Multi-branch deep radial basis function networks for facial emotion recognition

Fernanda Hernández-Luquin · Hugo Jair Escalante

Received: 29 January 2021 / Accepted: 17 August 2021 / Published online: 27 August 2021
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2021

Abstract
Emotion recognition (ER) from facial images is one of the landmark tasks in affective computing with major developments in the last decade. Initial efforts on ER relied on handcrafted features that were used to characterize facial images and then feed to standard predictive models. Recent methodologies comprise end-to-end trainable deep learning methods that simultaneously learn both, features and predictive model, where the most successful models are based on convolutional neural networks (CNNs). While these models have excelled at this task, they still fail at capturing local patterns that emerge in the learning process. We hypothesize these patterns could be captured by variants based on locally weighted learning. Specifically, in this paper we propose a CNN-based architecture enhanced with multiple branches formed by radial basis function (RBF) units that aims at exploiting local information at the final stage of the learning process (i.e., in the layers close to the output layer). Intuitively, these RBF units capture local patterns shared by similar instances using an intermediate representation, then the outputs of the RBFs are feed to a softmax layer that exploits this information to improve the predictive performance of the model. This feature could be particularly advantageous in ER as cultural / ethnicity differences may be potentially identified by the local units. We evaluate the proposed method in several ER datasets and show the proposed methodology achieves state-of-the-art performance in some of them, even when we adopt a pre-trained VGG-Face model as backbone. We show the proposed method outperforms consistently CNNs that do not have the proposed local learning component. Moreover, we found the proposed model is advantageous when training datasets are reduced and when merging images coming from different distributions (e.g., combining two ER datasets). We show it is the incorporation of local information what makes the proposed model competitive.

Keywords Locally weighted learning · Radial basis function networks · Emotion recognition · Convolutional neural network · Looking at people

1 Introduction
Automated emotion recognition (ER) is the ability to identify human emotional states by analyzing speech, facial expressions and body gestures [26]. ER has been proved to be very useful in areas as: affective computing, human–computer interaction, and as support tool for psychology, psychiatry, neurology, and related applications and sub fields, e.g., for pain assessment, deception detection, etc. (see, e.g., [12, 22, 32, 39]). Clearly, automated methods for ER have a great potential impact in a wide variety of fields.

One of the most studied and useful modalities for ER from an affective computing perspective is that comprising visual information, including still images and sequences. In particular, facial expression recognition (FER) focuses on the analysis of facial imagery with the aim of building predictive models to match faces with emotions. Traditional approaches to FER were based on standard machine learning methods (e.g., support vector machines, neural
The formulation of a variant of CNNs that aim at
• bone architecture (VGG-Face).

Such examples and samples that are difficult to classify with
ments in performance, in particular, for minority class
into account this local information may lead to improve-
hypothesize that enhancements to CNN models that take
of the subjects in the dataset under consideration. We
hypothesize that enhancements to CNN models that take
into account this local information may lead to improve-
ments in performance, in particular, for minority class
examples and samples that are difficult to classify with
such global models.

At this point, it is important to emphasize that our goal is
to exploit local information at the instance level, similarly
as $k$–NN and other instance-based learning methods, while
CNNs and attention mechanisms capture locally spatial
information at the feature level.

In this paper, we introduce an enhanced CNN-based
architecture equipped with multiple sets of radial basis
function (RBF) units that aim at capturing local patterns.
We call our proposed method a Multi-Branch Deep RBF
Network. The proposed method comprises several branches
of RBF units coupled with a standard CNN that acts as
feature extractor, the outputs of the multi-branch compo-
nent are then concatenated and feed to a softmax layer that
plays the role of classifier. The model is trainable end-to-
end and inherits all of the benefits of CNNs for FER, with
the additional advantage of exploiting local information.
The proposed methodology is evaluated in a number of
benchmark FER datasets and we show that the proposed
enhancement outperforms considerably baselines that
include a CNN without any locality component. In addi-
tion, the proposed method obtains state-of-the-art perfor-
ance in some of the considered datasets; this is
remarkable, given the simplicity of the considered back-
bone architecture (VGG-Face).

The contributions of this paper are threefold:

• The formulation of a variant of CNNs that aim at
  incorporating local information explicitly in the
model

As such, we are exploring the first steps of locally\(^1\)
weighted learning in the context of deep
learning. To the best of our knowledge, this is the first
work adopting a locally weighted learning scheme in
the task of FER. Please note that there are very few
efforts on locally weighted deep learning in general,
see, e.g., [37, 51, 56].

• The introduction of Multi-branch deep RBF net-
  works for emotion recognition

This is an enhancement to CNNs that successfully exploits
local information at the instances level, we show its effec-
tiveness in the FER task.

• An experimental evaluation

showing the proposed methodology coupled with a standard
architecture (VGG-Face) as backbone outperforms baseline models
and achieves state-of-the-art performance that is compa-
orable to more sophisticated and complex methodolo-
gies for some datasets. More importantly, we provide
evidence that the competitive performance of the
proposed model is due to the incorporation of local
information.

The findings and conclusions drawn from this paper
motivate further research on the study of locally weighted
learning in the context FER and in general machine
learning.

The remainder of this paper is organized as follows.
Section 2 reviews related work on FER and local learning.
Next, Sect. 3 introduces the proposed multi-branch deep
RBF network model in detail. Then, Sect. 4 presents an
experimental evaluation of the proposed method in
benchmark FER datasets. Finally, Sect. 5 outlines conclu-
sions and future work directions.

2 Related work

This section briefly reviews related work on FER and on
the intersection of locally weighted learning with deep
learning.

2.1 Emotion recognition

Automated ER is a task that has been studied for a while
now, where the definition of emotion and a characteriza-
tion of human emotions was inherited from the psychology
field. Although there are too many taxonomies and defi-
nitions, the most widely accepted categorization is that of
Paul Ekman who defined six universal emotions: anger,
disgust, fear, happiness, sadness and surprise [6];

\(^1\) The term local in this context is with respect to samples in the
dataset, and not to features like is the case in attention-based models
or even in standard CNNs that can capture locally spatial information.
subsequently contempt was also considered as a basic emotion. We adhere to this categorization in the remainder of the paper.

The FER task was initially faced with a standard machine learning models feed with handcrafted features that aimed to capture discriminative facial characteristics. Common feature extractors comprise histogram of oriented gradients (HoG) [2, 14, 49], local binary patterns (LBP) [15, 43, 59], among others [10, 11]. While the considered classification models include support vector machines (SVM) [8, 47], AdaBoost [3, 10] and decision trees among others [5, 54]. While competitive, most of the methods from the first wave relied on handcrafted features that not necessarily are representative or descriptive of the data [30].

The advances in deep learning have motivated a second wave of methodologies that rely on deep learning [22]. Contrary to the traditional approaches, these methods simultaneously learn the representation for the input and the predictive model. In this way, features are derived entirely from data and these are tied to the classification model being learned. The performance of these models has surpassed that from traditional models and they comprise the standard solution to approach the FER task. The most commonly used deep learning architectures that have been used for FER are convolutional neural networks (CNNs), see, e.g., [4, 23, 35, 40, 44, 46, 50]. These models take as input raw images and learn multiple layers of convolutional filters that are applied to the inputs of the layer and their outputs feed to the next one. These models are coupled with other types of layers including dense and softmax layers to learn the predictive part of the model. Other popular architectures comprise residual networks [7, 24, 42, 52, 53] and sequential models (e.g., LSTMs) [58]. Additionally, these methods are able to incorporate additional mechanisms and features into the learning process making them quite effective and self-contained, for instance, architectures with attention mechanisms [7, 20, 33, 53], ad hoc loss functions [42] and other complex procedures [25, 57, 60]. The author is referred to [22] for a comprehensive survey on FER with deep learning-based methodologies. In Sect. 4, we compare the performance of the proposed methodology with baseline CNN architectures and with some of the most recent and effective CNN-based methodologies for FER.

### 2.2 Deep learning models with local learning

Despite the effectiveness of deep learning-based solutions, there are still open questions that deserve attention from the community and that could have a great impact into the field. One of these questions has to do with the lack of specific mechanisms in CNNs for taking into account local information at the instance level. Local information has proven to be very helpful in classical models within machine learning. Consider for instance locally weighted regression [34], where the incorporation of samples close to a query sample is used to approximate a regression function locally. This feature enables a linear model (e.g., least squares regression) to approximate nonlinear decision surfaces. There are many other cases where local information has proven to be very useful including support vector machines [19], learning vector quantization [36], decision trees and even there is a variant of neural network that implements locally weighted learning: the radial basis function network [34]. The next section reviews related work on some efforts on locally weighted learning for deep learning.

In the context of deep learning, local learning has been scarcely studied in the context of CNNs and other deep neural networks (DNNs). Zadeh et al. introduced a deep RBF model that aimed to make robust predictions against adversarial attacks [56]. The model is formed by a CNN architecture tied with RBF units in the output layer (these were used to perform classification). A loss function tailored to be robust against adversarial attacks was proposed. Vindervová et al. presented a method where a DNN and RBF networks are concatenated to classify adverse examples correctly [51]. Although these efforts combine deep learning with locally weighted learning the aim is not to improve the predictive performance of the model in general, but for specific adversarial scenarios. Moreover in [51], the DNN and the RBF network are trained separately, which results in the combination of two independent models.

There are other efforts from the community that aim at taking advantage of other locally weighted learning mechanisms in the context of deep learning. For instance, Bahri et al. proposed a deep k-nearest neighbor (DkNN) model for detecting noisy examples [1]. Earlier, Papernot et al. introduced a DkNN model that estimated neighbors across layers aiming to have additional information for the predictive model (interpretability) [37]. Like the methods in [51, 56], the DkNN methods target adversarial or noisy examples. Other efforts like that of Yu et al. have applied the principle of local learning (a model is trained only on the most relevant data for a given input) for very specific scenarios [55].

### 2.3 Discussion

The FER task has been approached for a while and the most effective solutions are those based on deep learning methodologies. These methods have the appealing features that they can learn simultaneously features and predictive model. While these models have obtained outstanding
performance, there are several questions around these models that deserve to be explored. In this paper, we aim to explore the benefits of locally weighted learning (LWL) into deep learning models for approaching the FER task. LWL has been scarcely studied in the context of deep learning. There are few efforts in this direction and all of them target very specific scenarios, for instance, classification with adversarial examples and interpretability. We argue that LWL could be beneficial for FER because there are samples that in order to be correctly classified, the model should build a sub-classifier that considers only samples similar to the query point. Intuitively, consider ER datasets in which a minority group is underrepresented, and dominated by another group (see Sect. 4.4 for an example). Global models are prone to fail to correctly classify such instances and this type of issues could be alleviated with LWL. For these reasons, we propose in this paper a LWL deep learning model that does not target specific scenarios (like adversarial samples). We show the potential of this model in an experimental evaluation.

3 Multi-branch deep radial basis function networks

The working hypothesis of this work is that the incorporation of local information at the instance level into the learning process of CNN-based models improves the recognition performance of the enhanced model for the FER task. The intuition behind this hypothesis is that in FER there may exist groups of instances that share similarities to each other in one or more aspects (e.g., in terms of ethnicity or age). Therefore, when classifying a query sample, a prediction that is built by taking into account information of similar instances should improve the recognition performance.

We introduce in this section a model that implements such an idea, the so-called Multi-Branch Deep RBF network model. In a nutshell, the model uses a CNN as backbone (e.g., VGG-face) and it is enhanced with a new layer formed by multiple branches of RBF units. Such RBF layer receives as input feature maps from the preceding layers of the CNN and its outputs are concatenated and connected to a softmax layer that makes predictions over the considered classes. This enhancement allows a CNN to implicitly incorporate local information that can have a positive impact in recognition performance. A graphical diagram of the proposed model is depicted in Fig. 1; the remainder of this section describes the proposed model in detail.

3.1 Radial basis function networks

An RBF unit is type of neuron that is associated with a center and a radius, it can be considered as a prototype in the input space whose position is updated (learned) from data. These units are commonly used in RBF networks and LVQ-based models. In the case of the former models, a set of units defines a layer, and commonly there is an RBF unit per class associated with the problem at hand (i.e., RBF units are often used instead of softmax ones for the predictive part of the model). The output \( h_i \) of an RBF unit \( i \) given input \( x_i \) is computed as follows:

\[
h_i = \phi_i(x_i) = e^{-\left(\|x_i - \mu_i\|^2 / (2\sigma_i^2)\right)}
\]

where \( x_i \in \mathbb{R}^d \) is a \( d \)--dimensional feature vector, and \( \mu_i \in \mathbb{R}^d, \sigma_i \in \mathbb{R} \) are the center and radius of RBF unit \( i \).

3.2 Multi-branch RBF layer

As previously mentioned, the proposed model extends a backbone CNN with RBF units arranged into branches as shown in Fig. 1. We relied on VGG-Face [38] as backbone because it is a well known and generic enough CNN for facial analysis that has proven to be very helpful in FER and related tasks when used as pre-trained model. One should note that VGG-Face is not a state-of-the-art methodology as those that are being currently proposed in the context of FER (see Sect. 2). Our decision for relying in a generic model lies in that we wanted to prove the proposed extension could lead to improvements with a standard model. Relying on more complex or elaborated models would make it more complicated to assess the actual improvement due to our local modules.

VGG-Face is an architecture formed by a series of convolutional layers followed by fully connected layers that are in turn followed by a softmax layer in charge of the classification process [38]. We modify the last few layers of backbone architecture as follows. We dropped all of the fully connected layers, and instead connected multiple branches of RBF units (see Expression (1)) to the output of the last convolutional layer. Such convolutional layer (see Fig. 1) returns as output 512 feature maps of dimensionality \( 7 \times 7 \). We take the activation of these maps as inputs to the multiple RBF branches.

The motivation behind having multiple branches of RBFs is that if one would have a single RBF, the input would be very high dimensional (i.e., \( 7 \times 7 \times 512 = 25,088 \)) and potentially useful local information would get lost or would be very difficult to process. In contrast, having multiple branches of RBFs each taking as input a low-dimensional vector could lead to exploit local information easily. In fact, it would be expected that
each branch could capture a local pattern different from the rest (see Sect. 4.3). Therefore, we propose to process the outputs of the last convolution layer in such a way that each branch takes an input of manageable size. Specifically, we process the last convolutional layer with a set of filters that yield a 7×7 output and we use as many of these filters as branches are considered in the model (see Fig. 1). These outputs are flattened and feed to the RBF branches in a fixed order.

In the proposed model, the pre-trained convolutional layers of the VGG-Face model are used [38]. The RBF centers and their corresponding radius are initialized randomly. The model is then trained end-to-end using backpropagation and stochastic gradient descent (more details in Sect. 4) using the FER dataset at hand. We performed experiments on different ways of freezing and updating weights for the whole architecture, we observed there was no significant difference in performance when updating or not the (convolutional) weights inherited from the backbone architecture; therefore, we decided to froze the weights of VGG-Face and learn only the remainder of the parameters. One should note that even when the display in Fig. 1 shows a branch per feature map, a reduced number of branches could be used as well, see Sect. 4.

The additional parameters introduced by the RBF-based model to the VGG-Face backbone are the units and radius of the RBF units (μi and σi) in Equation (1). These parameters are adjusted during the learning process. Likewise, the additional hyperparameters are the number of branches and number of units per branch. The latter should be specified by the user and are dependent on the dataset at hand. In Sect. 4.2, we show an ablation study on these two hyperparameters that can serve as guideline for users. Intuitively, a large number of branches would result in more diverse units that could capture different patterns, while too many units in each branch may lead to capturing unbalanced information across units (see Sect. 4.3 for a qualitative analysis).

In the next section, we present an experimental evaluation that shows the proposed model outperforms strong baseline models. In particular, we compare the proposed model to a reference model that replaces the RBF branches by dense layers; this is illustrated in Fig. 2. This is important to mention as this comparison will allow us to determine the actual benefits of having local information instead of fully connected units as it is standard in CNN models.

3.3 Discussion

We just introduced the Multi-Branch Deep RBF Network model: an enhancement to CNNs in which a layer formed by several branches of RBF units is added before the softmax classification layer. The main novelty of this

![Fig. 1 Diagram of the proposed multi-branch deep radial basis function network architecture. A backbone architecture (VGG-Face in this case) is enhanced with multiple branches of RBF units which are then connected to a dense softmax layer.](image)

![Fig. 2 Illustration of units used in the proposed model (left) and the CNN variant, multi-branch CNN (right). Each module has a convolutional filter followed by a flatten layer. The difference of each module is the layer used after the flatten layer: a uses RBF layer while b uses dense layer with a ReLU activation function.](image)
proposal is the adoption of multiple RBF sub networks that allow the model to deal with the high dimensionality of the input space, likewise, having multiple branches allow the model to capture specific local information in each of these. Compared to alternative solutions from deep RBF networks, which use a single branch of RBF units, in our model the outputs of multiple branches are feed to a softmax layer that makes predictions, whereas in reference work (see, e.g., [51, 56]) RBF units are used to make the predictions directly. Additionally, one should note that the focus of previous work has been on using local information for under covering adversarial attacks (see, e.g., [37, 51, 56]), while in this paper our goal is to improve the overall classification process. Finally, to the best of our knowledge this is the first effort on trying to incorporate local information at the instance level into the task of FER. As shown in the next section, the proposed enhancement improves considerably the performance of reference models and performs favorably with state-of-the-art solutions that are based on much elaborated mechanisms and models (e.g., attention-based models).

4 Experimental evaluation

This section presents an experimental evaluation of the proposed model in the FER task; the goal is to show the competitiveness and benefits of the Multi-Branch Deep RBF Network model when compared to reference models and to state-of-the-art solutions. We first introduce the datasets and experimental settings; then, we present an ablation study that analyzes the performance of our model under different parameter settings; next we show some visualizations that aim at highlighting the benefits of our model; then we analyze the performance of our model when considering other backbones than VGG-Face; finally, we compare the performance of our model to reference and state-of-the-art methods and conclude with a discussion.

4.1 Experimental settings

For the experimental comparison, we used the following benchmark datasets that have been widely used in the literature (see [22]): Real-world Affective Faces Database (RAF-DB), RAF-DB Compound [21, 23], the Extended Cohn-Kanade Dataset (CK+) [27], the Japanese Female Facial Expression (JAFFE) Dataset [28, 29] and FER 2013 [13]. Additionally, we performed experiments in a challenging dataset combining both CK+ and JAFFE datasets. Samples from the considered datasets are shown in Figs. 3 and 4 and some statistics are presented in Table 1.

The considered datasets comprise a diversity in terms of the number of samples, background/recording conditions and complexity, where one should distinguish datasets under the standard ER setting from datasets of greater

![Fig. 3 Sample images associated with different emotions for the CK+, JAFFE, FER2013 and RAF-DB datasets. Please note that for the CK+ dataset we are showing an image of the Contempt emotion instead of the Neutral one, as it is the only dataset without the latter emotion](image)

![Fig. 4 Sample images associated with different emotions considered in the RAF-DB Compound dataset](image)

| Dataset | #Tr. | #Val. | #Test | # E |
|---------|------|-------|-------|-----|
| CK+ [27] | 877  | 94    | 123   | 7   |
| JAFFE [29] | 143  | 35    | 35    | 7   |
| CK+JAFFE | 1020 | 129   | 158   | 8   |
| FER 2013 [13] | 28,709 | 3589  | 3589  | 7   |
| RAF-DB [21] | 12,271 | 3068  | 3068  | 7   |
| RAF-DB Compound [23] | 3162 | 792   | 792   | 11  |

The number of classes (E) and training, validation and test samples are shown. Please note that for CK+JAFFE the number of classes is 8 as we include the Contempt emotion that is only present in CK+
difficulty. Standard datasets including CK+, JAFFE, FER2013 and RAF-DB comprise basic emotions and images coming from the same distribution. The challenging datasets are CK+-JAFFE and RAF-DB Compound, the former made up by merging images of the CK+ and JAFFE datasets, and the latter considering a fine-grained classification of emotions, see Fig. 4.

The intuition behind experimenting with the merged CK+-JAFFE dataset lies in that we wanted to assess the performance of our model when there are clear differences across samples from the same category.

On the other hand, the RAF-DB Compound is challenging because it considers compound emotion categories (e.g., Fearfully-surprised, sadly-angry, happily-disgusted, etc.); 11 categories are considered, see Fig. 4. The idea of considering this dataset is to show the benefits of incorporating local information into the recognition process for approaching a fine-grained FER task. It is expected that the proposed model is more advantageous in the two considered challenging datasets.

For all of the datasets, we used the top-1 accuracy on the test set as the evaluation measure of performance. This is in agreement with previous work using the same datasets. The same partitions for training and testing were used in datasets where these were available (RAF-DB, RAF-DB Compound and FER2013) and random splits of 80% for training and validation and 20% for testing were used CK+ and JAFFE. For the latter datasets, multiple partitions were generated and their results averaged in each experiment.

The models were trained/fine tuned using the Adam [17] optimizer with a batch size of 32 during 100 epochs. The performance in validation was used to monitor convergence of the model. We determined the value of $\sigma$ experimentally as $\sigma = 0.0528$. The model was trained in a laptop with a Nvidia GTX card 2080 with 8Gb of VRAM, and a processor I7 6700 K with 32 Gb of RAM.

### 4.2 Ablation study

In this section, we evaluate the performance of the proposed model when varying the number of branches and units. We present in Fig. 5 the results of this evaluation for the six considered datasets. Results are shown as heat maps (the darker the better), the number of branches is specified in the $x$ axis, and the number of units is shown in the $y$ axis.

As it can be seen from this figure, mixed results are obtained for the different datasets. Being the CK+-dataset the easiest and RAF-DB Compound the toughest in terms of recognition performance. The difference between the lowest and highest performance achieved for every dataset makes clear that it is necessary to adequately tune both of these parameters (e.g., compare the lowest and highest performance in Fig. 5b and f).

Although no general conclusion can be drawn on the values of parameters, a pattern that seems to be present in all of the datasets is that a larger the number of branches seems to result in better performance of the model. Also, it seems that a small number of RBF units combined with large number of branches is a somewhat robust combination of parameters. In general, the obtained performance in most datasets is competitive with the state-of-the-art (see Sect. 4.6). For the experiments reported in the next sections, the best configuration of parameters for each dataset was used.

### 4.3 Visualization of centers

We now present visualizations of learned RBF centers for two configurations of parameters of the proposed model for the CK+-JAFFE dataset. We chose this particular dataset because it is one formed by instances from two different datasets and we expect the local information to be particularly helpful (see Sect. 4.6). Also, one should note that performance for this dataset did not vary too much for the different choices of parameters as shown in Fig. 5c.

Figure 6 shows the centers for a configuration with 4 branches and 8 RBF units per branch; the reported performance for this configuration was 0.943. From this figure, it can be seen that centers across the branches are very different to each other. Corroborating the hypothesis that different centers are modeling different aspects of the input feature maps. It is only for branch 1 that there seem to be similarities among centers (column 1, rows 4–7 of the left plot). In general, it seems that the relevant information is located near the center of the image (blue values in the center, yellow for the background), which make sense given the approached task is ER. However, there are a few centers that are also giving large weights to the region surrounding the face (blue background).

Figure 7 shows the centers but for a different configuration: 8 branches and 4 RBF units per branch; the reported performance of 0.953. Again, centers seem to be visually different to each other, although the differences across RBF units of the same branch (rows, right plot) are less notorious. This could be reflecting the fact that branches are capturing local patterns with subtle differences across RBF units (except branch 5, fifth row in Fig. 7 that seems to be learning the same pattern in the 3 RBF units). In fact, this type of centers results in better performance for the approached dataset. Finally, it is worth to emphasize that in both cases the centers seem to converge to an useful
Fig. 5  FER recognition performance when varying the number of branches and RBF units per branch in the proposed model.
4.4 Comparison with reference models

Table 2 shows a comparison of performance of the proposed model with other variants of CNN that approach the same task. We report the average and standard deviation obtained from 10 experiments with different random initialization. As reference models, we considered: (1) the backbone model, VGG-Face, the pre-trained network was subject of a fine-tuning process with the new classes, the last layer was removed and replaced by a softmax one with as many units as ER classes; the fully connected layers were reinitialized and subject to the fine-tuning process too. We considered this reference model to evaluate the added value of incorporating local information on top of it with our method. (2) Multi-branch CNN is a model in which the branches of RBF units are replaced by dense layers (see Fig. 2). The motivation for using this model is to determine whether adding more parameters (RBF units in our model vs. dense layers in the multi-branch CNN) to the backbone is the cause of improvement. Overall, the goal of this experiment is to assess the benefits of the proposed model when compared to competitive models that do not incorporate local information.

From Table 2, it is clear that the proposed model outperforms both of the reference models. The differences in performance are significant for most datasets and there are also dramatic improvements in some cases. Compare, for instance, the performance of VGG-Face and the proposed model for the RAF-DB and RAF-DB Compound datasets. The differences in performance are impressive. This could be due to the mismatch between the datasets (both in terms of type of images and classes) used for training VGG-Face and the ones considered for evaluation, even when we fine-tuned the FC layers of the model, the mismatch seems to be too large as to be learned by the FC layers. Actually, the MB-CNN baseline outperforms VGG-Face in all but the JAFFE dataset. Showing evidence that the added layers to the standard VGG-Face architecture improved the recognition performance.

We further analyze the differences in performance between VGG-Face and MN-RBFN. Figure 8 shows the representation, starting from random numbers (left plots in Figs. 6 and 7).

Table 2 Top-1 accuracy classification for the considered datasets

| Datasets | VGG-Face | MB-CNN | MB-RBFN |
|----------|----------|--------|---------|
| CK+      | 0.8291 ± 0.003 | 0.8381 ± 0.051 | 0.9964 ± 0.0037 |
| JAFFE    | 0.6352 ± 0.012 | 0.5971 ± 0.032 | 0.9796 ± 0.0314 |
| CK+JA    | 0.8341 ± 0.0018 | 0.8594 ± 0.021 | 0.9872 ± 0.0024 |
| FER13    | 0.4731 ± 0.035 | 0.6751 ± 0.0082 | 0.6815 ± 0.0097 |
| RAF      | 0.4289 ± 0.058 | 0.7237 ± 0.041 | 0.810 ± 0.0014 |
| RAF-C    | 0.2330 ± 0.0012 | 0.4739 ± 0.0034 | 0.5768 ± 0.0074 |

We compare the performance of the proposed model (MB-RBFN) with reference models (VGG-Face and MB-CNN)
confusion matrices for VGG-Face and the proposed model on the challenging CK+-JAFFE dataset. It can be seen that VGG-Face makes considerably more mistakes for the anger, fear and sadness categories. Our model misclassified 4 images from the JAFFE dataset and 6 from CK+, whereas the VGG-Face model made 28 and 29 mistakes for JAFFE and CK+ images, respectively. This represents 58% of images from JAFFE in the test set and only 10% if CK+ images. This clearly illustrates the benefits of incorporating local information into the CNN model: underrepresented samples are better classified (similar behavior was observed for the other baseline model). We refer the reader to Appendix A for a comparison of confusion matrices for the three models in the RAF-DB Compound dataset.

In order to further analyze these errors, Fig. 9 shows sample images from the surprise and fear categories. The former being one of the best classified by both models\footnote{We included images for surprise instead of happy because for the latter category an single test image from JAFFE was included; hence, it is not an informative class.} and the latter the most difficult class for the VGG-Face model. It can be seen from this figure that samples for the surprise category share a notable pattern regardless of their origin: the mouth is open in all cases, this makes the generic VGG-Face model to correctly classify most of the test instances in the mixed dataset. However, for the fear category images coming from CK+ and JAFFE look visually different to each other, yet sharing similarities within each dataset. This makes this class particularly challenging to VGG-Face, while the proposed model is able to correctly classify every instance from this class. This could be due to the local information incorporated into the model, and we think this is the main distinctive feature of our proposal.

Regarding the MB-CNN baseline (column 3 in Table 2), it is also outperformed by the proposed model in every considered dataset, where the lowest improvement obtained was for the FER2013 dataset. We hypothesize this could be due to the large number of images available for training in this dataset (more than 28,000), that allow the MB-CNN model to find a competitive configuration of parameters with the extended dense layer added to the VGG-Face model. Still, the proposed model obtained the highest performance overall (see Appendix A). This difference in performance shows that is the local information, as captured by the proposed model, was the decisive factor.

\begin{figure*}[h]
\centering
\includegraphics[width=\textwidth]{cm.png}
\caption{Confusion matrices obtained by VGG-Face (top) and the proposed model (down) for the CK+-JAFFE dataset.}
\end{figure*}

\begin{figure*}[h]
\centering
\includegraphics[width=\textwidth]{images.png}
\caption{Images from the test set of CK+-JAFFE for the surprise and fear categories. In each sub-figure, the top row displays the CK+ images and the bottom one shows the JAFFE images.}
\end{figure*}
for obtaining better performance across the considered datasets.

4.5 Performance on different backbones

In this section, we analyze the performance of the proposed model when using other backbones in the MB-RBFN model. Our goal is to provide evidence that the improvement offered by the inclusion of the multi-branch RBF units is consistent with other pretrained models. For this study we considered FaceNet [41] and a ResNet18 [16] model. FaceNet is a CNN-based on an inception architecture [48]; the network we used has been pretrained on 10M facial images from the Celeb-1M dataset. Also, we used a ResNet18 model [16] that was trained on the datasets we are considering for evaluation.

Results of this experiment are shown in Tables 3 and 4 for the FaceNet and ResNet18 models, respectively. We report the performance obtained by the backbone, the MB-CNN and the MB-RBFN models as described in Sect. 4.4; we report the results of a single run with each of the models and for each of the datasets. For each of the datasets, we used the configuration of branches and units that was considered in Table 2.

From Tables 3 and 4, it can be seen that the proposed model still obtains the highest performance in most of the datasets. However, this time there are two datasets for each backbone in which our model does not perform that well. Regarding the FaceNet backbone (Table 3), the MB-RBFN obtained lower performance than the MB-CNN method in the RAF datasets. This could be due to the fact the FaceNet model was pretrained in a dataset that is 4X larger than that used for pretraining VGG-Face. On the other hand, when using the ResNet18 backbone (Table 4) the MB-RBFN obtains slightly lower performance than MB-CNN for the CK+-JAFFE dataset, although the difference is rather small. Also, the ResNet18 model obtained lower performance than the backbone itself for the RAF-DB Compound dataset.

Despite the improvements in performance with respect to VGG-Face are smaller, this preliminary study (recall we are reporting a single run of each model and we did not spend time tuning the hyperparameters of the model) shows that the proposed model is also competitive when changing the backbone. Interestingly, it is in large datasets where the improvement is lower, this reinforces the hypothesis that when small datasets are available our proposed model could be more advantageous.

Finally, since we are using the same number of branches and units for the MB-RBFN model in the three considered backbones, comparing the performance of the MB-RBFN model in Tables 2, 3 and 4 can give us insights on how dependent the number of branches and units is with respect to the backbone and the dataset used for pretraining them. It seems such dependence is negligible, as the performance across the different datasets is similar regardless of the considered backbone.

4.6 Comparison with the state of the art

In this section, we compare the performance obtained by the proposed model with state-of-the-art references that have used the same datasets. For this comparison, we use results obtained with the VGG-Face backbone, this is because these results are the average of 10 runs and therefore are more reliable than those obtained with the two other tested backbones. Table 5 shows the results of the comparison. For each of the considered datasets, we report the performance of recent references including the best result reported so far in each dataset to the best of our knowledge. One should note that for our model we report the average over 10 runs as reported in Table 2, while for the reference models we take the single best result in the corresponding references. We include the results obtained by the baseline models for completion.
From Table 5, it can be seen that it is only in two datasets, FER2013 and RAF-DB, out of the five considered for this evaluation that the proposed model does not achieve performance competitive with the state of the art. Interestingly, these are precisely the two datasets with the largest number of samples with 28,807 and 12,271, respectively. This result seems to indicate that the proposed model is particularly helpful for low mid-sized datasets. Also, the competitive performance in CK+ and JAFFE seems to suggest that the model performs better in small datasets recorded under controlled conditions. This is partially true as the MB-RBFN model achieves competitive performance in RAF-DB-C, which has been recorded in the wild (see below). Likewise, since the references under comparison are based on extremely complex models, tailored mechanisms and sophisticated procedures, it is not strange that they perform better when enough data is available.

As previously mentioned, the proposed model achieves very competitive performance in CK+, JAFFE and RAF-DB Compound datasets. In CK+ our work establishes a new reference result and in the JAFFE and RAF-DB Compound datasets the model achieves comparable performance. It is remarkable the performance obtained by the proposed model in the RAF-DB Compound dataset, as this features a problem of (very) fine-grained classification with overlap among classes (compare the classes Sadly-Disgusted and Sadly-Angry, see Fig. 10) and highly imbalanced (4 classes comprise 72% of the samples, and the 7 remaining classes with 6% less out of the total number of samples). This result provides further evidence that the model is particularly helpful for this type of problems.

It is important to emphasize that among the references considered in the comparison we are including all types of recent models and mechanisms and it is to some extent unfair to compare a model like ours, which uses a simple backbone (VGG-Face). For instance, for JAFFE the only method that obtains better performance than our model is based on a complex CNN equipped with attention mechanisms and taking advantage of both learned and hand-crafted features. Whereas for RAF-DB Compound, the method with the highest accuracy is a ResNet18 model with a so-called separate loss that enhances the initial architecture to consider intra and inter class information for the ER process. Clearly, our model is advantageous in terms of simplicity; besides, it is possible that if we rely on a more complex backbone models the performance of the multi-branch RBF model could be even superior. In fact, using a more elaborated backbone model could help the proposed model to reach competitive performance in the large-scale datasets FER and RAF-DB. Finally, please note that our model is not doing any ad hoc feature learning process: we are relying on the pre-trained VGG-Face model, while most other references comprise expensive learning-from-scratch or fine-tuning processes that often use additional external data.

### Table 5

Top-1 accuracy classification for the considered datasets

|       | CK+ | JAFFE | FER2013 | RAF-DB | RAF-DB C |
|-------|-----|-------|---------|--------|---------|
| Ref.  |     |       | Ref.    | Ref.   | Ref.    |
|       | Acc.|       | Acc.    | Acc.   | Acc.    |
| [4]   | 0.9806 | [20]  | 0.9852  | [25]   | 0.7830  | [45]   | 0.9055 | [24] | 0.5884 |
| [33]  | 0.9800 | [46]  | 0.9531  | [20]   | 0.7582  | [9]    | 0.8942 | [23] | 0.5795 |
| [40]  | 0.9732 | [33]  | 0.9280  | [33]   | 0.7002  | [57]   | 0.8690 | [23] | 0.5354 |
| [53]  | 0.9730 | [58]  | 0.9238  | [35]   | 0.6640  | [60]   | 0.8677 | [25] | 0.5020 |
| [31]  | 0.9537 | [50]  | 0.7810  | [44]   | 0.6617  | [25]   | 0.758  | [18] | 0.4830 |
| VGG-Face | 0.8295 |       | 0.6352  |       | 0.4731  |       | 0.4289 | –   | 0.2330 |
| MB-CNN | 0.8381 |       | 0.5971  |       | 0.6751  |       | 0.7237 | –   | 0.4739 |
| MB-RBF | **0.9964** |       | 0.9796  |       | 0.6815  |       | 0.81   | –   | 0.5758 |

Bold values represent the best result for each dataset.

We compare the performance of the proposed model with state-of-the-art references. The number between parenthesis is the position in the rank of the results when ranking performance from highest to lowest.
4.7 Discussion

We have presented an experimental evaluation of the proposed Multi-Branch Deep RBF network model. We reported experiments in six datasets widely used for ER, two of them were variants that presented particular challenges with which state-of-the-art methods struggle. Our experimental evaluation showed that the proposed model outperforms considerably to reference models that included a similar model formed by dense layers only and the backbone. The comparison with the reference models together with a visual inspection of the learned centers comprises evidence that local information as captured by the proposed model is useful for approaching the ER task.

On the other hand, the proposed model compared favorably with recent methodologies that are based on much more complex techniques and procedures. This is an outstanding result given that the proposed model relies on a very generic, yet effective, backbone model: VGG-Face. Interestingly, it was shown that the proposed model offers more advantageous in datasets with more challenging conditions, namely small–medium sample size, with high class-imbalance, class overlap and with images coming from two different distributions. The obtained results are thus encouraging and motivate further research on the incorporation of local information into deep learning.

5 Conclusions

We introduced the Multi-Branch Deep RBF network, a model that improves CNNs by a mechanism that allow it to incorporate local information in the recognition process. The proposed model relies on VGG-Face as backbone for feature extraction, where the last convolutional layer of this model is connected to multiple branches of RBF units. The outputs of these are concatenated and connected to a softmax layer. The proposed model is initialized with VGG-Face and the RBF layers are fine tuned. Experimental results are reported in six ER datasets.

The following summarize the main findings of this work:

- The inclusion of local information, via the multi-branch RBF units, improves significantly the performance of a CNN model. In fact, the proposed model outperforms a similar model extended with a dense layer, showing that the RBF units are responsible of the improvement in performance.

- The proposed model is competitive with state-of-the-art methods based on more complex architectures and mechanisms, even when we rely on a standard backbone (VGG-Face). The model achieved competitive results in 3 out of 5 datasets with a much simpler implementation.

- The proposed model proved to be more advantageous for datasets with challenging conditions that include small sample size, high overlap among classes, datasets with mixed distributions in the test set and with high imbalance ratios.

- The centers of RBF units from different branches capture local information and this information resulted very helpful for classifying samples coming from different distributions (e.g., with the CK+-JAFFE dataset).

The findings and results presented in this paper are encouraging and motivate further research. Specifically, in future work we are planning to explore the following:

- Applying the proposed method in other appropriate domains. We anticipate the proposed model could be very helpful in fine-grained classification problems.

- Alternative ways of incorporating local information in CNNs. Including the use of autoencoders and deep clustering methods, and combining them with the proposed MB-RBFN.

- Using information from RBF centers for interpretability. Trying to determine whether centers can be associated with certain type of samples, and trying to identify undesired biases in the model by analyzing the centers and radius.

Confusion matrices for RAF-DB Compound

In this section, we analyze the confusion matrices for the proposed and reference models in the RAF-DB Compound dataset. Figure 11 shows these confusion matrices. From this figure, it can be seen that the proposed model is advantageous over both reference models. Consider, for instance, the Angrily-Disgusted class: despite this is the majority class, it is a very difficult to predict for the three models; clearly, MB-RBF obtains the best performance, by reducing considerably the number of misclassifications made by VGG-Face and MB-CNN. Similar behavior can be observed for non-majority classes. This analysis
illustrates the benefits offered by incorporating local information into the proposed MB-RBF model.

**Funding** This work was supported by Project Grant CONACYT CB-S-26314. The first author is supported by scholarship No. CONACYT 731653.

**Declarations**

**Conflict of interest** The authors declare that they have no conflict of interest.

**References**

1. Bahri D, Jiang H, Gupta M (2020) Deep k-nn for noisy labels. In Proceedings of the 37th International Conference on Machine Learning, PMLR, 119: 540–550
2. Chen J, Takiguchi T, Ariki Y (2015) Facial expression recognition with multithreaded cascade of rotation-invariant hog. 636–642
3. Chen L, Wu M, Pedrycz W, Hirota K Adaboost-knn with direct optimization for dynamic emotion recognition. In Emotion Recognition and Understanding for Emotional Human-Robot Interaction, Systems, 41–55. Springer
4. Chen Y, Wang J, Chen S, Shi Z, Cai J (2019) Facial motion prior networks for facial expression recognition. 1–4
5. Hivi ID, Maiwan BA (2019) Facial expression classification based on svm, knn and mlp classifiers. 70–75
6. Paul E. Universals and cultural differences in facial expressions of emotion. Nebraska Symposium on Motivation, 19, 207–283, 2917
7. Farzaneh AH, Qi X Facial expression recognition in the wild via deep attentive center loss. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2402–2411
8. Georgescu M-I, Ionescu RT, Popescu M (2019) Local learning with deep and handcrafted features for facial expression recognition. IEEE. Access 7:64827–64836
9. Gera D, Balasubramanian S (2021) Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition. Pattern Recogn Lett 145:58–66
10. Ghimire D, Lee J (2013) Geometric feature-based facial expression similarity in image sequences using multi-class adaboost and support vector machines. Sensors 13(6):7714–7734
11. Ghimire D, Lee J, Li Z-N, Jeong S (2017) Recognition of facial expressions based on salient geometric features and support vector machines. Multimedia Tools Appl 76(6):7921–7946
12. Gonzalez-Lozoya SM, Calleja Jorge de la P, Hugo Jair El , Medina MA, Ruiz AB (2020) Recognition of facial expressions based on CNN features. Multim Tools Appl, 79(19-20):13987–14007
13. Goodfellow IJ, Erhan D, Carrier Pierre Luc, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee D-H et al
17. Kingma, Diederik P, Adam, Jimmy B (2014) A method for
stochastic optimization. arXiv preprint arXiv:1412.6980.
18. Kollias D, Sharmanska V, Zafeiriou S. Face behavior à la carte: Expressions, affect and action units in a single network. CoRR (2019) arXiv:abs/1910.11111
19. Ladický L, Torr PHS (2011) Locally linear support vector machines. In Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML’11, page 985–992, Madison, WI, USA. Omnipress
20. Li J, Jin K, Zhou D, Kubota N, Zhaojie J (2020) Attention mechanism-based cnn for facial expression recognition. Neurocomputing 411:340–350
21. Li S, Deng W (2019) Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transact Image Proc 28(1):356–370
22. Li S, Deng W (2020) Deep facial expression recognition: A survey. IEEE Transact Affective Comput
23. Li S, Deng W, JunPD (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. 2584–2593
24. Li Y, Yao L, Li J, Guangming L (2019) Separate loss for basic and compound facial expression recognition in the wild. 897–911
25. Liang L, Lang C, Li Y, Feng S, Zhao J (2020) Fine-grained facial expression recognition in the wild. IEEE Transact Inf Forensics Security 16:482–494
26. Liang Paul P, Zadeh A, Morency L-P (2018) Multimodal local-global ranking fusion for emotion recognition. 472–476
27. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. 94–101
28. Lyons M, Akamatsu S, Kamachi M, Hyoba J (1998) Coding facial expressions with gabor wavelets. 200–205
29. Lyons M, Kamachi M, Hyoba J (1998) The japanese female facial expression dataset (JAFFE) database
30. Malaiej A, Amor BB, Daoudi M, Srivastava A, Berretti S (2011) Shape analysis of local facial patches for 3d facial expression recognition. Pattern Recognition, 44(8):1581–1589
31. Meng Z, Liu P, Cui J, Han S, Tong Y (2017) Identity-aware convolutional neural network for facial expression recognition. In 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), 558–565. IEEE
32. Miao H, Zhang Y, Li W, Zhang H, Wang D, Feng S (2018) Chinese multimodal emotion recognition in deep and traditional machine learning approaches. 1–6
33. Minaee S, Abdolrashidi A (2019) Deep-emotion: facial expression recognition using attentional convolutional network. arXiv preprint arXiv:1902.01019
34. Mitchell TM (1997) Machine learning. International Edition. McGraw-Hill Series in Computer Science. McGraw-Hill
35. Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. 1–10
36. Nova D, Estévez PA (2014) A review of learning vector quantization classifiers. Neural Comput. Appl. 25(3–4):511–524
37. Papernot N, McDaniel P (2018) Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. arXiv preprint arXiv:1803.04765
38. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. British Machine Vision Conference
39. Pitaloka DA, Ajeng WT, Basaruddin, and Dewi YL. (2017) Enhancing cnn with preprocessing stage in automatic emotion recognition. Proc Comput Sci 116:523–529
40. Rahul Ravi SV, Yaduhkrisha, et al (2020) A face expression recognition using cnn & lbp. 684–689
41. Schroff F, Kalenichenko D, Philbin J, Facenet (2015). A unified embedding for face recognition and clustering. IEEE Comput Soc, 815–823
42. Sepas-Moghaddam A, Etemad A, Correia PL, Pereira F (2019) A deep framework for facial emotion recognition using light field images. 1–7
43. Shi J, Zhu S (2021) Learning to amend facial expression representation via de-albino and affinity. arXiv preprint arXiv:2103.10189
44. Shima Y, Omori Y (2018) Image augmentation for classifying facial expression images by using deep neural network pretrained with object image database. pp 140–146
45. Suk M, Prabhakaran B (2014) Real-time mobile facial expression recognition system-a case study. 132–137
46. Szegedy C, Liu W, Jia Y, Sermanet P, Reed SE, Anguelov D, ErhanDumitru, Vanhoucke V, Rabinovich A (2014) Going deeper with convolutions. CoRR, arXiv:abs/1409.4842
47. Turan C, Lam K-M (2018) Histogram-based local descriptors for facial expression recognition (fer): a comprehensive study. J visual Commun Image Represen 55:331–341
48. Videla LS, Ashok PM, Kumar. (2020) Facial expression classification using vanilla convolution neural network. 1–5
49. Vudrerova P, Neruda R (2018). Deep networks with rfb layers to prevent adversarial examples. Springer, pp 257–266
50. Wang K, Peng X, Yang J, Shijian L, Qiao Y (2020) Suppressing uncertainties for large-scale facial expression recognition. 6897–6906
51. Wang K, Peng X, Yang J, Meng D, Qiao Y (2020) Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transac Image Proc 29: 4057–4069
52. Wei W, Jia Q, Chen G (2016) Real-time facial expression recognition for affective computing based on kinect. 161–165
53. Aron Y, Grauman K (2017) Fine-grained comparisons with attributes. Visual Attributes. Springer, 119–154
54. Zadeh PH, Hosseini R, Sra S (2018) Deep-rbf networks revisited: Robust classification with rejection. arXiv preprint arXiv:1812.03190
55. Zeng J, Shan S, Chen X (2018) Facial expression recognition with inconsistently annotated datasets. 222–237
56. Zhang H, Huang B, Tian G (2020) Facial expression recognition based on deep convolution long short-term memory networks of double-channel weighted mixture. Pattern Recog Lett 131:128–134
57. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns. 1–5
58. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns. 1–5
59. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns. 1–5
60. Zhao X, Liang X Liu L, Li T, Han Y, Vasconcelos N, Yan S (eds) (2018). Deep networks with rbf layers to prevent adversarial examples. Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.