Facial Expression Classification Using Convolutional Neural Network and Support Vector Machine

Valfredo Pilla Jr, André Zanellato, Cristian Bortolini,
Humberto R. Gamba and Gustavo Benvenutti Borba
Graduate Program in Electrical and Computer Engineering
Federal University of Technology - Paraná
Email: vpillajr@mail.com

Henry Medeiros
Department of Electrical and Computer Engineering
Opus College of Engineering
Marquette University

***Abstract*—Perception of facial emotional expressions is an important element of human communication, supporting social interactions by means of their informative, evocative and incentive functions. Thus, computational approaches to automatically classify facial emotional expressions may contribute to improve human machine interfaces, among other applications. This work presents an algorithm for human emotional state classification from grayscale still images of faces. These images are from the Extended Cohn-Kanade public dataset and represent three emotional states: Aversion, Happiness and Fear. Preprocessing applied to the images are restricted to cropping the region around the eyes and mouth, image resizing, and intensity mean pixel to pixel subtraction. Images characteristics extraction are handled by a previously trained Alexnet Convolutional Neural Network. The classification system is a Support Vector Machine and has achieved an average accuracy of 98.52% on the aforementioned classes.**

I. Introduction

The study of facial expressions had its first significant result with Darwin's research [1] on its impact in the evolution of the species as a form of nonverbal communication. This form of communicating through identifying an emotion in the facial expression, substantially faster than verbal communication, would bring a great advantage to the human species, theorized Darwin. Under this evolutive perspective, the facial expressions would be universal to all humans. Recent studies [2] support this theory, with scientists identifying six universal facial expressions: Happiness, Sadness, Fear, Surprise, Anger and Aversion. Emotional perception is an important element of human communication, used to interpret events, social interactions and to human relations in general [3]. Emotional expression has several actions associated with it, such as face movements and body gestures, changes in voice tone, physiological changes in the skin resistance and facial flushing, just to cite a few. Emotional perception is a complex action that may involve several elements. Limiting the study of emotional expressions to the observation of the state of facial muscles [4], emotional perception understanding can be restricted to the analysis of sequential images or even single images.

Given the universality of these facial expressions, there is great interest in a computational mechanism to recognize them automatically.

In this context, techniques for expression recognition are usually based on assessments of the movements of the facial muscles [4] and the eyes [5], or on measures to establish a relation between the shape of parts of the face and the emotions. This information can be obtained through still images or through sequences of images that show the emotion going from neutral to its apex.

A system to classify emotional states from images can be comprised of a set of algorithms that extract characteristics (features) from these images and another algorithm that uses such characteristics to establish a class-label to each image.

Recently, computational intelligence techniques such as deep neural networks (DNN) [6], in particular convolutional neural networks (CNNs), have been used to extract features more succesfully than manually designed ad hoc extractors.

The architecture of CNNs requires a broad set of parameters that are learned from a large set of previously labeled data, in this case, a set of previously labeled images. In some situations, this can be a problem, since it may be difficult to find public datasets with a huge amount of images. A strategy to get around this limitation is the artificial enlargement of the database through label-preserving modifications in the data. This process is commonly known as *data augmentation*. In this work, we use a previously trained CNN [7] to extract the images descriptors. Therefore, data augmentation was not necessary. The descriptors generated by the CNN are then used as input to a support vector machine (SVM) which classifies the emotional states in images.

In the following sections we present a brief review of the related work, the architecture of the proposed classification system and the achieved results. Finally, results are discussed and future work suggestions are presented.

II. Related Studies

A wide range of approaches to preprocessing, feature extraction, and classification is available in the literature on

emotion detection based on facial expression. Some of these approaches using the Expanded Cohn-Kanade database (CK+) [8] are briefly described below.

In [9], for example, among various techniques, photometric standardization (normalization based on homomorphic filtering) and histogram manipulation to the normal distribution are found to achieve the best classification results when used as preprocessing for the purpose of removing the variance of illumination; the face features are extracted by the convolution of the preprocessed images with 40 Gabor masks. A PCA Kernel is then applied to the features which are then sent to a neural network, consisting of an input layer and two hidden layers trained with the layer-wise greedy strategy. Finally, a softmax classifier is applied. This work made use of images of the six classes of basic expressions from the CK+ database ("Surprise", "Fear", "Aversion", "Anger", "Happiness" and "Sadness"), achieving an average accuracy of 96.8%. For the subset of expressions "Aversion", "Surprise" and "Happiness" the accuracy reached 100%.

The approach in [10] aims to identify the same six basic expressions and a seventh one, "Contempt". This method draws discriminative features from still images combining holistic features and features based on local distances to the characterization of facial expressions. The distance-based features are subsequently quantized to form intermediate-level features using bag of words. The classifier used is the SVM. The average accuracy was 87.7%. For "Aversion", "Surprise" and "Happiness" the accuracies reached 91.53%, 96.39% and 95.65%, respectively.

In [11], Zhong *et al.* explore common information and specific information found in different expressions inspired by the observation that only a few parts of the face are active in the expression, such as regions around the eyes and mouth. The authors make use of a MTSL (Multitask Sparse Learning) framework, comprising a first stage in which expression recognition tasks are combined with common local patches. In a first step the dominant patches are found to each expression and in a second step are associated with two tasks, recognition of facial expression and verification of the face, in order to learn specific facial patches for individual expression. The best results for the CK+ database show an average accuracy of 91.53%, and for the classes "Aversion", "Surprise" and "Happiness" reached accuracies of 94.11%, 98.70% and 96.35%, respectively.

The authors in [12] designed a dedicated CNN to classify the six emotion expression in CK+. For image preprocessing, an eye locator algorithm is used to center the images and crop them in order to remove the background. Images are rotated so that the eye line is horizontal (spatial normalization). To train the dedicated CNN, it was necessary to expand the original database (data augmentation) by generating variations of every training image by rotating the original image. These images are submitted to intensity normalization and used for the training and evaluation of the network. The classification is made by layers of fully connected neurons. The average accuracy obtained for the best training was 93.74%, where the "Aversion" class achieved an accuracy of 96.55%, "Happiness" 98.06% and "Fear" 97.38%.

III. Architecture of the proposed system

The proposed system aims to detect the emotional state of subjects from photos in which they express these emotions. We restrict our classification task to three emotional classes: "Aversion", "Happiness" and "Fear". According to [4], these three emotions in addition to "Sadness" are the classes that can be distinguished when the unique available information is the still image.

A. Image Data Set

According to the classical model of Ekman [2], facial expressions may represent six different emotions: (1) "Happiness" (2) "Sadness" (3) "Fear" (4) "Surprise" (5) "Anger" and (6) "Aversion". However, recent studies by R. Jack *et al.* [4] claim that the distinction of emotions based exclusively on facial still images is feasible only on the following four classes: (1) "Happiness" (2) "Sadness" (3) "Fear" and/or "Surprise" and (4) "Anger" and/or "Aversion". R. Jack *et al.* [4] consider that the distinction between "Fear" and "Surprise" and between "Anger" and "Aversion" requires additional contextual information not present in the still images. In this context, the emotional classes considered in the present work are (a) "Aversion" (b) "Happiness" and (c) "Fear".

The images used in this work were extracted from the Expanded Cohn-Kanade dataset (CK+) [8]. This dataset consists of 593 sets of image sequences taken from 123 subjects while they express different emotions. Each sequence of images starts from a neutral or approximately neutral expression, continuing until the full expression of emotion is elicited. The images are in grayscale.

The CK+ dataset has a different number of images taken for each subject in each emotional class, and a different number of subjects for each class. For this reason, this work makes use of the last three images of an emotion expression sequence for each subject in each class, when the expression shape is already established. Subjects with less than three images for a given emotional expression were discarded.

The class "Aversion" of the CK+ database is the one that has the least number of subjects with at least three images. It was found that this class is limited to 51 subjects. For a uniform representation of the classes, "Happiness" and "Fear", we also limited the number of subjects to 51 in these classes. Thus, each of the classes were represented by three images of 51 subjects (the total number of images per class is 51 subjects $\times$ 3 images $=$ 153 images, and the total number of images is 153 images $\times$ 3 classes $=$ 459).

From each class composed of three images of 51 subjects, three training and three test subsets were created, the Data Groups 1 to 3 (Fig. 1). Each of these subsets have images from 36 of the 51 subjects (70.6% of the subjects) separated for training images and the remaining 15 subjects (29.4%) are separated for testing. There are no repetitions of subjects between the three subsets of test and training in the same class,

TABLE I
TRAINING AND TEST DATA SET SUMMARY (SECTION III-A).

Subjects per class	51
Images per subject	3
Total of images per class	153
Classes	"Aversion", "Happiness" and "Fear"
Training	36 subjects per class (108 images)
Test	15 subjects per class (45 images)

and the samples were selected randomly. Table I summarizes the organization of the training set and test set. Fig. 1 shows how the original data set of each class was decomposed into three subsets and Fig. 2 shows examples of training and test images for each class.

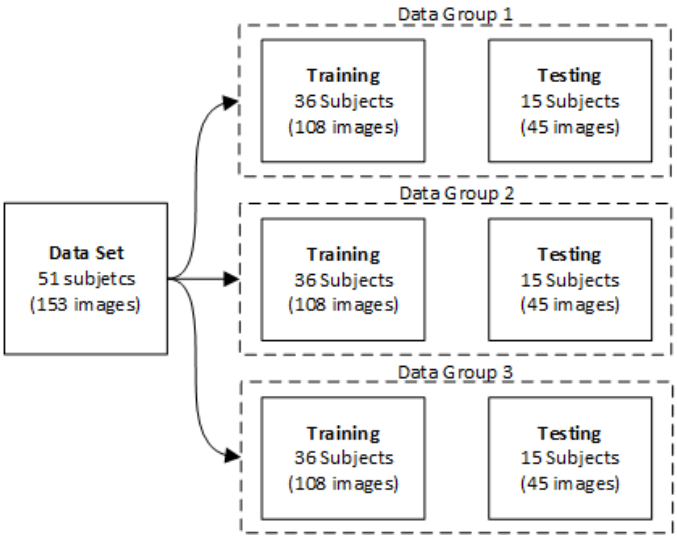

Fig. 1. Training and evaluation dataset organization for each class. Each class contains data from 51 subjects (153 images). The original set is split into three new subsets, Data Groups 1 to 3. These three Data Groups are used in three distinct processes of training and evaluation of the classifier system, as explained in detail in Section III-A. There is no subject overlap in test data among the groups.

Fig. 2. Examples of the three classes of images: (a) "Aversion" (b) "Happiness" and (c) "Fear". These are 176×156 pixels cropped images from the Extended Cohn-Kanade database (CK+) [8].

B. Classification System

Conventionally, classification systems rely on *ad hoc* representative characteristics extracted from the input data. These characteristics are used as inputs to a classification mechanism that relates a given input image to one of the available classes [13].

In this work, the system inputs are images described in Section III-A. The characteristics extracted from these images are obtained by a deep neural network [6], specifically a convolutional neural network previously trained and publicly available, known as Alexnet [7]. The classifier is an linear kernel SVM with an one-versus-one coding design [14]. Since the CNN is pretrained, only the SVM requires training. The training process is carried out from still images of subjects displaying emotions as described in Section III-A. The overall structure of the proposed system is shown in Fig. 3.

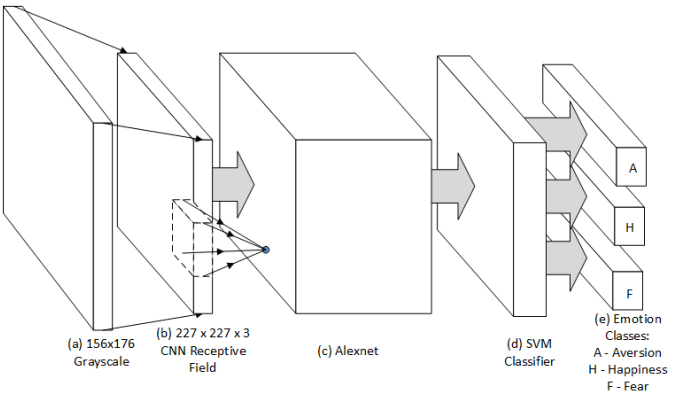

Fig. 3. Architecture of the system. (a) Input: greyscale image with width of 156 pixels and height of 176 pixels derived from the Extended Cohn Kanade dataset [8]. (b) Input image is resized to 227×227 pixels and replicated in two additional dimensions (composing substitute layers for the three layers R, G and B required by the Alexnet [7]). (c) Convolutional layers of the Alexnet deep neural network. (d) SVM classifier inputs are the features generated by the Alexnet. (e) Facial expressions classes: "Aversion", "Happiness" and "Fear".

Input images to the proposed architecture consist of grayscale cropped images with width of 156 pixels and height of 176 pixels taken from the CK+ database, which contains faces expressing the emotions "Aversion", "Happiness" and "Fear". Cropping the images is necessary to limit them to the region around the eyes, mouth and nose of the subject (Fig. 2). An automatic cropping algorithm based on Haar cascade face detection [15] was developed in order to facilitate this task. This algorithm determines the region of interest (ROI) using Haar cascade face detection, it finds the central point and crops the ROI (157×176 pixels) from the image. The only additional preprocessing applied to the images is the intensity mean pixel to pixel subtraction. The features extracted from these images by the CNN are used to train and test the SVM classifier, according to the description in Section III-A.

Input images must be resized to the required input dimensions of Alexnet. The images are resized to 227×227 pixels using bicubic interpolation. Alexnet requires three channel images at the input (RGB format). Since the original CK+ images

are grayscale, a three channel representation is obtained by simply replicating the image into the other two channels.

Images are then processed by the Alexnet deep neural network [7], which is composed of five layers of convolutional neural filters [6] structured as a hierarchical set of image feature filters. Beyond the convolutional layers, the original architecture has three layers of fully connected (FC) neurons, which perform the classification. Associated with the last layer there is a softmax structure. In the architecture presented in this work, the five convolutional layers extract characteristics from input images and the outputs of the last convolutional layer *Conv5* are used as inputs for an SVM classifier [16], [17]. The other layers of Alexnet network (FCs and softmax layers) are not used. Table II presents a summarized version of Alexnet.

The supervised learning module of the proposed system corresponds to an SVM classifier [13]. Its inputs are the characteristics obtained from the outputs of the layer *Conv5* of Alexnet. The specific architecture uses binary learners and one-versus-one coding design [18]. The output classes generated by the SVM correspond to the three the emotions expressed in the images: "Aversion", "Happiness", and "Fear". That is, the set of emotions that we are interested in detecting from the still images.

TABLE II
SUMMARY OF THE ALEXNET DEEP NEURAL NETWORK ARCHITECTURE [7].

Layer name	**Structure**
Input image	RGB, 227×227 pixels
Conv1	Convolutional layer: 96 filters of characteristics with receptive field of $11 \times 11 \times 3$
Relu1	ReLU
Norm1	Cross channel normalization with 5 channels per element
Pool1	Max Pooling with respective field of 3×3, stride 1×1 and padding 0×0
Conv2	Convolutional layer: 256 filters of characteristics with receptive field of $5 \times 5 \times 48$
Relu2	ReLU
Norm2	Cross channel normalization with 5 channels per element
Pool2	Max Pooling with respective field of 3×3, stride 1×1 and padding 0×0
Conv3	Convolutional layer: 384 filters of characteristics with receptive field of $3 \times 3 \times 256$
Relu3	ReLU
Conv4	Convolutional layer: 384 filters of characteristics with receptive field of $3 \times 3 \times 192$
Relu4	ReLU
Conv5	Convolutional layer: 256 filters of characteristics with receptive field of $3 \times 3 \times 192$
Relu5	ReLU
Pool5	Max Pooling with respective field of 3×3, stride 1×1 and padding 0×0
FC6	Layer with 4096 fully connected neurons
FC7	Layer with 4096 fully connected neurons
FC8	Layer with 1000 fully connected neurons
Prob	Softmax

IV. RESULTS

Since Alexnet contains a pretrained CNN, just the SVM module of the proposed system required training. The training trial (training-test cycle) was repeated ten times for each subset of images (Data Groups 1, 2 and 3 in Fig. 1), and the best performance among these ten trials is presented in the confusion matrices in Tables III, IV and V, for each Data Group of the three classes.

Table VI summarizes the accuracy obtained for each Data Group of the three classes. Table VII presents the global performance obtained by the proposed system. It is possible to note that the mean accuracy was 98.52%.

TABLE III
CONFUSION MATRIX - DATA GROUP 1. BEST RESULTS IN TEN TRAINNING TRIALS. AVERAGE ACCURACY: 100.00%.

		Prediction [%]		
		Aversion	Happiness	Fear
Actual Class	Aversion	100.00	0.00	0.00
	Happiness	0.0	100.00	0.00
	Fear	0.0	0.0	100.00

V. CONCLUSION

This work presented and evaluated a classifier for three emotional states expressed by subjects in still images. The adopted classification architecture consists of a CNN and an SVM classifier. The purpose of the CNN is to extract features of the input images. A pretrained Alexnet CNN was used, which allowed the use of a small dataset to train the SVM classifier only.

To the best of our knowledge, the application of a pretrained CNN for emotional state classification with CK+ has not been explored in previous works. Furthermore, a relevant contribution of the proposed architecture concerns the image preprocessing stage. While other architectures rely on several stages of preprocessing in order to achieve expressive results, the one proposed here comprises just an automatic cropping and an intensity mean pixel to pixel subtraction.

The results achieved in this work indicate that the proposed architecture is promising for the recognition of the emotional states. They also provide some insights into directions for future studies. Some topics that could be further explored are fine-tuning the CNN using data augmentation, using contextual information in order to obtain a classifier for the six emotions, and incorporating strategies that take into account the fact that regions of a face contain different amounts of information required to classify the emotional states [19].

TABLE IV
CONFUSION MATRIX - DATA GROUP 2. BEST RESULTS IN TEN TRAINNING TRIALS. AVERAGE ACCURACY: 100.00%.

		Prediction [%]		
		Aversion	Happiness	Fear
Actual Class	Aversion	100.00	0.00	0.00
	Happiness	0.0	100.00	0.00
	Fear	0.0	0.0	100.00

TABLE V
CONFUSION MATRIX - DATA GROUP 3. BEST RESULTS IN TEN TRAINNING TRIALS. AVERAGE ACCURACY: 95.56%.

		Prediction [%]		
		Aversion	Happiness	Fear
Actual Class	Aversion	100.00	0.0	0.00
	Happiness	13.33	86.67	0.00
	Fear	0.0	0.0	100.00

TABLE VI
AVERAGE ACCURACY [%] PER CLASS AND DATA GROUP.

Data Group	**Aversion**	**Happiness**	**Fear**
1	100.00	100.00	100.00
2	100.00	100.00	100.00
3	100.00	100.00	86.67
Average Accuracy [%]	100.00	100.00	95.56

TABLE VII
GLOBAL PERFORMANCE.

Data Group	**Average Accuracy [%]**
1	100.00
2	100.00
3	95.56
Average Accuracy [%]	98.52

REFERENCES

[1] C. Darwin, *The Expression of the Emotions in Man and Animals*. D. Appleton and Company, 1899.

[2] D. Keltner and P. Ekman, *Handbook of Emotions*, ch. 15 - Facial Expression of Emotion, pp. 151–249. Guilford Publications, Inc., 2nd ed., 2000.

[3] E. M. Provost, Y. Shangguan, and C. Busso, "Umeme: University of michigan emotional mcgurk effect data set," *IEEE Transactions on Affective Computing*, vol. 6, pp. 395–409, Oct 2015.

[4] R. E. Jack, O. G. Garrod, and P. G. Schyns, "Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time," *Current Biology*, vol. 24, no. 2, pp. 187 – 192, 2014.

[5] M. W. Schurgin, J. Nelson, S. Iida, H. Ohira, J. Y. Chiao, and S. L. Franconeri, "Eye movements during emotion recognition in faces," *Journal of Visualization*, vol. 14, no. 13, 2014.

[6] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning." Book in preparation for MIT Press, 2016.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

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

[9] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *2015 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–6, Sept 2015.

[10] F. S. Hsu, W. Y. Lin, and T. W. Tsai, "Automatic facial expression recognition for affective computing based on bag of distances," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pp. 1–4, Oct 2013.

[11] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Transactions on Cybernetics*, vol. 45, pp. 1499–1510, Aug 2015.

[12] A. T. Lopes, E. de Aguiar, and T. Oliveira-Santos, "A facial expression recognition system using convolutional networks," in *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*, pp. 273–280, Aug 2015.

[13] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC, 2006.

[14] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *J. Mach. Learn. Res.*, vol. 1, pp. 113–141, Sept. 2001.

[15] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1, pp. I–511–I–518 vol.1, 2001.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, June 2014.

[17] F. J. Huang and Y. LeCun, "Large-scale learning with SVM and convolutional for generic object categorization," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 1, pp. 284–291, June 2006.

[18] J. Zhou, H. Peng, and C. Y. Suen, "Data-driven decomposition for multi-class classification," *Pattern Recognition*, vol. 41, no. 1, pp. 67 – 76, 2008.

[19] M. W. Schurgin, J. Nelson, S. Iida, H. Ohira, J. Y. Chiao, and S. L. Franconeri, "Eye movements during emotion recognition in faces," *Journal of Vision*, vol. 14, no. 13, p. 14, 2014.