Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis

Journal:	*Transactions on Affective Computing*
Manuscript ID	TAFFCSI-2016-12-0217
Manuscript Type:	SI - Apparent Personality Analysis
Keywords:	Facial emotion recognition, Hierarchical committee, Convolutional neural networks

Supervised Committee of Convolutional Neural Networks in Automated Facial Expression Analysis

Gerard Pons and David Masip

Abstract—Automated emotion recognition from facial images is an unsolved problem in computer vision. Although recent methods achieve close to human accuracy in controlled scenarios, the recognition of emotions *in the wild* remains a challenging problem. Recent advances in Deep learning have supposed a significant breakthrough in many computer vision tasks, including facial expression analysis. Particularly, the use of Deep Convolutional Neural Networks has attained the best results in the recent public challenges. The current state-of-the-art algorithms suggest that the use of ensembles of CNNs can outperform individual CNN classifiers. Two key considerations influence these results: (i) The design of CNN involves the adjustment of parameters that allow diversity and complementarity in the partial classification results, and (ii) the final classification rule that assembles the result of the committee. In this paper we propose to improve the assembling of the committee by introducing supervised learning on the ensemble computation. We train a CNN on the posterior-class probabilities resulting from the individual members allowing to capture non-linear dependencies among committee members, and to learn this combination from data. The validation shows a 5% improvement with respect to previous state-of-the art results based on averaging classifiers, and a 4% to the majority voting rule.

Index Terms—Facial emotion recognition, Hierarchical committee, Convolutional neural networks.

1 Introduction

The human face is a relevant informational cue for emotion perception. After a few seconds, we form an accurate first impression of someone's emotions by just observing their face. The applications of facial expression analysis span from human computer interaction (HCI) [1], student engagement estimation [2], emotionally aware devices [3], to the improvement of expression production in autism disorder patients [4].

The state-of-the-art in automated facial expression analysis shows excellent performance in the controlled scenario, where images are acquired in studio environments. Nevertheless, the categorization of emotions *in the wild*, is still an unsolved problem. Hand crafted features (mainly Gabor Filters, Local Binary Patterns, Principal Component Analysis and Independent Component Analysis) fail to model the large degree of variability present in non-controlled scenes. Facial expression classifiers must generalize the learned features to recognize images from new unseen subjects, which might have a different facial morphology. Besides the strong intra-class variability, facial expression algorithms *in the wild* must also deal with strong local changes in the illumination conditions, out of plane rotations, large variations in pose and point of view, and low resolution imaging.

Recent advances on computer vision and particularly in object recognition suggested that new methods based on Deep Learning could improve facial expression classification tasks. Convolutional Neural Networks (CNNs) have represented a relevant breakthrough, especially since the last improvements on the ImageNet Challenge [5]. Similar improvements have been obtained with the application of deep CNNs to facial expression classification [6]. Currently, these end-to-end methodologies have been improved using ensembles of Deep Learning classifiers. An ensemble consists of a committee of classifiers (usually a set of CNNs) that aggregates the partial results of each classifier to produce a unified response in testing time. Committees of deep CNNs pose two related questions: (i) Which is the best way to generate complementary and diverse classifiers, and (ii) which is the best decision rule that combines the ensemble of classifiers. In this paper we will deal with the second one.

Current methods in the computer vision literature use averaging as a decision rule. Usually, each classifier in the ensemble receives a uniform weight to aggregate all classifiers' results. The ensemble decision can be improved using alternative non uniform hand-crafted weights [7]. In this paper we present a new method to weight each classifier in the committee. Using a predefined setting for training each classifier of the ensemble of neural networks (varying the input data and the parameters governing the filters of the network), we propose to use supervised learning hierarchically on the last fully connected output of the CNN. Particularly, we use a CNN to learn the non-linear relationships among the classifiers in the ensemble that better discriminate the validation data between the basic emotions. Our proposal automatically learns the best possible combination model from data. We hypothesize that a learned CNN can better disentangle the relationships between classifiers' outputs and capture common features among basic emotions. Experimental results show a 5% improvement in accuracy with respect to averaging the classifier's weights, and a 4% with respect to the majority voting rule.

- *G. Pons and D. Masip are with the Department of Computer Science, Universitat Oberta de Catalunya, Spain.*
E-mail: {gponsro, dmasipr}@uoc.edu

2 Related work

Automated emotion recognition algorithms from facial expressions are usually based on the emotions defined in the early works by Ekman and Friesen [8], where they identified six basic emotions that are shared among cultures, namely: *anger, disgust, fear, happiness, sadness and surprise*. Ekman at al. [9] also designed the Facial Action Coding System (FACS), currently one of the most used methodologies in the computer vision field. The FACS system describes 44 Action Units (AUs), and each AU defines the movement of a specific set of muscles that express facial emotions. Du et al. [10] extend this set of basic emotions to the compound emotion categories, i.e. those constructed combining the six basic emotions from the work of Ekman et al. [8].

Emotion classification algorithms from facial expressions can rely on static images [11], [12] or video sequences [13]. Depending on the features used for the classification task, we can identify two general methods: geometric-based methods and appearance-based methods. In the first case, a set of landmarks are placed on fiducial keypoints of the face, and the dynamic structure of these landmarks is used to extract discriminant features. Kotsia and Pitas [14] use 119 landmarks using a grid-tracking algorithm to infer emotions from sequences, and Jeni et al. [15] apply a similar approach using 117 keypoints to recognize emotions through a Procrustes transformation.

Appearance-based emotion recognition methods use descriptors from pixel images to train facial expression classifiers. In [1], Bartlett et al. use Gabor filters to train a combination of SVM and Adaboost classifiers. Buciu et al. [16] compare a similar SVM classfication on Gabor features with ICA projections. Gabor filters have also been used in conjunction with PCA and LDA [17] (under the Nearest Neighbor decision rule). Local Binary Patterns (LBP) [18] are another successful state-of-the-art approximation to appearance-based facial expression recognition. Shan et al. [19], [20] use LBP descriptors with multi-scaled sliding windows to classify emotions, using SVM on top of a feature selection step based on boosting. Feng et al. [21] use linear programming on the LBP features on the JAFFE face database [22].

One the one hand, hybrid algorithms constitute a robust solution to emotion recognition. Zhang et al. [23] reveal that a combination of features extracted using Gabor wavelets and a geometrical descriptor slightly improves the use of appearance-based features. On the other hand, emotions can be expressed as a combination of individual Action Unit detectors, both in terms of structure and appearance. In [14], the authors use the classification of AUs as intermediate features to categorize emotions on the Cohn-Kanade database [24]. Pantic and Rothkrantz [25] propose using an expert system for emotion inferring based on the presence/absence of AUs to perform a weighted prediction of the emotion labels. AU classifiers are used to learn the base of rules for emotion recognition. In [26], Sanchez et al. use AU classification as intermediate features for inferring subtle emotional cues from large streams of video.

Recent surveys on facial expression recognition can be found in [27], [28]. One of the recent trends in facial expression analysis is the use of Deep Learning methods. Meng et al. [29] propose a two stage network to first classify individual frames, and use a Time-Delay Neural Network (TDNN) at a second stage to model the temporal relationships among individual predictions. Kim et al. [30] use a set of Deep Belief Network models for the same task. In [31], Jung et al. propose a deep network that combines temporal appearance features and the temporal geometry of a specific set of facial landmarks. Deep Convolutional Neural Networks (CNN) have also been used for training robust facial expression classifiers [32]. Mollahosseini et al. [6] use a CNN with two convolutional and four inception layers with improved results on seven publicly available datasets. Ding et al. [33] train a two-step model, where they first adjust the CNN weights with regularization constraints, and finally add fully connected layers that learn the classification parameters on top of the pre-trained features from the initial stage.

Recent advances in deep learning suggest that the use of an ensemble of CNNs can yield improved accuracies on image classification tasks [34]. Yu and Zhang train a classifier ensemble of multiple CNNs, and pre-train each model on different samples of the FER Challenge database 2013 [35]. Similarly, Kim et al. [7] train a hierarchical committee of CNNs and combine their decisions using a hand crafted hierarchical decision rule. In this work, we hypothesize that the use of trainable committees of CNN ensembles could be beneficial with respect to pre-fixed committee models. Our proposal defines the combination of CNN classifiers as a classification problem itself, and learns the best combination rule of the basic classifiers from the training data .

3 Convolutional Neural Networks for emotion recognition

As previously commented, the latest findings in the emotion recognition field suggest that the use of a framework composed of a set of CNNs improves the accuracy of the recognition process. Figure 1 shows a representation of this framework. However, in order to achieve the best performance, the set of CNNs must be diverse to improve the recognition capability of the whole framework when being assembled. In this section we explain how this set of CNN was designed and carried out.

3.1 Training the individual CNNs

3.1.1 Baseline architecture

In this work, a total of 72 individual CNNs were trained. All of them using the same baseline architecture. It consists of two convolution and pooling stages, followed by a hidden layer and output fully-connected layer. The first convolutional layer uses 6 kernels of size 3×3, 5×5, or 7×7, depending on the configuration which results in 6 output feature maps. The strides and pads are set in order to maintain the same size as the input of the layer. The first max-pool layer sub-samples the input reducing its size by half. The second convolutional layer uses 12 filters of size 3×3, 5×5, or 7×7, where each of the 12 filters is applied to every feature map resulting from the first layer. As in the first layer, the strides and pads are set in order to maintain the same size as the input of the layer and the max-pooling

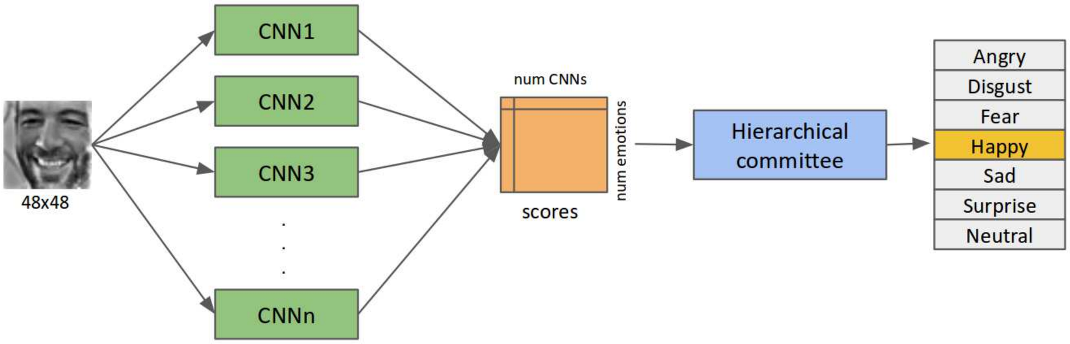


Fig. 1. Overall architecture of the emotion recognition system. A committee composed by 72 CNN classifiers produces a tensor, were the 7 classification scores for each image sample and CNN are hierarchically aggregated into a final set of scores per emotion.

reduces the output size by half. Depending on the configuration, the output is connected to a fully-connected layer of 256, 512 or 1024 neurons. A dropout of 0.5 is included in each layer during the training. The training of the CNN is implemented by dividing the training set into batches and iterating for a fixed number of epochs. In this work the batch size used was 105, using balanced batches including the same number of images from the different emotions which are selected randomly at each batch. Therefore, 15 images of each emotion are selected to be part of a training batch. In order to avoid overfitting, the early-stopping strategy is used. Since the validation examples are considered to be representative of future test examples and never used for gradient descent during the training process, they can be used to define whether the model ceases improving in the learning process. Thus, the validation set is assessed every 100 steps, and if the performance does not increase during 500 steps, the training is finished. The learning rate is set initially at 0.001 and exponentially decays during the training process.

3.1.2 CNNs configurations

To build a diverse set of CNNs, it is essential to improve the performance of the hierarchical committee. When dealing with committees of CNNs the diversity is given by the configuration of the different parameters and the initialization of the networks [36]. The CNNs were trained with different sizes of filters and a different number of neurons in the fully-connected layer. As a result, in this paper we describe the individual CNNs with filters of size 3×3 as *SMALL*, 5×5 as *MEDIUM* and 7×7 as *LARGE* networks. The different fully connected layers are 256, 512 and 1024, notated as *FC256*, *FC512* and *FC1024*, respectively. The initialization of the networks is given by the use of pre-trained models with other databases as initial point for the training of every CNN.

The databases used for this pre-training stage are: the Facial Expression Recognition 2013 database (FER) database [35], the MMI database [37] and the LFW [38]. The FER database consists of 28,709 examples for training, 3,589 for validation and 3,589 for testing. The MMI database consists of over 2,900 videos and high-resolution images of 75 subjects. It is fully annotated for the presence of Action Units (AU) in videos, but partially coded for the six basic emotions. From these 2,900 only 40 videos were annotated with basic emotions. In order to work with static images as in the other databases, the emotion peak frame of every video was extracted, along with the first frame for neutral images. Thus, a total of 80 images from this database were used for our experiments. Finally, the LFW is a face-recognition-aimed database which contains more than 13,000 images of faces from 1,680 subjects collected from the web. To be used in this experiment, the seven subjects with the highest number of images were selected. As a result, 1,288 images were used for this experiment. Figure 2 shows a sample of images from the different databases used in this work. In addition, combinations of these databases were used to obtain more pre-trained models. For instance, a model trained with the MMI database is used to initialize the training of a model trained with the database FER, which will then be used as a pre-trained model for the final training with the database SFEW2.0. In this paper we describe this configuration as *MMI+FER*. Thus, the pre-trained configurations are the following: *FER*, *LFW*, *LFW+FER*, *LFW+MMI*, *LFW+MMI+FER*, *MMI*, *MMI+FER*, *No pre-train*. As a result of all possible configurations we trained a set of 72 different CNNs.

4 Hierarchical committee

The main advantage of using different CNNs for solving a complex problem such as emotion recognition is the fact that some of these may complement one another. When training a large number of CNNs, some will be better than others at recognizing certain emotions. Therefore, it is important to combine their outputs in order to obtain the most accurate final decision.

The most used rules in committees to decide the importance of each CNN decision are the majority voting rule and the average rule. In the majority voting rule the class labels of the predictions obtained by each member of the committee are used to determine the class with the highest number of votes. Instead of using the labels, the average rule uses the class-related scores yielded from each classifier. Thus, the class with the highest average of scores obtained from the CNNs is selected as final output.

The most straightforward way to decide the importance of each CNNs decision is to assign a weight to each individual CNN defining how trustworthy their result is.

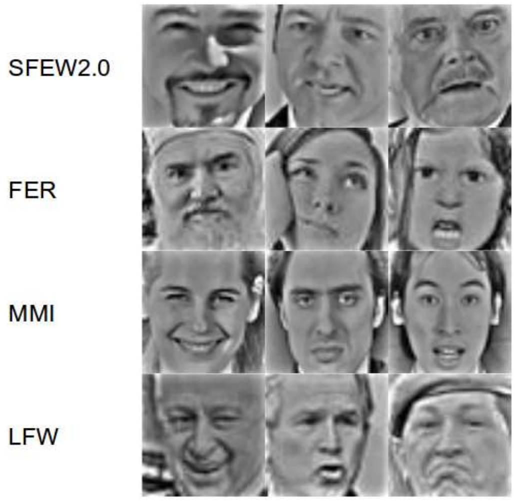


Fig. 2. Sample of images from the different databases used in this work. Images are resized and pre-processed with the isotropic diffusion based normalization.

The average rule is the basic case of this, where the same weight is assigned to each member of the committee. This motivated us to consider more complex methods for this process in order to increase the importance of the best performing CNNs. In this work, a Convolutional Neural Network was trained to decide the best class.

In comparison to non-linear strategies, such as SVM or simple Neural Networks in which only the data from one emotion can be used (leading to train one classifier per emotion), CNN can process the information of all the emotions simultaneously by using input tensors. Hence, the CNN can deal with input data of the following structure: $Nimatges \times NCNN \times Nemotions$, a 3-dimensional array with the scores of each individual CNN for all the images and emotions. This particular feature led us to the hypothesis that CNN may improve the decision making in the emotion recognition task.

4.1 Training the CNN for hierarchical committee

The architecture of the CNN for hierarchical committee was designed following the same baseline as for the emotion recognition problem. Thus, it consists of two convolution and max-pooling layers, followed by a hidden layer and output fully-connected layer. The first convolutional layer uses 6 kernels of size 3×3. The second convolutional layer uses 12 filters of size 3×3. In both convolutional layers, the strides and pads are set in order to maintain the same size as the input of the layer, and the max-pooling reduces the output size by the half. Finally the output is connected to a 100-neuron-fully-connected layer, without dropout. The batch size used was 42, with balanced batches including the same number of images from the different emotions which are selected randomly at each batch. The learning rate is set initially at 0.01 and exponentially decays during the training process.

The dataset for training and evaluation of the hierarchical committee is composed of the posterior class probabilities yielded from the 72 CNNs during the training stage. Thus, the input of the committee CNN is a 3-dimensional array of size $Nimatges \times NCNN \times Nlabels$. During the training stage, the scores obtained by the 858 images were used during the training process of the 72 CNNs, resulting as an array of $858 \times 72 \times 7$. Along with the scores for the training dataset, the ones for validation and test images were also stored for every CNN, resulting as inputs of size $100 \times 72 \times 7$ and $436 \times 72 \times 7$ for the committee, respectively.

5 Experiments

This work was assessed with the SFEW2.0 database [39], released for a sub-competition in the 3rd Emotion Recognition In the Wild 2015 (EmotiW2015) challenge. The database was created by extracting frames from emotional film clips to obtain images in close-to-real world conditions, which are labeled with 7 expressions (angry, disgust, fear, happy, sad, surprise, and neutral). The database consists of 958 images for training, 436 for validation, and 372 for testing. Since the labels of the test dataset are not provided, the validation set was used as test set for the experiments in this work, while 100 random images from the training dataset were selected to become part of our validation dataset. The database also provides the datasets after being processed with a face alignment algorithm. In this work, aligned images were used, which were resized to 48×48 pixels and pre-processed with the isotropic diffusion based normalization [40] in order to reduce the effect of illumination variation in images (Figure 2 depicts an example of several normalized images).

All the experiments in this work were developed using TensorFlow [41] on NVIDIA GeForce GTX Titan GPUs.

5.1 Results of individual CNNs

The classification rates for the 72 deep CNNs tested with the validation data from SFEW2.0 are reported in Table 1. Our best individual model with the highest accuracy of 37.8% was the configuration *LARGE_FC256_LFW+MMI+FER*. Regarding the different configurations used in this set of experiments, some trends can be highlighted.

The pre-trained models used to initialize the CNNs are the most influential variables in the different configurations. The pre-training with the best result only used the *FER* database, given the large number of images for emotion recognition. Networks initialized with *FER* models generalize easily given the large diversity of images present in this database. Without any kind of pre-training, the CNNs obtained on average better results than being initialized with the *MMI* database, indicating that using non-adequate data as initialization may lead to a worse training than without. *MMI* is composed of a reduced number of images, which after the experiments can be considered insufficient to generalize and initialize the CNNs correctly. Similarly, *LFW* obtained slightly better results than without pre-training but it can not be considered a significant increase in the CNNs performance. Although the large number of images in the *LFW* database, its aim is to be used for person recognition problems which may explain the poor performance when initializing an emotion recognition CNN.

Regarding the size of the convolutional filters, *SMALL* CNNs obtained the worst results showing that 3×3 filters

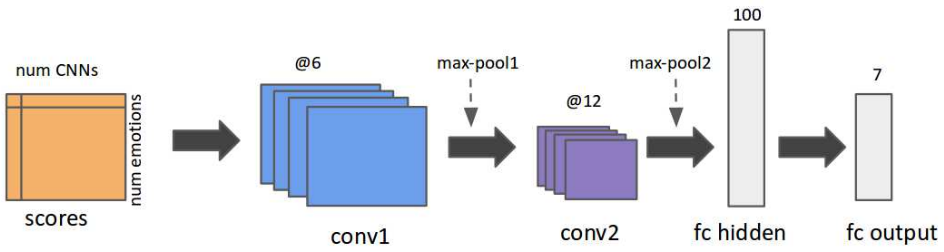


Fig. 3. Architecture of the hierarchical committee of CNNs.

TABLE 1
Test results of individual CNNs. Accuracy in % of the 72 evaluated networks.

Pre-training	SMALL			MEDIUM			LARGE			Mean(Std)
	FC256	FC512	FC1024	FC256	FC512	FC1024	FC256	FC512	FC1024	
FER	34.8	30.0	32.5	28.2	34.5	34.8	36.5	36.0	35.8	33.6(2.87)
LFW	30.5	32.2	32.8	29.8	31.0	28.5	33.2	29.2	29.0	30.6(1.72)
MMI	29.8	30.0	29.2	31.0	28.8	32.0	28.0	30.8	28.5	29.8(1.30)
MMI+FER	30.8	27.8	30.8	28.0	28.8	31.5	31.8	32.0	28.2	29.9(1.74)
LFW+FER	34.5	31.8	30.5	33.8	33.5	34.0	32.5	28.2	33.5	32.5(2.02)
LFW+MMI	30.2	33.2	26.8	31.0	28.5	32.8	31.8	28.0	29.0	30.3(2.18)
LFW+MMI+FER	30.0	30.8	31.2	32.8	34.0	36.0	**37.8**	35.8	33.5	33.5(2.63)
No pre-training	30.5	29.0	29.5	32.2	30.2	29.2	31.8	26.2	33.5	30.2(2.13)
FC Mean(Std)	30.6(1.76)	31.3(2.03)	30.4(1.93)	30.8(2.09)	31.1(2.5)	32.3(2.61)	32.9(3.04)	30.7(3.62)	31.5(2.88)	
Size Mean(Std)		30.8(1.88)			31.5(2.39)			31.7(2.88)		

can not capture sufficient local information of the image to recognize emotions. On the other hand, both *MEDIUM* and *LARGE* filters performed similarly achieving an average accuracy of 31.5% and 31.7% respectively. The size of the fully-connected neural network does not affect the result of the CNN, since the average values for the different configurations of neurons are similar. Particularly, 31.43%, 31.03% and 31.40%, for 256, 512 and 1024, respectively.

5.2 Results of the hierarchical committee

Table 2 shows the comparison between performance of our proposal with the other approaches. Specifically, the proposed approach results are compared to the performance of the following strategies: majority vote, average score rule, VAexpoWA, SVM and NN.

The average score rule and majority vote rule were selected as they are two of the most used strategies in committees, and were implemented as described in Section 4. One of the most important strategies in hierarchical committees of CNNs proposed by the winners of the EmotiW2015 challenge, is VAexpoWA [42]. This strategy uses the validation accuracy and scores to compute a weight for each member of the committee. To increase the difference of weight between good and bad classifiers, an exponentially-weighted average is performed, where the value of the exponential is scanned empirically over [-50:0.1:150].

In order to compare the CNN committee with other non-linear weighting strategies, a SVM and NN were implemented. In both strategies, one classifier per emotion was trained after which the class with the highest score was selected as a final decision. The SVM strategy was implemented using the Scikit-learn Toolbox [43] using a batch of size 40 and a parameter C of 100. Regarding the NN, it was designed as a simple one-layer network using also a batch of size 40, 100 epochs and learning rate of 0.001 which is exponentially decayed along the epochs.

The proposed committee obtained a 39.3% accuracy for the test dataset. This strategy outperformed the rest of the approaches, including the best individual CNN by almost 2%. This lead us to deduce that the non-linearity of CNN as a committee correctly results in a better decision of the contribution of each individual CNN when deciding the label for a given image. The accuracy for each emotion is depicted in the confusion matrix shown in Figure 4, where the good results obtained with the Neutral and Happy class are remarkable. On the other hand, the Disgust and Fear emotion were never classified correctly. However, this is consistent with results previously reported for some of the top-10 participants in the EmotiW 2015 challenge [44], [45]. This performance may be caused by the nature of the class, being inherently more challenging to classify, or by the small number of examples available for the class.

During these experiments we found that the disappointing performance of both SVM and NN are due to the nature of the problem we address in this work. In both cases, one classifier per emotion was trained, receiving an array of size $Nimatges \times NCNN$ as input. Using the scores from the training images, the classifier only dealt with positive examples with a high score from all the CNNs. Therefore, when testing the classifier with a new example, the scores obtained from the test stage in the CNNs were not as high and homogeneous, leading to incorrectly classify most of the positive examples. In order to solve this issue, both SVM

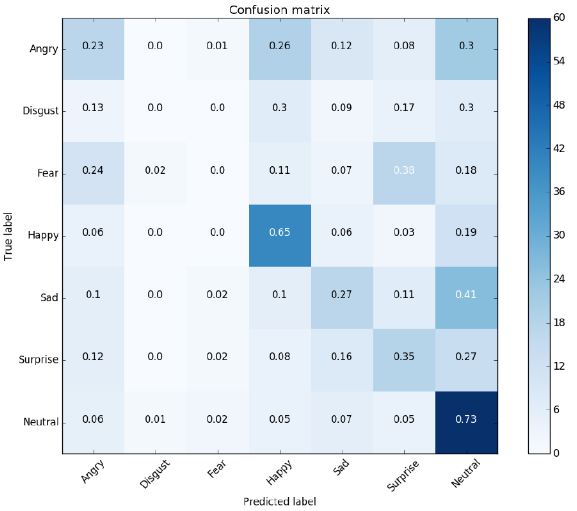


Fig. 4. Confusion matrix for the test dataset from SFEW database after the hierarchical committee decision.

and NN were trained with scores obtained from validation images. Although in this case the classifiers training and testing data were equivalent, the reduced number of validation images did not allow a proper training of the classifiers.

TABLE 2
Test results of hierarchical committee CNN compared with other committee strategies.

Method	Accuracy
Best individual CNN	37.8%
Majority vote	35.0%
Average	34.2%
VAexpoWA	37.5%
SVM	22.2%
Neural Network	23.0%
Proposed CNN	**39.3%**

6 Conclusion

In this paper we presented a framework based on a supervised hierarchical committee of deep CNNs for emotion recognition. Specifically, we proposed the use of a deep CNN to ensemble the outputs of the individual members of the committee. First we trained a set of 72 CNNs with diverse configurations including several parameters as well as different initialisations from pre-trained models using different databases. With the posterior-class probabilities of these individual networks, we formed a hierarchical committee based on a deep CNN.

Our proposed approach was assessed on the SFEW2.0 database for the EmotiW 2015 sub-challenge, and its performance was compared to the most used committee strategies, majority voting and average rule, the exponentially-weighted average VAexpoWA, and two non-linear methods, SVM and NN, obtaining the best results in terms of accuracy.

Future work will consist of studying and developing a framework capable of internally integrating the final classification CNN with the CNNs of the committee, training both steps in a end-to-end manner. Thus, both stages, which are now performed separately, will benefit from providing feedback to each other during the training process.

Acknowledgments

This research was supported by TIN2015-66951-C2-2-R grant from the Spanish Ministry of Economy and Competitiveness and NVIDIA Hardware grant program.

References

[1] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction." in *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, vol. 5. IEEE, 2003, pp. 53–53.

[2] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[3] M. Soleymani and M. Pantic, "Emotionally aware tv," in *Proceedings of TVUX-2013: Workshop on Exploring and Enhancing the User Experience for TV at ACM CHI*, 2013.

[4] J. Cockburn, M. Bartlett, J. Tanaka, J. Movellan, M. Pierce, and R. Schultz, "Smilemaze: A tutoring system in real-time facial expression perception and production in children with autism spectrum disorder," in *ECAG 2008 Workshop Facial and Bodily Expressions for Control and Adaptation of Games*. Citeseer, 2008, p. 3.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, p. 2012.

[6] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–10.

[7] B.-K. Kim, J. Roh, S.-Y. Dong, and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces*, pp. 1–17, 2016.

[8] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[9] ——, "Facial action coding system: A technique for the measurement of facial movement. palo alto," *CA: Consulting Psychologists Press. Ellsworth, PC, & Smith, CA (1988). From appraisal to emotion: Differences among unpleasant feelings. Motivation and Emotion*, vol. 12, pp. 271–302, 1978.

[10] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[11] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[12] M. Pantic and L. J. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 3, pp. 1449–1461, 2004.

[13] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[14] I. Kotsia and I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *Image Processing, IEEE Transactions on*, vol. 16, no. 1, pp. 172–187, 2007.

[15] L. A. Jeni, D. Takacs, and A. Lorincz, "High quality facial expression recognition in video streams using shape related information only," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2168–2174.

[16] I. Buciu, I. Pitas *et al.*, "Ica and gabor representation for facial expression recognition," in *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on*, vol. 2. IEEE, 2003, pp. II–855.

[17] H.-B. Deng, L.-W. Jin, L.-X. Zhen, and J.-C. Huang, “A new facial expression recognition method based on local gabor filter bank and pca plus lda,” *International Journal of Information Technology*, vol. 11, no. 11, pp. 86–96, 2005.

[18] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[19] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[20] ——, “Robust facial expression recognition using local binary patterns,” in *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, vol. 2. IEEE, 2005, pp. II–370.

[21] X. Feng, M. Pietikainen, and A. Hadid, “Facial expression recognition with local binary patterns and linear programming,” *Pattern Recognition And Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, vol. 15, no. 2, p. 546, 2005.

[22] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, “Coding facial expressions with gabor wavelets,” in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 200–205.

[23] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu, “Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron,” in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 454–459.

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

[25] M. Pantic and L. J. Rothkrantz, “Expert system for automatic analysis of facial expressions,” *Image and Vision Computing*, vol. 18, no. 11, pp. 881–905, 2000.

[26] D. Sanchez-Mendoza, D. Masip, and A. Lapedriza, “Emotion recognition from mid-level features,” *Pattern Recognition Letters*, vol. 67, pp. 66–74, 2015.

[27] E. Sariyanidi, H. Gunes, and A. Cavallaro, “Automatic analysis of facial affect: A survey of registration, representation, and recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[28] C. A. Corneanu, M. Oliu, J. F. Cohn, and S. Escalera, “Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 99, p. 1, 2016.

[29] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, “Time-delay neural network for continuous emotional dimension prediction from facial expression sequences,” *IEEE transactions on cybernetics*, vol. 46, no. 4, pp. 916–929, 2016.

[30] Y. Kim, H. Lee, and E. M. Provost, “Deep learning for robust feature generation in audiovisual emotion recognition,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 3687–3691.

[31] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, “Joint fine-tuning in deep neural networks for facial expression recognition,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2983–2991.

[32] V. Mayya, R. M. Pai, and M. M. Pai, “Automatic facial expression recognition using dcnn,” *Procedia Computer Science*, vol. 93, pp. 453–461, 2016.

[33] H. Ding, S. K. Zhou, and R. Chellappa, “Facenet2expnet: Regularizing a deep face recognition net for expression recognition,” *arXiv preprint arXiv:1609.06591*, 2016.

[34] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, “Convolutional neural network committees for handwritten character classification,” in *Proceedings of the 2011 International Conference on Document Analysis and Recognition*, ser. ICDAR ’11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 1135–1139. [Online]. Available: http://dx.doi.org/10.1109/ICDAR.2011.229

[35] “Facial Expression Recognition 2013 database,” https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data, 2013.

[36] D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3642–3649.

[37] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat, “Web-based database for facial expression analysis,” in *Proceedings of IEEE Int’l Conf. Multimedia and Expo (ICME’05)*, Amsterdam, The Netherlands, July 2005, pp. 317–321.

[38] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[39] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, “Video and image based emotion recognition challenges in the wild: Emotiw 2015,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 423–426.

[40] R. Gross and V. Brajovic, “An image preprocessing algorithm for illumination invariant face recognition,” in *International Conference on Audio-and Video-Based Biometric Person Authentication*. Springer, 2003, pp. 10–18.

[41] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” 2015, software available from tensorflow.org. [Online]. Available: http://tensorflow.org/

[42] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, “Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 427–434.

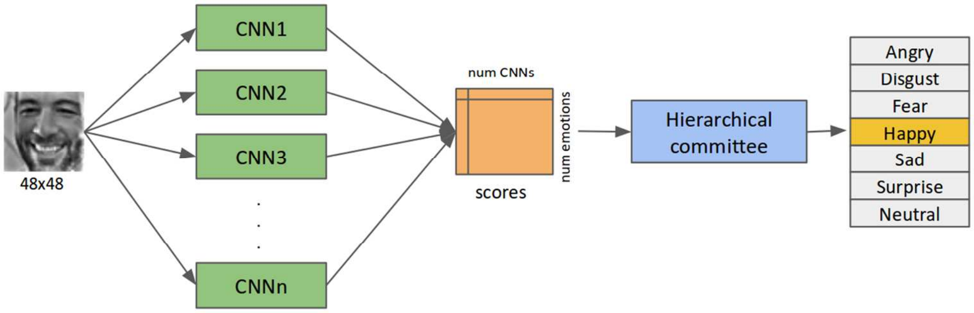
[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[44] W. Li, F. Abtahi, and Z. Zhu, “A deep feature based multi-kernel learning approach for video emotion recognition,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 483–490.

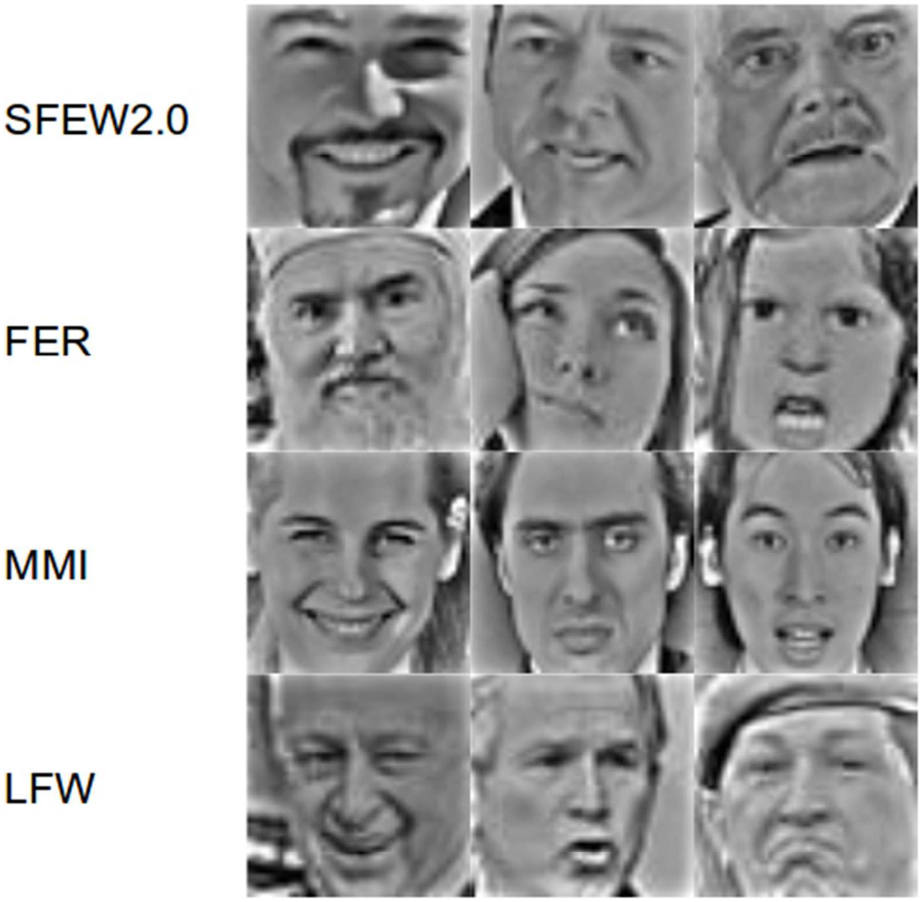
[45] G. Levi and T. Hassner, “Emotion recognition in the wild via convolutional neural networks and mapped binary patterns,” in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 503–510.

Gerard Pons is a Ressearch Assistant at Universitat Oberta de Catalunya (Spain). He graduated in Computer Science Engineering, and received the Master in Industrial, Automation, and Computer Systems degree, both at the University of Girona (Spain) in 2009 and 2010, respectively. His master thesis was done in collaboration with the Institute of Biomedical Engineering of the University of Oxford. He obtained the PhD from the University of Girona in 2014. Both Master and PhD thesis were conducted in breast imaging, particularly in detection and segmentation of breast tumors in ultrasonography. Collaborating in research since 2007, his main research interests include computer vision, image processing, artificial intelligence, pattern recognition, and machine learning.

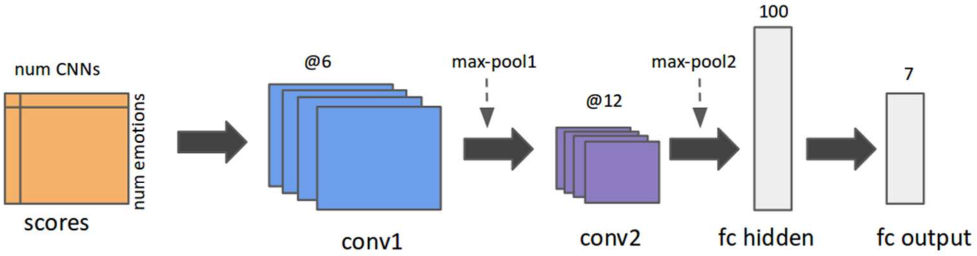
David Masip is associate professor in the Computer Science, Multimedia and Telecomunications Department, Universitat Oberta de Catalunya (Spain) since February 2007, and since 2015 is the director of the UOC Doctoral School. He is the director of the SUNAI (Scene Understanding and Artificial Intelligence) research group. He graduated in Computer Science in the Universitat Autonoma de Barcelona in 2001. He obtained a PhD degree in the Computer Vision Center (Spain) in September 2005. He obtained the best Thesis Award on Computer Science in the Universitat Autonoma de Barcelona. Previously, he worked as assistant professor in the Applied Mathematics Department, Universitat de Barcelona. His main research interests are computer vision, deep learning algorithms, and facial expression classification.



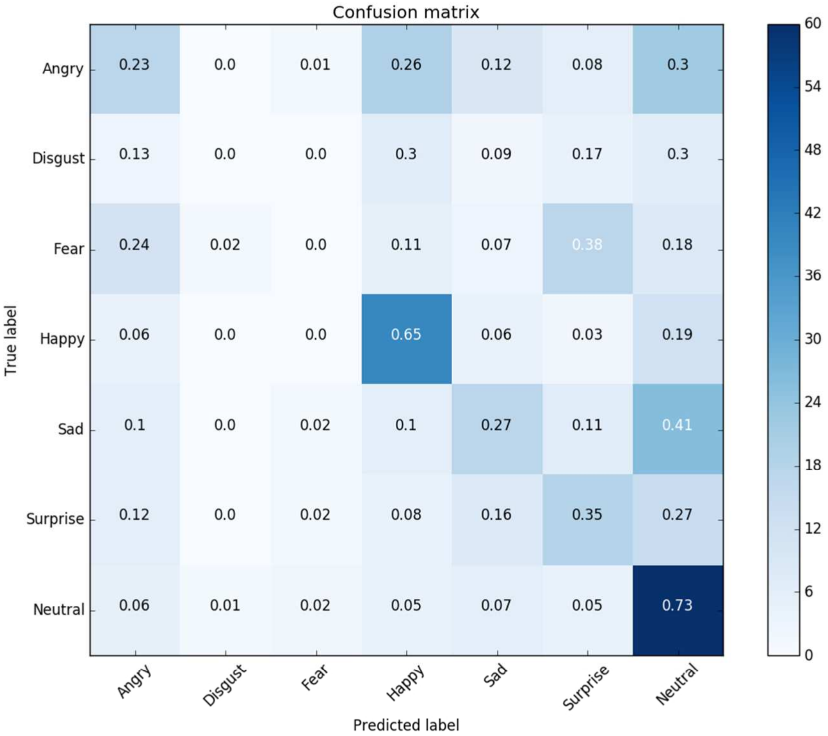
Overall architecture of the emotion recognition system. A committee composed by 72 CNN classifiers produces a tensor, were the 7 classification scores for each image sample and CNN are hierarchically aggregated into a final set of scores per emotion.



Sample of images from the different databases used in this work. Images are resized and pre-processed with the isotropic diffusion based normalization.



Architecture of the hierarchical committee of CNNs.



Confusion matrix for the test dataset from SFEW database after the hierarchical committee decision.