Multimodal Emotion Recognition using Deep Networks

C. Fadil, R. Alvarez, C. Martínez[1,3], J. Goddard[4] and H. Rufiner[1,2]

[1] Center for Signals, Systems and Computational Intelligence (SINC),
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral, Santa Fe, Argentina
[2] CONICET, Argentina
[3] Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Entre Ríos, Argentina
[4] Departamento de Ingeniería Eléctrica, Universidad Autónoma Metropolitana, Ciudad de México, México

Abstract— **In the last years, several efforts have been devoted to the automatic recognition of human emotions. On the one side, there are several works based on speech processing and on the other side, using facial expressions in still images. More recently, other modalities such as body gestures, biosignals and others have been started to be used. In this work we present a multimodal system that process audiovisual information, exploiting the prosodic features in the speech and the development of the facial expressions in videos. The classification of the video in one of six emotions is carried out by deep networks, a neural network architecture consisting of several layers that capture high-order correlations betwen the features. The obtained results show the suitability of the proposed approach for this task, improving the performance of standard multilayer Perceptrons.**

Keywords— **emotion recognition, autoencoders, deep networks, prosodic features, facial expressions**

I. Introduction

One of the most interesting topics today in the field of Human-Computer Interaction is the recognition of human emotions. The development of systems capable to identify the emotion of the person who is interacting with, would conduct to respond properly in a more natural way. Thus, these systems could lead to a number of applications for our everyday life (interactive games and entertainment industry), clinical studies for diagnosis of emotional state in psychiatric patients, addition of sensitivity to customer services or call centers, among others [1, 2].

During the last decade, most works on this area use one of two modalities, either speech or image. On the one hand, the prosodic features such as pitch, energy and linear prediction and cepstral coefficients were used [3]. On the other hand, the measurement of facial expressions have led to the main research line using the visual data. Here, the features extracted mainly consisted on holistic representations (discrete Fourier coefficients, PCA projections of the face), parametric flow models and facial landmarks [4]. More recently, other modalities such as body gestures, biosignals and others have been started to be used [5].

The classification schemes reported mainly consisted on standard methods for related tasks: Gaussian mixtures models, Bayesian classifiers, hidden Markov models and Support Vector Machines [6]. The neural networks architectures found in literature mainly consist of variants of a multilayer artificial networks [7, 8].

In this work we present a multimodal system that process audiovisual information, exploiting the prosodic information in the speech and the progress of the facial expressions through time in videos. The classification of the emotion contained in the video is carried out in one of six emotions by means of deep autoencoder networks. This archictures have been recently proposed and consist of several layers built up with the aim to capture high-order correlations between the features in a hierarchical manner [9].

The rest of the article is organized as follows. Section II details the feature extraction stage for speech and audio. Section III explains the training and classification by the deep autoencoder networks. Section IV shows the experiments and results obtained. Finally, Section V concludes the paper.

II. Feature extraction

The audio and video tracks are only processed during the development of the emotion, so the first step in this stage is to detect the voiced segment. This task is carried out by means of a voice activity detector [11]. After that, the initial and final silences are discarded.

The audiovisual features are extracted from the complete development of the emotion, with a fixed number of features in order to feed the autoencoders. Due to intrinsic differences in length of the videos, the mean number of frames (47) was calculated and afterwards used to uniformly sample each video.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
C. Fadil, R. Alvarez, C. E. Martínez, J. Goddard & H. L. Rufiner; "Multimodal Emotion Recognition using Deep Networks"
Anales del VI Congreso Latinoamericano de Ingeniería Biomédica (CLAIB 2014), pp. 355, oct, 2014.

A. Audio features

From the speech audio, the following features are extracted:

- Prosodic features: mean value and standard deviation of the energy and the fundamental frequency estimated with the PEFAC method [12], both calculated using windows of 10 ms. [13].
- Spectral characteristics: mean logarithmic spectrum (MLS) given by:

$$S(k) = \sum_{n=1}^{47} \log|(v(n,k))| \quad (1)$$

  where k is the frequency band and $v(n,k)$ is the Discrete Fourier Transform in the n-esim frame. 30 MLS coeficients were calculated in the frequency range (0-1200Hz), given that in [14] it obtained the better results in emotion separation.
- Cepstral features: mel frequency cepstral coefficients (MFCC) calculated with Hamming windowing of 1024 samples.

From these features, two different vectors were conformed:

- $\mathbf{fva}_{46}$: 46 features composed by 12 MFCC, 30 MLS, energy and pitch (mean and standard deviation).
- $\mathbf{fva}_{70}$: 70 features composed by 12 MFCC and their first and second derivatives, 30 MLS, energy and pitch (mean and standard deviation).

B. Video feature extraction

The steps applied in this stage are:

1. *Frame selection*: taking into account the mean number of frames for all the database used (47 frames), for longer videos the initial and end frames are discarded, while in shorter videos the central frames are repeated.
2. *Face finding and facial landmark detection*: the analyzed regions of the face correspond to the mouth and eyes areas. They were selected since they involve more emotional content in facial expressions [4]. To find them in the face, 8 points were obtained using the *flandmark* library [15], that allow to exactly locate the center of the face (ε_0), eyes (ε_5 and ε_1 –right eye–, ε_2 and ε_6 –left eye), mouth (ε_3 and ε_4) and nose (ε_7). Figure 1 shows a model of a face and these points.
3. *Segmentation of region of interest in mouth and eyes*: these areas encompass relevant information with emotional content, in eyebrows and small places around the mouth. Figure 3 show examples of these areas. They are then processed in grayscale at a fixed size of 70×35 pixels each, giving a total of 4900 values per frame.
4. *Images normalization and re-grouping per emotion*: the mouth and eyes are vectorized and concatenated, giving a vector of 1×230300 per video. Then, the vectors of each emotion are grouped together, preparing the next step.
5. *PCA calculation*: one PCA space for each emotion is obtained and then the patterns are projected and the required number of components are kept for the final feature vectors.

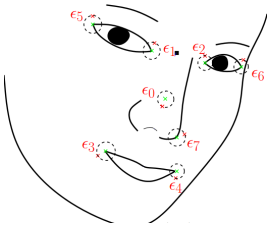

Fig. 1: Location of facial landmarks (adapted from [15]).

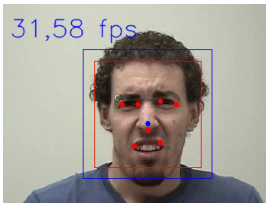

Fig. 2: Example of the landmarks calculated on a particular frame.

Fig. 3: Example of the segmented zones.

Thus, two feature vectors for each video are obtained:

- $\mathbf{fvv}_{70}$: first 70 components of PCA, comprising 85% of the variance.
- $\mathbf{fvv}_{85}$: first 85 components of PCA, comprising 90% of the variance.

III. Classification using deep networks

The classifier used in this work is a multilayer Perceptron (MLP), pre-trained by the technique of deep autoencoding. The neural network is constructed by stacking autoencoders optimizing the architecture layer by layer, to obtain the best results from autoencoding up to and final classification [9].

The architecture of the neural network is: **i** neurons in the input layer (given by the number of audio-video features used), two hidden layers ($\mathbf{h}_1$ and $\mathbf{h}_2$) with variable number of units and an output layer (**o**) of 6 units, one for each of the emotions considered. This classifier, called *deep classifier*, is

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
C. Fadil, R. Alvarez, C. E. Martínez, J. Goddard & H. L. Rufiner; "Multimodal Emotion Recognition using Deep Networks"
Anales del VI Congreso Latinoamericano de Ingeniería Biomédica (CLAIB 2014), pp. 355, oct, 2014.

the result of the concatenation of a *deep autoencoder* with the output layer.

Several preliminary experiments were carried out to optimize the number of neurons in each layer (n_{h1}, n_{h2}) and the parameters of the backpropagation training algorithm (learning rate and momentum). Also, these series of experiments allowed to select the sinousoidal activation function for all the layers, which obtained better perfomance than logistic, tangent hyperbolic and identity.

The main steps performed in training the deep autoencoder are summarized next, for a more thorough explanation please see [9]. First, we build the autoencoder that codifies the input layer in the first hidden layer, which has **i** neurons in the input layer, $\mathbf{h_1}$ in the hiddern layer ($\mathbf{h_1} < \mathbf{i}$) and **i** in the output layer. This autoencoder is trained with the backpropagation algorithm. Once the network is trained to reproduce the input layer in its output layer, the best architecture is selected. The output layer is discarded and the patterns are passed by the network $\mathbf{i} + \mathbf{h_1}$ to generate the inputs for the next autoencoder, which will be $\mathbf{h_1} + \mathbf{h_2} + \mathbf{h_1}$. Now, this autoencoder is trained with the backpropagation algorithm. Again, the best network is selected and the output layer is discarded. The two networks are inverted and joined together to achieve the a new autoencoder $\mathbf{i} + \mathbf{h}_1 + \mathbf{h}_2 + \mathbf{h}_1 + \mathbf{i}$. It is trained in the same manner, then the last two layers are discarded to obtain the network $\mathbf{i} + \mathbf{h}_1 + \mathbf{h}_2$. The patterns are passed through it to obtain the inputs for the last stage. A new network is built in the form $\mathbf{h}_2 + \mathbf{o}$, and trained in a supervised manner using backpropagation. The best network is selected and joined to the previous autoencoder, to form the final deep classifier $\mathbf{i} + \mathbf{h}_1 + \mathbf{h}_2 + \mathbf{o}$, which is now pre-trained. A final fine adjustment in all the weights of the network is carried out by training up to reach the best performance. The training is done using the statistical technique of cross-validation with 10 folds. Figure 4 shows an scheme of the final architecture.

IV. Experiments and results

The proposed technique was experimented with the *RML Emotion Database*. It has 720 videos of 8 individuals from different gender and culture, in 6 different languages [10].

The classification experiments were designed to test the performance of the four multimodal feature vectors resulting from the audio sets $\mathbf{fva}_{46}$ - $\mathbf{fva}_{70}$ and the video sets $\mathbf{fvv}_{70}$ - $\mathbf{fvv}_{85}$. The four combinations were built as:

- $\mathbf{fv}_{116} = \mathbf{fva}_{46} + \mathbf{fvv}_{70}$
- $\mathbf{fv}_{131} = \mathbf{fva}_{46} + \mathbf{fvv}_{85}$
- $\mathbf{fv}_{140} = \mathbf{fva}_{70} + \mathbf{fvv}_{70}$
- $\mathbf{fv}_{155} = \mathbf{fva}_{70} + \mathbf{fvv}_{85}$

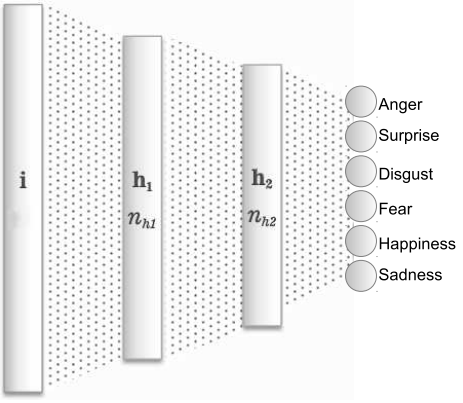

Fig. 4: Final deep classifier.

For each set, different neural network architectures were tested. To compare the performance of the proposed method, a standard MLP was also trained. Also, its architecture was experimented up to obtain the best performance for each feature set. For brevity reasons, here only the best architecture for each feature set will be detailed, which resulted to be $\mathbf{fv}_{116}$.

Table 1 resumes the results obtained with the best architecture for each deep classifier and MLP experimented. For the deep networks, the best architecture resulted $CP2_{116}$ with a 79.72% of correct classification. In this case, the improvement over the standard MLP was about 3%.

Monomodal experiments were also designed to test how the audio and video features contribute to the overall performance of the proposed method. The four monomodal feature vectors were tested with several neural network architectures (MLP). With audio, the best set resulted $\mathbf{fva}_{46}$ with a 77.08% of correct classification. With video, 75.41% was achieved with the set $\mathbf{fvv}_{85}$. These results show that the multimodal approach improves the classification performance.

There were also conducted multimodal experiments with a single language. In this case, the classification performance was similar to the multilanguage tests. The best result obtained was a 78.33% of correct classification.

Table 2 shows the average confusion matrix for the 10 partitions of cross-validation with the best deep classifier. Here it can be seen that, in the one hand, the most difficult to distinguish emotion was disgust. The classifier confuses this emotion, in most cases, with fear. On the other hand, the most distinguishable emotion was anger (classification rate of 88.33%). The rest of emotions were well distinguished by the classifier with a correct classification rate near to 80%.

The election of the speech feature vectors without the MFCC derivatives could be given by the fact that these derivatives help to reduce the effects of noise in the signals.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
C. Fadil, R. Alvarez, C. E. Martínez, J. Goddard & H. L. Rufiner; "Multimodal Emotion Recognition using Deep Networks"
Anales del VI Congreso Latinoamericano de Ingeniería Biomédica (CLAIB 2014), pp. 355, oct, 2014.

Table 1: MSE error and performance for feature set $\mathbf{fv}_{116}$

Feature set	Architecture				MSE	MSE	MSE	% Classification
	i	**h_1**	**h_2**	**o**	Encoder 1	Encoder 2	Encoder 3	
$CP1_{116}$	116	70	35	6	0.36	0.24	0.72	77.92
$CP2_{116}$	**116**	**100**	**40**	**6**	**0.09**	**0.53**	**0.65**	**79.72**
$CP3_{116}$	116	100	50	6	0.09	0.41	0.55	78.33
$CP4_{116}$	116	100	60	6	0.09	0.32	0.47	78.66
$CP5_{116}$	116	105	60	6	0.08	0.31	0.47	77.64
MLP_{116}	**116**	**50**		**6**	-	-	-	**76.81**
MLP_{116}	116	138		6	-	-	-	76.81

Table 2: Confusion matrix for deep classifier with architecture $CP2_{116}$.

Emotion	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	**88.33%**	1.67%	1.67%	0.83%	0	7.50%
Disgust	0.83%	**70%**	10%	5%	8.33%	5.84%
Fear	7.50%	2.50%	**82.50%**	1.67%	3.33%	2.50%
Happiness	4.17%	3.33%	4.17%	**80.83%**	3.33%	4.17
Sadness	0	4.17%	7.50%	7.50%	**78.33%**	2.50%
Surprise	10%	3.33%	5%	2.50%	0.84%	**78.33%**

They are also more important in cases where the noise is noticeable. In the case of the experimented data, noise is not important in the audio signals, so the inclusion of the derivatives only results in greater complexity of the classifier, without adding meaningful information for discriminating the emotions.

The statistical significance of the results was evaluated considering the probability that the classification error of a given classifier ε (deep network) is smaller than the one of the reference system ε_{ref} (MLP) [16]. The statistical independence of the errors for each pattern was assumed and the binomial distribution of the errors was modeled by means of a Gaussian distribution. Therefore, comparing the results obtained for $CP2_{116}$, a $Pr(\varepsilon_{ref} > \varepsilon) > 96.77\%$ was obtained.

V. Conclusions

We presented an automatic emotion recognition system on multimodal data. The feature extraction was centered on prosodic information from speech and visual information from eyes and mouth in video. The classification was carried out using a novel neural network approach, the deep classifier built from stacked autoencoders.

In the experiments, our approach was compared with the standard multilayer Perceptron (MLP) technique, with different combinations of audio and visual features. For all the cases, the deep classifier outperform the results for the MLP near 5% in classification rate, confirming the feasibility of these networks for the task.

Future works could be devoted to incorporate an automatic clustering of emotion by the spectral characteristics of the speech, in a hierarchical way to separate the most hard to classify emotions, as presented in [14]. Other extension of this work would be to classify the emotion through time during its development, in the so-called *emotional profile*.

Acknowledgements

The authors wish to thank: the *Universidad Nacional de Litoral* (with CAI+D 2011 #58-511, #58-519, #58-525), and the SEP and CONACyT from México (with Program SEP-CONACyT CB-2012-01, No.182432), for their support.

References

1. Fragopanagos N, Taylor John G. Emotion recognition in human–computer interaction *Neural Networks.* 2005; 18:389–405.
2. Karray F., Alemzadeh M., Saleh J. A., Arab M. N.. Human-computer interaction: Overview on state of the art *International Journal on Smart Sensing and Intelligent Systems.* 2008; 1:137–159.
3. El Ayadi M., Kamel M. S., Karray F.. Survey on speech emotion recognition: Features, classification schemes, and databases *Pattern Recognition.* 2011; 44:572–587.
4. Bettadapura, V. Face expression recognition and analysis: the state of the art, arXiv preprint arXiv:1203.6722, 2012.
5. Jang E.-H., Park B.-J., Kim S.-H., Eum Y., Sohn J.-H. A Study on Analysis of Bio-Signals for Basic Emotions Classification: Recognition Using Machine Learning Algorithms in *2014 International Conference on Information Science and Applications (ICISA)*, 2014; 1–4.
6. Marrero-Fernández P., Montoya-Padrón A., Jaume-i-Capó A., Buades Rubio J.M. Evaluating the Research in Automatic Emotion Recognition, IETE Technical Review, Taylor & Francis, 2014; 31:3, 220–232.
7. Gunes H., Schuller B., Pantic M., Cowie R. Emotion representation, analysis and synthesis in continuous space: A survey in *2011 IEEE International Conference onAutomatic Face & Gesture Recognition and Workshops (FG 2011)*, 2011; 827–834.
8. Kukla E., Nowak P. Facial Emotion Recognition Based on Cascade of Neural Networks in *New Research in Multimedia and Internet Systems*, Springer, 2015; 67–78.
9. Hinton G., Salakhutdinov R. Reducing the dimensionality of data with neural networks *Science.* 2006; 313:504–507.
10. De Silva L. C., Hui S. C. Real-time facial feature extraction and emotion recognition in *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proc of the 2003 Joint Conf of the Fourth Intn'l Conf on*, 2003; 3:1310–1314.
11. Sohn J., Nam Soo K., Wonyong S. A statistical model-based voice activity detection, *IEEE Signal Processing Letters*, 1999; 6:1, 1-3.
12. Gonzalez S., Brookes M. A pitch estimation filter robust to high levels of noise (PEFAC), in *Proc. EUSIPCO*, 2001; 451–455.
13. Deller J., Hansen J., Proakis J. *Discrete Time Processing of Speech Signals*. Macmillan Publishing, 1993.
14. Albornoz E., Milone D., Rufiner H. Spoken emotion recognition using hierarchical classifiers. *Computer Speech & Language.* 2011; 25:556–570.
15. Michal U., Franc V., Václav H., Detector of Facial Landmarks Learned by the Structured Output SVM in *VISAPP '12: Proc of 7th Int'l Conf on Computer Vision Theory and Applications*, SciTePress, 2012, 1547-556.
16. Toutenburg H. *Statistical analysis of designed experiments*. Springer, 2002.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
C. Fadil, R. Alvarez, C. E. Martínez, J. Goddard & H. L. Rufiner; "Multimodal Emotion Recognition using Deep Networks"
Anales del VI Congreso Latinoamericano de Ingeniería Biomédica (CLAIB 2014), pp. 355, oct, 2014.