Multi-subject Continuous Emotional States Monitoring by Using Convolutional Neural Networks

Dennis Núñez Fernández
Universidad Nacional de Ingeniería
Lima, Peru
dnunezf@uni.pe

***Abstract*—Tracking of emotional states is very important for building intelligent systems, to have an efficient human-machine interaction and have a better understanding of human behavior. Nevertheless, most of the state-of-the-art works in emotion recognition employ complex algorithms, which are difficult to implement in real-time on devices with low computational resources. This work emphasizes on fast and effective techniques in order to efficiently recognize facial expressions. In this regard, we propose a fast face recognition model based on Local Binary Pattern operator and a straightforward Convolutional Neural Network for emotion classification. The results of our recognition system show an accuracy of 66.5% in the FER2013 wild dataset, which is near to state-of-the-art techniques but uses by far less computational resources.**

***Index Terms*—Emotional States, Intelligent Systems, Human-Machine Interaction, Convolutional Neural Networks.**

I. Introduction

Emotions are an important way to interact with other people independently from cultural diversities and ethnicity. Therefore, implementation of emotion recognition systems will take a more important role in future human-machine interaction systems, cognitive robotics, intelligent systems and more. However, emotion recognition is a difficult task since each person expresses their emotions in a different way. In this regard, the capability to recognize human emotions based on facial expressions has been an essential research area in Computer Vision during the last decades.

In recent years, advances in Microelectronics have allowed to exponentially increase the computational power of processors. This has allowed the implementation complex Artificial Intelligence (AI) algorithms in several areas like data mining, medical diagnosis and others. In fact, Deep Learning, has shown promising results in emotional state recognition.

In this paper, we implement an application of Computer Vision and Deep Learning. Our proposed system monitors emotional states of pre-recorded subjects by using a custom CNN model which balances a high accuracy rate and the most simple CNN architecture. To accomplish this, we combine two Computer Vision and Deep Learning algorithms: a Local Binary Patterns Histograms (LBPH) technique for fast face recognition and a lightweight Convolutional Neural Networks (CNNs) for quick emotional state recognition.

In Section 2, we outline the state-of-the-art on emotional state recognition. Later, our approach is described in detail in Section 3 and the obtained experimental results are reported in Section 4. Finally, the paper concludes in Section 5 with some comments and remarks.

II. Related Work

Recognition of emotional states based on facial features has been deeply studied in several areas from Psychology and Neuroscience to Computer Science. State-of-the-art techniques summarized in [1] gives an updated understanding about this topic and compares several techniques.

Almost all of the works related to emotion recognition are based on complex algorithms, which makes difficult to be implemented in real-time applications. In [2], the authors develop a CNN model that recognizes only three emotional states (happy, neutral and talking) with an accuracy of 97.6%. The recognition system proposed in [3] utilizes Local Directional Number Pattern (LDN) to find micropatterns and then extracting the directional information, obtaining 92.9% but based on a small database.

A more recent work [4], employes a CNN with 7 hidden layers and minimization of hinge loss. This model achieves an accuracy of 61.29% for six emotions. In [5], the authors develop an embedded recognition system based on K-Nearest Neighbor (K-NN) algorithm for Regression Modelling and LBP features to recognize six emotional states. However, this method shows an accuracy rate of only 47.44%. Additionally, [6] proposes an SVM based on facial landmarks, but only achieve an average accuracy of 70.65% for three different emotions.

III. Proposed Method

A. Overview

Our system is intended to recognize previously recorded subjects, track their emotional states and plot the results of each person over time. Furthermore, this recognition system was designed to work in real-time on devices with standard computing resources. To accomplish with these computational demands, our system processes grayscale images captured with a constant time interval.

978-1-7281-3643-1/19/$31.00 ©2019 IEEE

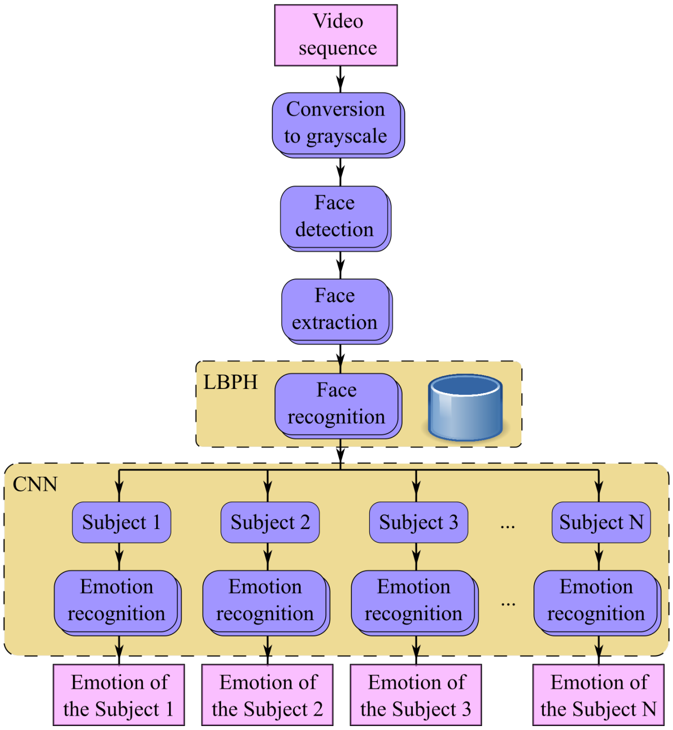

Fig. 1. Diagram of the proposed recognition system.

The diagram of the proposed system is presented in Fig. 1. The system starts converting the input RGB image to grayscale since facial features are independent of skin color. Another important reason to use grayscale images is due Local Binary Pattern Histogram (LBPH) and Deep Learning techniques are quite complex, therefore, conduct these algorithms on original RGB images demand more computational resources and a longer response time. After conversion of images to grayscale format, the system performs face detection to find all faces in the image. Face detection employs a LBP-based classifier, a Machine Learning technique that uses LBP features to conduct recognition. Later, face region is extracted and scaled to 48x48 pixels. Next, face recognition identifies pre-recorded subjects by using LBPH algorithm. Finally, emotion recognition employs a CNN to distinguish emotional states for each pre-identified subject. As well, these states are plotted in real-time for each subject to register its variation over time.

B. Face Detection

For face detection we deploy a LBP feature-based cascade classifier [7]. We prefer LBP over Haar features due LBP does all calculations in integers, and therefore much faster. This makes of LBP features an excellent choice for devices with standard computing resources. The LPB-based classifier works as follows: first, since LBP features of 24x24 pixels result in about 160,000 different features, we only need the most important features for face detection. The selection the these features are conducted by an Adaboost algorithm (short for Adaptive Boosting). This constructs a strong classifier as a linear combination of these important features. It is obvious that an image will be composed of a large number of sub-windows without faces, therefore, the algorithm for face detection should be focused on a fast rejection of non-face regions. To accomplish this target, the algorithm employs cascades of strong classifiers, see Fig. 2.

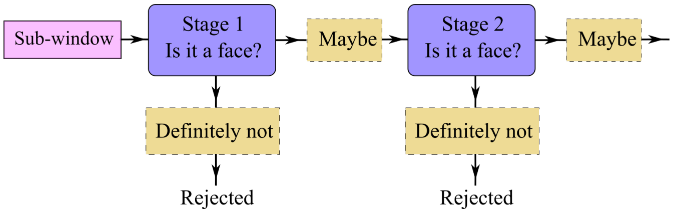

Fig. 2. Stages of the LBP feature-based cascade classifier.

C. Face Recognition

Some popular algorithms for face recognition are based on Linear Discriminant Analysis (LDA), which uses training labels (e.g. FisherFaces), Principal component analysis (PCA) (e.g. EigenFaces), and Local Binary Patterns Histograms (LBPH). In this work, we implemented an LBPH algorithm [8] since it demands low computational resources and its robustness against monotonic gray scale transformations and different resolutions [9]. The LBPH algorithm works as follows: first, the algorithm is trained on different samples for each person, then, LBP operation is applied. After, the transformed image is divided by regions of the same size in order to find local features. Finally, the histogram of each region is extracted and all histograms are concatenated into a single histogram, which represents each person. Fig. 3 depicts this procedure. In a nutshell, given an input image, the LBPH algorithm will extract its LBPH, compare with the stored histograms and return the label corresponding to the closest histogram.

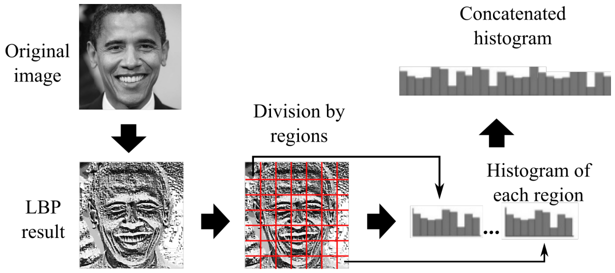

Fig. 3. LBP histogram obtained from a single person.

D. Emotional States Recognition

CNNs have several advantages compared to other techniques such as high accuracy reported on similar recognition tasks, focuses on local spatial coherence, independence from prior knowledge. The Caffe framework [10] was used due to its expression, speed, and modularity to design and train deep learning models. But the main reason to use Caffe in this work is because it is faster than Tensorflow, Theano and Keras for inference [11], therefore, Caffe will contribute us to implement a recognition system with a fast response time.

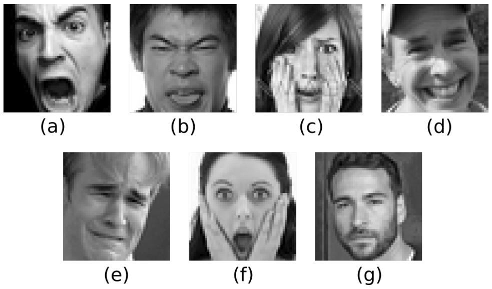

Fig. 4. Samples of images used for training and testing our neural network. (a) angry, (b) disgust, (c) fear, (d) happy, (e) sad, (f) surprise, (g) neutral

E. FER2013 Wild Dataset

A number of datasets for emotional state recognition can be found on the Internet. However, Facial Expression Recognition 2013 (FER2013) dataset [12] was selected for this work because its more diverse and faces without posing are more closely to reflect the distribution of faces in real-time video capturing. This dataset is composed of 14,553 grayscale images of 48x48 pixels. Furthermore, the dataset is divided into seven classes: angry, disgust, fear, happy, sad, surprise and neutral, see Fig. 4. In order to increase the number of images for training and testing, we implemented data augmentation. Therefore, shifting of 4 pixels on horizontal/vertical axis, rotation of 15° on clockwise/counterclockwise direction and horizontal reflection were applied to the original images. After data augmentation, (14,553)*5*3*2 = 436,590 images were obtained. To deploy our CNN model, the dataset is split into ∼90% for training (388,597 images) and ∼10% for testing (48,007 images).

F. Convolutional Neural Network

The recognition system has two main targets: obtain a low latency on standard computers without GPU support and get a high accuracy for emotion recognition. To achieve these targets we work with small images of 48x48 pixels, and use a CNN with few learnable parameters. This CNN is formed by two convolutional layers followed by one non linearity (ReLU) activation function and one max-pooling layer, two full-connected (FC) layers of 500 neurons each one and a 7-way softmax layer, see Fig. 5. However, since cascades of linear convolutions addresses to a linear system, a ReLU activation function is employed after each convolutional layer to add non linearity to the network. As well, ReLUs offers a mechanism for improving translation invariance, which helps with position independence. To sum up, the proposed model is composed by only 1M learnable parameters, which is significantly less than other CNNs like AlexNet (60M parameters) or GoogleNet (6.8M parameters) [13].

G. Training Process

Training was done by 10,000 forward-propagation and error back-propagation iterations as well as updating the weights according these mechanisms. In this regard, learning rate is used to control how much we set the weights of our CNN with respect to loss gradient as follows: First, the learning rate (lr) starts with an initial value called base learning rate (0.01), then the learning rate is adjusted over time according to the ’Inverse Decay’ learning policy. In other words, the learning rate is defined by: $lr = 0.01 * (1 + 0.0001 * iter)^{-0.75}$.

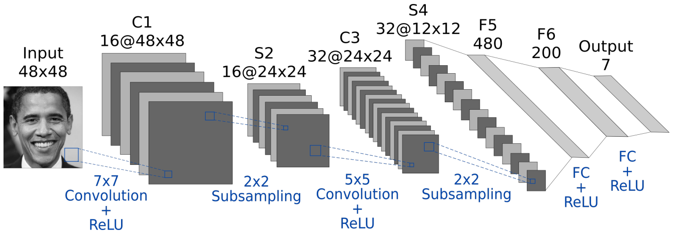

Fig. 5. Architecture of the CNN for emotional state recognition.

As explained above, training phase is conducted by a series of forward-propagation and back-propagation steps, which are computationally expensive, executed thousands of times before achieving a desirable response. To satisfy these computational demands we used a GPU NVIDIA GeForce GTX 1050 Ti.

IV. Experimental Results

A. Results of the CNN model

The performance of our CNN model is evaluated on the FER2013 wild dataset by calculating the accuracy, standard deviation, max. accuracy and plotting the confusion matrix. The results show a max. accuracy of 66.9%, an average accuracy of 66.5% and a standard deviation of 1.03 (average of 20 consecutive iterations). Other metrics to evaluate the performance of our recognition model is the confusion matrix. This allows us visualize the number of classification errors per classes, see Fig. 6. Misclassifications are mainly produced by similarities in face expressions, for instance, ’surprised’ is usually confused with ’angry’ and ’sad’ with ’annoyed’. Achieved accuracy apparently is low, however state-of-the-art works on FER2013 wild dataset are around 73% [14]. Indeed, precisely recognition of facial expressions is a difficult task, even for a person. Reference [12] indicates that human accuracy on FER2013 is around 65%. A pitfall with this dataset is that it suffers from some label errors and blank images due to the way it was collected [12]. So, our results could have been affected by these mistakes.

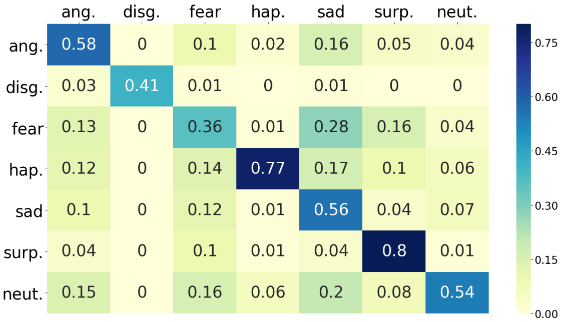

Fig. 6. Confusion matrix of the proposed model.

Table I compares our results with previously published results. Despite several advantages of state of the art SVM models, these achieves a low accuracy rate for emotion recognition. On the other hand, the proposed CNN has a smaller architecture and less learnable parameters, while maintaining the same accuracy compared to other CNN models. VGG, Inception and ResNet architectures were not considered since these have a higher response time [10] compared with the architectures included in Table I.

TABLE I
ACCURACY FOR THE PROPOSED CNN, AND RESULTS FROM SIMILAR LIGHT MODELS (I: INCEPT., C: CONV., P: POOL., F: FULL.)

Reference	Accuracy	Architecture	Params.
SVM [15]	43.2%	-	-
SVM+HOG [15]	45.9%	-	-
AlexNet CNN [16]	61.1%	CPCPCCCPFF	60M
Inception CNN [16]	66.4%	CPCPIIPIPFF	16M
CNN [15]	66.7%	CCPCPFFF	24M
Proposed CNN	**66.5%**	**CPCPFF**	**1M**

B. Results of the complete system

The recognition system was developed in C++ language and implemented on a personal computer: Intel Core 7 Octa-Core CPU @3.8 GHz, 12 GB RAM. Fig. 7 shows the experimental results of the recognition system on a recorded video (https://youtu.be/ieWPPhJV3e0). This shows a press conference given by the former president B. Obama and the current president V. Putin. As we can see in the video, most of the time the former president B. Obama answers some questions from the press while president V. Putin remains silent. So, the graphic of emotions of B. Obama shows more variations due facial expressions presented during his speech. At the moment indicated by the black arrow, B. Obama and V. Putin smile, which is reflected in the emotions monitored. Misclassifications happen due face identification and emotion recognition are affected by face rotations and lighting conditions. Additionally, the response time is 0.23 seconds (average of 20 consecutive iterations).

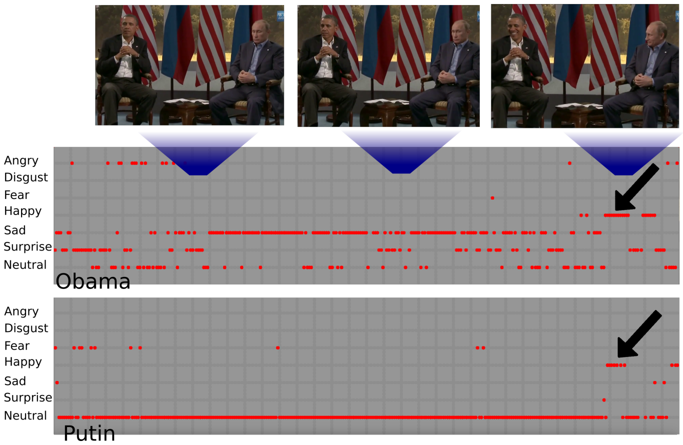

Fig. 7. Results of the recognition system on a recorded video.

V. CONCLUSIONS

In this paper, we introduce a system for multi-subject emotions monitoring. In this regard, this system is able to recognize seven basic emotions on personal computers with limited processing resources. We make use of techniques that demand low processing power: the Local Binary Pattern Histogram (LBPH) algorithm for facial recognition system and a Convolutional Neural Network (CNN) with few learnable parameters for emotion recognition. In fact, the experiments show promising results, obtaining an accuracy of 66.5% and a response time of 0.23 seconds. Therefore, the system can be used in several applications like human-machine interaction, audience response, tele-assessment system, and more. In addition to this, the results make of this paper a reference point for future works in similar recognition tasks.

REFERENCES

[1] Mehta, D.; Siddiqui, M.F.H.; Javaid, A.Y. Facial Emotion Recognition: A Survey and Real-World User Experiences in Mixed Reality. Sensors 2018, 18, 416.

[2] Matsugu, M.; Mori, K.; Mitari, Y.; Kaneda, Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw. 2003, 16, 555559

[3] Rivera, A.R.; Castillo, J.R.; Chae, O.O. Local directional number pattern for face analysis: Face and expression recognition. IEEE Trans. Image Process. 2013, 22, 17401752.

[4] Yu, Z.; Zhang, C. Image based static facial expression recognition with multiple deep network learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 913 November 2015; ACM: New York, NY, USA, 2015; pp. 435442.

[5] S. Turabzadeh, H. Meng, R. M. Swash, M. Pleva and J. Juhar, ”Real-time emotional state detection from facial expression on embedded devices,” 2017 Seventh International Conference on Innovative Computing Technology (INTECH), Luton, 2017, pp. 46-51.

[6] B. T. Nguyen, M. H. Trinh, T. V. Phan and H. D. Nguyen, ”An efficient real-time emotion detection using camera and facial landmarks,” 2017 Seventh International Conference on Information Science and Technology (ICIST), Da Nang, 2017, pp. 251-255.

[7] P. Viola and M. Jones, ”Rapid object detection using a boosted cascade of simple features,” Proceedings of the 2001 IEEE Computer Society Conf. on Comp. Vision and Pattern Recogn. CVPR 2001, pp. I-511-I-518 vol.1.

[8] Ahonen T., Hadid A., Pietikinen M. (2004) Face Recognition with Local Binary Patterns. In: Pajdla T., Matas J. (eds) Computer Vision - ECCV 2004. Lecture Notes in Computer Science, vol 3021. Springer, Berlin.

[9] P. Jaturawat and M. Phankokkruad, ”An evaluation of face recognition algorithms and accuracy based on video in unconstrained factors,” 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Batu Ferringhi, 2016, pp. 240-244.

[10] Y. Jia, et al.. Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the 22nd ACM international conference on Multimedia (MM 14). ACM, New York, NY, USA, 675-678

[11] S. Shi, Q. Wang, P. Xu and X. Chu: Benchmarking State-of-the-Art Deep Learning Software Tools, arXiv:1608.07249v7, 2017.

[12] Goodfellow I.J. et al.. Challenges in Representation Learning: A Report on Three Machine Learning Contests. In: Lee M., Hirose A., Hou ZG., Kil R.M. (eds) Neural Information Processing. ICONIP 2013. Lecture Notes in Computer Science, vol 8228. Springer, Berlin, Heidelberg

[13] F. Altenberger, and C. Lenz: A Non-Technical Survey on Deep Convolutional Neural Network Architectures, arXiv:1803.02129v1, 2018.

[14] Pramerdorfer, C., Kampel, M.: Facial Expression Recognition using Conv. Neural Networks: State of the Art, arXiv:1612.02903v1, 2016.

[15] M. Quinn, G. Sivesind, G. Reis, Real-time Emotion Recognition From Facial Expressions, Stanford University, Tech. Rep., 2017.

[16] A. Mollahosseini, D. Chan and M. H. Mahoor, ”Going deeper in facial expression recognition using deep neural networks,” 2016 IEEE Winter Conf. on Applications of Comp. Vision (WACV), Lake Placid, NY, 2016.