Neural Classifier Trained with Virtual Samples Generated by Concurrent Self-Organizing Maps, *Proc. 11th WSEAS Int'l. Conf. Signal Processing, Computational Geometry and Artificial Vision (ISCGAV '11)*, Florence, Italy, 2011, pp. 266-271.

[17] V. Neagoe, A. Ciotec, Virtual Sample Generation Using Concurrent-Self-Organizing Maps and Its Application for Facial Expression Recognition, *Proc. Int'l IEEEAM Conf. Mathematical Models for Engineering Science (MMES'10)*, Tenerife, Spain, Nov. 30-Dec. 2, 2010, vol. I, pp. 167-181.

[18] A. Rao, N. Thiagarajan, Recognizing Facial Expressions from Videos using Deep Belief Networks, *Stanford CS 229 Machine Learning Final Projects,* Autumn 2010.

[19] F. Y. Shih, C.-F. Chuang, P. S.P. Wang, Performance Comparison of Facial Expression Recognition in JAFFE Database, *Int'l J. Pattern Recognition and Artificial Intelligence*, Vol 22, No. 3, 2008, pp. 445-459.

[20] R. Valenti, N. Sebe, T. Gevers, Facial Expression Recognition: A Fully Integrated Approach, *14th Int'l Conf. Image Analysis and Processing Workshops (ICIAPW 2007)*, Modena, Italy, 10-13 Sept. 2007, pp. 125-130.

[21] Z. Zeng, M. Pantic, G.I. Roisman, T.S. Huang, A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 31, No. 1, Jan. 2009, pp. 39-58.