A Real-Time Recognition System for User Characteristics Based on Deep Learning

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

***Abstract*—This paper describes an implementation of a novel real-time recognition system which is capable to identify important information from a single user such as gender, age, emotions and hand gestures. The key of this recognition system is the classification process. This is carried out by using several convolutional neural networks that were designed to achieve a high accuracy rate and acceptable response time making use of low computational resources. As a result, this recognition system could be useful in numerous applications like human-computer interaction, person identification, security control and others.**

***Index Terms*—Recognition System, Convolutional Neural Networks, Human-Machine Interaction.**

I. Introduction

Features extracted from user's face such as age, gender, emotions, as well as hand gestures are important in our daily activities and are essential for a diverse range of applications from automotive user interfaces, to entertainment, to human-computer interaction, to person identification, to security control, and more. For this reason, recognition of these features has been deeply studied during the last decades in areas such as Computer Vision and Machine Learning.

However, few works related to recognition of these features on a single real-time system has been implemented. A work proposed in [1] recognizes age, gender and emotions using convolutional neural networks (CNNs) and its own private dataset. In addition to this, [2] proposes a similar system for gender, smile, and age recognition using an all-in-one CNN, but requires high computational resources.

In this paper, a novel recognition system is proposed to recognize gender, age, emotions and hand gestures. In order to accomplish these targets, we designed and trained four small CNNs, which were focused on balance high accuracy, time response and power consumption. Also, the posposed system deals with an important problem presented in the previous papers: the lack of human-machine interaction.

Unlike comercial recognition systems, the system that we develop in this work does not use cloud computing, this means that doesn't need internet connection and user information is not exposed to security leaks or commercial purposes.

The structure of this paper is composed of three sections as follows: Second section discusses the methodology used to extract interested regions and classify the different user charatceristics. Third section gives an analysis of the overall system performance. Finally, the last section presents some conclusions and remarks.

II. Methodology

A. Overview

As mentioned before, our proposed system recognizes four user characteristics. Classification is carried out in real-time on a personal computer, processing images frame by frame in this way: First, the camera captures RGB images of size 640x480 pixels, later face and hand regions are extracted using LBP cascades. Then, pre-processing is perfomed in order to set the image as input for each CNN. Finally, the corresponding results are shown in the screen. The diagram of the whole system is presented in the Fig. 1. This system is composed by four CNNs since each characteristic such as gender or hand gesture depends of different regions of the body and different pre-processing techiques. The four CNNs were trained using the Caffe framework [3] since its expression, speed, and modularity make of this a good choice for our project. The entire system was implemented in C++ programming language and using OpenCV 3.3.0 libraries to make use of the computational resources in an optimal way.

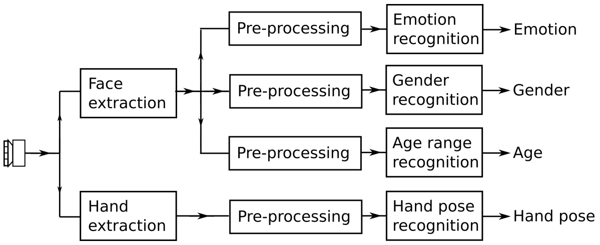

Fig. 1. Block diagram of the proposed recognition system.

B. Face and Hand Extraction

Face and hand detection are fundamental steps in the proposed system. In this regard, we implemented fast face and hand detectors [4] based on local binary pattern (LBP) classifiers. This technique combines four key concepts:

- LBP-like Features: simple rectangular features. Each face or hand image can be considered as a composition of small patterns which can be effectively perceive by the LBP operator.
- Integral Image: for rapid features detection. The value at any point (x, y) in the summed-area table is the sum of all the pixels above and to the left of (x, y).

978-1-5386-5491-0/18/$31.00 ©2018 IEEE

- AdaBoost or Adaptive Boosting: a machine-learning method. This helps to select small features that facilitates fast and easy computation. Unlike other methods, AdaBoost gives desired region of the object discarding unnecessary background.
- Cascade Classifier: to combine many features efficiently. Eliminates candidates by making stricter requirements in each stage with later stages being much more difficult for a candidate to pass. Candidates exit the cascade if they pass all stages or fail any stage.

C. Emotion Recognition

Facial expresions is a crutial way to display emotions and takes an important role in human interaction, for this reason, the capability to recognize human emotions based on face images has been an essential research area in Computer Vision during the last decades. In this sense, several advances in this area has been performed using different methods, for instance, some techniques are based on muscle features and machine learning [5]. However, recent works [6] [7] use CNNs, achieving up to 96% of accuracy on CK+ dataset. In this project, we proposed a custom CNN model that balance the highest accuracy rate and the most simple CNN architecture.

a) Dataset: This work utilizes a part of the Facial Expression Recognition 2013 (FER-2013) dataset [8] The entire FAR-2013 is composed of 35,887 gray-scale pictures of 48x48 pixels. However, this work uses 27,192 images into five categories (Angry, Happy, Sad, Surprise, Neutral), see Fig. 2. For training purposes, the dataset was divided into 24,181 images for training and 3,011 for testing.

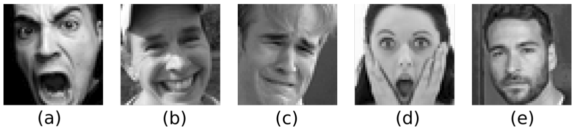

Fig. 2. Sample pictures for facial emotions used during training and testing. (a) angry, (b) happy, (c) sad, (d) surprise, (e) neutral

b) Model: Concerning to recognize emotions, we implemented a CNN based on the LeNet network [9], which is composed by local connections (instead of fully-connected layers), weight sharing, and spatial or temporal sub-sampling. Fig. 3 shows the architecture of the proposed CNN. This is fed by gray-scale images of 48x48 pixels, also utilizes two convolutional and subsampling layers for feature extraction, and one full-connected (FC) layer for classification.

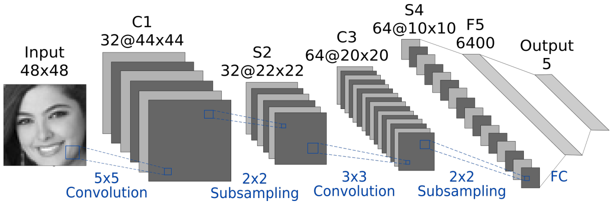

Fig. 3. Architecture of the CNN for emotion recognition.

c) Experimental Results: The performance of the recognition system shows an accuracy rate of 86.5% since our CNN has few parameters, training images are small, and the gray-scale images used for training don't give relevant information for feature extraction. Confusion matrix, see Table I, reveals which emotions are misclassified by the trained network. For instance, angry faces are sometimes confused with surprised and sad faces because these emotions are expressed by a different facial expression for each person. Even for humans, some emotions are difficult to distinguish.

TABLE I
CONFUSION MATRIX FOR THE EMOTION RECOGNITION MODEL

	Angry	**Happy**	**Sad**	**Surprise**	**Neutral**
Angry	**0.72**	0.18	0.02	0.06	0.02
Happy	0.01	**0.90**	0.04	0.00	0.05
Sad	0.00	0.01	**0.89**	0.00	0.10
Surprise	0.00	0.01	0.00	**0.67**	0.32
Neutral	0.02	0.00	0.02	0.00	**0.94**

D. Age and Gender Recognition

Age and gender are important components in human interaction, for example, language, grammar rules and vocabulary vary from men to women, and from elders to young people. In this sense, the problem of ager and gender classification has been deeply studied during the last decades. Some works like [10] works with geometric-based facial descriptors for age classification, while others [11] utilizes LBP histogram features for age, gender and ethnicity classification. Nevertheless, novel advances in Deep Learning has allowed to implement effcient age and gender recognition systems. For example, [12] achieves an accuracy rate up to 92.93% for age estimation using a complex CNNs on the MORPH dataset.

a) Dataset: This project employs the Adience benchmark [13]. This consists of 26,580 non-standardized RGB images of 2,284 subjects in real-world conditions captured by The Open University of Israel. These face pictures were divided by age in eigth categories (0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53, 60+), see Fig. 4. Testing was performed using five-fold, subject-exclusive cross-validation.

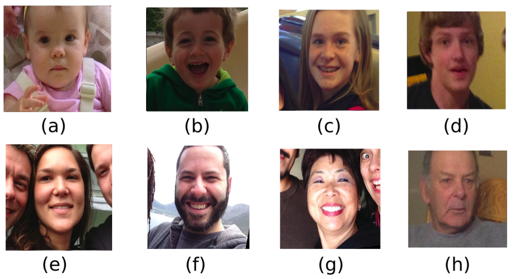

Fig. 4. Some samples of face images used over training and testing steps. (a) 0-2, (b) 4-6, (c) 8-13, (d) 15-20, (e) 25-32, (f) 38-43, (g) 48-53, (h) 60+

b) Model: Age and gender CNN-based classifiers were designed using the same architecture, see Fig. 5. Since recognize facial features related to age is a difficult classification task, this CNN utilizes RGB images of 227x227 pixels and a more complex architecture compared with the emotion recognition model. In this regard, our architecture for age and gender recognition is composed by eight layers as follows: three convolutional layers, subsampling layers after each convolution layer, and two full-connected layers of 512 neurons length each one.

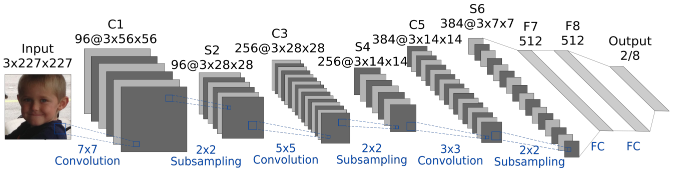

Fig. 5. Architecture of the CNN for gender and age recognition.

c) Experimental Results: The recognition system achieves an accuracy rate of 53.8% for gender estimation and 84.7% for age estimation. This quite good accuracy is because the large size of RGB images in the dataset allows to extract useful facial features. Table II provides the confusion matrix for age recognition and shows classification errors. Gender and age missclasification are produced by labeling mistakes, low resolution and occlusions. Other reason for gender misclassification is the difficulty to distinguish this characteristic in very young children.

TABLE II
CONFUSION MATRIX FOR THE AGE RECOGNITION MODEL

	0-2	**4-6**	**8-13**	**15-20**	**25-32**	**38-43**	**48-53**	**60+**
0-2	**0.67**	0.18	0.01	0.02	0.02	0.04	0.02	0.04
4-6	0.16	**0.60**	0.14	0.04	0.02	0.01	0.00	0.03
8-13	0.01	0.21	**0.54**	0.14	0.02	0.03	0.02	0.03
15-20	0.01	0.03	0.08	**0.23**	0.51	0.06	0.05	0.03
25-32	0.01	0.01	0.14	0.06	**0.61**	0.16	0.01	0.01
38-43	0.02	0.01	0.02	0.22	0.14	**0.34**	0.01	0.24
48-53	0.00	0.01	0.00	0.01	0.15	0.32	**0.36**	0.15
60+	0.02	0.08	0.05	0.05	0.16	0.05	0.13	**0.46**

E. Hand Gesture Recognition

Hand gestures is a very important way to reinforce human communication, but also an easy and attractive means for human-machine interaction. For example, hand gestures could allow to us to control and interact with screeens or robots without physical contact. For this reason, in the last decades, CNNs has been the state-of-the-art technique for hand gesture recognition, for instance, [14] proposed a CNN for hand pose classification using a simple Gaussian skin color model to obtain skin regions. Also, [15] utilizes 3D CNNs to recognize 19 hand gestures and achieves a correct classification rate of 77.5% on the VIVA dataset.

a) Dataset: The dataset for hand gestures classification was provided by an open database from AGH University of Science and Technology [16]. This is composed by 73,124 gray-scale images of size 48x48 pixels divided into ten different hand gestures. But, for proposes of this porject, only six hand gestures (56,694 pictures) of the entire dataset were selected. From these images, 42,027 were used for training and 14,667 for testing. Fig. 6 depicts one sample picture for each hand gesture.

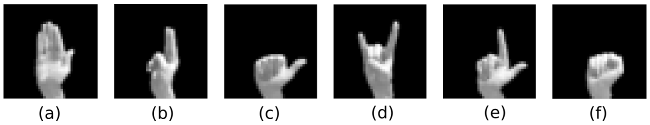

Fig. 6. Sample images of hand gestures used in training and testing steps. (a) gest. 1, (b) gest. 2, (c) gest. 3, (d) gest. 4, (e) gest. 5, (f) gest. 6

b) Model: The CNN architecture for hand pose recognition was designed based on the basic LeNet model [9]. In this sense, our model consists of two convolutional layers, one max pooling layer after each convolutional operation for subsamplig, and one full-connected layer of 500 neurons lenght. Furthermore, since hand gesture features can be easy recognized from basic single-channel images, our network utilizes binary images of 48x48 pixels as input. The proposed architecture is presented in Fig. 7.

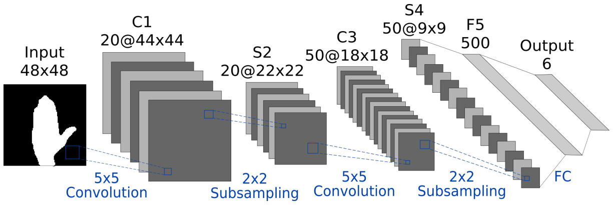

Fig. 7. Architecture of the CNN for hand gesture recognition.

c) Experimental Results: The recognition system shows an accuracy of 98.8% in testing step. This is an outstanding result despite the fact that the proposed CNN is small and uses only binary images for training. Table I shows the confusion matrix, which discloses which hand gestures are confused by the CNN. These missclassifications are mainly produced by similarities in some hand gestures. Other factor for recognition errors is the distortion of the hand region, which is affected by light conditions and differences in skin tones.

TABLE III
CONFUSION MATRIX FOR THE HAND GESTURE RECOGNITION MODEL

	Gest. 1	**Gest. 2**	**Gest. 3**	**Gest. 4**	**Gest. 5**	**Gest. 6**
Gest. 1	**0.98**	0.02	0.00	0.00	0.00	0.00
Gest. 2	0.00	**0.99**	0.00	0.01	0.00	0.00
Gest. 3	0.00	0.00	**0.98**	0.00	0.02	0.00
Gest. 4	0.00	0.00	0.00	**1.00**	0.00	0.00
Gest. 5	0.00	0.00	0.00	0.00	**1.00**	0.00
Gest. 6	0.00	0.00	0.00	0.00	0.04	**0.96**

III. System Results

In previous subsections we described the design and structure of each CNN for every user characteristic detection. Thereby, after training each CNNs with its corresponding datasets by using the Caffe framework, we obtained a file for each CNN with their architecture and their learned weights. Later, these files are employed into the whole real-time system in order to perform the most important part of the system, the classification tasks. Finally, the performance of this system is evaluated in real-time under various conditions such as different backgrounds, distances, and different light conditions. This evaluation was conducted on a standard laptop computer with the following computational resources: Intel Core 7 Octa-Core CPU @3.8 GHz and 12 GB RAM.

As shown in Fig. 8, the system correctly recognizes gender, age, emotions and hand gestures for two different subjects in most number of tests under different backgrounds, light conditions, positions, distinct distances, and small rotations. As well, evaluation step shows that response time of the recognition system is about 144.40 milliseconds (average from ten consecutive frames) for capturing and processing a single image. The duration of this response time depends mostly of the age recogntion model since working with large images, wich are used to recognize important facial features, demands more computational resources. This response time can be reduced by decreasing size of images used for age recognition model or reducing the complexity of the CNN.

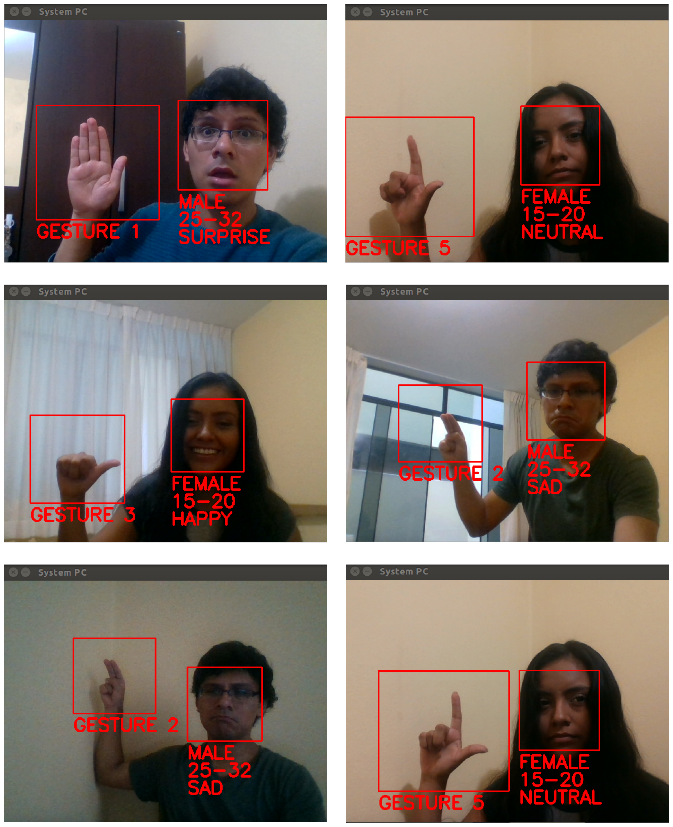

Fig. 8. Results of the recognition system for two subjects.

IV. Conclusions

In this paper, we proposed a novel real-time recognition system based on convolutional neural networks that detects user characteristics such as emotions, gender, age and hand gestures. In order to perform classification with high accuracy rate and fast time response, a straightforward architecture for each CNN was designed. The recognition system, which was impelented on a personal computer, shows a successfully recognition in most cases under real enviroments and a fast response time of about 144.40 milliseconds. These results allows their implementation in several areas like medicine, entertainment, security, advertisements and more.

References

[1] A. Dehghan, E.G. Ortiz, G. Shu, and S.Z. Masood, DAGER: Deep Age, Gender and Emotion Recognition Using Convolutional Neural Networks, arXiv:1702.04280v2, 2017.

[2] R. Ranjan, S. Sankaranarayanan, C. D. Castillo and R. Chellappa, ”An All-In-One Convolutional Neural Network for Face Analysis,” 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, 2017, pp. 17-24.

[3] 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.

[4] Liao S., Zhu X., Lei Z., Zhang L., Li S.Z. (2007) Learning Multi-scale Block Local Binary Patterns for Face Recognition. In: Lee SW., Li S.Z. (eds) Advances in Biometrics. ICB 2007. Lecture Notes in Computer Science, vol 4642. Springer, Berlin, Heidelberg.

[5] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel and J. R. Movellan. Automatic Recognition of Facial Actions in Spontaneous Expressions. Journal of Multimedia 1 (2006): 22-35.

[6] Duncan, Dan L., Gautam Shine and Chris English. Facial Emotion Recognition in Real Time. (2016).

[7] A. T. Lopes, E. de Aguiar, A. F. De Souza, Th. Oliveira-Santos, Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order, Pattern Recognition, Volume 61, 2017, Pages 610-628.

[8] 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

[9] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, ”Gradient-based learning applied to document recognition,” in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov 1998.

[10] Young Ho Kwon and N. da Vitoria Lobo, ”Age classification from facial images,” 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, 1994, pp. 762-767.

[11] Yang Z., Ai H., Demographic Classification with Local Binary Patterns. In: Lee SW., Li S.Z. (eds) Advances in Biometrics. ICB 2007. Lecture Notes in Computer Science, vol 4642. Springer, Berlin, Heidelberg.

[12] S. Chen, C. Zhang, M. Dong, J. Le and M. Rao, ”Using Ranking-CNN for Age Estimation,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 2017, pp. 742-751.

[13] G. Levi and T. Hassncer, ”Age and gender classification using convolutional neural networks,” IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, 2015, pp. 34-42.

[14] M. Han, J. Chen, L. Li and Y. Chang, ”Visual hand gesture recognition with convolution neural network,” 2016 17th IEEE/ACIS International Conf. on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), Shanghai, 2016, pp. 287-291.

[15] P. Molchanov, S. Gupta, K. Kim and J. Kautz, ”Hand gesture recognition with 3D convolutional neural networks,” 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, 2015, pp. 1-7.

[16] Núñez Fernández D., Kwolek B., Hand Posture Recognition Using Convolutional Neural Network. In: Mendoza M., Velastn S. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2017. Lecture Notes in Computer Science, vol 10657. Springer, Cham