1. Souha Ayadi[1], 2. Zied LACHIRI[2]

University of Tunis el Manar,Tunisia (1)(2), Signal Image and Information Technology Laboratory , SITI,
National Engineering school of Tunis
ORCID. 1. 0000-0002-6389-893

doi:10.15199/48.2022.03.20

Visual Emotion sensing using Convolutional Neural Network

***Abstract**. The objective of this article is to present a CNN architecture relevant to the Interactive Emotional Dyadic Motion Capture (IEMOCAP). Since the database showed some issues during the training phase, we are using frames as inputs instead of video recorder to minimize the error and increase the accuracy.* We apply the methodology of transfer learning by adjust the number of layers and the weight of the database. The results of the female and male genders are 91% and 89% respectively .

Streszczenie.** Celem tego artykułu jest przedstawienie architektury CNN odpowiedniej do interaktywnego emocjonalnego przechwytywania ruchu (IEMOCAP). Ponieważ baza danych wykazała pewne problemy w fazie uczenia, używamy klatek jako danych wejściowych zamiast rejestratora wideo, aby zminimalizować błąd i zwiększyć dokładność. Stosujemy metodologię transferu uczenia się dostosowując liczbę warstw i wagę bazy danych. Wyniki dla płci żeńskiej i męskiej wynoszą odpowiednio 91% i 89%.* (Wizualne wykrywanie emocji za pomocą splotowej sieci neuronowej***)

Keywords: Visual Emotion recognition, CNN.
Słowa kluczowe: Wizualne rozpoznawanie emocji, CNN.

Introduction

Emotion recognition is a very high topic over the last years.To automatically recognize emotions, the neural network architectures have been developed to improve the results to better imitate the human ability to recognize emotions. However, researchers spot the light on some issues encountered while dealing with this subject such as Chung-Hsien et al.[12] proved the difficulty of data collection and annotation for naturel scene. Pei et al.[14] worked on removing the degradation of the image to improve the classification results. Hossein et al.[15] worked on both audio and visual data using only some frames from the video. Wei et al.[16] presented a keyframe extraction algorithm based on CNN and SVM to avoid the influence of the silent frames on the results. Shambharkar et al.[17] used deep convolutional neural network (DCNN) on video sequences along with deer hunting optimization (DHO) for emotion classification from a movie trailer. Hong-Wei et al. [8] worked on proving the accuracy of their model by following a transfer learning for deep convolutional neural network (CNN). While khorrami et al.[10] apply it to recognize emotions in videos from Audio/Visual Emotion Challenge (AVEC2015). Moreover, Bradly et al.[4] used the CNN to train a single frame to predict the output label and used the pre-trained as a frame-wise feature extractor in order generate an input for RNN. Keya et al.[7] proposed a system for multimodal expression recognition in the wild by using deep convolutional neural network via transfer learning and presenting an approach of combining audio and visual features based on score fusion. Noroozi et al. [9] applied a CNN in order to summarize videos into key-frames for a multimodal emotion recognition.

However, there are some issues when processing some databases [12] that still have not yet been processed such as having more than one person in the video recorder, the angle from which the video was taken. hold, the distance between the person and the camera. This is the case with our chosen database (IEMOCAP database) which contains dyadic interactive emotions between two actors of the opposite sex.

This article is dedicated to solving the problem of the database of having more than one person in the frame, making the emotion detection difficult. And create a model capable of maintaining three main tasks, face detection, emotion classification, and gender detection.

To achieve our goal, we propose to apply a convolutional neural network (CNN) for emotion recognition.

Our work is an inspiration between two main approaches, Arriaga et al. that proposed a CNN model on fer2013 dataset and the accuracy was up to 96\% [1], while Bargal et al. [3] present an approach to recognize emotions using images in videos in order to improve the accuracy results.

This paper analyzes the visual information to recognize emotions such as {happy, sad, angry, neutral, surprised, disgust and fear} using IEMOCAP database. The visual information existed in the database in the form of video. The data is treated as frames instead of videos to facilitate the training phase and improve the learning ability of the model. We adopt transfer learning as methodology on a tested model and adjust the number of layers for the learning phase. The model is based on two-dimensional convolution (conv2D) because it is very robust for image processing.

In section two, We present a general overview about visual emotion recognition approaches.

In section three, We describe the CNN architecture by explaining the content of each layer.

In section four, we present the accuracy results and comparing it with the rsults in [1] and [3].

By the end, we will recapitulate our work, leading to our achievement.

Visual Emotion Expression

Visual emotion recognition considered a popular topic over the last decade. The main steps to recognize automatically the emotions are: Face Acquisition, Facial data extraction and representation, and Facial expression recognition [1]. Face acquisition is a processing stage to detect the region where the face exist, it can be by detecting the face for the first frame and track it in all the video, or detect it in each frame, and detect, of course, the head position. After face detection, there is the extraction of facial changes caused by a set of emotions, using two methods, which are “geometric features based methods” by extracting the facial features points or landmarks [13] to form a vector that represent the face geometry, and calculate the difference between the original point, when there is no emotion, and the moving point when there is an emotion detected. Then “appearance based methods” applied on all the face or a part of it to extract a feature vector. The last step is facial expression recognition, which is a set of emotions appears on the face can be identified as facial action units or prototypic emotional expressions [7].

Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) are both dedicated for emotion recognition. The RNN is mostly dedicated for sequential data, which is more related to speech recognition applications [6] because it preserve the information along with the time duration. Also, it goes backward and forward to extract the information, which took time more than we need. The RNN model presented in [11], seems to achieve good results for voice recorder for IEMOCAP database. When it comes to video recorders in [11] the results do not pass 50\%.

Moreover, for vision related applications, convolutional neural network (CNN) [14] consider as a powerful tool. CNN is good for extracting position-invariant features, beside, is easier to train and have many fewer parameters than fully connected networks with the same number of hidden units. It consists of a number of convolutional and subsampling layers optionally followed by fully connected layers, where the CNN take advantage of the fact that the input consists of images and they constrain the architecture in a more sensible way. Which bring the idea of training our dataset as images [2] for minimizing the error for better recognition and classification. In addition, combine this idea with transfer learning, which mean to use a trained network that have been trained before [1] and apply our dataset. Therefore, we freeze the weight of all the layers and remove only the last one and change it with our own classifier to fit our dataset, and train the network normally (Freezing the layers means not changing the weights during gradient descent/optimization).

Model

Our proposed model as shown in Fig1 based on feeding the input data to a sequence of layers. The real size of the input frame is [480x680x3]. To ease the calculation, the entry is resized to 48 height, 48 width, and three R, G, B color channels.

We apply two-dimensional convolution (conv2D) and Batch-Normalization on the input twice to speed up the process and improve the learning rate. The purpose of using conv2D is for feature extraction and better classification due to the parameters that characterize Conv2D and the purpose of using Batch-Normalization is to ease and preserve the stability of the learning process.

The convolution starts from the left sliding to the right by applying a filter (or a kernel) (5x5) all over the entire input image and the region applied to is receptive field, where the depth of the filter is the same as the depth of the input. As the filter is sliding or convolving, there is a multiplication of the values of the filter with the original pixel values and all summed up to finally have a single number. This process is repeated for each location on the input volume.

There are four group of layers consists of conv2D and Batch-Normalization twice , then a Rectified linear unit (ReLU) function is added to enhance the training process, and dropout which is a regularization technique used to prevent the model from errors that appeared during the learning phase. The parameters within the conv2D layer are the filters of size 16 for the first layer , 32 for the second layer, 64 for the third layer and 128 for the last layer, the kernel size (5x5) for the first two layers and (3x3) for the last two layers, and and use 'same' for padding to keep the same spatial dimensions as the input data . The parameters of the Average pooling are the pool-size (2x2) and 'same' for padding. We use "Average-pooling" when the length and the width of the input volume change where the depth not changing, that serves two main purposes, the first purpose is, the cost of the computation reduces, the second is it will control over-fitting where the model so destroyed and cannot generalize well for the validation and test sets and gives 100% on the training set and only 50% on the test data. To deal with over-fitting, the layer drop out a random set of activations by setting them to zero. By this way, the net will be able to provide the right classification even though it drop out some activations.The best dropout value working in this case is 0.5.

Fore layers was enough to achieve our goal and could be duplicated depends on the complexity of the database. The final stage of our model contain conv2D, Batch-Normalization, conv2D, Global average pooling because is more meaningful and interpretable and it allows the input image to be in any size. By the end, the resulting vector fed into the Softmax layer for classification and recognition.

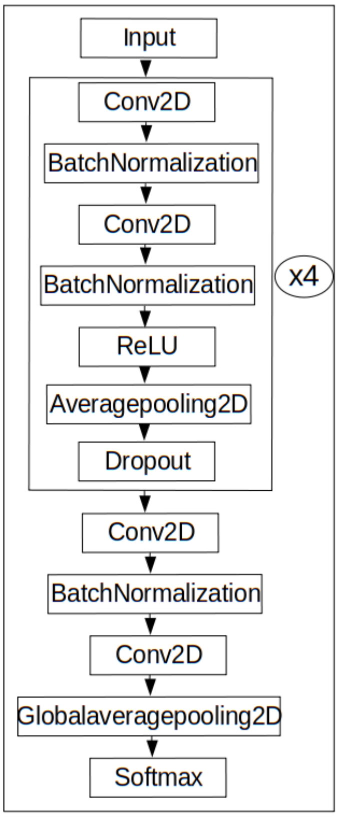

Fig.1. Description of the following process with the content of each layer of our CNN model.

Results and discussions

Database

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) [5] is a dataset dedicated for emotions recognitions that contain an audio-visual recording of dyadic mixed-gender pairs of actors performing a seen. There is 28 seens, every seen provide some emotions. In each recording, only one actor wears motion-capture MoCap markers on, basically the one on the left and the others being recorded by microphones and cameras, which means two persons appears in the video recording. The emotions provided in all the videos recorders are {"angry", "happy", "sad", "neutral", "frustrated", "excited", "surprised", "disgusted", "other"}[5]. The visual emotion expression data collected by asking subject, mixed in gender and ages, to perform screpted and unscrepted scenes leads to a set of emotions. Some obstacles comes, for face analysis, while performing in front of the camera such as presence of more than one person and the entourage of the subject that we interested by, head rotation, lightning,it influences accuracy of face detection, feature training and expression recognition.

Fig.2. IEMOCAP Database.

Most of the database are prepared with one person in front of the camera with no other person around, which is not the case for the IEMOCAP database [5],

Results

We attempt to train our dataset as images as shown in Fig.1 to focus more on the actor that wears the Mokap sensor, and not getting confused by the second actor, which leads us to ameliorate the results, giving to the difficulty that we've been deeling with, and accelerate the process because the weight of the images is much more lighter than the video recorder.

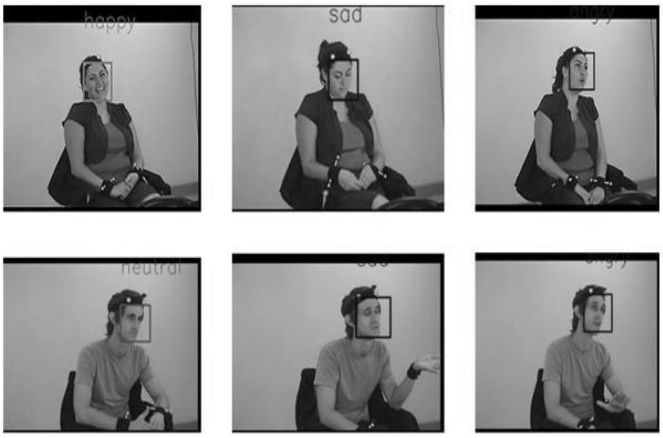

Fig.3. Face detection and emotion recognition.

Fig.3 present an emotion detection from facial expressions by detect the face area. The visualisation of the detection by drowing a rectangular on the face and the detected emotion is writed on top of the rectangle.

Table 1. Accuracy Results for IEMOCAP male and female gender.

	Female Accuracy	Male Accuracy
Happy	92%	91%
Sad	89%	93%
Neutral	93%	93%
Angry	85%	94%
Frustration	88%	91%
Surprise	97%	92%
Excited	89%	81%
Other	100%	84%

Table 2. Results for normalized confusion matrix in [1] and [3].

	Results in [1]	Results in [3]
Angry	60%	72.45%
Disgust	55%	0%
Fear	41%	37.14%
Happy	87%	81.94%
Sad	53%	42.50%
Surprise	77%	0%
Neutral	65%	74.09%

Moreover,each video recorder in IEMOCAP database contain two or three different emotions, we did focuse on detecting the main and essential emotions in every recorder, which lead us to the accuracy presented in the Table 1.

Comparing with both works presented in Table 2, we achieve good results for IEMOCAP database by 89% as shown in Table 1.

The confusion matrix presented in Fig.4 and the ROC curves presented in Fig.5 , are an example of one tested video of the IEMOCAP database for each gender. We tested the model on each video recorder without needing to remove any part of the video. The main detected emotions in the tested video , as shown in Fig.4 and Fig.5 , are sad and neutral.

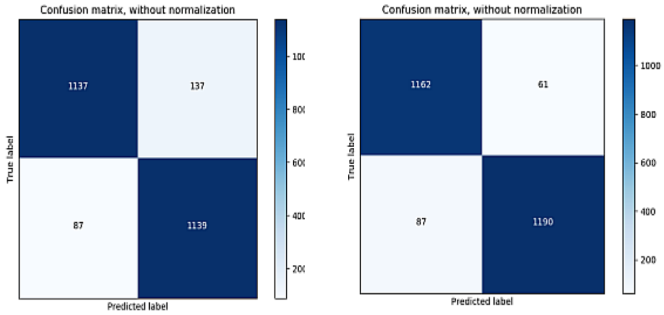

Fig.4. Example of the achieved confusion matrix for male and woman.

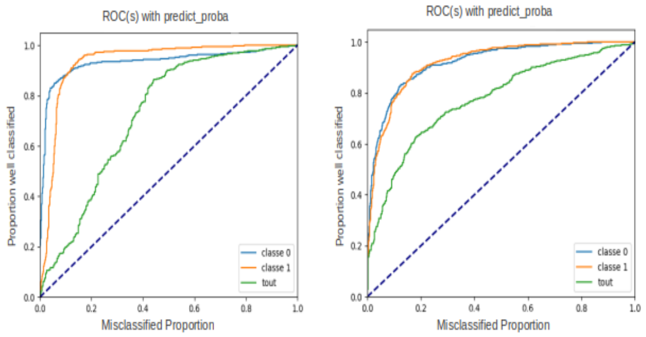

Fig.5. Example of the achieved ROC curves for male and woman.

The accuracy results for the commun emotions is ameliorated by at least 5%. For the emotion "Happy" we achieved 91.5% which is better by 5% comaring to [1] and by 9.56% comparing to [3], while for emotion "Sad", we got 91% which means ameliorated by 8% for the first work and by 48.5% for the second work. For the "Angry" emotion, we got 89.5%, 29.5% comparing to [1] and 17.05% comparing to [3]. For the "surprise" emotion, we achieved 94.5%, which means 17.5% comparing to [1] and 94.5% comparing to [3]. For the "Neutral" emotion, we had 93%, 28% comparing to [1] and 18.91% comparing to [3].

Conclusion

Because visual emotion recognition is a very high topic over the few last years and a very remarkable achievement was done in this field, the dyadic interactive emotions are a posed problem and still not been solved yet, especially when the database is surrounded by different problems that makes the emotion detection very difficult. In this article, an emotion recognition approach is applied on the IEMOCAP database. We were able to maintain three main tasks , face detection , emotion detection and gender recognition by training our database as images to minimize loss for better recognition and better results. We achieve our primary goal by improving a CNN model to become able to recognize emotions for dyadic interactive emotions and improving the accuracy results by reaching 91% for female and 89% for male. For a future continuation, we propose to add Long Shirt Term Memory (LSTM) for memorizing the results and not losing the information during the entire scene.

***Authors**: Souha AYADI, Signal Image and Information Technology(SITI) Laboratory, Department of Electrical Engineering, National Engineering School of Tunis, Campus Universitaire Farhat Hached el Manar BP 37, Le Belvedere 1002 TUNIS, E-mail: souha.ayadi@enit.utm.tn; Zied LACHIRI, Signal Image and Information Technology(SITI) Laboratory, Department of Electrical Engineering,National Engineering School of Tunis, Campus Universitaire Farhat Hached el Manar BP 37, Le Belvedere 1002 TUNIS, E-mail: zied.lachiri@enit.utm.tn.*

REFERENCES

[1] Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. Real-time convolutional neural networks for emotion and gender classification. arXiv preprint arXiv:1710.07557, 2017.

[2] Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, and Cha Zhang. Emotionrecognition in the wild from videos using images. In Proceedings of the 18th ACM International Conference on Multimodal Interaction, pages 433–436, 2016.

[3] Marco Bellantonio, Mohammad A Haque, Pau Rodriguez, Kamal Nasrollahi, Taisi Telve, Sergio Escalera, Jordi Gonzalez, Thomas B Moeslund, Pejman Rasti, and Gholamreza Anbarjafari. Spatio-temporal pain recognition in cnn based super-resolved facial images. In Video Analytics. Face and Facial Expression Recognition and Audience Measurement, pages 151162. Springer, 2016.

[4] Kevin Brady, Youngjune Gwon, Pooya Khorrami, Elizabeth Godoy, William Campbell, Charlie Dagli, and Thomas S Huang.Multimodal audio, video and physiological sensor learning for continuous emotion prediction. In Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, pages 97–104, 2016.

[5] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4):335, 2008.

[6] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. Representation learning for speech emotion recognition. In Interspeech, pages 3603–3607, 2016.

[7 Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image and Vision Computing, 65:66–75, 2017.

[8] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on international conference on multimodal interaction, pages 443–449, 2015.

[9] Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing, 10(1):60–75, 2017.

[10] Siyang Song, Enrique Sánchez-Lozano, Mani Kumar Tellamekala, Linlin Shen, Alan Johnston, and Michel Valstar. Dynamic facial models for video-based dimensional affect estimation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 0–0, 2019.

[11] Samarth Tripathi, Sarthak Tripathi, and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. arXiv preprint arXiv:1804.05788, 2018.

[12] Chung-Hsien Wu, Jen-Chun Lin, and Wen-Li Wei. Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. APSIPA transactions on signal and information processing, 3, 2014.

[13] Mira Jeong, Byoung Chul Ko, Sooyeong Kwak, and Jae-Yeal Nam. Driver facial landmark detection in real driving situations. IEEE Transactions on Circuits and Systems for Video Technology, 28(10):2753–2767, 2017.

[14] Yanting Pei, Yaping Huang, Qi Zou, Xingyuan Zhang, and Song Wang. Effects of image degradation and degradation removal to cnn-based image classification.IEEE transactions on pattern analysis and machine intelligence, 2019.

[15] M Shamim Hossain and Ghulam Muhammad. Emotion recognition using deep learning approach from audio–visual emotional big data. Information Fusion, 49:69–78, 2019.

[16] Jie Wei, Xinyu Yang, and Yizhuo Dong. User-generated video emotion recognition based on key frames. Multimedia Tools and Applications, 80(9):14343–14361, 2021.

[17] Prashant Giridhar Shambharkar and MN Doja. Movie trailer classification using deer hunting optimization based deep convolutional neural network in video sequences. Multimedia Tools and Applications, 79(29):21197–21222, 2020.