Memory Integrity of CNNs for Cross-Dataset Facial Expression Recognition

Dylan C. Tannugi[1], Alceu S. Britto Jr.[2] and Alessandro L. Koerich[1]

Abstract—**Facial expression recognition is a major problem in the domain of artificial intelligence. One of the best ways to solve this problem is the use of convolutional neural networks (CNNs). However, a large amount of data is required to train properly these networks but most of the datasets available for facial expression recognition are relatively small. A common way to circumvent the lack of data is to use CNNs trained on large datasets of different domains and fine-tuning the layers of such networks to the target domain. However, the fine-tuning process does not preserve the memory integrity as the CNN has the tendency to forget patterns they have learned. In this paper we evaluate different strategies of fine-tuning a CNN with the aim of assessing the memory integrity of such strategies in a cross-dataset scenario. A CNN pre-trained on a source dataset is used as the baseline and four adaptation strategies have been evaluated: fine-tuning its fully connected layers; fine-tuning its last convolutional layer and its fully connected layers; retraining the CNN on a target dataset; and the fusion of the source and target datasets and retraining the CNN. Experimental results on four datasets have shown that the fusion of the source and the target datasets provides the best trade-off between accuracy and memory integrity.**

I. INTRODUCTION

Automatic facial expression recognition has been investigated in the last years due to the great number of applications, ranging from human-computer interaction, emotion analysis [2] to detection of driver fatigue [24]. Several approaches based on handcrafted features [3], [14], [24], [25] such as textural, eigenfaces, etc. have been shown to be very effective. However, the most recent approaches are based on deep convolutional neural networks (CNNs) [9], [11], [13], [17], [20], [21], [22], [23]. However, one of the main drawbacks is the amount of data required to train such deep networks. One way to circumvent the lack of data is to use CNNs pre-trained in large datasets even if the datasets and the problem lies in other domains. For instance, CNNs pre-trained in datasets such as ImageNet have been used in many different problems such as histopathologic image classification [4], environmental sound classification based on 2D representations [6], logo classification [19] and many others. This strategy has also been used for facial expression recognition given the small size of datasets. Li and Deng *et al.* [10] present a comprehensive survey on deep facial expression recognition (FER) that describes the standard pipeline of a deep FER system, the available datasets that are widely used in the literature. They also review existing novel deep neural networks and related training strategies that are designed for FER based on both static images and dynamic image sequences. Lopes *et al.* [11] present a CNN model for facial expression classification. A large amount of gray-scale face images are provided as well as the class and location of the eyes on each image. Several pre-processing steps such as spatial normalization, data augmentation by generating synthetic samples, resizing and intensity normalization are use eliminate the bias related to the brightness of the image. The experiments carried out on the CK dataset achieved a percentage of 93.74% of correct classification. Furthermore, the authors show that making changes to the images before providing them to a CNN makes improves the performance. Sun *et al.* [16] explore audio, visual and physiological signals to continuously predict the value of the emotion dimensions (arousal and valence). They have evaluated a variety of handcrafted and deep visual features such as Dense SIFT features (MSDF), and some types of Convolutional neural networks (CNNs) features to recognize the expression phases of the frames. Yu and Zhang *et al.* [22] present an approach for facial expression recognition for the Emotion Recognition in the Wild Challenge 2015. The proposed approach is based on an ensemble of three state-of-the-art face detectors, followed by a classification module with the ensemble of multiple deep convolutional neural networks (CNN). Each CNN model is initialized randomly and pre-trained on FER dataset. The pre-trained models are then fine-tuned on the training set of SFEW 2.0 and achieves 55.96% and 61.29% respectively on the validation and test set of SFEW 2.0, surpassing the baseline of 35.96% and 39.13% respectively.

Zhang *et al.* [26] present a comprehensive review of transfer-learning methods for cross-dataset visual recognition. In particular, we are interested in homogeneous feature and label spaces, where the feature spaces and label spaces of the source and target datasets are identical, but domain divergence exists across the source and target datasets. In this problem, a small number of labelled data in target domain is available. However, the labelled target data are generally insufficient for learning an effective classifier. This is also called supervised domain adaptation or few-shot domain adaptation in the literature [26]. Another group of methods transfer the parameters of discriminative classifiers across datasets [7], [8], [18]. Mousavi *et al.* [13] use deconvolution to visualize the characteristics learned by a CNN trained with the Cohn-Kanade (CK) dataset and tested on the same

[1]Dylan C. Tannugi and Alessandro L. Koerich are with Department of Software and IT Engineering, École de Technologie Supérieure, University of Québec, H3C 1K9, Montréal, QC, Canada. `dylan.cohen-tannugi.1@ens.etsmtl.ca`, `alessandro.koerich@etsmtl.ca`
[2]Alceu S. Britto Jr. is with Pontifical Catholic University of Paraná, Curitiba, PR, 80215-901, Brazil. `alceu@ppgia.pucpr.br`

dataset and on the JAFFE dataset. Zavarez *et al.* [23] propose a CNN based on the VGG net [15] which uses two different initialization methods: randomly initialized weights and weights from a pre-trained VGG-Face model. Experiments were carried out with different combinations of training and test sets of seven datasets: AR Face, CK+, BU3DFE, JAFFE, MMI, RaFD, and KDEF. For each combination, the training set is composed of the merging of six datasets leaving one dataset for the test set. The fine-tuned VGG net performed better than the VGG initialized with random weights and achieved a accuracy of 88.58% on CK+ as the test set. Indeed, only the combination having JAFFE as test set obtained a worse result for the fine-tuned VGG net. This shows that fine-tuning a CNN is often effective as well as more stable than one initialized with random weight. The fine-tuned model also outperformed the state-of-the-art for four out of five datasets evaluated. Mayer *et al.* [12] present a system for facial expression recognition that is evaluated on multiple datasets. Comparing classifiers across datasets determines the classifiers' capability to generalize more reliable than traditional self-classification. Li and Deng *et al.* [9] present a deep Emo-transfer network (DETN) to deal with the problem of cross-dataset facial expression recognition which embeds maximum mean discrepancy in the deep architecture to reduce dataset bias. The experimental results on both lab-controlled and real-world facial expression datasets have shown competitive performances across various facial expression transfer tasks. Wang *et al.* [17] introduce an unsupervised domain adaptation method, which is especially suitable for small unlabelled target datasets. They have trained a generative adversarial network (GAN) on the target dataset and use the GAN generated samples to fine-tune the model pre-trained on the source dataset. In the process of fine-tuning, the unlabelled GAN generated samples distributed pseudolabels dynamically according to the current prediction probabilities. They demonstrated the effectiveness of the method on four facial expression recognition datasets with two CNN structures. Yang *et al.* [21] present a method based on transfer learning and sparse coding to learn a dictionary and transfer it to facial expression to obtain a feature representation by sparse coding. The experimental results in CK+, JAFFE and NVIE show that the transfer learning based on sparse coding method can effectively improve the expression recognition rate in the cross-domain facial expression recognition task. Yan *et al.*[20] proposed a transductive deep transfer learning architecture based on VGGface16-Net to jointly learn a common optimal nonlinear discriminative features from the non-frontal facial expression samples between the source and target datasets and cross-dataset non-frontal facial expression classification task. Extensive cross-dataset experiments on BU-3DEF and Multi-PIE datasets are presented and the experimental results show that the transductive deep transfer network outperforms the state-of-the-art cross-database facial expression recognition methods.

Most of the works focus on adapting an existing CNN to a new dataset (target), maximizing the performance of such a model on the target dataset and leaving aside (forgetting) the performance on the source dataset. In this paper we evaluate different adaptation approaches with the aim of obtaining a model that provides the best trade-off in terms of performance for both the source and target datasets. In case of new subjects are added to the data, which is the best way to integrate such new users? Should we modify or not the learned model? And if we decide to modify it, should we fully train the model and on which data? Should we rebuild it from scratch or preserve the already learned parameters and just slightly adapt it to the new data? Therefore, in such a case, the main question that we are trying to answer in this paper is whether transfer learning is useful of not. How useful is transfer learning when we have several datasets for the same problem, and therefore we can combine these datasets and design a customized architecture for tackling the problem. Furthermore, are the current transfer learning strategies able to deal with memory integrity and plasticity as CNNs have the tendency to forget patterns they have learned

In this paper we show that transfer learning or fine-tuning of CNNs is not so effective to preserve memory integrity. Instead, fully training the network is a much better option as it usually leads to the best performance on both the source and the target datasets. The contributions of this paper are twofold: (i) characterize different transfer learning strategies subject to the constraint of preserving memory integrity; (ii) fine-tuning or retraining a CNN? Which action leads to the best performance if we look for a good balance between the performance on the source dataset and on a target dataset.

This paper is organized as follows: Section II describes some transfer-learning strategies for cross-dataset recognition. Experimental results on cross-dataset transfer learning and recognition on four datasets are presented in Section **??**. Finally, the conclusions and perspectives of future work are stated in the last section.

II. TRANSFER LEARNING STRATEGIES FOR CROSS-DATASET RECOGNITION

In facial expression recognition, most of the datasets have a small number of labelled data. However, such datasets are generally insufficient for learning an effective model. One way to circumvent such a lack of data is to use cross-dataset strategies where there are at least two datasets, one of them referred to as the source dataset is used to train a model and the other, referred to as target dataset is the dataset to which the model has to be adapted. Deep networks can generally learn more transferable features [1], [5].

Considering that the source and the target datasets belong to the same domain, namely facial expression recognition, according to Zhang *et al.*[26], this can be characterized as homogeneous feature spaces and label spaces because observation spaces $\mathcal{X}$ as well as the label spaces $\mathcal{Y}$ are both identical between source and target datasets. Hence the target and the source datasets are different in terms of data distributions ($P(\mathcal{X}, \mathcal{Y})$).

Therefore, in our approach for cross-dataset facial expression recognition, we consider a pre-trained CNN on
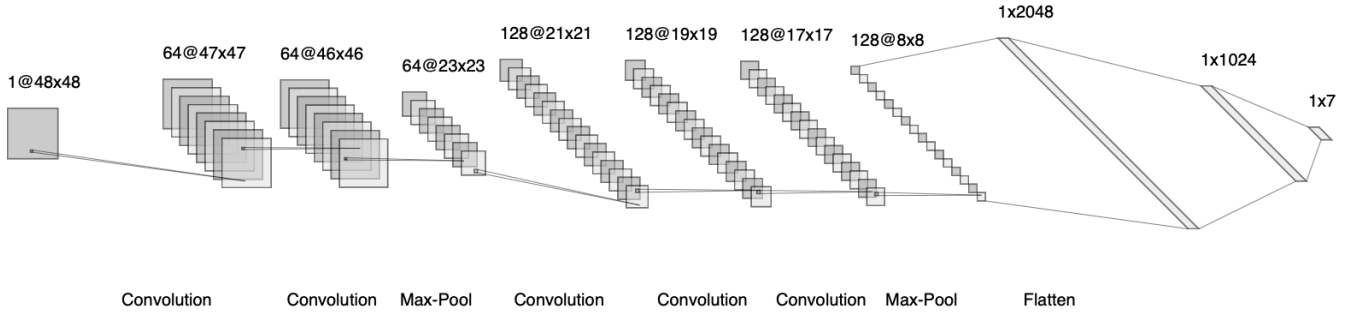
Fig. 1: Architecture of the baseline CNN for facial expression recognition.

a dataset that belongs to the same domain and for the same task. The architecture of the proposed CNN is shown in Fig. 1. This architecture is a simplified version of a VGG network [15], with less convolutional layers, given the amount of data available for training. However, the number of parameters is 19.2M which is about seven times less than a VGG network, which has 138M of trainable parameters. The proposed CNN takes a 48×48 gray-level image as input and it has five convolutional layers interchanged with two max-pooling layers and three fully connected layers. Rectified linear units (ReLU) and batch normalization are used after each convolutional layer. The first-two convolutional layers have 64 filters of size 2×2 and the last three convolutional layers have 128 filters of size 3×3. Furthermore, the spatial resolution of the second and fifth convolution layers are reduced by twice with two max-pooling layers to limit the number of parameters and the computation cost. The first fully connected layer is made up of 2,048 units, followed by a second layer of 1,024 and the output layer with seven units. Drop-out of 50% is used after the two-first fully connected layers to reduce the over-fitting. We use such a CNN in our study to compare four different approaches of crossing different datasets. For each of these approaches, only two datasets are crossed at each time, but it is also possible to cross more than two datasets.

A. Fine-tuning the Fully Connected (FC) Layers

For this first approach we consider a pre-trained CNN on dataset that belongs to the same domain and for the same task. The CNN is therefore trained and optimized on a source dataset (FER dataset) as illustrated in Fig. 2. A second dataset named target dataset is then used to fine-tune only the fully connected (FC) layers of the pre-trained CNN as illustrated in Fig. 3a. This new training considers the previous weights as the initial ones and it looks to minimize the error of the network on the target dataset. The parameters of the other layers are not modified. In this case, our assumption is that the representation learned in the convolutional layers on the source dataset is reasonable and general and it does not need to change because the source and target tasks are in the same domain. Usually, the higher level features computed by the last few layers are usually task-specific and are not transferable to new target tasks.
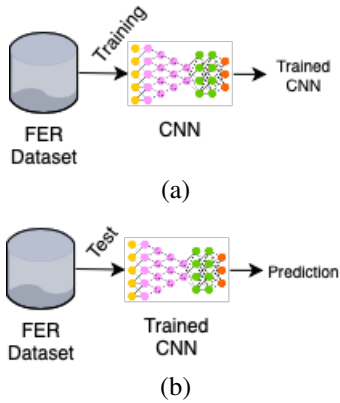

(a)

(b)

Fig. 2: The baseline CNN is: (a) trained on the training split of the source dataset (FER) and; (b) evaluated at test partition of the same dataset (source dataset).
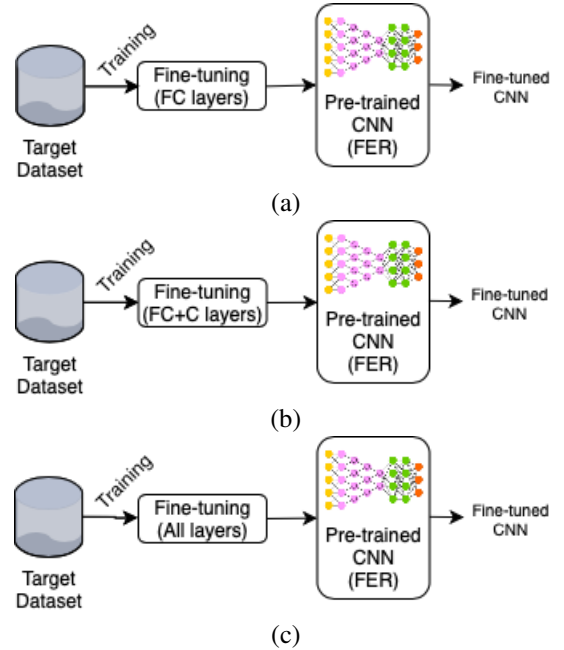

(a)

(b)

(c)

Fig. 3: The baseline CNN is fine-tuned using three strategies: (a) fully connected layers (FC); (b) fully connected layers and the last convolutional layer (FC+C); (c) all layers.

B. Fine-tuning the Fully Connected (FC) and the Last Convolutional Layer

This second method also requires the use of an already trained CNN. As with the previous method, CNN was trained by the FER database. This method is also a method of adjusting some CNN layers by a second dataset as illustrated in Fig. 3b. This time, the adjusted layers are the last convolutional layer and the three fully connected layers. This second method is very similar to the first one. The difference is that by adjusting the last CL layer, we are also modifying the learned representation. The convolutional layers are responsible for extracting features that best characterizes the classes. Modifying the convolutional layers therefore modifies the features that the model considered important. In this case, the assumption is that the representation learned in the convolutional layers is reasonable, but it can be adapted to the target dataset.

C. Fine-tuning All Layers

The third approach also requires a pre-trained CNN. As for the two previous cases, the CNN was pre-trained on the FER database. For this approach, the CNN is fully-trained a second time on a target dataset as illustrated in Fig. 3c. However, the CNN is initialized with the parameter values learned previously on the source dataset. Therefore, this approach involves training the CNN twice, with two different datasets while keeping the parameters learned on the source dataset and the initial parameters for learning on the target dataset. In this case, the assumption is that the whole network should be adapted to the target task and dataset, but it still will keep some "memory" of the source dataset.

D. Train the CNN on Merged Datasets

The last approach does not require a pre-trained CNN. Indeed, it consists of merging the source and target datasets into a single dataset and use such a dataset to train a CNN initialized with random weights. For this approach, the FER dataset was always used as the source dataset and it was merged with one of the other three target datasets. The CNN is then trained in both datasets simultaneously. This is the most time-consuming approach as it requires the fully training of the network.
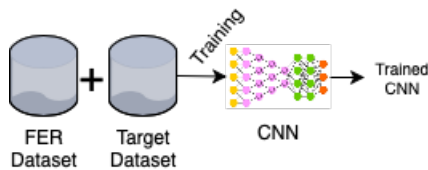

Fig. 4: The baseline CNN is randomly initialized and trained on merged source and target datasets.

Given the four proposed approaches for crossing datasets, which of them is the most appropriate in a situation where our aim is to minimize the error rate on both source and target datasets?
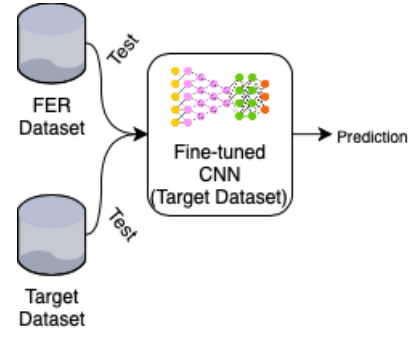

Fig. 5: The fine-tuned CNNs are evaluated on both source and target test datasets.

III. EXPERIMENTAL RESULTS

We have evaluated the four proposed approaches of crossing different datasets on four publicly available datasets: FER, JAFFE, TFEID and MUG. The Facial Expression Recognition (FER) Challenge dataset is made up of 35,888 48×48 gray-level images split into 28,709 images for training (80%), 3,589 images for validation (10%) and 3,589 images for test (10%). We have tried to keep the same proportion of data into each subset for the three other datasets. Besides, the data splitting also considered that a subject should be only in one of the subsets to avoid biasing the CNNs. The Japanese Female Facial Expression (JAFFE) dataset is made up of only 213 gray-level images with 256×256 pixels from 10 Japanese females. This dataset was split into 131 images for training (60%), 15 images for validation (10%) and 67 (30%) for test. The Taiwanese Facial Expression Image dataset (TFEID) is made up of 510 colour images of 40 Taiwanese subjects with different dimensions, ranging from 600 to 606 of height and from 480 to 595 in width. This dataset was split into 405 images for training (79.4%), 47 images for validation (9.2%) and 58 (11.4%) for test. Finally, the Multimedia Understanding Group (MUG) Facial Expression dataset is made up of 919 color videos with between 60 and 100 frames of 896×896 of resolution from 86 Caucasian adults. We need to explain here the process of frame extraction and labelling. Therefore, we ended up with 7,352 images which were split into 5,877 (80%) for training, 734 (10%) for validation and 741 (10%) for test. Table I summarizes the datasets and the amount of data in each partition.

Dataset	Total	Number of Samples and (%)		
		Training	Validation	Test
FER	35,888	28,709 (80.0)	3,589 (10)	3,589 (10.0)
JAFFE	213	131 (61.5)	15 (7.0)	67 (31.5)
TFEID	510	405 (79.4)	47 (9.2)	58 (11.4)
MUG	7,352	5,877 (80.0)	741 (10)	734 (10.0)

TABLE I: The four datasets and the three subsets.

Figures 6(a), 6(b) and 6(c) show that the accuracy on the test set of JAFFE, TFEID and MUG respectively. It
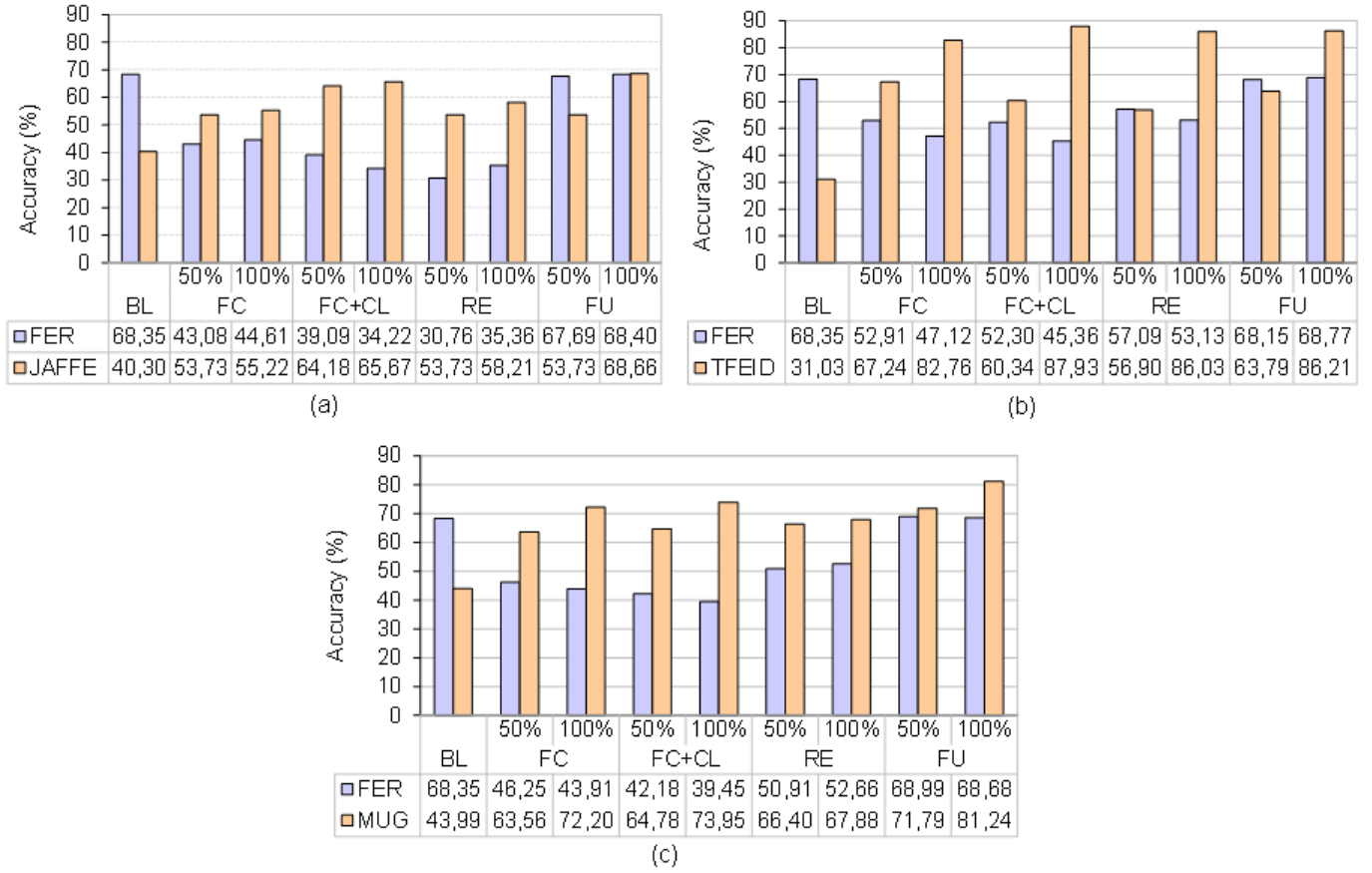
Fig. 6: Cross-dataset results of training with 50% and 100% of the dataset. Baseline model (BL), fine tuning the fully connected layers (FC), fine tuning the fully connected plus the last convolutional layer (FC+CL), model retraining of all layes (RE), and the fusion of training sets (FU) considering: a) FER dataset as source and JAFFE dataset as target; b) FER dataset as source and TFEID dataset as target; and c) FER dataset as source and MUG dataset as target.

is possible to see that the error percentage is lower for the experiments made with the whole of the dataset than learning with half of this dataset. This suggests that the larger the dataset used for learning, the more precise the CNN classification will be for the data in this dataset. However, by observing the results for the FER test set, it is not possible to see a trend. It is therefore conceivable that the amount of data in the target training dataset does not really affect the effectiveness of the CNN classification for the source dataset.

1) Performance on the source test set after fine-tuning the CNN:: The three fine-tuning approaches have reduced significantly the accuracy of the resulting CNN on FER test set between 20.94% and 28.67% in average. On the other hand, fusing FER dataset and a target dataset has kept or even in improved the accuracy in 0.27% in average.

2) Performance on the target test sets after fine-tuning the CNN:: The three fine-tuning approaches have improved significantly the accuracy of the resulting CNN on the target test sets between 20.57% and 37.41% in average. Fusing FER dataset and a target dataset has improved kept or even in improved the accuracy in 40.26% in average.

Tables **??**, **??** and **??** also show that only the merging approach does not have a higher error percentage than

the baseline model. In some cases, such an approach even achieves slightly lower error rates. Therefore, it seems that the merging approach is the most effective regarding the performance of the resulting CNN on the source dataset. Therefore, we can conclude that the dataset fusion approach provides the best trade-off because it is the only approach that does not have a negative effect on the source dataset, and it is the approach that gives the best performances for the target datasets.

CONCLUSION

In this paper we have investigated several approaches to supervised domain adaptation of CNNs, considering the problem of facial expression recognition. The main drawback of fine-tuning pre-trained CNNs is that it implicitly imposes a forgetting mechanism which degrades their performance on the source dataset. The availability of pre-trained networks has attracted several researchers to simply fine-tune these pre-trained networks on their target datasets. However, one has to take into account that "transfer learning" is not always useful, especially when we are working on the same domain. Therefore, one of the main contributions of this paper is to show that the massive use of pre-trained networks not

always leads to the best performance. It is preferable to use a simpler architecture, more adapted to the problem than using a complex pre-trained network and transfer learning.

Since our goal is quite different from other researchers which aims to create the most effective facial expression recognition approach for a target domain, comparing the results achieved in this work with the results of some of the works reviewed earlier, does not seem relevant. Among the five datasets used in [23], only JAFFE was also used in our experiments. For such a dataset, Zavarez *et al.* [23] achieved 44.32% of accuracy with their fine-tuned model while our approach achieved performances between 40.30% and 68.66% with no fine-tuning and with the fusion of training set and fully training of the CNN respectively. Therefore, we may say that the proposed approaches for dataset crossing are more efficient. Finally, Yang *et al.*[21] use JAFFE as the target dataset and the recognition rates achieved in cross-domain datasets are always lower than 40%.

REFERENCES

[1] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In I. Guyon, G. Dror, V. Lemaire, G. Taylor, and D. Silver, editors, *Proc. of ICML Workshop on Unsupervised and Transfer Learning*, volume 27, pages 17–36, Bellevue, Washington, USA, 02 Jul 2012. PMLR.

[2] P. Cardinal, N. Dehak, A. L. Koerich, J. Alam, P. Boucher, A. Lameiras, and P. Boucher. ETS System for AV EC 2015. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge - AVEC '15*, pages 17–23, 2015.

[3] M. J. Cossetin, J. C. Nievola, and A. L. Koerich. Facial expression recognition using a pairwise feature selection and classification approach. In *Intl Joint Conf on Neural Networks (IJCNN)*, pages 5149–5155. IEEE, jul 2016.

[4] J. de Matos, A. de S. Britto Jr., L. E. S. Oliveira, and A. L. Koerich. Double transfer learning for breast cancer histopathologic image classification. In *Intl Joint Conf on Neural Networks (IJCNN)*, pages 1–6. IEEE, jul 2019.

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 647–655, Bejing, China, 22–24 Jun 2014. PMLR.

[6] M. Esmaeilpour, P. Cardinal, and A. L. Koerich. Unsupervised feature learning for environmental sound classification using cycle consistent generative adversarial network. *Applied Soft Computing*, To appear:1–15, 2019.

[7] D. V. Jiaolong Xu, Sebastian Ramos and A. M. Lopez. Domain adaptation of deformable part-based models. *IEEE Trans Patt Anal. Mach. Intell.*, 36(12):23672380, 2014.

[8] R. Y. Jun Yang and A. G. Hauptmann. Cross-domain video concept detection using adaptive SVMs. In *Proc. ACM Intl Conf on Multimedia*, pages 188–197. IEEE, 2007.

[9] S. Li and W. Deng. Deep Emotion Transfer Network for Cross-database Facial Expression Recognition. In *24th Intl Conf on Patt Recognition (ICPR)*, volume 2018-Augus, pages 3092–3099. IEEE, aug 2018.

[10] S. Li and W. Deng. Deep Facial Expression Recognition: A Survey. pages 1–22, apr 2018.

[11] A. T. Lopes, E. De Aguiar, and T. Oliveira-Santos. A Facial Expression Recognition System Using Convolutional Networks. *Brazilian Symposium of Computer Graphic and Image Processing*, 2015-Octob:273–280, 2015.

[12] C. Mayer, M. Eggers, and B. Radig. Cross-database evaluation for facial expression recognition. *Patt Recognition and Image Analysis*, 24(1):124–132, 2014.

[13] N. Mousavi, H. Siqueira, P. Barros, B. Fernandes, and S. Wermter. Understanding how deep neural networks learn face expressions. In *Intl Joint Conf on Neural Networks (IJCNN)*, pages 227–234. IEEE, jul 2016.

[14] L. E. S. Oliveira, M. Mansano, A. L. Koerich, and A. S. Britto Jr. 2D Principal Component Analysis for Face and Facial-Expression Recognition. *Computing in Science & Engineering*, 13(3):9–13, may 2011.

[15] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. pages 1–8, 2014.

[16] B. Sun, S. Cao, L. Li, J. He, and L. Yu. Exploring Multimodal Visual Features for Continuous Affect Recognition. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, pages 83–88, New York, New York, USA, 2016. ACM Press.

[17] X. Wang, X. Wang, and Y. Ni. Unsupervised Domain Adaptation for Facial Expression Recognition Using Generative Adversarial Networks. *Computational Intelligence and Neuroscience*, 2018:1–10, jul 2018.

[18] S.-F. C. Wei Jiang, Eric Zavesky and A. Loui. Cross-domain learning methods for high-level visual concept classification. In *Proc. IEEE Intl Conf on Image Processing*, pages 161–164. IEEE, 2008.

[19] K. L. Wiggers, A. S. Britto Jr., L. Heutte, A. L. Koerich, and L. E. S. Oliveira. Document Image Retrieval Using Deep Features. *Proc Intl Joint Conf on Neural Networks*, 2018-July, 2018.

[20] K. Yan, W. Zheng, T. Zhang, Y. Zong, and Z. Cui. Cross-database non-frontal facial expression recognition based on transductive deep transfer learning. pages 1–8, nov 2018.

[21] Y. Yang, W. Zhang, and Y. Huang. Cross-domain expression recognition based on sparse coding and transfer learning. In *AIP Conf Proc*, volume 1839, page 020188, 2017.

[22] Z. Yu and C. Zhang. Image based Static Facial Expression Recognition with Multiple Deep Network Learning. In *Proc ACM on Intl Conf on Multimodal Interaction*, pages 435–442, New York, New York, USA, 2015. ACM Press.

[23] M. V. Zavarez, R. F. Berriel, and T. Oliveira-Santos. Cross-Database Facial Expression Recognition Based on Fine-Tuned Deep Convolutional Network. *30th Conf on Graphics, Patterns and Images (SIBGRAPI)*, pages 405–412, 2017.

[24] T. H. H. Zavaschi, A. S. Britto Jr., L. E. S. Oliveira, and A. L. Koerich. Fusion of feature sets and classifiers for facial expression recognition. *Expert Systems with Applications*, 40(2):646–655, feb 2013.

[25] T. H. H. Zavaschi, A. L. Koerich, and L. E. S. Oliveira. Facial expression recognition using ensemble of classifiers. In *IEEE Intl Conf on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1489–1492. IEEE, may 2011.

[26] J. Zhang, W. Li, P. Ogunbona, and D. Xu. Recent advances in transfer learning for cross-dataset visual recognition: A problem-oriented perspective. *ACM Comput. Surv.*, 52(1):7:1–7:38, Feb. 2019.