Emotion Recognition Expressed on the Face By Multimodal Method using Deep Learning

Abdoul Matine Ousmane, Tahirou Djara, Médésu Sogbohossou, Antoine Vianou

Abstract: ***Emotional recognition plays a vital role in the behavioral and emotional interactions between humans. It is a difficult task because it relies on the prediction of abstract emotional states from multimodal input data. Emotion recognition systems operate in three phases. A first that consists of taking input data from the real world through sensors. Then extract the emotional characteristics to predict the emotion. To do this, methods are used to exaction and classification. Deep learning methods allow recognition in different ways. In this article, we are interested in facial expression. We proceed to the extraction of emotional characteristics expressed on the face in two ways by two different methods. On the one hand, we use Gabor filters to extract textures and facial appearances for different scales and orientations. On the other hand, we extract movements of the face muscles namely eyes, eyebrows, nose and mouth. Then we make an entire classification using the convolutional neural networks (CNN) and then a decision-level merge. The convolutional network model has been training and validating on datasets.***

Keywords: CNN, deep learning, emotion recognition, facial expressions.

I. INTRODUCTION

Emotions help to improve communication between individuals, to ensure a better understanding of the message conveyed and to adapt to a given situation. Emotions play a key role in decision making. They also influence behavior and shape the personality.

Studies by Ekman (1992) [1] suggest that there are six basic emotions which are universal among different cultures, namely happiness, surprise, fear, sadness, anger and disgust. These emotions are associated with specific facial expressions. Facial expressions also play a major role in communication of feelings and attitudes, among other important cues such as postures, gestures, verbal and vocal expressions [2].

Work on automating facial expression analysis has become more and more common in recent years. Already in 1978, Suwa et al. [5] compared all facial muscle movements with prototypical movement patterns for different facial expressions. Thus, many databases of facial images were published in [4] from 1990 and the field became very active [3]. The evolution of research in this area has helped human experts to interpret relevant information extracted from the real world. Specifically, these systems are also implemented in the field of health to detect psychological distress for example [6], in education by estimating student engagement and in gaming for improvement of the players' experience [7].

One of the challenges of emotions recognition is to consider the complexity of the scenarios in the real world. This consumes the appearance of the subjects, the variations of lighting, the properties of the sensors, the variations of lighting, etc ... The demand for large scale datasets is overflowing With the success of deep learning methods [8] able to learn hierarchical representations from data.

The exploration of these techniques in emotions recognition is a major challenge because in the latter, the number of labeled data proximate to reality is limited compared to the numbers of object to recognition. The other challenge is to find a way to leverage additional information from different extraction methods. It would also be important to be able to build an integrated system, such as a neural network architecture, formed as a single system, from input data to predictions.

This article is subdivided into three parts. In the first part, we presented the architecture of an emotion recognition system in general and the methods and tools we used. In the second part, we presented our approach then results and discussion in the last part.

II. EMOTION RECOGNITION SYSTEM ARCHITECTURE

An emotion recognition system has three phases: capture phase, analysis phase and finally the interpretation phase (see Fig.1). Each phase fulfills a specific function. The information is captured from the real world by peripherals (camera, microphone, etc.) at the capture phase. Then this information is then analyzed to extract the emotionally relevant characteristics at the analysis phase. The characteristics extracted from the previous phase are interpreted to determine an emotion in the interpretation phase.

Revised Manuscript Received On December 08, 2019.

***** Correspondence Author

Abdoul Matine Ousmane*, Laboratoire D'electrotechnique De Télécommunication Et D'informatique Appliquée (Letia/Epac), Université D'abomey-Calavi (Uac). Institut D'innovation Technologique (Iitech), Email: Matines1@Yahoo.Fr

Tahirou Djara, Laboratoire D'electrotechnique De Télécommunication Et D'informatique Appliquée (Letia/Epac), Université D'abomey-Calavi (Uac). Institut D'innovation Technologique (Iitech), Email: Csm.Djara@Gmail.Com

Médésu Sogbohossou, Laboratoire D'electrotechnique De Télécommunication Et D'informatique Appliquée (Letia/Epac), Université D'abomey-Calavi (Uac).

Antoine Vianou, Laboratoire D'electrotechnique De Télécommunication Et D'informatique Appliquée (Letia/Epac), Université D'abomey-Calavi (Uac). Institut D'innovation Technologique (Iitech), Email : Avianou@Yahoo.Fr

Fig.1. Different phases of an emotion recognition system

III. METHODS AND TOOLS

A. *Convolutionnal neural network (CNN)*

CNNs have been used successfully to achieve better performance in several vision activities. The CNN principle is the observation that the visual cortex has cells that respond only to stimuli in the subfields of the visual field. The same types of stimuli are detected by several cells. Each of them only responds to a different subregion, serving the entire visual field. They behave in the same way by using shared weights across the entire input image and detecting the same local patterns [9].

In a CNN, a convolutional layer consists of several cores that are applied separately to the input. At the exit, a multi-channel feature map. Here, the number of channels corresponds to the number of cores as illustrated in Fig.2.

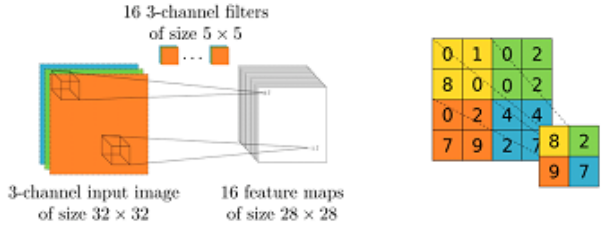


Fig.2. convolutional layer: an input (RGB) image to which a set of 16 x 3 x 5 x 5 x 5 x 5 filters is applied, creating 16 28 x 28 size feature cards. Maximum two by two pooling areas without overlap [9].

An activation function can be applied to the output. The latter can be transmitted to another convolutional layer or to a connected MLP layer. The output is reconfigured into a single vector if the next layer is fully connected.

B. *Facial Expression Databases*

Facial expression recognition system must in practice identify the emotions of a large number of different faces, which requires several sequences of images.

To develop and evaluate face analysis applications, tests on large collections of image sequences are required. Since image sequence recordings of the moving face are necessary to study the temporal dynamics of facial expressions, static images are important for information on the configuration of facial expressions that are essential, so the sequences Facial video and also static images are important in the facial expression classification process.

We use the following bases: Le dataset FER2013, Cohn-Kanade(ck+), FEEDTUM, JAFFE.

C. *Technical tools*

We performed a comparative study between three programming languages recommended for machine learning and chose python.

we used Open Source Computer Vision (OpenCV) library that is written in C and C ++ and can be used on Linux, Windows and MacOS X. Interfaces have been developed for Python, Ruby, Matlab and other languages. Open CV is oriented towards real-time applications. One of the goals of OpenCV is to help people quickly build sophisticated vision applications using simple computer vision infrastructure.

we used Keras: is a high-level neural network API, written in Python and capable of operate on TensorFlow or Theano. It was developed with a focus on rapid experimentation. It was developed as part of the research effort of the Open-ended Neuro Electronic Intelligent Robot Operating System (ONEIROS) project, and its main author and maintainer is François Chollet, a Google engineer.

IV. OUR APPROACH

In our approach, at the sensor level, we first proceed to face detection. Then at the analysis level, we extract the emotional characteristics of the face in two ways: one by appearance method and the other by geometric method. Finally, at the interpretation level, we use a convolved neuron network model trained and validated on datasets. The Fig.3. shows this complete final architecture.

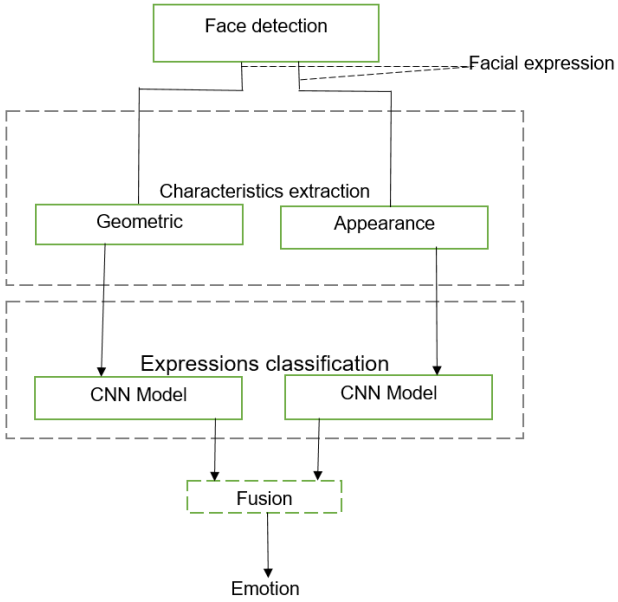


Fig.3. steps of our approach

Step 1: Face detection

Step 2: extraction of the optimized characteristics by the methods of appearance and geometry.

Step 3: Classification of the Emotion Issue of CNN Training and Validation

Step 4: Merge the results of the classification

These steps are best explained in the following sections

A. *Face Detection*

Several face detection methods have been proposed in recent years. The treatment here consists in looking in an image of faces position and extracting them in the form of a set of thumbnails in order to facilitate their subsequent processing. A face is

considered correctly detected if the extracted image size does not exceed 20% of the actual size of the facial region and essentially contains the eyes, nose and mouth.

We used the haar descriptors under opencv to detect the face. It was originally developed by P. Viola and M. Jones.

B. *characteristics Extraction by appearance method*

Gabor filters are among the most robust texture extraction techniques for different scales and orientations.

To do this, we first normalize the faces on the same scale based on the eyes position. Thus, we first detect the each eye position, then we calculate the center between the two eyes and the angle ω between eyes axis and horizontal axis. We calculate the scale according to which we decrease the size of the image. Finally, we apply an affine transformation to the face using the matrix M.

$$M = \begin{pmatrix} a & b & -aC_x - bC_y + \alpha l \\ -b & a & bC_x - aC_y + \beta h \end{pmatrix} \quad (1)$$

with a = s cosω and b = s sinω, α and β are empirically selected

Once the normalized face, we convoluons with a bank of 40 Gabor filters using the following gabor function.

$$G_k(z) = \frac{\|k\|^2}{\sigma^2} \exp\left(-\frac{\|k\|^2\|z\|^2}{2\sigma^2}\right)\left(exp(ik^t z) - exp\left(\frac{\sigma^2}{2}\right)\right) \quad (2)$$

C. *Characteristics Extraction by Geometric Method*

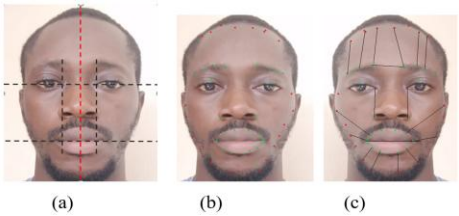
At this level, information is extracted on the shapes, positions and movements of facial features, namely the eyes, eyebrows, nose and mouth through a number of defined points. Facial deformities are coded using the distances between these points.

The location of these points is based on the calculation of the position of three axes: horizontal axis of eyes, horizontal axis of the mouth and axis of the face symmetry. Thus, we make a horizontal Projection of the gradients for the detection of the eyes axis. The Segmentation of the mouth by the color and the saturation makes it possible to obtain mouth axis. The vertical projection of the gray levels of the nose region limited by the two axes makes it possible to obtain symmetry axis (see Fig.4 (a)) .

Based on the three axes, the coordinates of 30 points are defined. Fig.4 (b) illustrates the positions of these points.

Two points type are defined. On the one hand, expression-invariant points are located in the most stable areas of the face. Represented in Fig.4. (b) by dots in red, they are detected on the face contour, the outer corners of the eyes and the top of the nose. On the other hand, moving points are located at the lower corners of the nose, the mouth contour, on the eyelids and on the eyebrows. They are represented by green in Fig.4. (b). Finally the distances between invariant points and moving points are calculated. Fig.3. (c) illustrates all these distances. The expression is encoded by the distances calculated in the current image (Di) and normalized by the distances in the neutral image D0i.

$$\Delta D = \left(\frac{D_1}{D_{01}}, \ldots, \frac{D_i}{D_{0i}}, \frac{D_{21}}{D_{021}}\right) \quad (3)$$



(a) (b) (c)

Fig.4. steps of geometric method

D. *Expression classification*

After characteristics extracting by these different methods, it is advisable to predict the emotion. The task is to rank each face according to the emotion expressed in the facial expression in one of the seven categories (0 = anger, 1 = disgust, 2 = fear, 3 = happy, 4 = sadness, 5 = surprise, 6 = neutral). A convolutional neural networks model (CNN) trained and validated in data sets has been used. it includes a set of fully connected layers at the end. Two layers are special: the input layer that receives the raw data and the output layer that produces the results. Between the two layers is 4 intermediate layers (see Fig.5).

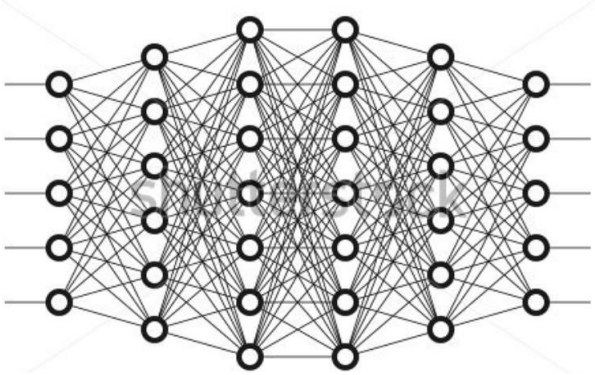


Fig.5. deep learning neural network

V. RESULTS AND DISCUSSIONS

Our main achievements as well as our actions taken to set up the internal simulation environment are as follows.

The FER2013 database was mainly used to improve our model. Then we extended it to other database. Comparing the precision of the training of our model with others allowed us to deduce that our model offers satisfactory results (see Table-I).

DeXpression is a CNN architecture independent of any handmade feature extraction, the main component of which is the FeatEx block, which allows the extraction of features in parallel, while the architecture proposed by Dachapally, which is a little closer to ours, is an 8-layer CNN with three convolution layers, three consolidation layers and two fully connected layers.

This table tells us that with less training, our model provides a better recognition.

On the other dataset namely: CK + (5876 images), FEEDTUM and JAFFE, it was possible to obtain an accuracy of nearly 99%.

Table-I: The FER2013 dataset has been very well analyzed and the best possible recognition accuracy is obtained with our model [10].

Models	Characterisation	Training accuracy	Test accuracy
SVM(OVO)	Scaled pixels	43.36 %	40.17 %
CNN	Dachapally	53.88 %	52.38 %
CNN	DeXpression	72.25 %	61.63 %
CNN	our model	65 %	66

A. *Confusion Matrix*

Table-II and Table-III are respectively the confusion matrices of appearance method and geometric method calculated on CK + dataset. We note that the average recognition rate of appearance method is 3% higher than that of geometric method.

However, the recognition rates of fear, surprise and neutral expression calculated with the geometric method are higher than those calculated with the appearance method. We find that each method is advantageous for a set of emotions.

Table-II : Confusion matrix of the appearance method on CK+ dataset

%	angry	disgust	fear	happy	sad	surprise	neutral
angry	0.93	0.01	0.03	0.02	0.05	0.01	0.11
disgust	0.00	0.97	0.04	0.02	0.03	0.01	0.02
fear	0.20	0.22	0.87	0.00	0.17	0.01	0.41
happy	0.03	0.00	0.01	1	0.02	0.01	0.00
sad	0.12	0.00	0.40	0.02	1	0.00	0.33
surprise	0.06	0.00	0.17	0.04	0.33	0.94	0.40
neutral	0.06	0.03	0.01	0.00	0.09	0.00	0.87

Table-III : Confusion matrix of the geometric method on CK+ dataset

%	angry	disgust	fear	happy	sad	surprise	neutral
angry	0.90	0.01	0.03	0.02	0.08	0.01	0.04
disgust	0.40	0.84	0.04	0.22	0.03	0.01	0.02
fear	0.00	0.22	0.91	0.01	0.07	0.01	0.06
happy	0.00	0.09	0.01	0.96	0.02	0.01	0.00
sad	0.12	0.00	0.40	0.02	0.88	0.00	0.13
surprise	0.01	0.20	0.17	0.04	0.33	0.97	0.00
neutral	0.06	0.03	0.01	0.00	0.09	0.00	1

Table-IV and Table-V present respectivelly confusion matrices of appearance method and geometric method on FEEDTUM dataset. The average recognition rate of appearance method is 86%, while that of the geometric method is 50%. We note that both methods have lower recognition rates than those obtained for simulated CK + base emotions.

This decrease is due to the low intensities of spontaneous expressions. We also note that the average recognition rate of the appearance method is higher than the average recognition rate of the geometric method.

Table-IV: Confusion matrix of the appearance method on FEEDTUM dataset

%	angry	disgust	fear	happy	sad	surprise	neutral
angry	0.87	0.01	0.00	0.02	0.08	0.01	0.00
disgust	0.09	0.84	0.04	0.22	0.04	0.01	0.02
fear	0.00	0.22	0.81	0.04	0.07	0.08	0.00
happy	0.20	0.02	0.01	0.99	0.02	0.01	0.01
sad	0.12	0.00	0.06	0.02	0.75	0.00	0.11
surprise	0.01	0.20	0.17	0.04	0.33	0.91	0.00
neutral	0.06	0.03	0.01	0.00	0.09	0.00	89

Table-V: Confusion matrix of the geometric method on FEEDTUM dataset

%	angry	disgust	fear	happy	sad	surprise	neutral
angry	0.58	0.21	0.00	0.08	0.01	0.08	0.08
disgust	0.37	0.15	0.04	0.28	0.12	0.10	0.04
fear	0.26	0.22	0.27	0.10	0.05	0.40	0.00
happy	0.06	0.02	0.01	0.82	0.02	0.11	0.04
sad	0.41	0.02	0.08	0.16	0.18	0.00	0.22
surprise	0.02	0.00	0.17	0.14	0.03	0.79	0.00
neutral	0.06	0.03	0.01	0.00	0.19	0.00	80

Some interesting observations on the predictions of the model are to be underlined after study of the confusion matrix obtained on FER2013 (Table-VI): Higher accuracy, 74% of recognition obtained with the emotion joy (joy). Lower accuracy, 33% recognition achieved with fear emotion.

One can draw conclusion that the datasets used to form the model do not contain enough fear emotion. It is therefore important to reform the model with a much larger database containing too much fear emotion label. For this purpose, the Affecnet database is better indicated.

Table-VI: Confusion matrix obtained on FER2013 dataset

%	angry	disgust	fear	happy	sad	surprise	neutral
angry	0.65	0.01	0.06	0.02	0.09	0.01	0.16
disgust	0.45	0.37	0.05	0.02	0.05	0.01	0.05
fear	0.20	0.00	0.33	0.02	0.17	0.06	0.22
happy	0.07	0.00	0.03	0.74	0.02	0.01	0.13
sad	0.16	0.00	0.09	0.02	0.39	0.00	0.33
surprise	0.06	0.00	0.17	0.04	0.02	0.62	0.10
neutral	0.10	0.00	0.03	0.03	0.09	0.00	0.73

Table-VII Recognition rate of fusion methods on CK + dataset

fusion methods	RR	happy	Angry	Fear	Disgust	sad	Surprise	Neutral
average	99.5	100	100	98	97	100	100	100
product	99	100	98.8	98	97	100	100	100
max	99.6	100	100	98	97	100	100	100
linear	97.9	100	97	98	99	100	100	95
Gaussian	98.7	100	97	98.1	97	100	100	98.3

We note in Table-VII that the Gaussian merger gives the best recognition rate among the classification merge methods. However, it remains slightly lower than the merger

recognition rate by the product.

We also note that disgust remains emotion with the lowest recognition rate for all statistical rule-based merge methods and the Gaussian merge method. On the other hand, linear fusion improves the recognition of disgust.

Table-VIII Recognition rate of fusion methods on FEEDTUM dataset

fusion methods	RR	happy	Angry	Fear	Disgust	sad	Surprise	Neutral
average	83.7	99	67.2	79.2	83	73.1	90.1	93
product	82.8	99	67	83.4	78.2	73.1	87.8	91
max	82.2	99	71.6	75	72.4	73.1	90.3	94.4
linear	86.2	99	88.2	72.6	84.5	80.1	90	88.7
Gaussian	87.5	99	90.1	78.2	84	78.3	92	88.7

The recognition rates of the methods of fusion by statistical rules are higher than that obtained by the geometric method, however they remain slightly lower than the recognition rate of the method of appearance. Average melting, however, improves the recognition rates of fear and neutral expression with respect to the appearance method and the geometric method.

Classification-based merge methods have very interesting recognition rates. Gaussian and linear fusion methods have higher average recognition rates.

B. *Training and test processes*

The maximum accuracy achieved is nearly 65% for training and 62.2% for validation. It is therefore this model that has been recorded and used for the tests (Fig.6).

The finding here is that the loss during training is less than the loss during validation. We can therefore say that our model works better on the learning data than on the validation data that are unknown to it (see Fig.7). This over-adjustment is normal because it is not high [10].

Fig.6. Training and validation phase in the FER2013 database: evaluation of the accuracy during two phases.

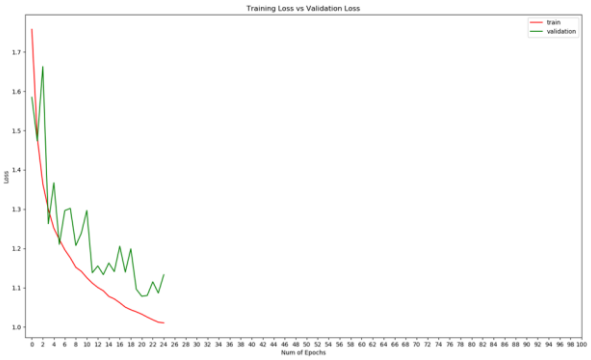


Fig.7. Training and validation phases in the FER2013 database: evaluation of the loss observed during the two phases

The model is formed around 28 epochs and we find that it is very sensitive to initialization and generally learns slowly.

VI. CONCLUSION

For this work, we are interested in the expression of the face to determine emotion. We extracted emotional characteristics in two different ways. First, the texture of the face under different scales and different orientations. Second, the movements of the face muscles, namely eyes, eyebrows, nose and mouth. Then, a separate classification is done using a convolutional network model that has been formed and validated on datasets. Then a merger at the decision level is made.

Comparing our model with others based on CNN and SVM, we found that with less training, our model offers better recognition in FER2013 database. On another set of data, namely: CK +, FEEDTUM and JAFFE, it was possible to obtain an accuracy of nearly 99%.

According to the results, based on CK +, we retain that each method is advantageous for a set of emotions. We also note that disgust remains emotion with the lowest recognition rate for all statistical fusion methods and the Gaussian fusion method. In contrast, linear fusion enhances the recognition of disgust.

On FEEDTUM database, we note that the average recognition rate of the method of appearance is higher than the average recognition rate of the geometric method. Also, classification-based merge methods have very interesting recognition rates. Gaussian and linear fusion methods have higher average recognition rates.

On FER2013 database, highest accuracy is obtained with joy (joy) 74% of recognition. The inferior Precision is obtained with 33% fear emotion. It can be concluded that the data sets used to form the model do not contain enough fear emotion. It is therefore important to reform the model with a much larger database containing too many expressions of fear emotion. For this, the Affecnet database is better indicated.

REFERENCES

1. Ekman, Paul (1992). Are there basic emotions? Psychological Review, 550–553.

2. Mehrabian, Albert (1971). Silent messages. Wadsworth.
3. Bartlett, Marian Stewart and Hager, Joseph C and Ekman, Paul and Sejnowski, Terrence J (1999). Measuring facial expressions by computer image analysis. Psychophysiology, 36 (02), 253–263.
4. Dahl, George E and Sainath, Tara N and Hinton, Geoffrey E (2013). Improving deep neural networks for lvcsr using rectified linear units and dropout. Proc. ICASSP. 8609–8613.
5. Suwa, Motoi and Sugie, Noboru and Fujimora, Keisuke (1978). A preliminary note on pattern recognition of human emotional expression. International joint conference on pattern recognition. vol. 1978, 408–410.
6. Scherer, Stefan and Stratou, Giota and Mahmoud, Mohamed and Boberg, Jill and Gratch, Jonathan and Rizzo, Alessandro and Morency, Louis-Philippe (2013). Automatic behavior descriptors for psychological disorder analysis. Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. IEEE, 1–8.
7. Shaker, Noor and Asteriadis, Stylianos and Yannakakis, Georgios N and Karpouzis, Kostas (2011). A game-based corpus for analysing the interplay between game context and player experience. Affective Computing and Intelligent Interaction, Springer. 547–556.
8. Bahdanau, Dzmitry and Cho, Kyunghyun and Bengio, Yoshua (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
9. S. Chetlur & al. cudnn : E-cient primitives for deep learning. arXiv preprint arXiv :1410.0759. 2014.
10. Abdoul M. OUSMANE, Tahirou DJARA, Faizath J. ZOUMAROU and Antoine VIANOU : automatic recognition system of emotions expressed through the face unsing machine learning : application to police interrogation simulation, 3rd international conference on Bio-engineering for Smart Technologies, Paris, April 2019.

AUTHORS PROFILE

Abdoul Matine OUSMANE is a PhD student at University of Abomey-Calavi, Benin. His research focuses on: biometrics, signal processing and images, computer intelligence, industrial applications and symbolic programming. He is a member of the research laboratory: Laboratory of Electronics, Telecommunications and Applied Data Processing (LETIA / EPAC). He graduated research master from the Institute of Training and Research in Computer Science (IFRI) at the University of Abomey-Calavi in 2012. He is a consultant in the field of computer analysis, web an mobile developer.

Tahirou Djara is a Senior Lecturer at the Polytechnic School of Abomey-Calavi located in the University of Abomey-Calavi, Bénin. His research interests include: biometrics, signal and image processing, computational intelligence, industrial applications and symbolical programming. He is member of the research laboratory: Laboratory of Electronics, Telecommunications and Applied Data Processing Technology (Laboratoire d'Electrotechnique de Télécommunication et d'Informatique Appliquée– LETIA/EPAC). He received the PhD degree in signals and image processing from the University of Abomey-Calavi, in 2013. He is a consultant in quality assurance in higher education and consultant in the field of science and engineering technology.

Médésu SOGBOHOSSOU was born in Cotonou, Rep. of Benin, on July 14, 1975. He obtained a doctorate in Control and Applied Computer Science prepared jointly from Université d'Abomey-Calavi (Rep. of Benin) and Université de Nantes (France), in 2009. He received the Electronic Ing. degree from Institut National Polytechnique - Houphouët-Boigny (Côte d'Ivoire), in 2002. He is currently an Assistant professor at Ecole Polytechnique d'Abomey-Calavi (Rep. of Benin). His research activities mainly focus on embedded systems and formal modelling of the discrete event systems.

Antoine Vianou is a PhD Engineer in Energy and Electricity sciences. He has been graduated through many universities as the University of Dakar and the University of Evry Val d'Essonne. He is a Full Professor in Engineering Sciences and Technologies (E.S.T.). Pr. VIANOU is currently Chairman of the Sectoral Scientific Committee of E.S.T. of the Scientific Council of UAC in Benin and is also Director of the Laboratory of Thermophysic Characterization of Materials and Energy Mastering. He is the Director of the Doctoral School of Engineering Sciences in UAC. During his academic career, Professor VIANOU taught in several African Universities and in several French ones. He is author of over hundred articles in the fields of Engineering Sciences and Technologies. In addition, he received several honors in recognition for his professional career.