Facial expression recognition via a jointly-learned dual-branch network

Original Scientific Paper

Yamina Bordjiba

University of Badji Mokhtar,
Faculty of Technology,
Department of Computer Science
BP 12, Annaba, Algeria
bordjiba.yamina@univ-guelma.dz

University of 8 mai 1945,
Faculty of Mathematics, computer science and sciences of matter,
Department of Computer Science, Labstic Laboratory
BP 401, Guelma, Algeria

Hayet Farida Merouani

University of Badji Mokhtar,
Faculty of Technology, Department of Computer Science, LRI Laboratory
BP 12, Annaba, Algeria
hayet.merouani@univ-annaba.org

Nabiha Azizi

University of Badji Mokhtar,
Faculty of Technology, Department of Computer Science, Labjed Laboratory
BP 12, Annaba, Algeria
azizi@labged.net

Abstract *– Human emotion recognition depends on facial expressions, and essentially on the extraction of relevant features. Accurate feature extraction is generally difficult due to the influence of external interference factors and the mislabelling of some datasets, such as the Fer2013 dataset. Deep learning approaches permit an automatic and intelligent feature extraction based on the input database. But, in the case of poor database distribution or insufficient diversity of database samples, extracted features will be negatively affected. Furthermore, one of the main challenges for efficient facial feature extraction and accurate facial expression recognition is the facial expression datasets, which are usually considerably small compared to other image datasets. To solve these problems, this paper proposes a new approach based on a dual-branch convolutional neural network for facial expression recognition, which is formed by three modules: The two first ones ensure features engineering stage by two branches, and features fusion and classification are performed by the third one. In the first branch, an improved convolutional part of the VGG network is used to benefit from its known robustness, the transfer learning technique with the EfficientNet network is applied in the second branch, to improve the quality of limited training samples in datasets. Finally, and in order to improve the recognition performance, a classification decision will be made based on the fusion of both branches' feature maps. Based on the experimental results obtained on the Fer2013 and CK+ datasets, the proposed approach shows its superiority compared to several state-of-the-art results as well as using one model at a time. Those results are very competitive, especially for the CK+ dataset, for which the proposed dual branch model reaches an accuracy of 99.32, while for the FER-2013 dataset, the VGG-inspired CNN obtains an accuracy of 67.70, which is considered an acceptable accuracy, given the difficulty of the images of this dataset.*

Keywords*: facial expression recognition, deep learning, CNN, VGGnet, transfer learning, EfficientNet, dual branch network, features fusion.*

1. INTRODUCTION

Based on a cross-cultural study, Ekman et al. [1] defined six basic emotional expressions: disgust, anger, fear, happiness, sadness, and surprise. Since these expressions are universal among human beings, they demonstrate that certain basic emotions are perceived in the same way among human beings, independently of their culture. While recent advanced research in neuroscience and psychology has indicated that the six basic emotions model is not universal [2], but culture-specific, most studies in the field of facial expression recognition focus on this model.

During the last decades, facial expression recognition has emerged as an important and challenging topic in several fields such as computer vision, artificial intelligence, and human-computer interaction. Generally, traditional works on facial expression recognition were conducted in two steps: first, expression features

are extracted to represent the given image/video, and then a classification stage is carried out to recognize the different expressions from the extracted features. Most conventional methods rely on handcrafted features or shallow learning, such as Neural Network [3], Bayesian Network (BN) [4], Support Vector Machine (SVM) [5], Adaboost [6], and Random Forest [7].

For conventional facial expression recognition systems, the extraction of facial features is a very crucial step, and it affects the later classification decision. Generally, it should be noted that the employed methods to extract these handcrafted features use labeled data in the context of supervised learning. In addition, these handcrafted features such as the representation of LBP and Gabor wavelets capture low-level information on facial images, except the high-level representation of facial images [8]. In addition to that, conventional approaches require relatively less computational and memory power than approaches based on deep learning. For these reasons, these approaches are still under study for use in real-time embedded systems because of their low computing complexity and high accuracy [9]. By investigating several approaches for facial expression recognition, deep neural networks generally offer better classification performance and achieve very good results in terms of accuracy in facial expression recognition compared to conventional approaches; this is due to their automatic and intelligent feature extraction. Researches show that facial expression recognition is significantly and efficiently influenced by extracted facial features. The challenge in the process of training such networks is the limit of available samples in the facial expression recognition datasets. Focusing on performance enhancement of facial expression recognition systems (FER), we propose a hybrid model combining CNN's-based extracted features to ensure complementarity and diversity, and transfer learning advantages in classification for FER applications.

Our contributions for this paper are as follows:

1. A dual-branch model based on a novel simple and efficient CNN inspired by VGGnet architecture and a pre-trained CNN, which is an efficient network is proposed to compensate for the lack of training samples, by merging their extracted features and to enhance recognition accuracy.
2. A joint training strategy is designed for the proposed dual-branch model.
3. Two datasets which are Fer-2013 and CK+ are employed to validate the effectiveness of our architectures. CK+ is a classic facial expression dataset and FER2013 provides samples of faces captured in the real world.

The rest of this document is organized as follows. Section 2 details the proposed FER approach. Section 3 presents and discusses the obtained results of the experiment. Finally, conclusions are presented in Section 4.

2. PROPOSED METHOD

This paper proposes a novel dual-branch system architecture for leveraging both types of CNN learning, namely learning from scratch and transfer learning, to diversify the feature maps fed to the classifier. The proposed FER system, as illustrated in Fig. 1, involves the following steps: two convolutional branches, generating each one its own feature maps. Those outputs will be combined to represent the responsible feature vector for classification step. The first branch extracts features from the input image through the convolutional layers of a VGG-inspired CNN; while the second branch extracts features from the same image through a pre-trained network. Finally, feature maps are merged using concatenation, and classification is performed using a fully connected layer with a Softmax activation function to recognize expressions. These steps will be presented in detail in the next sections.

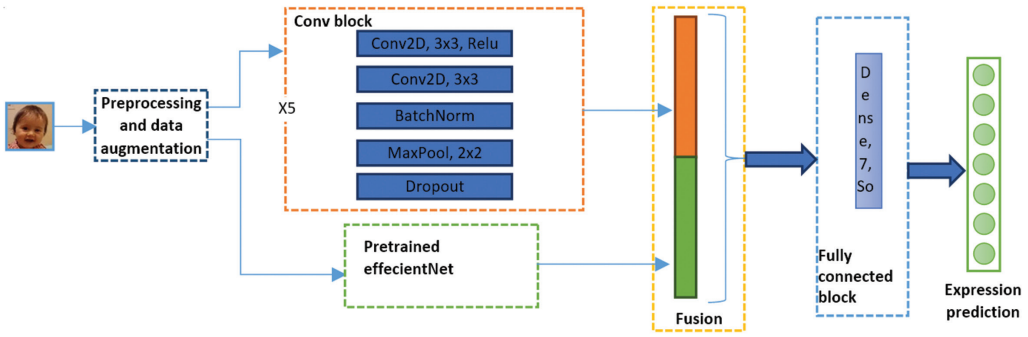

Fig. 1. Proposed method overview

2.1. VGG INSPIRED BRANCH

In 2014, Simonyan and Zisserman [10] proposed a very deep network called VGG16. In their work, they evaluate networks, increasing the depth and with very small convolution filters (3x3). This architecture won first and second place in the location and classification tracks, respectively, at the ImageNet 2014 Challenge. Based on this architecture, and after testing and experimentation with several configurations, we propose two simple and deep models, VGGinspiredCNN1 and VGGinspiredCNN2, enriched by batch normalization to improve generaliza-

tion and optimization and dropout layers. Their architecture is composed of five convolutional blocks and a fully connected block. Each convolution block is composed of two convolution layers (all used filters are 3x3 size like the VGGNet models) followed by batch normalization and a max-pooling layer (with a kernel size of 2x2) and a dropout layer. Each convolution layer is equipped with a non-linear rectification (Relu). The fully connected block of the first model is composed of two fully connected layers with 512 and 7 outputs respectively. The fully connected layer of the second model is only one layer with seven outputs. The VGG-inspired CNN architectures are described in Fig. 2 and Table 1.

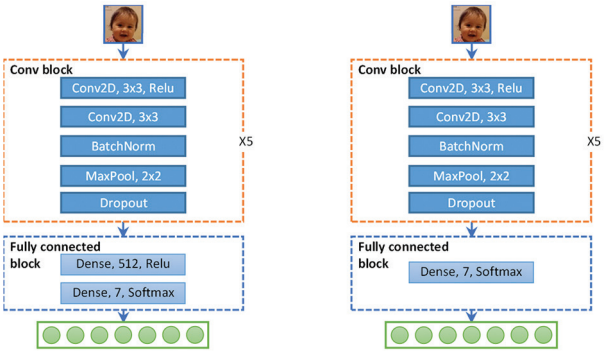

Fig. 2. The proposed VGGinspiredCNN architecture.

Table 1. The convolutional layer structure of the two proposed models inspired by VGGnet.

	Input	Conv+Relu	Pooling	Conv+Relu	Pooling	Conv+Relu	Pooling	Conv+Relu	Pooling	Conv+Relu	Pooling
Kernel size	48 x 48	3	2	3	2	3	2	3	2	3	2
Stride		1	2	1	2	1	2	1	2	1	2
Pad		0	0	0	0	0	0	0	0	0	0
# filters		64		128		256		512		512	
#Replications		2	1	2	1	2	1	2	1	2	1

2.2 PRETRAINED BRANCH

To enable CNNs to learn and extract features and achieve high accuracy, millions of samples must be used in their training base, however, existing facial expression data sets contain only just a few hundred or thousands of samples. This insufficient size is one of the main problems in CNN-based FER. To overcome this limit, the use of transfer learning will be a possible solution; it is a common practice where the network is first initialized with a set of pre-formed weights (and biases) based on a large-scale data set from one task and these parameters are then recycled to another new target task.

In order to obtain better accuracy than traditional CNNs, authors in [11] have proposed a family of models, EfficientNets, which can be systematically scaled according to available resources. A balance between network dimensions is obtained by simply scaling up them with a constant ratio. EfficientNets models transfer well to data sets such as CIFAR-100 [11], fruits [12], etc. with fewer parameters. Eight models of EfficientNetB0-EfficientNetB7 were examined for their efficiency and performance. In the transfer branch, preformed weights from the ImageNet dataset are used because it contains a large number of person images [13], about 952K images and this is very relevant for classifying the Fer-2013 and Ck+ facial expression datasets used in the evaluation. So, these pre-trained network parameters are used for initialization. Then, the model is trained, and fine-tuning will be performed to extract more specific features. Fig. 3 shows the architecture of the pre-trained CNN network.

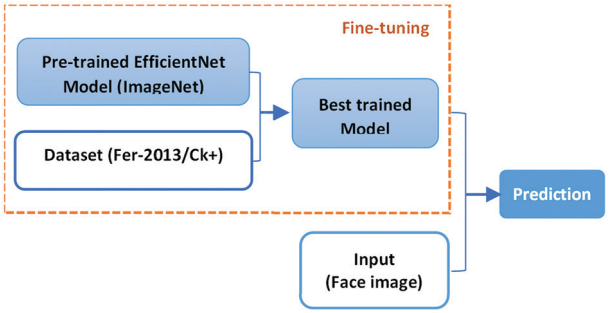

Fig. 3. The pre-trained CNN network architecture.

The base model EfficientNet-B0 consists of 18 convolution layers (with a kernel size of 3x3 or 5x5). Then, a flatten layer follows the max pooling as described in Table 2, its main building block is mobile inverted bottleneck MBConv, to which they also add squeeze-and-excitation optimization [11]. The other EfficientNet configurations, i.e., B1 - B7, are scaled from the basic configuration EfficientNet-B0 with different compound coefficients. A new classification layer replaces the last fully connected layers with seven classes (corresponding to seven expressions).

Table 2. EFFICIENTNET-B0 baseline network [11]

Stage i	Operator Fi	Input Resolution Hi x Wi	Output Channels Ci	Layers Li
1	Conv3x3	224 x 224	32	1
2	MBConv1, k3x3	112 x 112	16	1
3	MBConv6, k3x3	112 x 112	24	2

Stage i	Operator Fi	Input Resolution Hi x Wi	Output Channels Ci	Layers Li
4	MBConv6, k5x5	56 x 56	40	2
5	MBConv6, k3x3	28 x 28	80	3
6	MBConv6, k5x5	14 x 14	112	3
7	MBConv6, k5x5	14 x 14	192	4
8	MBConv6, k3x3	7 x 7	320	1
9	Conv1x1 & Pooling & FC	7 x 7	1280	1

2.3 FEATURE MAP FUSION MODULE

The feature vector concatenation is commonly used to merge and integrate multiple channels or branches in several architectures [14], [15]. The operation that combines features extracted from the VGGinspired branch and features extracted from the pre-trained model is defined as the following formula:

$$(x_1^V, x_2^V, \ldots)^T \oplus (x_1^P, x_2^P, \ldots)^T = (x_1^V, x_2^V, \ldots, x_1^P, x_2^P, \ldots)^T \quad (1)$$

Where: '$\oplus$' denotes vector concatenation operator, $(x_1^V, x_2^V, \ldots)^T$ denotes features extracted from VGGinspired branch and $(x_1^P, x_2^P, \ldots)^T$ denotes features extracted from the pretrained model.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1 USED DATASETS AND PREPROCESSING

To demonstrate the performance of the proposed models in facial expression recognition, two widely used facial expression recognition databases, i.e., the Fer-2013 database [16] and the Cohn Kanade database [17], are used.

The Facial Emotion Recognition 2013 (FER-2013) dataset was created by Pierre Luc Carrier and Aaron Courville and was introduced in the ICML 2013 workshop's facial expression recognition challenge [16]. The data set is composed of 35887 facial images, most of them in wild environments. It consists of three parts: the original training data (OTD), which consists of 28709 images, the public test data (PTD), which includes 3589 images, and the final test data (FTD), which includes 3589 images used to score the final models.

The extended Cohn Kanade database (CK+) [17] is the most widely used laboratory-controlled database for the evaluation of FER systems. It consists of 593 video sequences obtained from 123 subjects. Among them, 327 sequences from 118 subjects are labeled as one of seven expressions. For each sequence, only the last frame is labeled. The last three frames are extracted from each sequence in the CK+ dataset, which contains 981 facial expressions. The distribution of samples of the two datasets used in experiments are shown in Table 3.

Table 3. Number of images per each expression in FER-2013 and CK+ datasets.

Expression	Images number	
	FER-2013	CK+
anger	4953	135
disgust	547	177
fear	5121	75
happiness	8989	207
sadness	6077	84
surprise	4002	249
neutral	6198	/
contempt	/	54

All face images are resized to 48×48 pixels, then normalized to have zero mean and unit variance. To make the proposed model more robust to slight transformations and noise, data augmentation is applied using different linear transformations which are: rotation, horizontal flipping, zooming, and skewing of the central area.

3.2. EXPERIMENT SETTINGS

In order to provide evidence of the performance of the proposed dual-branch model, three different learning experiments are conducted: the classical CNN model based on the VGG architecture, transfer learning of all EfficientNet models, and joint learning of the dual-branch model.

For training, the images from the CK+ datasets are randomly shuffled and are split as follows: 85% training, 15% test. For the FER-2013 dataset, the entire training set (28,709) and the public test set (3,589) are used for training and validation, respectively.

The total loss function is optimized during the backpropagation using the Adam optimizer, it should be noted that different optimizers were tested, even stochastic gradient descents, and Adam appeared to perform better.

The implementation is based on the Keras library [18] with TensorFlow backend [19]. OpenCV [20] is used for all image operations. All the experiments have been executed with PyTorch and trained using Google Colaboratory[21]. In light of the limitations of a free Google Colab account, such as a maximum of 12 hours per training session, the number and type of GPUs, or VRAM capacity, the training phase has been carried out using several parameters as shown in Table 4 for both datasets Fer-2013 and Ck+.

Table 4. Experimental configurations for FER-2013 and CK+datasets.

	Network	Fer-2013			CK+		
		epochs	Batch size	Learning rate	epochs	Batch size	Learning rate
Learning from scratch	VGGinspiredCNN1	60	64	0,001	60	8	0,001
	VGGinspiredCNN2	60	64	0,001	100	16	0,001
Transfer learning	EfficientNet-B0	80	32	0.00001	80	16	0.0001
	EfficientNet-B1	100	32	0.00001	80	16	0,0001
	EfficientNet-B2	100	32	0.00001	80	8	0.0001
	EfficientNet-B3	100	32	0.00001	80	8	0.0001
	EfficientNet-B4	80	32	0.00001	80	8	0.0001
	EfficientNet-B5	80	32	0.00001	80	16	0.0001
	EfficientNet-B6	80	32	0.00001	80	16	0.0001
	EfficientNet-B7	80	32	0.00001	80	16	0.0001
Dual branch	CNN+EfficientNet-B0	60	64	0.001	120	8	0.001
	CNN+EfficientNet-B1	60	64	0.001	100	8	0.0001
	CNN+EfficientNet-B2	50	64	0.001	80	8	0.0001
	CNN+EfficientNet-B3	40	64	0.0002	100	8	0.0001
	CNN+EfficientNet-B4	40	128	0.0001	120	8	0.0001
	CNN+EfficientNet-B5	40	64	0.0001	120	16	0.0001
	CNN+EfficientNet-B6	30	64	0.0001	120	8	0.0001
	CNN+EfficientNet-B7	40	64	0.0002	120	16	0.0001

3.3 RESULTS AND DISCUSSION

In this section, details of the achieved results are provided. As previously mentioned, experiments were conducted to determine the effectiveness of the proposed dual-branch CNN model compared to CNN by training from scratch and EfficientNet models by transfer learning. So, the two CNN models inspired by VggNet, the transfer learning of the EfficientNet with its eight configurations, and the proposed dual branch model are tested on two widely used FER datasets: CK + and FER-2013. The FER-2013 dataset includes the expressions of seven labels: anger, disgust, fear, happiness, sadness, surprise, and neutral, while the CK+ dataset includes the same expressions except the neutral expression, and also includes the expression contempt.

1- Fer-2013dataset evaluation:

Experimental results of the CNN models trained from scratch, the pre-trained EfficientNet models as well as the proposed dual branch model on the FER-2013 dataset are given in Fig. 4 to Fig 6 and Table 5.

Accuracy of all proposed models compared to other state-of-the-art models is listed in Table 5, and Fig. 4 to Fig. 6 show the corresponding normalized confusion matrices. For the Fer-2013 dataset, the VGG-inspired CNN models give competitive performance compared to the results of state-of-the-art models and surpass human-level accuracy, based on the accuracy rates achieved by the experimentation. As indicated by the study conducted by [22] on deep learning-based facial expression recognition, the most accurate tests on the FER-2013 dataset using a single CNN network are in the range of 67-71% and the models that accomplish it are very performant.

Table 5. Proposed models versus other models' performance on the Fer-2013 dataset.

Network		Accuracy rate
Learning from Scratch	VGGinspiredCNN1	66.71
	VGGinspiredCNN2	67.70
Transfer learning	EfficientNet-B0	56.23
	EfficientNet-B1	57.48
	EfficientNet-B2	57.43
	EfficientNet-B3	58.32
	EfficientNet-B4	57.13
	EfficientNet-B5	57.70
	EfficientNet-B6	57.17
	EfficientNet-B7	60.10
Dual Branch	CNN+EfficientNet-B0	63.36
	CNN+EfficientNet-B1	62.88
	CNN+EfficientNet-B2	62.77
	CNN+EfficientNet-B3	63.80
	CNN+EfficientNet-B4	62.25
	CNN+EfficientNet-B5	62.09
	CNN+EfficientNet-B6	62.31
	CNN+EfficientNet-B7	62.22
State of art models	[23]	66.4 (Top-1)
	[25]	66
	[26]	65.2
	[24]	71.14

It can be observed from the confusion matrices that the expressions 'happy' and 'surprise' are easier to recognize, with an accuracy of more than 80%, while the expression 'fear' and the expressions 'sad' and 'disgust' are the most difficult to recognize with our best model VGGinspiredCNN2, with an accuracy of 44%, 51%, and 52% respectively. It is important to be mentioned that sometimes, as a human being, it is difficult to recognize whether an expression of sadness or fear, this is due to the fact that people do not all express their emotions in the same manner, and the low accuracy rate of the expression 'disgust' is due to the small number of samples of this expression in the Fer-2013 dataset.

On the other hand, the pre-trained EfficientNet models achieved acceptable accuracy rates for this dataset, with the EfficientNet -B7 model achieving its highest accuracy of 60.10%. Note that across all EfficientNet models, the expression 'happy' is always the most recognized with the highest accuracy (more than 76%), while the expression 'disgust' is always the less recognized by the different EfficientNet models, with an accuracy between 21% and 35%, except for the EfficientNet-B7 model, where this class reaches an accuracy of 53%

Table 5. Proposed models versus other models' performance on the Fer-2013 dataset.

Network		Accuracy rate
Learning from Scratch	VGGinspiredCNN1	93.40
	VGGinspiredCNN2	98.48
Transfer learning	EfficientNet-B0	99.32
	EfficientNet-B1	97.30
	EfficientNet-B2	97.97
	EfficientNet-B3	97.97
	EfficientNet-B4	97.30
	EfficientNet-B5	96.62
	EfficientNet-B6	96.62
	EfficientNet-B7	98.65
Dual Branch	CNN+EfficientNet-B0	99.32
	CNN+EfficientNet-B1	93.92
	CNN+EfficientNet-B2	94.59
	CNN+EfficientNet-B3	95.94
	CNN+EfficientNet-B4	95.94
	CNN+EfficientNet-B5	97.30
	CNN+EfficientNet-B6	98.65
	CNN+EfficientNet-B7	95.94
State of art models	[23]	93.2(Top-1)
	[27]	72.1
	[28]	96.8
	[24]	95.29
	[29]	93.24

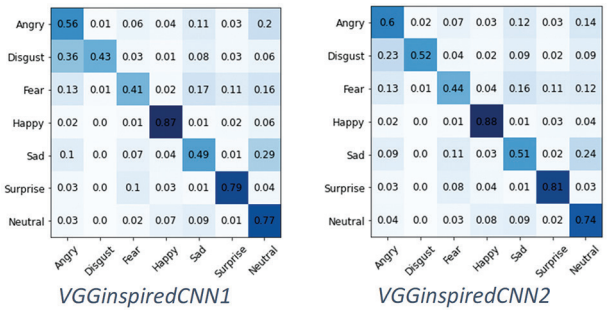

Fig. 4. Confusion matrices for VGGinspiredCNN models on the Fer-2013 database.

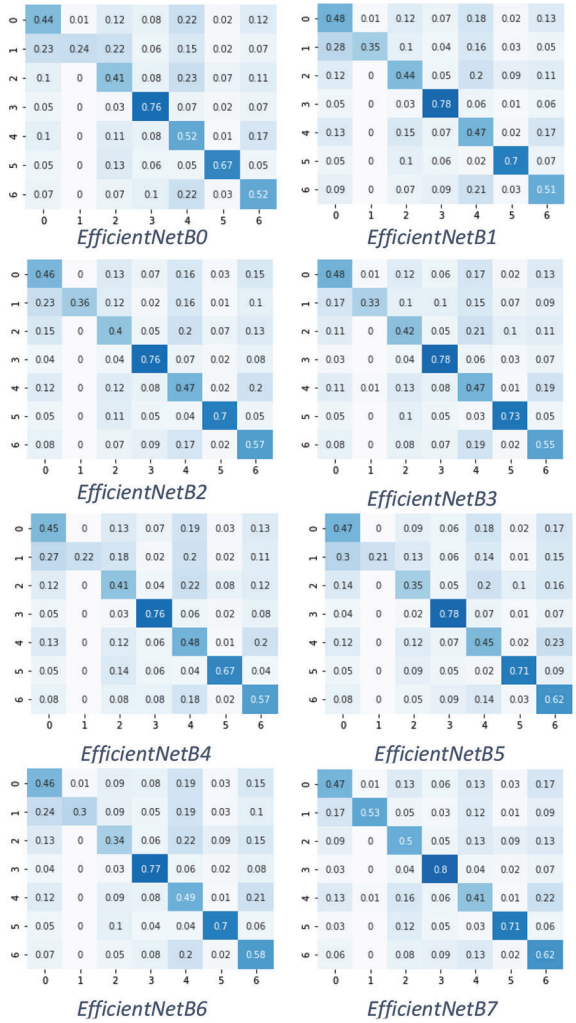

Fig. 5. Confusion matrices for EfficientNet models on the Fer-2013 database.

While the proposed dual-branch model provides a significant improvement for all EfficientNet configurations, especially for the EfficientNet B0 model, which gains 7.13% in accuracy, but the best performance is achieved by the EfficientNet-B3 model, with an accuracy of 63.80%. Confusion matrices of the dual-branch models show an improvement in the recognition rate of the expression 'disgust', which ranges from 47% to 70%, achieved by the best dual-branch model with EfficientNet-B3 for the Fer-2013 dataset, while no improvement is recorded on the recognition rate of the expression 'happy'.

Certainly, the proposed dual-branch model has improved the results for this dataset, however, Table 5 reveals also that the achieved recognition accuracy is not performing as well as the state-of-the-art models [43, 45] which were designed specifically for unconstrained facial expression recognition. In the proposed dual branch model, only one fully connected layer is utilized in order to obtain an efficient network, thus limiting its performance when handling the unconstrained FER task.

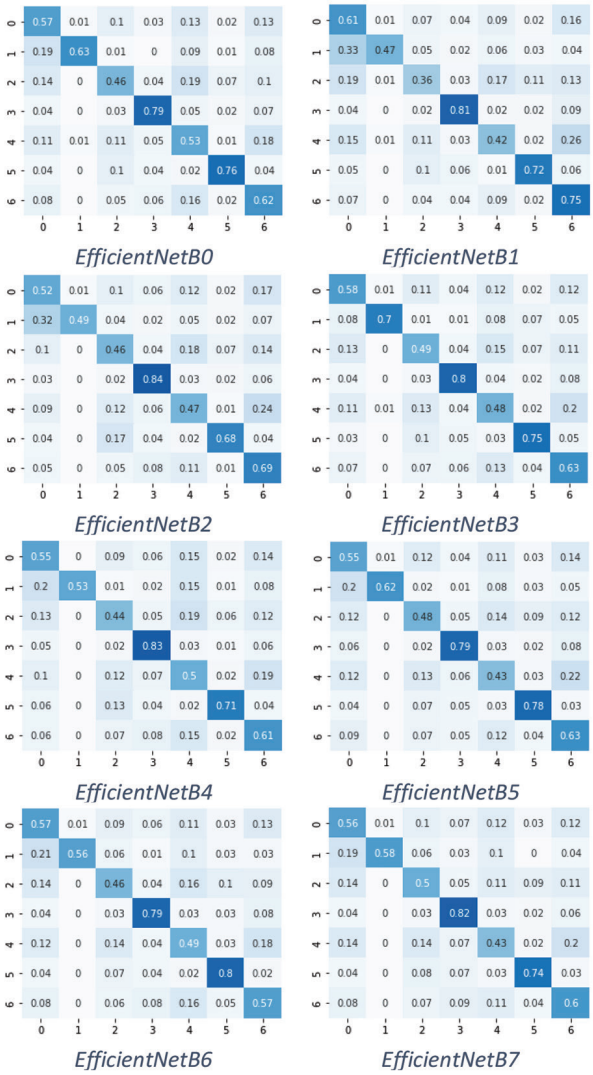

Fig. 6. Confusion matrices for the proposed dual branch model on Fer-2013 database.

2 - Ck+ dataset evaluation:

In the same way, as for the evaluation of the Fer-2013 dataset, confusion matrices for each expression and each of the proposed models on the CK+ dataset are presented in figures Fig. 7, Fig 8, and Fig.9, and the result of the comparison with other competing models is given in Table 6. CNN models inspired by VGG obtain very interesting and competitive results for the CK+ dataset, compared to the state-of-the-art models, especially, the VGGinspiredCNN2 model achieves a 98.48% accuracy rate which is better than all the reference works cited above.

Fig. 7. Confusion matrices for VGGinspiredCNN models on Ck+ database.

According to the normalized confusion matrix, 4 of 7 expressions ('anger', 'fear', 'happy', 'sadness') are recognized at 100%, two other expressions ('disgust' and 'surprise') are recognized at 98%, and only 'contempt' expression is recognized at 89%, which is confused with the expression 'sad'.

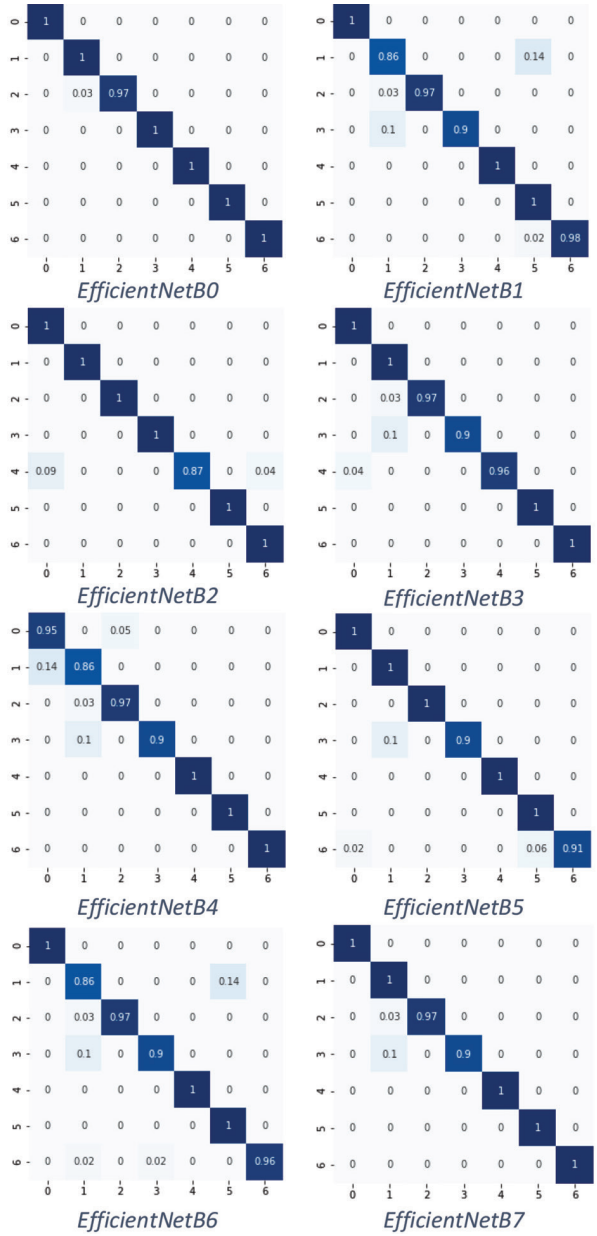

Fig. 8. Confusion matrices for EfficientNet models on CK+ database.

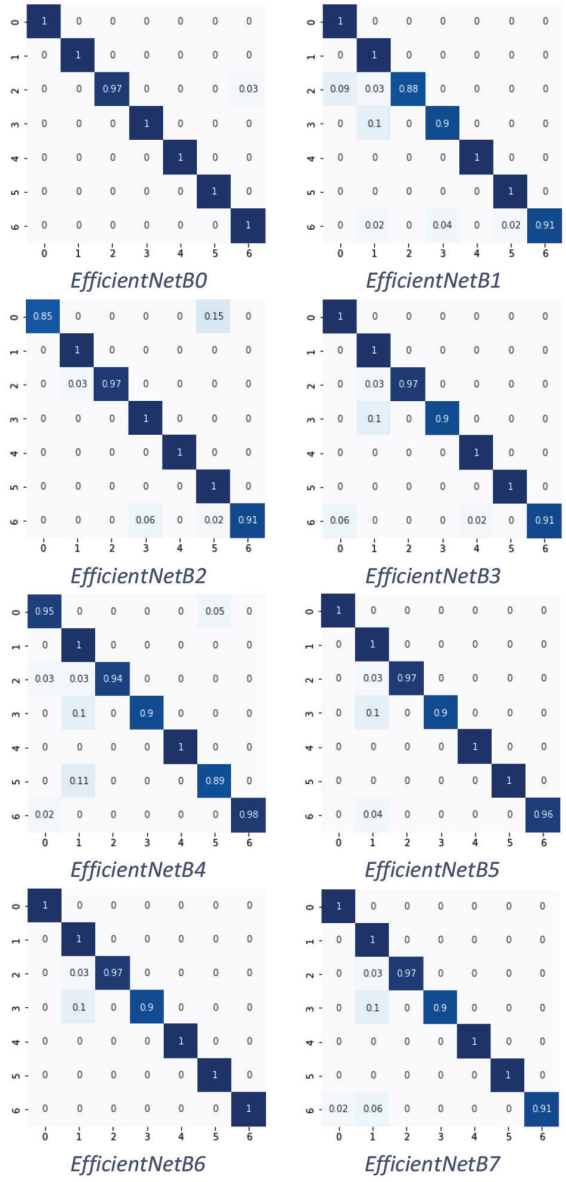

Fig. 9. Confusion matrices for the proposed dual branch model on the CK+ database

According to the tables, Table 5 and Table 6, comparing the results obtained by the proposed models and some reference models, the use of the dual-branch approach improves the results for the CK+ dataset, but not as much as for the Fer-2013 dataset. While the accuracy for the seven expressions in the CK+ dataset is high, the Fer2013 dataset has low classification accuracy due to mislabelling in the test set, except for the "happy" category. According to table 6, it can be seen that all EfficientNet models achieve an accuracy of more than 96%. EfficientNet-B5 and EfficientNet-B6 achieve the lowest accuracy of 96.62%, while EfficientNet-B0 realizes the best performance with an accuracy of 99.32%. The same accuracy rate is obtained in the proposed dual branch model.

Nevertheless, the Fer-2013 dataset is the most commonly used dataset for facial expression recognition. It should be noted that the human eye can hardly distinguish the appropriate emotion for some of them.

4. CONCLUSION

In this paper, the proposed dual-branch model is designed to take advantage of the commonly used learning approaches like learning from scratch and transfer learning, to recognize human facial expressions in the wild and under controlled laboratory conditions. Experiments and evaluation of the models using two reference datasets, Fer-2013 and CK+, showed very interesting and motivating results for both datasets. These results are competitive and outperform existing works. The most important challenge with the FER datasets is the limit of its size, and the unbalanced images of different classes, which do not favor deep learning. To overcome this limitation, two types of learning in the same model are employed simultaneously. In this approach, the training of the EfficientNets models is refined; in fact, those models are already trained for the Imagenet dataset, on the FER-2013 and CK+ datasets, and then the feature vector obtained is concatenated to that obtained from a classical CNN based on a very well-known and robust VGG architecture.

The proposed dual-branch model improved the accuracy of all expressions except "disgust", which is a bit weak like other methods on CK+. In the case of Fer-2013, the proposed dual branch model has a significant improvement in "disgust", which improves the accuracy to 70%. This demonstrates the efficiency and effectiveness of the proposed approach.

In future work, and to surmount some limitations of the proposed system, an investigation based generative adversarial network for data augmentation will be done; to improve the outcome of transfer learning, a two-step fine-tuning approach will be studied.

5. REFERENCES:

[1] P. Ekman, W. V. Friesen, "Constants across cultures in the face and emotion", Journal of personality and social psychology, Vol. 17, No. 2, 1971, pp. 124-129.

[2] R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, P. G. Schyns, "Reply to Sauter and Eisner: Differences outweigh commonalities in the communication of emotions across human cultures", Proceedings of the National Academy of Sciences, Vol. 110, No. 3, 2013, pp. E181-E182.

[3] H. Boughrara, M. Chtourou, C. B. Amar, L. Chen, "Facial expression recognition based on a mlp neural network using constructive training algorithm", Multimedia Tools and Applications, Vol. 75, No. 2, 2016, pp. 709-731.

[4] I. Cohen, N. Sebe, F. G. Gozman, M. C. Cirelo, T. S. Huang, "Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data", Proceedings of the IEEE Computer

Society Conference on Computer Vision and Pattern Recognition, 18-20 June 2003, Vol. 1, pp. 595-604.

[5] M. Szwoch, P. Pieniążek, "Facial emotion recognition using depth data", Proceedings of the 8th International Conference on Human System Interaction, 25-27 June 2015, pp. 271–277.

[6] Y. Wang, H. Ai, B. Wu, C. Huang, "Real time facial expression recognition with adaboost", Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004, Vol. 3, pp. 926-929.

[7] A. Dapogny, K. Bailly, S. Dubuisson, "Pairwise conditional random forests for facial expression recognition", Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 7-13 December 2015, pp. 3783-3791.

[8] X. Zhao, X. Shi, S. Zhang, "Facial Expression Recognition via Deep Learning", IETE technical review, Vol. 32, No. 5, 2015, pp. 347-355.

[9] M. Suk, B. Prabhakaran, "Real-time mobile facial expression recognition system-a case study", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23-28 June 2014, pp. 132–137.

[10] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv:1409.1556v6, 2014.

[11] M. Tan, Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks", arXiv:1905.11946v5, 2019.

[12] L. T. Duong, P. T. Nguyen, C. Di Sipio, D. Di Ruscio, "Automated fruit recognition using EfficientNet and MixNet", Computers and Electronics in Agriculture, Vol. 171, 2020, p. 105326.

[13] ImageNet, http://www.image-net.org/about-stats (accessed: 2020)

[14] J. A. Aghamaleki, V. A. Chenarlogh, "Multi-stream CNN for facial expression recognition in limited training data", Multimedia Tools and Applications, Vol. 78, No. 16, 2019, pp. 22861-22882.

[15] W. Zou, D. Zhang, D.-J. Lee, "A new multi-feature fusion based convolutional neural network for facial expression recognition", Applied Intelligence, Vol. 52, No 3, 2022, pp. 2918-2929.

[16] I. J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests", proceedings of the 20th International Conference on Neural Information Processing, Daegu, Korea, 3-7 November 2013, pp. 117-124.

[17] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression", Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, USA, 13-18 June 2010, pp. 94-101.

[18] Keras, https://github.com/keras-team/keras (accessed: 2020)

[19] M. Abadi et al. "TensorFlow: A System for Large-Scale Machine Learning", Proceedings of the 12th USENIX Conference Operating Systems Design and Implementation, Savannah, GA, USA, 2-4 November 2016, pp. 265–283.

[20] OpenCV, https://opencv.org/ (accessed: 2020)

[21] Google Colaboratory, https://colab.research.google.com/notebooks/intro.ipynb (accessed: 2020)

[22] S. Li, W. Deng, "Deep Facial Expression Recognition: A Survey", IEEE transactions on affective computing, pp. 1–1, 2020.

[23] A. Mollahosseini, D. Chan, M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks", Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Placid, NY, USA, 7-10 March 2016, pp. 1-10.

[24] J. Shao, Y. Qian, "Three convolutional neural network models for facial expression recognition in the wild", Neurocomputing, Vol. 355, 2019, pp. 82-92.

[25] O. Arriaga, M. Valdenegro-Toro, P. Plöger, "Real-time Convolutional Neural Networks for Emotion and Gender Classification", arXiv:1710.07557v1, 2017.

[26] P. Giannopoulos, I. Perikos, I. Hatzilygeroudis, "Deep Learning Approaches for Facial Emotion Recognition: A Case Study on FER-2013", Advances in hybridization of intelligent methods. Springer, 2018, pp. 1-16.

[27] R. Breuer, R. Kimmel, "A deep learning perspective on the origin of facial expressions", arXiv:1705.01842v2, 2017.

[28] H. Ding, S. K. Zhou, R. Chellappa, "FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition", Proceedings of the 12th IEEE International Conference on Automatic Face Gesture Recognition, Washington, DC, USA, 30 May-3 June 2017, pp. 118-126.

[29] D. K. Jain, P. Shamsolmoali, P. Sehdev, "Extended deep neural network for facial emotion recognition", Pattern Recognition Letters, Vol. 120, 2019, pp. 69-74.