Deep Architectures for Automatic Emotion Recognition Based on Lip Shape

Branislav Popović, Stevan Ostrogonac, Vlado Delić
Faculty of Technical Sciences
University of Novi Sad
Novi Sad, Serbia
bpopovic@uns.ac.rs

Marko Janev
Mathematical Institute
Serbian Academy of Sciences and Arts
Belgrade, Serbia

Igor Stanković
European Center for Virtual Reality
Brest National Engineering School
Brest, France

Abstract — **Deep belief networks are used in this paper in order to initialize a feed-forward neural network, used for classification of emotions based on lip shape. Deep architectures are trained in a layer-wise greedy manner during the unsupervised training phase, whereas the neural network is trained during the supervised training phase, using the images from the reference database. Four distinctive, universally recognizable emotional states with the highest recognition rates have been selected from the reference database and further processed. Even with a small amount of training data, encouraging results were achieved in comparison to the results obtained by human classification, as well as the results obtained by using previously developed algorithms.**

Keywords - emotion recognition; deep learning; Boltzmann machines; neural networks

I. INTRODUCTION

Deep learning is a machine learning paradigm inspired by the human brain [1]. It is based on learning several levels of representations, corresponding to a hierarchy of features, factors or concepts. In order to learn high-level representations of data, a hierarchy of intermediate representations is exploited [2]. Deep belief networks (DBN) consist of multiple layers of stochastic, latent variables, often called hidden units or feature detectors. Training is conducted in a greedy manner, making the locally optimal choice at each stage, training one layer at a time, exploiting an unsupervised learning algorithm for each layer. Layers are represented as the Restricted Boltzmann Machines (RBM) [3]. Deep belief networks have been successfully applied on a number of issues in different areas, such as e.g. the artificial intelligence, natural language processing, pattern recognition, dimensionality reduction etc. [4]. In this paper, deep belief architecture has been used as a pre-training step, in order to initialize a neural network trained during the supervised training phase, using the images from the reference database [1].

Our task is to find the efficient algorithm for recognition of emotional states referring only to human lip shape. In order to achieve a reasonably successful classification rate, we have exploited and/or developed several different algorithms, as reported in [5]. The color and the shape of human lips are volatile, as well as lightning conditions, pose and several other effects. For the purposes of comparison, an adapted version of JAFFE database (*the Japanese Female Facial Expression Database*) has been used [6]. The original JAFFE database consists of 213 images of 7 facial expressions (6 basic facial expressions and one neutral) posed by 10 Japanese female models. The images were taken under different lighting conditions, which influenced the brightness, but also the contrast. The database contains facial expressions not only in the apex phase, but also from the initial moments of formation of facial expressions for a specified emotion, all the way to the end. The selected images have been cropped and additionally processed in order to be applicable for the purpose of emotion recognition based on lip shape.

In the first phase of our experiment, in order to evaluate the results and to examine the possible differences in perceiving the emotional states relying only on the shape of human lips between the European and Japanese subjects, four human subjects from Serbia classified images from the reference database into one of seven predefined categories, without the previous knowledge about the database, or any related technical knowledge. Based on the overall results of human classification, four most prominent, universally accepted emotional states have been selected from the reference database and additionally processed. In order to achieve the best possible results, we have utilized the algorithm as proposed in [1] and compared the results with our previously developed algorithms [5].

In section 2, the theory behind the deep belief networks will be presented in brief. In section 3, the image processing method will be described. Section 4 illustrates the architecture. The results will be given in comparison to the previously

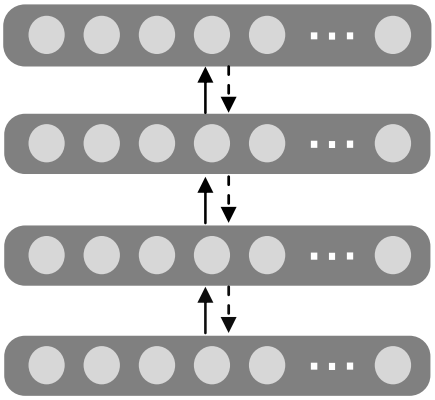

Figure 1. A three-layer DBN

developed algorithms. Several conclusions will be drawn in section 5.

II. DEEP ARCHITECTURES

DBN consists of building blocks represented as RBMs that can be stacked and efficiently trained in a greedy manner. RBM is an energy-based generative model. It consists of a layer of binary visible units representing the data, followed by a set of layers of hidden units that learn to represent features that capture higher-order correlations in the data. Each hidden unit creates a two region partition of the input space with a linear separation. Each intersection between the half-planes generated by units corresponds to a region in the input space associated with the same hidden configuration. The two layers are connected by a matrix of symmetrically weighted connections, and there are no connections within a layer [2]. RBMs can represent any discrete distribution if enough hidden units are used, and if the weights and offsets are set properly. DBNs learn to extract a deep hierarchical representation of the training data by modeling the joint distribution between the observed vector and the number of hidden layers [4], given as

$$P(x, h^1, ..., h^l) = (\prod_{k=0}^{l-2} P(h^k \mid h^{k+1}))P(h^{l-1}, h^l) \quad (1)$$

where $x = h^0$ is a visible layer, $P(h^{k-1}, h^k)$ is a conditional distribution for the visible units conditioned on the hidden units of the RBM at level k, and $P(h^{l-1}, h^l)$ is the joint distribution between the visible units and the hidden units in the top-level RBM. A three-layer DBN structure is illustrated in Fig. 1 (based on [7]).

Computational theories of deep learning are inspired and closely related to a class of theories of human brain that could be observed as a huge deep architecture with a temporal dimension. A given input percept is represented hierarchically at multiple levels of abstraction. Each level corresponds to a different area of cortex. Higher levels represent more abstract and invariant features [8]. Similarly, training of each RBM is the same as training of all of the previous RBMs, except that the training data is mapped through all of the previous RBMs before being used as training samples. In order to use the DBN for classification, a sample is applied to the lowest visible layer, and then filtered through the entire DBN until it reaches the last hidden layer. Therefore, each layer of RBMs models more abstract features. The learning algorithm is unsupervised, but the DBN could be used in order to initialize a feed-forward neural network (FFNN), as proposed in [1].

Figure 2. The adapted JAFFE database (happy expression)

III. IMAGE PROCESSING

In the initial image processing phase, the rough lip area in the images from the reference database was selected (Fig. 2) using Face Detector, as presented in [9] and [10]. In order to improve the contrast of the grayscale images, and therefore to compensate the complexion differences, as well as the differences in shooting conditions, histogram equalization procedure had to be applied, which means the redistribution of intensity levels for every single image, in order to achieve a uniform histogram. All of the images were additionally filtered using the average and the median filter in order to get smoother images by removing smaller irregularities and scratches. Filters size was determined in accordance with the recognition rates and the visually determined effectiveness of the filtering procedure.

In order to binarize the images, a suitable threshold has been applied per each and every image. The value of the threshold has been determined by practice. After binarization process has been completed, the area around the lips, as well as parts of nose and chin, shadows and other remaining irregularities had to be removed. This was achieved by cutting out the frames and removing the largest connected areas beside the lips. Remaining pixels were used in order to calculate the boundaries of lips contours in the original image more accurately.

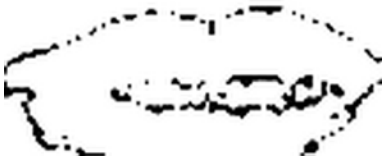

Figure 3. Contours of lips (example)

Based on these calculations, knowing the positions of lips in the original image, the images were cropped again and the gamma correction was applied. Finally, the edges of lips contours were determined by using the appropriate filter and the remaining images were resized for the purpose of comparisons (Fig. 3). In order to be able to use these images during the unsupervised training phase, the images had to be vectorized, which means that every column of the matrix representing the image had to be attached into a single vector.

TABLE I. THE CONFUSION MATRICES FOR HUMAN SUBJECTS S1 TO S4

Subjects S1 to S4 / Annotated and recognized emotional states											
S1	***Happy***	***Fear***	***Disgust***	***Surprise***	***Acc. [%]***	***S2***	***Happy***	***Fear***	***Disgust***	***Surprise***	***Acc. [%]***
Happy	28	2	0	1	90	***Happy***	29	0	2	0	94
Fear	2	4	13	13	13	***Fear***	2	6	17	7	19
Disgust	2	2	17	6	63	***Disgust***	3	10	13	1	48
Surprise	1	13	0	16	53	***Surprise***	3	6	0	21	70
S3	***Happy***	***Fear***	***Disgust***	***Surprise***	***Acc. [%]***	***S4***	***Happy***	***Fear***	***Disgust***	***Surprise***	***Acc. [%]***
Happy	30	0	1	0	97	***Happy***	28	1	1	1	90
Fear	0	6	15	11	19	***Fear***	0	8	13	11	25
Disgust	0	6	19	2	70	***Disgust***	0	3	19	5	70
Surprise	2	2	1	25	83	***Surprise***	1	5	1	23	77

TABLE II. THE CONFUSION MATRICES FOR DEEP BELIEF ARCHITECTURE

Number of epoch – learning rate / Annotated and recognized emotional states											
100–0.1	***Happy***	***Fear***	***Disgust***	***Surprise***	***Acc. [%]***	***100–0.6***	***Happy***	***Fear***	***Disgust***	***Surprise***	***Acc. [%]***
Happy	23	0	5	3	74	***Happy***	26	0	2	3	84
Fear	0	20	5	7	63	***Fear***	0	20	4	8	63
Disgust	3	8	9	7	33	***Disgust***	2	5	13	7	48
Surprise	6	11	4	9	30	***Surprise***	4	10	3	13	43
200–0.1	***Happy***	***Fear***	***Disgust***	***Surprise***	***Acc. [%]***	***200–0.6***	***Happy***	***Fear***	***Disgust***	***Surprise***	***Acc. [%]***
Happy	24	1	4	2	77	***Happy***	29	0	1	1	94
Fear	0	21	7	4	66	***Fear***	0	29	2	1	91
Disgust	4	7	13	3	48	***Disgust***	0	3	24	0	89
Surprise	3	5	2	20	67	***Surprise***	1	3	0	26	87

IV. DESIGN OF EXPERIMENTS

A. *Human Classification Rates*

Bearing in mind the constraints imposed by the database, the differences in perceiving and expressing emotional states among different ethnic groups, as well as the fact that the images from the reference database have been cropped in order to observe only the lower part of the subject's face (lips, part of nose and chin), the adapted version of the reference database had to be reevaluated in order to determine classification rates for seven universally accepted emotional states present in the original database (happy, surprise, sad, annoyed, fear, disgust and neutral). Four subjects, thirty to sixty-five years old, independently classified images into one of seven predefined categories, based on their subjective evaluation. The subjects had no previous knowledge about the original database, or any related technical knowledge about the issue. The images were presented randomly in order to avoid the influence from images that belong to the same expression series to affect the final decision. The results were highly compatible among the different subjects.

In accordance with the procedure presented in [5], in order to be able to compare the results between the different algorithms, the reference database has been stripped down to include only four most prominent emotional states, based on the overall results of human classification. The adapted database therefore contains 120 out of 213 images, with four basic emotional states (happiness, fear, disgust and surprise). The experiments were repeated and the final recognition rates, with the full confusion matrices for all four subjects are presented in Table I. Clearly, a large confusion exists among the expressions of emotions fear and disgust, as well as fear and surprise, in both directions. The results could be explained by the fact that the subjects were shown only the area around the lips, whereas no other information was included.

B. *Deep Belief Network*

Deep learning architecture has the ability to learn from unlabelled data in order to achieve a fine representation of input data during the unsupervised pre-training phase. The parameters learned are used in order to initiate a model, that will be trained in a supervised fashion, in order to additionally adjust the parameters for a given classification task. Each layer will be trained in an unsupervised fashion, depending of the activations from the layer below, one after another. It is

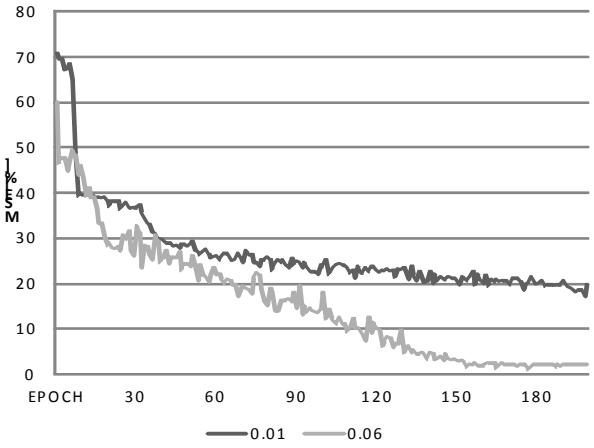

Figure 4. MSE for different learning rates (example)

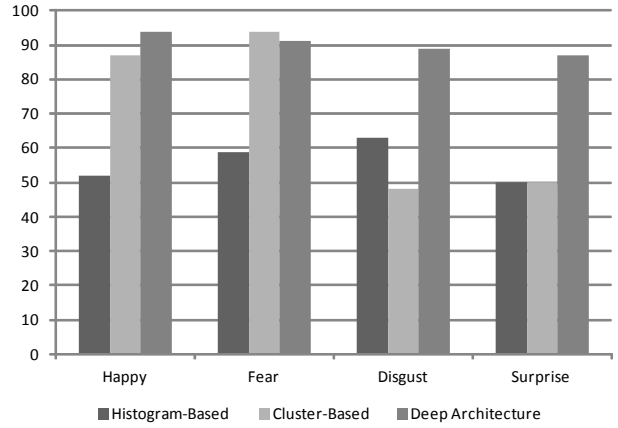

Figure 5. The accuracy of the proposed algorithms

supposed that the unsupervised pre-training moves the parameters to a region in parameter space closer to a global optimum in order to achieve more natural representation of data, whereas the global supervised learning rarely changes the pre-trained parameters much.

Using Deep Learning Toolbox [1] DBN were constructed, consisting of three layers implemented in a form of RBMs. Each RBM contained 100 hidden neurons, and it was trained in a layer-wise greedy manner. Considering the size of the database and the lack of training data, leave-one-out cross-validation was used in order to compare the results of classification by human subjects with the recognition rates obtained by using deep learning methods, and by previously reported histogram-based and cluster-based algorithms [5]. Each RBM was therefore trained using a set of 119 images (excluding a test image at each iteration), using at each epoch randomly selected mini-batches of size 15, with a fixed learning rate of 0.01 for 100 epochs, as in [1]. One epoch could be observed as one iterative cycle (119 images).

In the supervised training phase, DBN were used in order to initialize a feed-forward neural network, used for the final classification of emotional expressions based on lip shape. The FFNN consists of 4 layers of sizes 100-100-100-4, where the last 4 neurons represent the output label units. Smaller number of units was used because of smaller amount of data available for training and testing purposes than the one reported in [1]. Training was done using randomly selected mini-batches of size 15 for 100 and 200 epochs, using a fixed learning rate of 0.1 and 0.6, respectively. Maximum output unit was chosen as the label for each testing sample. Number of epoch could be determined as a trade-off between the recognition accuracy and the possibility of over-fitting the neural network that has to be avoided. Example of the mean squared error (MSE) on training set for the first 200 epochs is given in Fig. 4 for learning rates of 0.01 and 0.06. Considering the size of the reference database, the more aggressive approach (higher learning rate) gave somewhat better results, as it was expected. The code ran for about 36 hours in total and the results are presented in Table II. Promising results were achieved, although in a certain extent, this could be explained by the fact that the images were taken under controlled environmental conditions. The images were also a part of a series of expressions from the initial moments of formation of facial expression, until the end.

The experiments showed a large benefit from using deep architectures in comparison to the previously reported algorithms, even for a small amount of training data. Comparative results for the algorithm used in this paper (200 epochs, learning rate of 0.06) and the histogram-based and cluster-based methods proposed in [5] are presented in Fig. 5. Unlike the results presented in this paper, larger confusion could be observed between the emotions happy and disgust (histogram-based algorithm), as well as between emotions disgust and fear (cluster-based algorithm), although the overall recognition rate was higher than the overall recognition rate in case of classification by human subjects. In case of the algorithm used in this paper, confusion exists between emotions fear and disgust and surprise and fear, which was expected because the lips were either sealed or slightly opened in both cases, in most of the relevant images, and no other clues were taken into account.

V. CONCLUSION

In a usual everyday dialogue, humans are accustomed to observe not only the face, but the whole body, combining different audio-visual channels in order to recognize the different emotional expressions. Pose, gesture, voice and other clues are used in order to conclude about the emotional state of our interlocutors. In laboratory conditions where subjects have to make a decision about the emotional state observing only the area around the lips, there could be a great confusion between different emotional states, as it was proven by the experiments presented in this paper. Comparisons were made based on the samples from the reference database which contains emotional expressions in various stages, where some of those stages could hardly be distinguished by humans.

Machines recognize emotions by employing different mathematical operations. The conclusion is usually made by calculating a kind of similarity measure between the input sample and the models or features in our training database. If we provide sufficient amount of annotated data, classification rates could be higher than the rates obtained through classification by humans. However, the results would be

limited to the specified conditions, and therefore several different clues have to be joined together. Based on the experimental results, we conclude that deep architectures could be used even in a case where there is only a small amount of data available. Howerer, there is always a possibility of overfitting, and further research has to be made in order to acquire efficient and reliable system. The results presented in this paper could easily be extended in order to include other emotional states. The algorithm is flexible in a sense that the database contained the images with different lighting and contrast levels, as well as the images in which emotional expressions were not clearly expressed.

ACKNOWLEDGMENT

This research work has been supported by the Serbian Ministry of Education, Science and Technological Development, and it has been realized as a part of "Development of Dialogue Systems for Serbian and Other South Slavic Languages" research project (id TR 32035).

REFERENCES

[1] R. B. Palm, "Prediction as a candidate for learning deep hierarchical models of data", Technical University of Denmark, Palm 2012.

[2] Y. Bengio, "Learning deep architectures for AI", Foundations and Trends in Machine Learning, vol. 2, no. 1, 2009.

[3] G. E. Hinton, S. Osindero and Y.-W. Teh, "A fast learning algorithm for deep belief nets", Neural Computation, vol. 18, no. 7, 2006, pp. 1527-1554.

[4] G. E. Hinton, "Deep belief networks", Scholarpedia, vol. 4, no. 5, 5947, 2009.

[5] B. Popović, V. Crnojević, V. Delić, "Histogram based and cluster based methods for automatic emotion recognition based on lip shape", DOGS 2010, Iriški Venac, Serbia, 2010, pp. 160-163.

[6] Y. Yorozu, M. Hirano, K. Oka and Y. Tagawa, Japanese Female Facial Expressions (JAFFE), Database of digital images, 1997.

[7] N. Lopes, B. Ribeiro, J. Goncalves, "Restricted Boltzmann machines and deep belief networks on multi-core processors", Neural Networks (IJCNN), 2012, pp. 1-7.

[8] J. M. H. Lobato and D. H. Lobato, "The new generation of neural networks", Universidad Autonoma de Madrid, Computer Science Department, 2008.

[9] P. Aldrian, U. Meier, A. Pura, "Extract feature points from faces to track eye's movement", University of Leoben, Austria, 2009.

[10] I. Stanković, M. Karnjanadecha, V. Delić, "Improvement of Thai speech emotion recognition using face feature analysis", International Review on Computers and Software (IRECOS), vol. 7, no. 5, September 2012, pp. 2003-2015.