Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 175 (2020) 689–694

Procedia Computer Science

www.elsevier.com/locate/procedia

The 2nd International Workshop on the Future of Internet of Everything (FIoE)
August 9-12, 2020, Leuven, Belgium

Facial emotion recognition using deep learning: review and insights

Wafa Mellouk[a*], Wahida Handouzi[a]

[a]*Laboratoire d'automatique de Tlemcen (LAT), Tlemcen university, BP 320, Chetouane Tlemcen 1300, Algeria*

Abstract

Automatic emotion recognition based on facial expression is an interesting research field, which has presented and applied in several areas such as safety, health and in human machine interfaces. Researchers in this field are interested in developing techniques to interpret, code facial expressions and extract these features in order to have a better prediction by computer. With the remarkable success of deep learning, the different types of architectures of this technique are exploited to achieve a better performance. The purpose of this paper is to make a study on recent works on automatic facial emotion recognition FER via deep learning. We underline on these contributions treated, the architecture and the databases used and we present the progress made by comparing the proposed methods and the results obtained. The interest of this paper is to serve and guide researchers by review recent works and providing insights to make improvements to this field.

Keywords: facial emotion recognition, deep neural networks, automatic recognition, database

1. Introduction

Automatic emotion recognition is a large and important research area that addresses two different subjects, which are psychological human emotion recognition and artificial intelligence (AI). The emotional state of humans can obtain from verbal and non-verbal information captured by the various sensors, for example from facial changes [1], tone of

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: wafa.melloukaii@gmail.com

10.1016/j.procs.2020.07.101

voice [2] and physiological signals [3]. In 1967, Mehrabian [4] showed that 55% of emotional information were visual, 38% vocal and 7% verbal. Face changes during a communication are the first signs that transmit the emotional state, which is why most researchers are very interested by this modality.

Extracting features from one face to another is a difficult and sensitive task in order to have a better classification. In 1978 Ekman and Freisen [5] are among the first scientific interested in facial expression which are developed FACS (Facial Action Coding System) in which facial movements are described by Action Units AUs, they are broken down the human face into 46 AUs action units each AU is coded with one or more facial muscles.

The automatic FER is the most studied by researchers compared to other modalities to statistics which made by philipp et al.[6], but it is task that is not easy because each person presents his emotion by his way. Several obstacles and challenges are present in this area that one should not neglect like the variation of head pose, luminosity, age, gender and the background, as well as the problem of occlusion caused by Sunglasses, scarf, skin illness…etc.

Several traditional methods exist are used for the extraction facial features such as geometric and texture features for example local binary patterns LBP [7], facial action units FAC [5], local directional patterns LDA [8], Gabor wavelet [9]. In recent years, deep learning has been very successful and efficient approach thanks to the result obtained with its architectures which allow the automatic extraction of features and classification such as the convolutional neural network CNN and the recurrent neural network RNN; here what prompted researchers to start using this technique to recognize human emotions. Several efforts are made by researchers on the development of deep neural network architectures, which produce very satisfactory results in this area.

In this paper, we provide a review of recent advances in sensing emotions by recognizing facial expressions using different deep learning architectures. We present recent results from 2016 to 2019 with an interpretation of the problems and contributions. It is organized as follows: in section two, we introduce some available public databases, section three; we present a recent state of the art on the FER using deep learning and we end in section four and five with a discussion and comparisons then a general conclusion with the future works.

2. Facial available databases

One of the success factors of deep learning is the training the neuron network with examples, several FER databases now available to researchers to accomplish this task, each one different from the others in term of the number and size of images and videos, variations of the illumination, population and face pose. Some presented in the Table.1 in which we will note its presence in the works cited in the following section.

Table 1. A summary of some FER databases.

Databases	Descriptions	Emotions
MultiPie [10]	More than 750,000 images captured by 15 view and 19 illumination conditions	Anger, Disgust, Neutral, Happy, Squint, Scream, Surprise
MMI [11]	2900 videos, indicate the neutral, onset, apex and offset	Six basic emotions and neutral
GEMEP FERA [12]	289 images sequences	Anger, Fear, Sadness, Relief, Happy
SFEW [13]	700 images with different ages, occlusion, illumination and head pose.	Six basic emotions and neutral
CK+ [14]	593 videos for posed and non-posed expressions	Six basic emotions, contempt and neutral
FER2013 [15]	35,887 grayscale images collect from google image search	Six basic emotions and neutral
JAFFE [16]	213 grayscale images posed by 10 Japanese females	Six basic emotions and neural
BU-3DFE [17]	2500 3D facial images captured on two view -45°, +45°	Six basic emotions and neutral
CASME II [18]	247 micro-expressions sequences	Happy, Disgust, Surprise, Regression and others
Oulu-CASIA [19]	2880 videos captured in three different illumination conditions	Six basic emotions
AffectNet [20]	More than 440.000 images collected from the internet	Six basic emotions and neutral
RAFD-DB [21]	30000 images from real world	Six basic emotions and neutral

RaFD [22]	8040 images with different face poses, age, gender, sexes	Six basic emotions, contempt and neutral

3. Facial emotion recognition using deep learning

Despite the notable success of traditional facial recognition methods through the extracted of handcrafted features, over the past decade researchers have directed to the deep learning approach due to its high automatic recognition capacity. In this context, we will present some recent studies in FER, which show proposed methods of deep learning in order to obtain better detection. Train and test on several static or sequential databases.

Mollahosseini et al. [23] propose deep CNN for FER across several available databases. After extracting the facial landmarks from the data, the images reduced to 48x 48 pixels. Then, they applied the augmentation data technique. The architecture used consist of two convolution-pooling layers, then add two inception styles modules, which contains convolutional layers size 1x1, 3x3 and 5x5. They present the ability to use technique the network-in-network, which allow increasing local performance due to the convolution layers applied locally, and this technique also make it possible to reduce the over-fitting problem.

Lopes et al. [24] Studied the impact of data pre-processing before the training the network in order to have a better emotion classification. Data augmentation, rotation correction, cropping, down sampling with 32x32 pixels and intensity normalisation are the steps that were applied before CNN, which consist of two convolution-pooling layers ending with two fully connected with 256 and 7 neurons. The best weight gained at the training stage are used at the test stage. This experience was evaluated in three accessible databases: CK+, JAFFE, BU-3DFE. Researchers shows that combining all of these pre-processing steps is more effective than applying them separately.

These pre-processing techniques also implemented by Mohammadpour et al. [25]. They propose a novel CNN for detecting AUs of the face. For the network, they use two convolution layers, each followed by a max pooling and ending with two fully connected layers that indicate the numbers of AUs activated.

In 2018, for the disappearance or explosion gradient problem Cai et al. [26] propose a novel architecture CNN with Sparse Batch normalization SBP. The property of this network is to use two convolution layers successive at the beginning, followed by max pooling then SBP, and to reduce the over-fitting problem, the dropout applied in the middle of three fully connected. For the facial occlusion problem Li et al. [27] present a new method of CNN, firstly the data introduced into VGGNet network, then they apply the technique of CNN with attention mechanism ACNN. This architecture trained and tested in three large databases FED-RO, RAF-DB and AffectNet.

Detection of the essential parts of the face was proposed by Yolcu et al. [28].They used three CNN with same architecture each one detect a part of the face such as eyebrow, eye and mouth. Before introducing the images into CNN, they go through the crop stage and the detection of key-point facial. The iconic face obtained combined with the raw image was introduced into second type of CNN to detect facial expression. Researchers show that this method offers better accuracy than the use raw images or iconize face alone (See Fig.1.a).

In 2019, Agrawal et Mittal [29] make a study of the influence variation of the CNN parameters on the recognition rate using FER2013 database. First, all the images are all defined at 64x64 pixels, and they make a variation in size and number of filters also the type of optimizer chosen (adam, SGD, adadelta) on a simple CNN, which contain two successive convolution layers, the second layer play the role the max pooling, then a softmax function for classification. According to these studies, researchers create two novel models of CNN achieve average 65.23% and 65.77% of accuracy, the particularity of these models is that they do not contain fully connected layers dropout, and the same filter size remains in the network.

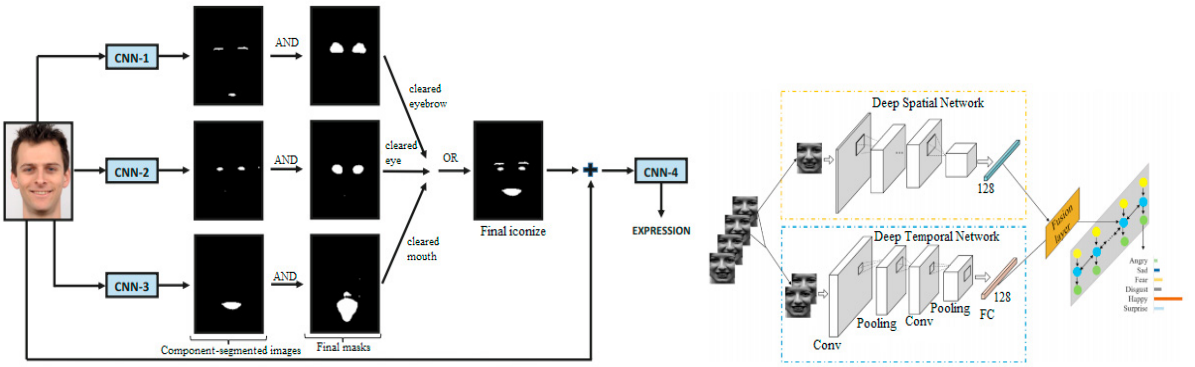
Deepak jain et al. [30] propose a novel deep CNN witch contain two residual blocks, each one contain four-convolution layer. These model trains on JAFFE and CK+ databases after a pre-processing step, which allows cropping and normalizing the intensity of the images.

Kim et al. [31] studies variation facial expression during emotional state, they propose a spatio-temporal architect with a combination between CNN and LSTM. At first time, CNN learn the spatial features of the facial expression in all the frames of the emotional state followed by an LSTM applied to preserve the whole sequence of these spatial features. Also Yu et al. [32] Present a novel architecture called Spatio-Temporal Convolutional with Nested LSTM (STC-NLSTM), this architecture based on three deep learning sub network such as: 3DCNN for extraction spatio-temporal features followed by temporal T-LSTM to preserve the temporal dynamic, then the convolutional C-LSTM for modelled the multi-level features.

Deep convolutional BiLSTM architecture was proposed by Liang et al. [33], they create two DCNN, one of which

is designated for spatial features and the other for extracting temporal features in facial expression sequences, these features fused at level on a vector with 256 dimensions, and for the classification into one of the six basic emotions, researchers used BiLTSM network. For the pre-processing stage, they used the Multitask cascade convolutional network for detecting the face, then applied the technique of data augmentation to broaden database (See Fig.1.b).

All of the researchers cited previously classifying the basic emotions: happiness, disgust, surprise, anger, fear, sadness and neutral, Fig 3. Present some different architecture proposed by the researchers who mentioned above.



(a) (b)

Fig. 1. Different deep learning methods proposed by Yoclu et al. [28] and (b). Liang et al. [33]

4. Discussion and comparison

In this paper, we clearly noted the significant interest of researchers in FER via deep learning over recent years. The automatic FER task goes through different steps like: data processing, proposed model architecture and finally emotion recognition.

The preprocessing is an important step, which was present in all the papers cited in this review, that consist several techniques such as resized and cropped images to reduce the time of training, normalization spatial and intensity pixels and the data augmentation to increase the diversity of the images and eliminate the over-fitting problem. All these techniques are well presented by lopes et al. [24].

Several methods and contributions presented in this review was achieved high accuracy. Mollahosseini et al. [23] showed the important performance by adding inception layers in the networks, . Mohammadpour et al. [25] prefer to extract AU from the face than the classification directly the emotions, Li et al. [27] is interested in the study the problem of occlusion images, also for to get network deeper, Deepak et al. [30] propose adding the residual blocks. Yolcu et al. [28] shows the advantage of adding the iconized face in the input of the network, enhance compared with the training just with the raw images. For Agrawal et Mittal.[29] offers two new CNN architecture after an in-depth study the impact of CNN parameters on the recognition rate. Most of these methods presented competitive results over than 90%. (See Table.2)

For extraction the spatio-temporal features researchers proposed different structures of deep learning such as a combination of CNN-LSTM, 3DCNN, and a Deep CNN. According to the results obtained, the methods proposed by Yu et al. [32] and Liang et al. [33] achieve better precision compared to the method used by Kim et al. [31]. With a rate higher than 99%.

Researchers achieve high precision in FER by applying CNN networks with spatial data and for sequential data, researchers used the combination between CNN-RNN especially LSTM network, this indicate that CNN is the network basic of deep learning for FER. For the CNN parameters, the Softmax function and Adam optimization algorithm are the most used by researchers. We also note in order to test the effectiveness of the proposed neural network architecture, researchers trained and tested their model in several databases, and we clearly see that the recognition rate varies from one database to another with the same DL model (See Table.2).

Table2 summarize all the articles cited above, lists the architecture, the database and the recognition rate.

Table 2. Comparison between presented works.

Authors	databases	The architecture used	Recognition rate

Mollahosseini et al. [23]	MultiPie[10], MMI[11], DISFA[34], FERA[12], SFEW[13], CK+ [14], FER2013[15]	CNN	94.7%, 77.9%, 55%, 76.7%, 47.7%, 93,2%, 61.1%
Lopes et al.[24]	CK+[14], JAFFE[16], BU-3DFE[17]	CNN	96.76% for CK+
Mohammadpour et al. [25]	CK+[14]	CNN	97.01%
Cai et al.[26]	JAFFE[16], CK+[14]	SBN-CNN	95.24%, 96.87%
Li et al.[27]	RAF-DB[21], AffectNet[20]	ACNN	80.54%, 54.84%
Yolcu et al.[28]	RafD[22]	CNN	94.44%
Agrawal et Mittal.[29]	FER2013[15]	CNN	65%
Deepak jain et al.[30]	JAFFE[16], CK+[14]	CNN	95.23%, 93.24%
Kim et al.[31]	MMI[11], CASME II [18]	CNN-LSTM	78.61%, 60.98%
Yu et al.[32]	CK+[14], Oulu-CASIA[19], MMI[11], BP4D[35]	STC-NLSTM	99.8%, 93.45%, 84.53%
Liang et al.[33]	CK+[14], Oulu-CASIA[19], MMI[11]	DCBiLSTM	99.6%, 91.07%, 80.71%

5. Conclusion and future work:

This paper presented recent research on FER, allowed us to know the latest developments in this area. We have described different architectures of CNN and CNN-LSTM recently proposed by different researchers, and presented some different database containing spontaneous images collected from the real world and others formed in laboratories (SeeTable.1), in order to have and achieve an accurate detection of human emotions. We also present a discussion that shows the high rate obtained by researchers that is what highlight that machines today will be more capable of interpreting emotions, which implies that the interaction human machine becomes more and more natural.

FER are one of the most important ways of providing information about the emotional state, but they are always limited by learning only the six-basic emotion plus neutral. It conflicts with what is present in everyday life, which has emotions that are more complex. This will push researchers in the future work to build larger databases and create powerful deep learning architectures to recognize all basic and secondary emotions. Moreover, today emotion recognition has passed from unimodal analysis to complex system multimodal. Pantic et Rothkrantz [36] show that multimodality is one of the condition for having an ideal detection of human emotion. Researchers are now pushing their research to create and offer powerful multimodal deep learning architectures and databases, for example the fusion of audio and visual studied by Zhang et al. [37] and Ringeval et al. [38] for audio-visual and physiological modalities.

Acknowledgements

This work supported by the Directorate General of Scientific Research and Technological Development DGRSDT.

References

[1] E. Sariyanidi, H. Gunes, et A. Cavallaro, « Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition », *IEEE Trans. Pattern Anal. Mach. Intell.*, oct. 2014, doi: 10.1109/TPAMI.2014.2366127.

[2] C.-N. Anagnostopoulos, T. Iliou, et I. Giannoukos, « Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011 », *Artif. Intell. Rev.*, vol. 43, nº 2, p. 155-177, févr. 2015, doi: 10.1007/s10462-012-9368-5.

[3] L. Shu *et al.*, « A Review of Emotion Recognition Using Physiological Signals », *Sensors*, vol. 18, nº 7, p. 2074, juill. 2018, doi: 10.3390/s18072074.

[4] C. Marechal *et al.*, « Survey on AI-Based Multimodal Methods for Emotion Detection », in *High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet*, J. Kołodziej et H. González-Vélez, Éd. Cham: Springer International Publishing, 2019, p. 307-324.

[5] M. H. Alkawaz, D. Mohamad, A. H. Basori, et T. Saba, « Blend Shape Interpolation and FACS for Realistic Avatar », *3D Res.*, vol. 6, nº 1, p. 6, janv. 2015, doi: 10.1007/s13319-015-0038-7.

[6] P. V. Rouast, M. Adam, et R. Chiong, « Deep Learning for Human Affect Recognition: Insights and New Developments », *IEEE Trans. Affect. Comput.*, p. 1-1, 2018, doi: 10.1109/TAFFC.2018.2890471.

[7] C. Shan, S. Gong, et P. W. McOwan, « Facial expression recognition based on Local Binary Patterns: A comprehensive study », *Image Vis. Comput.*, vol. 27, nº 6, p. 803-816, mai 2009, doi: 10.1016/j.imavis.2008.08.005.

[8] T. Jabid, M. H. Kabir, et O. Chae, « Robust Facial Expression Recognition Based on Local Directional Pattern », *ETRI J.*, vol. 32, nº 5, p.

784-794, 2010, doi: 10.4218/etrij.10.1510.0132.
[9] S. Zhang, L. Li, et Z. Zhao, « Facial expression recognition based on Gabor wavelets and sparse representation », in *2012 IEEE 11th International Conference on Signal Processing*, oct. 2012, vol. 2, p. 816-819, doi: 10.1109/ICoSP.2012.6491706.
[10] R. Gross, I. Matthews, J. Cohn, T. Kanade, et S. Baker, « Multi-PIE », *Proc. Int. Conf. Autom. Face Gesture Recognit. Int. Conf. Autom. Face Gesture Recognit.*, vol. 28, nº 5, p. 807-813, mai 2010, doi: 10.1016/j.imavis.2009.08.002.
[11] M. Pantic, M. Valstar, R. Rademaker, et L. Maat, « Web-based database for facial expression analysis », in *2005 IEEE International Conference on Multimedia and Expo*, juill. 2005, p. 5 pp.-, doi: 10.1109/ICME.2005.1521424.
[12] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, et K. Scherer, « The first facial expression recognition and analysis challenge », in *Face and Gesture 2011*, mars 2011, p. 921-926, doi: 10.1109/FG.2011.5771374.
[13] A. Dhall, R. Goecke, S. Lucey, et T. Gedeon, « Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark », in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, nov. 2011, p. 2106-2112, doi: 10.1109/ICCVW.2011.6130508.
[14] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, et 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*, juin 2010, p. 94-101, doi: 10.1109/CVPRW.2010.5543262.
[15] I. J. Goodfellow *et al.*, « Challenges in Representation Learning: A Report on Three Machine Learning Contests », in *Neural Information Processing*, Berlin, Heidelberg, 2013, p. 117-124, doi: 10.1007/978-3-642-42051-1_16.
[16] M. Lyons, M. Kamachi, et J. Gyoba, « The Japanese Female Facial Expression (JAFFE) Database ». Zenodo, avr. 14, 1998, doi: 10.5281/zenodo.3451524.
[17] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, et M. J. Rosato, « A 3D facial expression database for facial behavior research », in *7th International Conference on Automatic Face and Gesture Recognition (FGR06)*, avr. 2006, p. 211-216, doi: 10.1109/FGR.2006.6.
[18] W.-J. Yan *et al.*, « CASME II: An Improved Spontaneous Micro-Expression Database and the Baseline Evaluation », *PLoS ONE*, vol. 9, nº 1, janv. 2014, doi: 10.1371/journal.pone.0086041.
[19] G. Zhao, X. Huang, M. Taini, S. Z. Li, et M. Pietikäinen, « Facial expression recognition from near-infrared videos », *Image Vis. Comput.*, vol. 29, nº 9, p. 607-619, août 2011, doi: 10.1016/j.imavis.2011.07.002.
[20] A. Mollahosseini, B. Hasani, et M. H. Mahoor, « AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild », *IEEE Trans. Affect. Comput.*, vol. 10, nº 1, p. 18-31, janv. 2019, doi: 10.1109/TAFFC.2017.2740923.
[21] S. Li, W. Deng, et J. Du, « Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild », 2017, p. 2852-2861.
[22] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, et A. van Knippenberg, « Presentation and validation of the Radboud Faces Database », *Cogn. Emot.*, vol. 24, nº 8, p. 1377-1388, déc. 2010, doi: 10.1080/02699930903485076.
[23] A. Mollahosseini, D. Chan, et M. H. Mahoor, « Going deeper in facial expression recognition using deep neural networks », in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, mars 2016, p. 1-10, doi: 10.1109/WACV.2016.7477450.
[24] A. T. Lopes, E. de Aguiar, A. F. De Souza, et T. Oliveira-Santos, « Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order », *Pattern Recognit.*, vol. 61, p. 610-628, janv. 2017, doi: 10.1016/j.patcog.2016.07.026.
[25] M. Mohammadpour, H. Khaliliardali, S. M. R. Hashemi, et M. M. AlyanNezhadi, « Facial emotion recognition using deep convolutional networks », in *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, déc. 2017, p. 0017-0021, doi: 10.1109/KBEI.2017.8324974.
[26] J. Cai, O. Chang, X. Tang, C. Xue, et C. Wei, « Facial Expression Recognition Method Based on Sparse Batch Normalization CNN », in *2018 37th Chinese Control Conference (CCC)*, juill. 2018, p. 9608-9613, doi: 10.23919/ChiCC.2018.8483567.
[27] Y. Li, J. Zeng, S. Shan, et X. Chen, « Occlusion Aware Facial Expression Recognition Using CNN With Attention Mechanism », *IEEE Trans. Image Process.*, vol. 28, nº 5, p. 2439-2450, mai 2019, doi: 10.1109/TIP.2018.2886767.
[28] G. Yolcu *et al.*, « Facial expression recognition for monitoring neurological disorders based on convolutional neural network », *Multimed. Tools Appl.*, vol. 78, nº 22, p. 31581-31603, nov. 2019, doi: 10.1007/s11042-019-07959-6.
[29] A. Agrawal et N. Mittal, « Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy », *Vis. Comput.*, janv. 2019, doi: 10.1007/s00371-019-01630-9.
[30] D. K. Jain, P. Shamsolmoali, et P. Sehdev, « Extended deep neural network for facial emotion recognition », *Pattern Recognit. Lett.*, vol. 120, p. 69-74, avr. 2019, doi: 10.1016/j.patrec.2019.01.008.
[31] D. H. Kim, W. J. Baddar, J. Jang, et Y. M. Ro, « Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition », *IEEE Trans. Affect. Comput.*, vol. 10, nº 2, p. 223-236, avr. 2019, doi: 10.1109/TAFFC.2017.2695999.
[32] Z. Yu, G. Liu, Q. Liu, et J. Deng, « Spatio-temporal convolutional features with nested LSTM for facial expression recognition », *Neurocomputing*, vol. 317, p. 50-57, nov. 2018, doi: 10.1016/j.neucom.2018.07.028.
[33] D. Liang, H. Liang, Z. Yu, et Y. Zhang, « Deep convolutional BiLSTM fusion network for facial expression recognition », *Vis. Comput.*, vol. 36, nº 3, p. 499-508, mars 2020, doi: 10.1007/s00371-019-01636-3.
[34] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, et J. F. Cohn, « DISFA: A Spontaneous Facial Action Intensity Database », *IEEE Trans. Affect. Comput.*, vol. 4, nº 2, p. 151-160, avr. 2013, doi: 10.1109/T-AFFC.2013.4.
[35] M. F. Valstar *et al.*, « FERA 2015 - second Facial Expression Recognition and Analysis challenge », in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, mai 2015, vol. 06, p. 1-8, doi: 10.1109/FG.2015.7284874.
[36] M. Pantic et L. J. M. Rothkrantz, « Toward an affect-sensitive multimodal human-computer interaction », *Proc. IEEE*, vol. 91, nº 9, p. 1370-1390, sept. 2003, doi: 10.1109/JPROC.2003.817122.
[37] S. Zhang, S. Zhang, T. Huang, et W. Gao, « Multimodal Deep Convolutional Neural Network for Audio-Visual Emotion Recognition », in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, New York, NY, USA, 2016, p. 281–284, doi: 10.1145/2911996.2912051.
[38] F. Ringeval *et al.*, « Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data », *Pattern Recognit. Lett.*, vol. 66, p. 22-30, nov. 2015, doi: 10.1016/j.patrec.2014.11.007.