Real Time Facial Expression Recognition System Based on Deep Learning

Jose Carlos Bustamante, Ciro Rodriguez, Doris Esenarro

***Abstract*— The automatic detection of facial expressions is an active research topic, since its wide fields of applications in human-computer interaction, games, security or education. However, the latest studies have been made in controlled laboratory environments, which is not according to real world scenarios. For that reason, a real time Facial Expression Recognition System (FERS) is proposed in this paper, in which a deep learning approach is applied to enhance the detection of six basic emotions: happiness, sadness, anger, disgust, fear and surprise in a real-time video streaming. This system is composed of three main components: face detection, face preparation and face expression classification. The results of proposed FERS achieve a 65% of accuracy, trained over 35558 face images..**

Keywords: Real Time, Affective Computing, Facial Expression Recognition System, Deep Learning, Emotion Classification, Convolutional Neural Networks.

I. INTRODUCTION

Emotions have a main role in human life, people not only communicate with others verbally, instead we use different nonverbal means such as hand and body gestures, tone of voice or facial expressions, which are used to express feelings. For that reason, Psychology has played an important role in understanding how people express their emotions and developing concepts which are used by technology researchers to design systems which can automatically recognize feelings [1].

Also, it is wide accepted from psychological theory that human emotions can be classified into six basic emotions: surprise, disgust, fear, sadness, happiness and anger. The major role in expressing these emotions are facial motion or expressions and speech, though tone of voice is more probably to be intentionally modified to communicate different feelings [2]. Hence, facial expression is taken in this paper as the main expression channel to recognize human emotions.

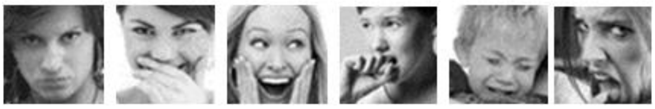

Figure 1. Six basic emotions: anger, happiness, surprise, fear, sadness and disgust.

Supporting the idea to use facial expressions as the main channel to express emotions, a research made by psychologist Mehrabian shows that the greatest proportion of the way to express information in human communication is facial expression – up to 55%, and another 38% is auxiliary language. The remaining 7% is surprisingly expressed in oral language [3].

Some of the applications of this topic include automated security, interactive robots, human-machine interaction computers, diagnosing mental diseases, games, education, entertainment, among others [4, 5].

Fortunately, thanks to advances in technology, there is currently computer equipment capable of processing large amounts of information with a speed much higher than previous years, as well as new Artificial Intelligence techniques which can be used to solve the emotion recognition problem.

At the beginning of the automatic facial expressions recognition research, most studies and inputs were performed in cleaner datasets under well-controlled laboratory environments, which is not a rely way to ensure the accuracy of the emotion recognition, because in real world scenarios there could be a lot of noise clogging human faces [6].

Facial Expression Recognition (FER) methods can be divided into two categories: video sequence-based methods (dynamic recognition from video) and image-based methods (static recognition from a single image). The majority of previous studies in FER were using image-based methods, which cannot capture the temporal variability in consecutive frames in video sequences. This is a problem because there is important information in frames sequence that is not taken advantage of [4].

Deep learning as a technique for solving FER problems has recently become used in state of art proposals, most particularly convolutional neural networks (CNN) model because you don’t have to manually extract the features since the network does it for itself by using convolutions and pooling operations.

In this paper a real time face expression recognition system using deep learning is proposed, which takes image sequences from a video stream and automatically detects the human emotion, also a report of how the emotions have changed during the entire duration of the video is shown.

Also, we want to design and implement the proposed FERS in a simple and open access way, in order to be easy to understand, use and deploy by future works. Therefore, the proposed FERS will have three components: face detection, facial preparation and facial expression classification.

Revised Version Manuscript Received on 16 September, 2019.

Jose Carlos Bustamante, National University of San Marcos, Peru, South America, E-Mail: jose.bustamante9@unmsm.edu.pe

Ciro Rodriguez, National University of San Marcos, Peru, South America, E-Mail: crodriguezro@unmsm.edu.pe

Doris Esenarro, National University Federico Villarreal, Peru, South America, E-Mail: desenarro@unfv.edu.pe

II. RELATETED WORKS

Using deep neural networks (DNN), in [4] they propose a new FER method based on a hybrid deep learning model. This model contains three deep models, the first two are CNN, one for spatial processing and the other for time processing, while the last one is a deep belief network (DBN) model. Finally, they use support vector machine (SVM) to classify the corresponding facial expression.

A FERS optimized for being used as a mobile application was proposed by Myung [6], in which the architecture consists in face detection, feature extraction and facial expression classification. The features are extracted using active shape model (ASM) and then both SVM model and mouth status are used for detecting neutral expressions, this is made for generating dynamic features only when a nonneutral expression is found. Finally, the SVM classifier shows the result of the expression as one of the six basic emotions.

A multimodal FER and speech recognition system was proposed in [7], which combines the speech detection and the facial expressions of the user to generate text outputs with emoticons according to the emotion they express. For this purpose, they use a CNN that has nine layers, 3 maxpooling layers and 3 fully-connected layers. They use the OpenCV open-access library to detect the human face region from webcam streaming, then crop the face, convert it to gray scale and use the result as input for the CNN model to predict the emotion and transform it to an emoji. However, as this is a multimodal proposal, it’s not optimized for facial expressions.

Authors in [8] proposed three video-based models for FERS. The first model is a differential geometric fusion network (DGFN), the second model is the deep-facial-sequential network (DFSN) and finally the DFSN-I model is proposed, which is a combination of DGFN and DFSN to achieve better performance. The third model is proven to achieve better accuracy, though is hard to implement and understand.

Since mobile applications have to deal with use of limited resources, an emotion recognition system is proposed in [9], which captures video from the embedded camera of a smart phone and then representative fames are selected from the images. Later, the face areas are detected by the Viola-Jones algorithm since it works fast and its suitable for a real-time implementation. Finally, the bandlet transform is applied to the detected face area to feed the gaussian mixture model (GMM) classifier and show the results.

A novel frame for automatic facial expression recognition system is proposed in [10] which extract a facial feature from the image sequence at the apex period. It is composed of two components: peak expression frame detection and facial expression feature extraction from this frame, for this purpose they use double local binary pattern (DLBP) and Taylor feature pattern (TFP), however under uncontrolled environments like variant illumination, face poses or noise the performance of the system would be affected.

The problem of determining whether the current FER result is reliable or not is focused in [11]. They determine if the result is not reliable, and if that is the case, they search for images with similar emotions using the result image as query. Finally, the result image and the similar ones feed the classifier to enhance the prediction of emotion.

Since noise and occlusion are major problems when trying to extract facial features from a face image, authors in [12] propose a CNN with attention mechanism (ACNN) which emulates how people recognize facial expressions, basically when a face is blocked we may judge the expression according to the symmetric ¬part of the face or related face regions. Hence, in this proposal the ACNN pays more attention to unblocked and informative regions.

In [13], an efficient algorithm to improve the recognition accuracy is proposed, which uses a hierarchical deep neural network structure which can re-classify the results of the first classification step. For the feature extraction step, they use appearance and geometric features.

IV. OBJECTIVES

The main objective of the study is to recognize human emotions in real time.

We found convenient to specify the following secondary objectives in order to achieve the main objective:

Determine the best channel to recognize human emotions in real time from a machine.

Identify the techniques of recognition in real time according to the best channel.

Choose the techniques which best suits real time recognition on human emotions according to complexity, accuracy, time costing and related works.

Use the selected technique to develop the proposed real time FERS.

4. REAL TIME FACIAL EXPRESSION RECOGNITION SYSTEM

Facial expression recognition is a classification problem with a finite number of results, which are the six basic emotions. The major problem is to recognize the emotion given in each particular frame of a video stream, and also store them. We found it convenient to decompose our proposed FERS in three main components which work together in the task of recognizing a facial expression in a single frame, in that way we can apply that procedure in every frame is taken from the video sequence.

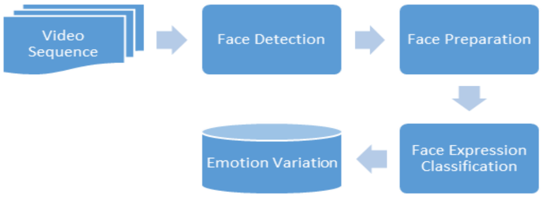

Figure 2. Proposed FERS architecture

The implementation of the proposed architecture FERS, shown in Fig. 2, has been made in Anaconda Python 3 Distribution, using OpenCV [14], Keras API and TensorFlow as backend.

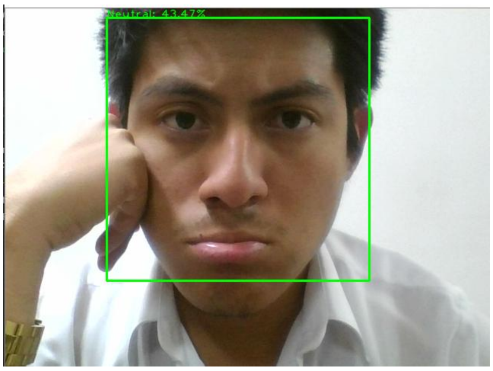

Figure 3. Proposed Real Time FERS detecting sadness.

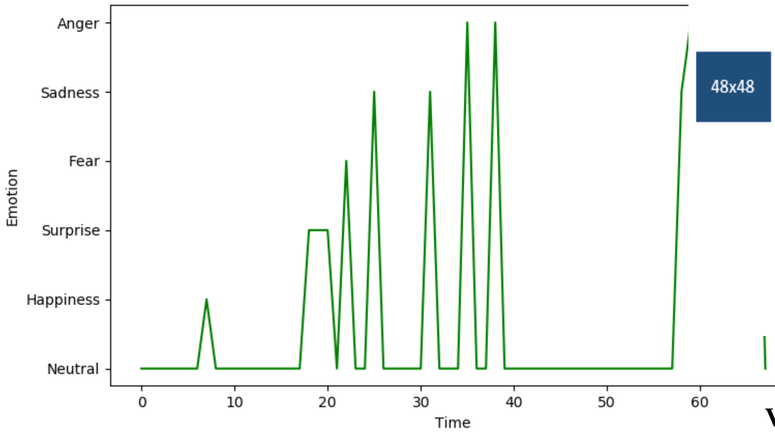

Figure 4. Emotion Variation Graph in an interval of 70 seconds.

Next, the main components will be explained:

4.1 Face Detection

The initial input of the system is a frame taken from the video sequence, then the frame feeds the face detection component, which detects zero, one or many faces in the frame and returns the coordinates of them.

For face detection component we used Viola-Jones algorithm [15], which is a machine learning based approach where a cascade function (HAAR cascade) is trained from a lot of positive and negative images, it also uses Adaboost for selection of best features extracted.

In this paper we didn't train the face detection component, since it is already trained in OpenCV.

The above explained algorithm only detect frontal faces in a frame because it was trained in this way.

4.2 Face Preparation

The set of coordinates taken from the previous component are used in this stage to extract the sub frame that contains the face or faces detected.

Then the image is converted to grayscale, reshaped and normalized, since the CNN was trained with this kind of images.

The necessary shape of the image is a 48 x 48 pixel image, with only one color channel.

The normalization is made by dividing the pixel by 255.0, which is the max value a pixel can reach.

4.3 Facial Expression Classification

We used deep learning for FER, the CNN model is trained with the open source dataset FER2013 provided by the FER Challenge 2013, which is composed of 35899 facial expression samples.

Once the model is trained, we can feed the CNN with the images taken from the previous component and predict the emotion.

Finally, we will collect all these predictions and make a graph in which the variation of emotions is printed

The CNN model proposed in this paper is a modification of CNN proposed in [16], decreasing the number of convolutional layers for less complexity and processing.

The CNN model proposed in this paper is composed as follows:

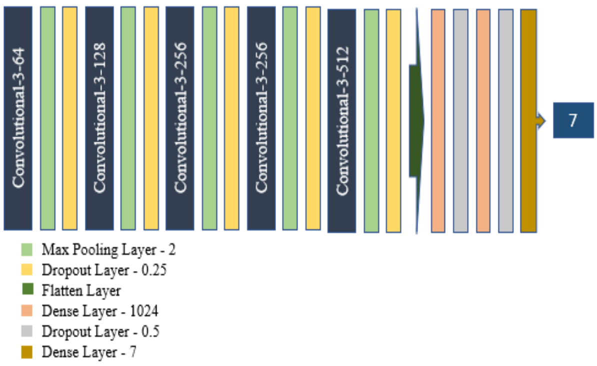

Figure 5. Proposed CNN Model

V. EXPERIMENTS AND VALIDATION

The dataset samples are divided into two groups: training samples and validation samples. The data distribution is as follow in Table 1. and Table 2.:

Table 1. Train data distribution

Emotion Labelled	Total Images
Angry	4462
Disgust	492
Fear	4593
Happy	8110
Sad	5483
Surprise	3586
Neutral	5572
Total	32298

Table 2. Validation data distribution

Emotion Labelled	Total Images
Angry	491
Disgust	55
Fear	528
Happy	879
Sad	594
Surprise	416
Neutral	626
Total	3589

We trained our model in Google Colab workspace because it has a 12GB GDDR5 VRAM GPU, which is used for better performance and training speed.

The training data is gray scaled, and each pixel value is between 0 and 255, so we normalize the data by dividing the pixel by 255.0.

Also, we had to transform labels array to categorical representation which is a 7 dimensional array.

We trained our model with the following parameters, as shown in Table 3.:

Table 3. Training parameters

Parameter	Value
Epochs	30
Batch Size	32
Learning Rate	0.0001

The training time was approximately 25 minutes with the data and parameters detailed above.

For validation we take into account two main metrics: accuracy rate and loss rate.

Accuracy rate measures how well the model predict the emotions in the given data, there are two different accuracy rates: accuracy obtained from training data and accuracy obtained from validation data, the comparative over the epochs is as follows:

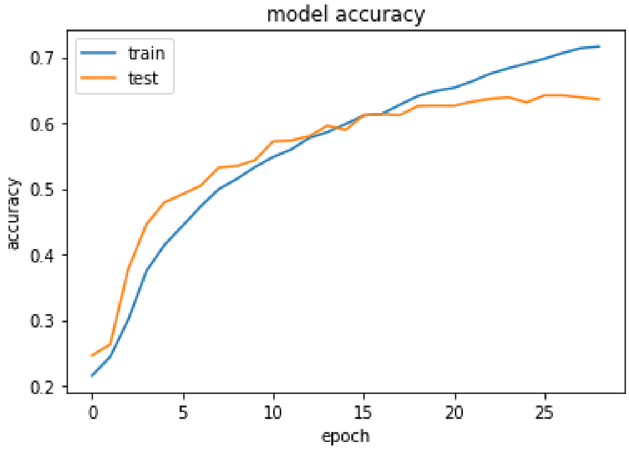

Figure 6. Accuracy rate comparison

Loss rate measures the value of cost function for the given data, there are two different loss rates: loss obtained from training data and loss obtained from validation data, the comparative over the epochs is as follows:

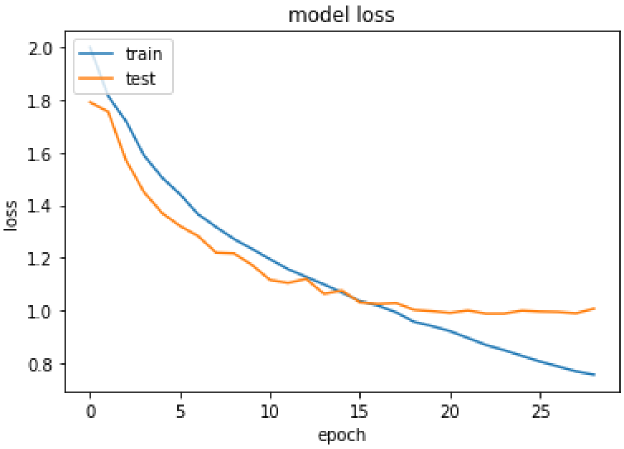

Figure 7. Loss rate comparison

As we can see in the above Figures 6 and 7, both the accuracy function and loss function are exponential functions, which reflects that the proposed model is responding and learning as expected.

The confusion matrix shown below in Table 4., reflects the performance of our trained CNN model:

Table 4. Confusion matrix

Predictions	Real Values						
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	253	17	33	33	74	11	70
Disgust	7	41	1	1	2	2	1
Fear	57	11	172	31	109	67	81
Happy	15	1	12	769	36	17	29
Sad	49	10	42	52	305	11	125
Surprise	9	2	13	22	8	347	15
Neutral	34	4	15	62	77	12	422

The normalized accuracy matrix shown below, reflects the accuracy of our trained CNN model by dividing the values by the total quantity of images labeled with the row label:

Table 5. Normalized confusion matrix

Predictions	Real Values						
	Angry	Disgust	Fear	Happy	Sad	Surprise	Neutral
Angry	**0.52**	0.03	0.07	0.07	0.15	0.02	0.14
Disgust	0.13	**0.75**	0.02	0.02	0.04	0.04	0.02
Fear	0.11	0.02	**0.33**	0.06	0.21	0.13	0.15
Happy	0.02	0.00	0.01	**0.87**	0.04	0.02	0.03
Sad	0.08	0.02	0.07	0.09	**0.51**	0.02	0.21
Surprise	0.02	0.00	0.03	0.05	0.02	**0.83**	0.04
Neutral	0.05	0.01	0.02	0.10	0.12	0.02	**0.67**

VI. FINDINGS & RESULTS

We can observe that the model accuracy rate in the test data didn't improve too much between epoch 20 and 30.

This also happens in the model error rate, which didn't decrease between the same epochs.

The model proposed has the highest accuracy rate in detecting happiness according to normalized confusion matrix.

Also, our model has the lowest accuracy rate in detecting fear.

In the normalized confusion matrix obtained, the main diagonal has the biggest values which means the detection accuracy is as expected.

VII. APPLICATIONS AND IMPROVEMENTS

The proposed model and FERS could be applied to many fields, such as sentiment analysis of students from a e-learning environment, a customer relationship management component to capture client's satisfaction on service, games using facial expressions to interact with characters, in hospitals or psychology centers to study patients' emotions.

Improvements in proposed model could include applying a service architecture to be more accessible for potential users, re-train the proposed CNN or applying data mining to the emotion variation storage.

VIII. CONCLUSIONS AND FUTURE WORKS

For future works, we encourage researchers to use the system and find issues in order to improve either system architecture, algorithms for a component or library used. Also, we encourage the use of novel methods in one or more of our proposed system components. We also encourage researchers to retrain the proposed CNN model with other datasets to improve its accuracy rate.

IX. ACKNOWLEDGEMENTS

We thank to the university, particularly our faculty for teaching and forming us as professional system engineers applying all the knowledge for the benefit of our society and contributing to science in general. We also thank our families and friends for always supporting us in this research.

REFERENCES

1. Ann, J. P Cherian & J. J Kizhakkethittam, «Overview on Emotion Recognition System» from 2015 International Conference on Soft-Computing and Network Security (ICSNS -2015), Coimbatore, (2015)
2. Busso, Z. Deng, S. Yildirim, M. Bulut, C. Min Lee, A. Kazemzadeh, S. Lee, U. Neumann & S. Narayanan, «Analysis of emotion recognition using facial expressions, speech, and multimodal information» from Proceedings of the 6th international conference on Multimodal interfaces, California, (2004)
3. Mehrabian, «Communication without words» Psychology Today, vol. 2, pp. 53-56, (1968)
4. S. Zhang, X. Pan, Y. Cui, X. Zhao & L. Liu, «Learning Affective Video Features for Facial Expression Recognition via Hybrid Deep Learning» IEEE Access, vol. VII, pp. 32297 - 32304, (2019)
5. A.Nicolai & A. Choi, «Facial Emotion Recognition Using Fuzzy Systems» from 2015 IEEE International Conference on Systems, Man, and Cybernetics, Hong Kong, (2015)
6. M. Hoon Suk, «Study of real-time facial expression recognition on noisy images and videos» The University of Texas at Dallas, Texas, (2018)
7. H. Wang & F. Shih, «Detect and Transmit Emotions in Online Chat using Affective Computing» Taiwan, (2018)
8. Y. Tang, X. Zhang & H. Wang, «Geometric-Convolutional Feature Fusion Based on Learning Propagation for Facial Expression Recognition» IEEE Access, vol. 6, pp. 42532 - 42540, (2018)
9. S. Hossain & M. Ghulam, «An Emotion Recognition System for Mobile Applications» IEEE Access, vol. 5, pp. 2281 - 2287, (2017)
10. Y. Ding, Q. Zhao, B. Li & X. Yuan, «Facial Expression Recognition from Image Sequence Based on LBP and Taylor Expansion» IEEE Access, vol. 5, pp. 19409 - 19419, (2017)
11. T. Doan, S. Kim, Y. Lu, S. Jung & C. Won, «Facial Action Units-Based Image Retrieval for Facial Expression Recognition» IEEE Access, vol. 7, pp. 5200 - 5207, (2017)
12. Y. Li, J. Zeng, S. Shan & X. Chen, «Occlusion aware facial expression recognition using CNN with attention mechanism» IEEE Transactions on Image Processing, vol. 8, pp. 2439 - 2450, (2018)
13. J. Kim, B. Kim, P. Roy & D. Jeong, «Efficient Facial Expression Recognition Algorithm Based on Hierarchical Deep Neural Network Structure» IEEE Access, vol. 7, pp. 41273 - 41285, (2019)
14. G. Bradski, «The OpenCV Library» Dr. Dobb's Journal of Software Tools, (2000)
15. P. Viola & M. Jones, «Rapid Object Detection using a Boosted Cascade of Simple Features» Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 511-518, (2001)
16. E. Barsoum, C. Zhang, C. Canton & Z. Zhang, «Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution» from International Conference on Multimodal Interaction, Tokyo, (2016)
17. H. Ribes Gil, «Desarrollo de un sistema de reconocimiento de emociones faciales en tiempo real» Barcelona, (2017)

AUTHORS

Jose Carlos Bustamante Falcón

Member of the artificial intelligence research group of the faculty of systems engineering at the National University Mayor de San Marcos, working as System Analyst / Programmer in Software Enterprise Services and as freelancer.

Ciro Rodriguez Rodriguez

Professor at the School of Software Engineering at the National University Mayor de San Marcos and also at the Computer Science School and Graduate School of the National University Federico Villarreal, with science studies at the Abdus Salam International Center for Theoretical Physics (ICTP) and the United States Particle Accelerator School (USPAS)

Doris Esenarro Vargas

Professor at the Faculty of Environmental Engineering and Graduate School of the National University Federico Villarreal, with studies in System Engineering, Architecture and Environmental Engineering.