Facial expression recognition using the bilinear pooling

Marwa Ben Jabra, Ramzi Guetari, Aladine Chetouani, Hedi Tabia, Nawres Khlifa

Facial Expression Recognition using the Bilinear Pooling

Marwa Ben Jabra[1], Ramzi Guetari[2], Aladine Chetouani[3], Hedi Tabia[4] and Nawres Khlifa[1]
[1]*Université de Tunis El Manar, Institut Supérieur des Technologies Médicales de Tunis, Laboratoire de Biophysique et Technologies Médicales, 1006 Tunis, Tunisie*
[2]*Université de Tunis El Manar, Institut Supérieur d'Informatique de Tunis, Laboratoire LIMTIC, Tunisie*
[3]*Université d'Orléans, (Loire Valley University), Poltytech' Orléans, France*
[4]*IBISC, Univ. Evry, Université Paris-Saclay, 91025, Evry, France*

Keywords: Facial Expression Recognition, Image Classification, Deep Learning, Bilinear Pooling, Bilinear-CNN.

Abstract: Emotions taint our life and allow expressing the different facets of the personality. Among the expressions of the human body, facial ones are the most representative of the mindscape of a person. Several works are devoted to it and applications have already been developed. The latter, based on computer vision, are nevertheless facing some limitations and difficulties that are related to the point of view, lighting, occlusions, etc. Artificial Neural Networks (ANN) have been introduced to solve some of these limitations. The latter give satisfactory results, but still have not solved all the problems such as camera angle, the position of the head and, the occlusions, etc. In this paper, we review models of neural networks used in the field of recognition of facial emotions. We also propose an architecture based on the bilinear pooling in order to improve the results obtained by previous works and to provide solutions to solve these recurring constraints. This technique greatly improves the results obtained by architectures based on conventional CNNs.

1 INTRODUCTION

The automatic recognition of humans' facial expressions is an emerging field in scientific research. The latter allows new applications in other areas such as human-computer interaction, security, medicine, etc. The goal of a facial expressions recognition system is to identify simple or combined universal emotions by analyzing the characteristics of facial deformities of the mouth, eyes and eyebrows. Reliable recognition of facial expressions under natural conditions is therefore necessary, but remains an area where many problems are still unresolved (variations in lighting, position of the head, occlusions, etc).Artificial Intelligence has made a quantum leap in recent years, with the advent of neural networks and machine learning. These techniques, combined with computers, sensors and cameras, have produced increasingly intelligent and autonomous systems. They have also made it possible to significantly change the artificial vision that is at the base of the recognition of individuals and their facial expressions. In the literature, several approaches have been proposed. Recently a special attention has been given to the methods based on the deep learning and particularly CNNs. The latter showed a real potential to overcome certain difficulties in the process of recognition and have thus significantly increased the classification rate.

Several works are devoted to this field but differ considerably by the CNN architectures adopted, the learning and test protocols applied, the parameters used (filter size, number of filters, etc.), etc. Indeed, the results that we studied in the literature and experimented with different architectures differ considerably from one model to another, depending on the choice of parameters. In addition, the same model can produce different results on the same set of data with two different initializations.

The fact that two CNNs give different results or that the same CNN with different initialization can give different results has spurred the idea of using the bilinear pooling (Liu et,2014) that consists in using two CNNs in parallel as described in § C. We believe that bilinear CNNs drastically improve the recognition performance comparing to previous works. This is what we are trying to demonstrate in this paper.

The contributions we bring through this study can be summarized as follows: first of all, we propose the

use of a pre-trained model (such as VGG16, Inception or ResNet50) with a transfer learning to classify the basic emotions. We then propose our own shallow CNN architecture built from scratch for the same purpose. This architecture has fewer layers than predefined architectures and a smaller number of parameters. Finally, we propose the use of the bilinear pooling to improve the results.

The rest of this paper is organized into four sections. In Section 2, we present the related work on the recognition of facial emotions with the CNNs and the bilinear CNN models, as well as the relevant literature on the combination of features. Then, In Section 3, we describe our contribution and our proposed deep neural network architectures for emotion recognition. In Section 4, we present our experimental results. The last section is devoted to the conclusion.

2 RELATED WORK

The recognition of facial emotions is based on facial landmarks. The most significant facial references that have a real impact on the facial analysis and recognition are the corners of the eyes, the eyebrows, the lobes of the ears, the tip of the nose, nostrils, chin, mouth, etc. While humans are naturally able to understand these emotions, facial landmarks identification is a challenging task in automatic emotions recognition. In fact, the performance of an automated approach depends on the selection of a discriminate feature set and the use of an efficient learning technique. Recent researches have shown the potential and effectiveness of the use of CNNs. Unlike other ANNs that require a pre-processing phase to set the learning features, CNNs are deep artificial neural networks in which, the convolution blocks automatically select the most relevant filters to extract relevant features.

Behzad Hasani et al. (Behzad 2017) proposed a 3D CNN architecture named 3D Inecption-ResNet (3DIR) for the extraction of spatial relationships in facial images and the temporal relationships between different video frames. The approach extracts 66 landmarks from the face using a regression method of local binary characteristics (Dhananjay 2014). These landmarks are used as inputs to the 3DIR model during the training phase. Toevaluate the proposed method they used four facial expression databases, which are MMI (Pantic 2005), CK+ (Lucey 2010), GEMEP-FERA (Banziger 2010), and DISFA (Mavadati 2013). They obtained a recognition rate of 67.52% on CK+ dataset.

Zhan Wu et al (Zhan 2017) proposed a three-stream 3D CNN that automatically extracts both spatial and temporal characteristics. The purpose is to fuse local and global facial expression features by using different methods that could produce different recognition rates. The concatenation method has given the best recognition rate of 78.42%.

Octavio Arriaga et al. (Arriaga 2017) proposed two real-time CNN models to simultaneously perform the face detection, the gender recognition and the emotion classification tasks. In the first model, they removed the fully connected layers and used a Global Average Pooling, which reduces each feature map to a scalar value. The second model is inspired by the Xception (Chollet 2017) architecture in which fully connected layers are eliminated and residual modules (Kaiming 2016) have been included. They also proposed to use depth-wise separable convolutions that separate the processes of features extraction to depth-wise convolution layers and point-wise convolution layers to separate the spatial cross-correlations from the channel cross-correlations. The accuracy rate for emotion recognition and classification reached 66% using the FER-2013 dataset.

Liu et al. (Liu 2014) proposed a CNN model named 3DCNN-DAP for the emotion recognition, based on Deformable Action Parts (DAP). This work focuses on the deformable facial action part, which is the most important facial information when analyzing emotions and dynamically changes according to the facial expressions. They proposed a technique for learning deformable parts of action and for detecting specific parts of the action of the face under structured spatial constraints. To evaluate the proposed method, they used CK+ and MMI posed expression datasets and, the FERA spontaneous dataset. The best accuracy rate they have obtained is 87.5% on CK+ dataset.

In their paper, Khorrami et al. (Khorrami 2015) proposed a CNN architecture for the selection of the most relevant parts of the human face allowing the recognition of facial expressions. They built a zero-bias CNN network, which consists on ignoring the biases of the convolutional layers in order to reduce the number of parameters. To evaluate the performance of this method, they used the CK + and Toronto Face Dataset (TFD) datasets. They obtained an accuracy rate of 81.8% with the CK + dataset and an accuracy rate of 79.4% with the TFD dataset.

Li et al. (Li 2018) proposed a Patch-Gated CNN (PG-CNN) model that breaks down the face into facial regions. They extracted an intermediate features map of the facial image using VGGnet

model. After, this intermediate features map is decomposed by the PG-CNN into 24 sub-feature-maps corresponding to 24 local patches. They then calculated the weight of each patch to perceive facial occlusions. Finally, the representation of the occluded face is obtained by the concatenation of the weighted local features. They achieved an accuracy of 80.28% using the CK + database.

In the paper of Xie and Hu (Xie 2019), the Deep Comprehensive Multi-patches Aggregation CNN (DCMA-CNN) for the recognition of facial expression has been introduced. This architecture is composed of two branches of a CNN. One branch extracts local features depicting the expression details of image patches, while the other extracts global features characterizing the high-level semantic information of an expression derived from the entire expression image. The local and global features are fused for the classification task. To evaluate the model they used CK+ and JAFFE facial expression datasets, and they obtained a rate accuracy of 88.67%.

Zhou et al. (Zhou 2018), proposed a method that uses bilinear-CNN models. They transformed facial expression analysis from a classification problem into a regression one to predict the facial expression updates based on local appearance features. They explored two different deep CNN architectures, VGG16 and ResNet50. To improve these methods by a bilinear pooling, they modified them accordingly to serve the regression task. Specifically, they removed all the layers after the global pooling layer that aggregated all features spatially; these layers include the last convolution layer and the loss layer. Over the global pooling layer, they stacked a new convolution layer with randomly initialized weights to output final embedding vectors for input images. They evaluated their models on the FER-2013 dataset and they obtained a rate accuracy of 81,79% using global pooling VGG16 and 83,78% using bilinear pooling VGG16. The proposed method recognize the basic emotion and a full range of other emotions.

Zhang et al. (Zhang 2019) proposed a Bilinear CNN model based on factorized bilinear pooling (FBP) that aims to use the audio and video features for the emotion recognition. In this proposed method, they extracted the emotion features from video and audio using two different parallel streams of CNN. Then, they fused the outputs feature maps using factorized bilinear pooling. Finally, they classify the emotion using a fully connected layer followed by a Softmax function. To validate their method, they used the audio-video AFEW dataset and they obtained a rate accuracy of 62,48%.

3 PROPOSED METHOD

The main contribution of this work is the use of bilinear pooling by combining a standard CNN (with generally performs well, but suffers from a huge number of parameters and are time consuming) and a shallow CNN built from scratch, which has a reduced number of parameters, and an acceptable recognition rate. Bilinear pooling is considered as a Second-order aggregation of CNN activations, which can provide improvements over classical representations using first-order aggregation (e.g., sum or max). Actually, the Bilinear CNN (B-CNN) is considered as a state-of-the-art network architecture for texture and fine-grained recognitions. A bilinear CNN model consists of two parallel CNNs, each of them extracts the features map from an input image. Features maps produced by each of the two CNNs are multiplied using the outer product generating second-order characteristics. The latter are then pooled to form high dimensional bilinear features to obtain an image descriptor. Bilinear features maps are normalized using sqrt (square root) and L2 normalization. Finally, fully connected layers and a Softmax function are used for the classification task. Such bilinear architectures have been proposed for the recognition and classification of facial emotions. Bilinear CNN models have proven effective for fine recognition, scene categorization, texture recognition, and visual question-and-answer tasks, among others. They are able to distinguish the subtle differences between cars, birds and planes (Lin et al, 2017) (Lin et al, 2015). In the following, we present the transfer learning process, the architecture of the shallow CNN as well as the bilinear architecture.

3.1 Transfer Learning for Emotion Recognition

First, we proposed to explore standard CNN architectures such as (Densenet, Mobilenet V1 & V2, InceptionResnetV2, NasnetMobile, Xception, ResNet50, VGG16 and VGG19), which are widely studied and exploited in practical applications. To train these architectures to the classification of the basic facial expression, we combined the transfer learning and the fine tuning. The transfer learning is a machine learning technique that consists of using a model trained in a particular task to perform another one. The fine tuning consists in using an already trained model and training it on a specific dataset. This technique requires that the last fully connected layer be replaced by another one adapted to the new task to be performed (classifying the seven basic

facial expressions in our case). These deep models have performed well, but they use a huge number of parameters for the training.

3.2 Shallow CNN Architecture

To obtain a good performance with small number of parameters, we proposed to introduce a new CNN architecture with a small number of layers. After several experimentations, we built a shallow CNN with 7 convolution layers, 3 max-pooling layers and a single fully connected layer. The table 1 summarizes the architecture of the proposed shallow CNN. To train this architecture to the classification of the basic facial expression, we used transfer learning and the fine tuning. Our architecture has been trained on facial expressions datasets. The classification of the facial expressions with the shallow architecture has given the accuracy rate of 67.78% (on the CK+ dataset), which is quite good considering the depth of the CNN and its number of parameters.

Table 1: Architecture of the shallow CNN.

Layer	Output shape	Params. count
Input Layer	224, 224, 3	0
Convolution	224, 224, 64	1792
Convolution	224, 224, 64	36928
Max-pooling	112, 112, 64	0
Convolution	112, 112, 128	73856
Convolution	112, 112, 128	147584
Max-pooling	56, 56, 128	0
Convolution	56, 56, 256	295168
Convolution	56, 56, 256	590080
Convolution	56, 56, 256	590080

3.3 Bilinear Pooling for Emotion Recognition

A bilinear CNN model consists of two parallel CNNs, each of them extracting the features map from an input image. Features maps produced by each of the two CNNs are multiplied using the outer product producing second-order characteristics. The latter are then pooled to form high-dimensional bilinear features to obtain an image descriptor. Bilinear features maps are normalized using sqrt (square root) and L2 normalization. Finally, fully connected layers and a Softmax function are used for the classification task. Such bilinear architectures have been proposed for the recognition and classification of facial emotions. Bilinear CNN models have proven effective for fine recognition, scene categorization, texture recognition, and visual question-and-answer tasks, among others. They are able to distinguish the subtle differences between cars, birds and planes (Lin 2017) (Lin 2015).

The bilinear architecture nevertheless poses some performance problems in terms of temporal complexity. The learning phase of both networks is time consuming. Even the recognition phase can often be time consuming. This is due to the depth of the architectures used and the number of parameters required for each branch of the bilinear architecture. Our goal is to reduce the processing time of a bilinear architecture while keeping a very good accuracy rate. The combination of a standard architecture with our shallow architecture seemed to us a good compromise that needed to be tested.

We tested a number of bilinear configurations. We combined two instances of standard architectures as well as combined with our shallow CNN. All the CNNs have been modified to become a bilinear architecture. All layers after the last convolution layer have been removed, namely the last pooling layer, the fully connected layer, and the loss layer. The weights of each CNN have also been fixed. The two CNNs were used in parallel with a different weight initialization for each of them in order to extract the feature maps from the same input images. These feature maps were merged using the Khroneker product (Paumard 2018). Two normalization layers follow the features extraction (sqrt and L2 have experimented with more or less the same performance). Finally, a fully connected layer with 7 outputs followed by a Softmax function was used for the final classification. The bilinear architecture that combines VGG16 and our shallow CNN is presented in Fig. 1.

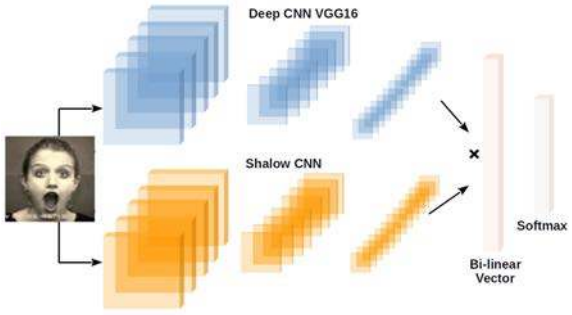

Figure 1: The proposed Bilinear CNN architecture.

4 RESULTS AND DISCUSSION

In this section, we present the facial expression databases used in our experiments and we detail the experimental results.

Table 2: Results obtained using standard configurations and bilinear ones on CK+ dataset.

	Standard CNNs			Bilinear pooling		
ARCHITECTURES	Training Accuracy	Test accuracy	Number of parameters	Training Accuracy	Test accuracy	Number of parameters
DenseNet (1)	53,10%	35,96%	8,062,504	91,08%	34,09%	7,040,538
DenseNet (2)	52,68 %	33,40%				
MobileNet-V1 (1)	53,75%	31,09%	4,253,864	96,53%	37,38%	14,568,903
MobileNet-V1 (2)	56,06 %	28,46%				
MobileNet V2 (1)	60,51 %	29,59%	3,538,984	100,00%	42,82%	13,721,671
MobileNet V2 (2)	61,32 %	31,87%				
NASNetMobile (1)	62.10%	30,90%	5,326,716	75,43%	41,18%	8,526,964
NASNetMobile (2)	63,30 %	34,09%				
InceptionResNetV2 (1)	61,5 7%	36,13%	55,873,736	82,62%	39,20%	49,758,951
InceptionResNetV2 (2)	59,77 %	33,41%				
ResNet50 (1)	58,38%	40,07%	25,636,712	86,90%	41,50%	44,363,911
ResNet50 (2)	59,03 %	42,98%				
Xception (1)	51,68%	42,31%	22,910,480	77,48%	36,77%	50,221,615
Xception (2)	53,99 %	44,01%				
VGG16 (1)	62,21%	69,16%	138,357,544	100,00%	84,30%	16,549,703
VGG16 (2)	65,89 %	70,60 %				
VGG19 (1)	63,91%	68,91%	143,667,240	100,00%	83,12%	21,859,399
VGG19 (2)	64,88 %	69,33%				
Shallow (1)	65,21%	67,78 %	3,140,423	100,00%	78,65%	3,929,735
Shallow (2)	62,97 %	65,77%				
VGG16 (2)	65,89 %	70,60%	143,667,240	100,00%	86,98%	9,470,279
Shallow (1)	65,21%	67,78%	3,140,423			

Table 3: Results obtained using standard configurations and bilinear ones on FEI dataset.

ARCHITECTURES	Training Accuracy	Validation accuracy	Number of parameters	Training Accuracy	Validation accuracy	Number of parameters
DenseNet (1)	60.34%	25,32 %	8,062,504	80,52%	31,09%	7,040,538
DenseNet (2)	59,63 %	23,99 %				
MobileNet-V1 (1)	60.62%	29,63 %	4,253,864	68,55%	31,62%	14,568,903
MobileNet-V1 (2)	61,08 %	30,10 %				
MobileNet V2 (1)	64,75 %	32,23 %	3,538,984	75,65%	42,33%	13,721,671
MobileNet V2 (2)	65,03 %	34,73 %				
NASNetMobile (1)	54.03%	22,91 %	5,326,716	65,43%	32,53%	8,526,964
NASNetMobile (2)	57,88 %	25,83 %				
InceptionResNetV2 (1)	61,5 7%	36,13%	55,873,736	81,43%	45,20%	49,758,951
InceptionResNetV2 (2)	60,55 %	34,11 %				
ResNet50 (1)	66,85%	50,49%	25,636,712	86,90%	51,31%	44,363,911
ResNet50 (2)	63,36 %	48,77 %				
Xception (1)	51,68%	32,31%	22,910,480	57,58%	45,46%	50,221,615
Xception (2)	49,65 %	36,98 %				
VGG16 (1)	70,09%	80,89 %	138,357,544	100,00%	83,52%	16,549,703
VGG16 (2)	69,83 %	79,89 %				
VGG19 (1)	68,39%	78,21%	143,667,240	100,00%	82,77%	21,859,399
VGG19 (2)	66,97 %	76,88%				
Shallow (1)	68,20%	67,20 %	3,140,423	97,67%	77,15%	3,929,735
Shallow (2)	65,37 %	64,20 %				
VGG16 (2)	70,09 %	80,89 %	143,667,240	100,00%	85,35%	9,470,279
Shallow (1)	67,21%	57,33 %	3,140,423			

4.1 Databases

In our experiments, we used CK+ and FEI facial expression datasets, which consist of seven basic expressions (anger, disgust, fear, happiness, sadness and surprise). The CK+ dataset (Cohn-Kanade extended dataset) is the extension of the CK one and it is composed of 327 video sequences. We extracted the last three images from each sequence of the CK+ video dataset to build a new image dataset with 981 mainly grey images. The FEI facial expression dataset has been created in the Artificial Intelligence Laboratory of FEI Brazil and it is composed of 252 images.

4.2 Results

We used the same CNN standard architecture with different initializations and we combined them to form the bilinear CNN model. The accuracy rate using the bilinear architecture has been improved compared the initial results. The bilinear VGG16 model gives the best accuracy rate but it still uses a huge number of parameters. A bilinear architecture using two instances of the shallow CNN has also been tested. It resulted in the third best accuracy while using a reduced number of parameters. Finally, we implemented a bilinear architecture by combining VGG16 with the shallow CNN to see the impact on the accuracy rate while reducing the number of parameters and then reduce the processing time.

Table 2 and table 3 summarize the results obtained using standard configurations and bilinear ones on respectively CK+ and FEI dataset. For the CK+ dataset, the fine-tuned VGG16 achieved the best accuracy rate, while the shallow CNN has achieved the third performance with an accuracy rate of 67,78%. In a bilinear configuration, VGG16 has also achieved the best performance with an accuracy rate of 84.30 %, and the shallow CNN has realized an acceptable performance with an accuracy rate 78,65%.

The combination of VGG16 and the shallow CNN achieved the best performance of all standard and bilinear configurations tested. It achieved the accuracy rate of 86,98%.For the, FEI Dataset, the VGG16 in a standard configuration reached the accuracy rate of 80.89%, which is the best performance of all the tested architectures. Our shallow CNN has given the third performance with an accuracy rate of 67,20%. The VGG16 in a bilinear configuration has given the best performance of all the standard architectures used in the same configuration. It, indeed reached the accuracy rate of 83.52%. Our shallow model used in a bilinear configuration has given the third accuracy rate that is 77,15%. Finally, the bilinear configuration in which we combined the VGG16 model with our shallow model gives the best performance of all the bilinear configurations as well as he standard configuration with an accuracy rate of 85,35%.

4.3 Comparison with the State of the Art

In Table 4, we compare the performance of the proposed Bilinear model on CK+ dataset, with different other methods, including Hand-crafted-based methods (TMS (Lin 2017), CDA (Jain 2011), MSR (Jain 2011) and ITBN (Rifai 2012)) and deep-based methods (3DCNN (Kaiming 2016), Zero-bias CNN (Liu 2014b), PG-CNN (Khorrami 2015) and LFCNN (Zhao 2011)). The recognition accuracy of some methods that are involved in comparison is far lower than the proposed model, which validates the effectiveness of our model.

Table 4: Results of hand-crafted and CNN models on (CK+) Dataset.

	Method	Validation accuracy
Hand-crafted features	TMS (Lin 2017)	91.80%
	CDA (Jain 2011)	85.00%
	MSR (Jain 2011)	91.40%
	ITBN (Rifai 2012)	86.30%
CNN Architecture	3DCNN-DAP (Kaiming 2016)	87.90%
	Zero-bias CNN (Liu 2014b)	81.80%
	PG-CNN (Khorrami 2015)	80.28%
	LFCNN (Zhao 2011)	88.67%
Our CNN models	VGG16	70.60%
	Shallow CNN	67.78%
	Bilinear VGG16	84.30%
	Bilinear Shallow CNN	78.65%
	Proposed Bilinear model	86.98%

In Table 5, we compare our proposed CNN models with (PCA (Wang 2013), LDA methods (Thomaz 2010) using FEI dataset. We notice that the handcrafted methods implemented FEI dataset give an accuracy rate better than the proposed model.

For the fine-grained facial expression analysis, we experimented various CNN architectures such as VGG16, ResNet50, Mobilnet, as well as our own shallow model. We evaluated these architectures on a standard CK + and FEI databases. The results have shown that deep CNN models achieve very promising results. Moreover, the bilinear CNN models have significantly outperformed their respective counterparts.

Table 5: Results of hand-crafted and CNN models on FEI dataset.

	Architectures	CNN	Proposed Bilinear CNN
CNN Architecture	VGG16	70,0%	85,35%
	Shallow CNN	67,2%	
Handcrafted features	PCA (Wang 2013)		94,00%
	LDA (Thomaz 2010)		92,20%

5 CONCLUSIONS

Advances in modern techniques in artificial intelligence have led to considerable advances in the field of facial recognition. Nevertheless, there are still a number of problems to solve in order to achieve performances close to those of humans. Our

contribution is, in our opinion, a contribution to advancing research in this area in the right direction.

Normally, the use of CNNs imposes the introduction of a large number of parameters with more or less efficient classification rates. The most powerful architectures often use the largest number of parameters. We tested these architectures individually and combined using the bilinear pooling. We found that a bilinear architecture improves performance, but the number of parameters remains too large.

To measure the impact of the number of parameters in the classification performance, we developed our own shallow architecture which, despite the small number of parameters, still gave good results when it's used individually.

We then used our shallow model with the standard VGG16 architecture that gave us the best performance. We evaluated our proposed bilinear model on the CK + and FEI datasets. We found that this model allowed us to obtain a good precision rate, namely 86.98% using the CK+ dataset, and 85.35% on the FEI dataset while using a reduced number of parameters. We can therefore conclude that the bilinear model CNN has clearly exceeded the performance of simple deep CNN models.

We compared the results with other deep models and handcrafted methods. We found that we performed better than some methods. We also found that methods that use local and global characteristics exceeded our method in terms of performance.

In perspective, we will propose a CNN model that extracts local and global characteristics and merges them with bilinear pooling to improve the performance of our method.

REFERENCES

Alizadeh S. and Fazel A. 2017. Convolutional Neural Networks for Facial Expression Recognition. *In Computer Vision and Pattern Recognition, arXiv: 1704.06756v1*.

Arriaga O., Valdenegro-Toro M. and Plöger P, 2018. Real-time convolutional neural networks for emotion and gender classification. *arXiv: 1710.07557T*.

Banziger T. and Scherer K. R., 2010. Introducing the geneva multimodal emotion portrayal (gemep) corpus. *Blueprint for affective computing: A sourcebook, pages 271–294*.

Behzad H. and Mahoor MH, 2017. Facial Expression Recognition Using Enhanced Deep 3D Convolutional Neural Networks. *arXiv: 1705.07871v1 [cs.CV]*.

Chollet F, 2017. Xception: Deep Learning with Depthwise Separable Convolutions. *In the proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Dhananjay R., Vinay A, Shylaja SS. and Natarajan S, 2014. Facial Landmark Localization – A Literature Survey. *International Journal of Current Engineering and Technology E-ISSN 2277 – 4106, P-ISSN 2347 – 5161 Vol.4, No.3*.

Jain S., Hu C., and J. K. Aggarwal, 2011. Facial expression recognition with temporal modeling of shapes. *In ICCV Workshops, pages 1642–1649*.

Kaiming H., Zhang X., Ren S., and Sun J., 2016. Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778*.

Khorrami P. and Huang S., 2015. Do Deep Neural Networks Learn Facial Action Units When Doing Expression Recognition? *In ICCV Workshop page 19-27*

Laptev D., Savinov N., Buhmann JM. and PollefeysM., 2016. Ti-pooling: Transformation-invariant pooling for feature learning in convolutional neural networks. *In Proc. Int. Conf. Comput. Vis. Pattern Recognit., pp. 289-297*.

Li Y., J. Zeng, S. Shan and X. Chen. Patch-Gated CNN for Occlusion-aware Facial Expression Recognition. 2018 24th International Conference on Pattern Recognition (ICPR) 1051-4651 29 November 2018

Lin T.Y., RoyChowdhury A., and Maji S., 2017. Bilinear Convolutional Neural Networks for Fine-grained Visual Recognition. *In IEEE Trans. Pattern Anal. Mach. Intell*.

Liu, P. (a), Han, S., Meng, Z., and Tong, Y., 2014. Facial Expression Recognition via a Boosted Deep Belief Network. *In the proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu M. (b), Li S., Shan S., Wang R., and Chen X., 2014. Deeply learning deformable facial action parts model for dynamic expression analysis. *In Asian Conference on Computer Vision, pages 143–157. Springer*.

Lucey P., Cohn JF., Kanade T., Saragih J., Ambadar Z., and Matthews I., 2010 The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. *In Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society Conference on, pages 94–101. IEEE.*

Mavadati SM., Mahoor MH, Bartlett K., Trinh P., and Cohn JF., 2013. Disfa: A spontaneous facial action intensity database. *In IEEE Transactions on Affective Computing, 4(2):151–160.*

Pantic M., Valstar M., Rademaker R., and Maat L., 2005. Webbased database for facial expression analysis. *In IEEE international conference on multimedia and Expo, pages 5– pp*.

Paumard M.M, Picard D. and Tabia H., 2018. Image Reassembly Combining Deep Learning and Shortest Path Problem, he European Conference on Computer Vision.

Rifai S., Bengio Y., Courville A., Vincent P., and Mirza M., 2012. Disentangling factors of variation for facial expression recognition. *ECCV 2012, pages 808–822*.

Thomaz CE. and Geradli GA., 2010. A new ranking method for principal components analysis and its application to face image analysis. *Image and Vision Computing Volume 28, Issue 6, pp 902-913*.

Wang Z., S. Wang, and Q. Ji. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In CVPR, pages 3422–3429, 2013

Xie S. and Hu H., 2019. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks. *In IEEE Transactions on Multimedia, Volume 21, Issue 1, pp 211 - 220.*

Zhan W., Chen T., Chen Y., Zhang Z. and Yuan G, 2017. NIRExpNet: Three-Stream 3D Convolutional Neural Network for Near Infrared Facial Expression Recognition. *Appl. Sci. 7, 1184*.

Zhang Y., Wang Z., Du J., 2019. Deep Fusion: An Attention Guided Factorized Bilinear Pooling for Audio-video Emotion Recognition. *arXiv: 1901.0488*

Zhao G., X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, Facial expression recognition from near-infrared videos. Image & Vision Computing, vol. 29, no. 9, pp. 607–619, 2011

Zhou F., Kong S., Fowlkes C., Chen T., and Lei B., 2018. Fine-Grained Facial Expression Analysis Using Dimensional Emotion Model. *arXiv: 1805.010242*