The Visual Computer
https://doi.org/10.1007/s00371-018-1585-8

ORIGINAL ARTICLE

Fast facial expression recognition using local binary features and shallow neural networks

Ivan Gogić[1] · Martina Manhart[1] · Igor S. Pandžić[1] · Jörgen Ahlberg[2]

Abstract
Facial expression recognition applications demand accurate and fast algorithms that can run in real time on platforms with limited computational resources. We propose an algorithm that bridges the gap between precise but slow methods and fast but less precise methods. The algorithm combines gentle boost decision trees and neural networks. The gentle boost decision trees are trained to extract highly discriminative feature vectors (local binary features) for each basic facial expression around distinct facial landmark points. These sparse binary features are concatenated and used to jointly optimize facial expression recognition through a shallow neural network architecture. The joint optimization improves the recognition rates of difficult expressions such as fear and sadness. Furthermore, extensive experiments in both within- and cross-database scenarios have been conducted on relevant benchmark data sets for facial expression recognition: CK+, MMI, JAFFE, and SFEW 2.0. The proposed method (LBF-NN) compares favorably with state-of-the-art algorithms while achieving an order of magnitude improvement in execution time.

Keywords Facial expression recognition · Neural networks · Decision tree ensembles · Local binary features

1 Introduction

Facial expression recognition (FER) is one of the basic challenges in the field of affective computing with potential applications in entertainment, marketing research, retail, psychology, and other fields. It has been widely expected that affect-sensitive applications may change the way we interact with computers [56], yet it still remains a challenge to build such systems. Facial expression recognition is an especially important part of these systems since a large part of human interactions are conveyed non-verbally [38]. Therefore, extensive efforts have recently been invested by the research community to produce methods that can robustly extract expressions from images or videos.

However, numerous challenges still lay ahead primarily due to the complex nature of the problem at hand in the form of large cultural and personal variations as well as variations in imaging conditions (face pose, lighting, occlusions etc.). With the proliferation of mobile and other low-powered "smart" devices within the Internet of Things (IoT) framework, the computing efficiency of computer vision algorithms becomes an increasingly important parameter along with standard accuracy measurements. Therefore, an accurate yet highly efficient algorithm is needed.

Traditional FER systems consist of three steps: face detection, feature extraction, and classification. However, with recent advances in deep learning algorithms, end-to-end convolutional neural networks have become prevalent in many computer vision fields. Their distinct, competitive advantage is the joint optimization of both feature extraction (through convolution filters' weights) and classification (through fully connected layers' weights). The largest obstacle, however, is the need for extremely large data sets in order to prevent over-fitting of deep networks. To continue, FER data sets are especially hard to collect due to the ethical issues of eliciting negative emotions (fear, anger, sadness) and the difficulty to act and annotate accompanying expressions. Recently, transfer learning [34,42] has been successfully used to solve such problems [9,30,54]. Nevertheless, an algorithm that can learn to extract custom task-specific features from a limited number of samples per expression would be beneficial.

✉ Ivan Gogić
ivan.gogic@fer.hr

[1] Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

[2] Department of Electrical Engineering, Computer Vision Laboratory, Linköping University, Linköping, Sweden

Published online: 22 August 2018

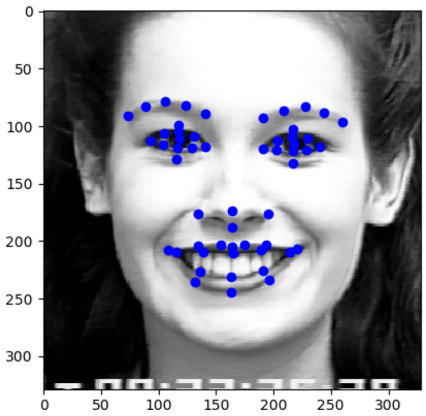

Fig. 1 The detected landmark points used for LBF extraction regions

As recently demonstrated, appearance features extracted around facial landmarks (i.e., mouth, eyes, and nose) greatly contribute to the classification accuracy [20,61]. It is, therefore, important to accurately locate the facial landmarks, a process commonly referred to as face alignment [28,46,62] (Fig. 1). Given the positions of important facial regions, extracting features from local patches can help reduce the extremely large pool of possible features and focus the algorithm on discriminative regions of the face.

With the above-mentioned concerns in mind, we present a fast facial expression recognition algorithm which uses simple and efficient pixel difference features (PDF) coupled with ensembles of decision trees [2] to train and extract highly discriminative shape-indexed local binary features. The extracted features represent relevant patterns for each expression which are used together with a shallow neural network to model their nonlinear interactions. The overview of the whole system is depicted in Fig. 2.

The main contributions of our work are as follows:

- We propose expression-specific feature extraction training framework using pixel difference features and ensembles of decision trees to produce highly discriminative sparse local binary features.
- We jointly optimize expression classification using a shallow neural network in order to model dependencies between classes.
- We demonstrate state-of-the-art recognition rates on the most widely used CK+ data set and competitive accuracy on other data sets (JAFFE, MMI, SFEW 2.0). State-of-the-art generalization ability is also demonstrated.
- An order of magnitude improvement in execution time (1 ms) while running on a single CPU core.

The rest of the paper is organized as follows. Section 2 introduces a research background on the topic of facial expression recognition. Our proposed method is described in Sect. 3. Section 4 presents the experimental validation of the contributions on benchmark data sets. Finally, conclusions are drawn in Sect. 5.

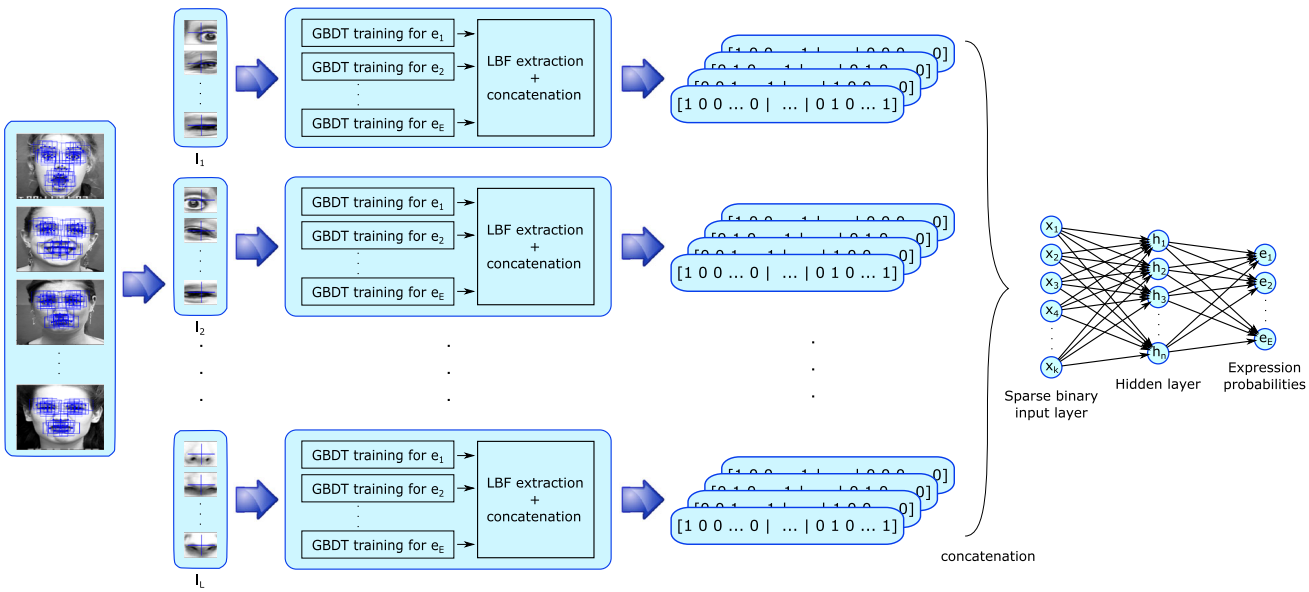

Fig. 2 The proposed method takes an image of a face with detected landmark points. Local patches are used to train the gentle boosted decision trees for each expression in a one-vs-all manner. The tree ensembles are encoded into local binary features which are concatenated into a single sparse binary feature vector. The sparse feature vector is used as an input into a simple 2-layer neural network which outputs the expression probabilities

2 Related work

In order to automatically recognize emotions and their related expressions, an investigation on how to define those terms needed to be done first. In [11], Ekman and Friesen discovered six basic or prototypic emotions (anger, disgust, fear, happiness, sadness, and surprise) whose facial expressions are culturally and racially invariant and are, therefore, great candidates for automatic systems which need clear categories. However, one important drawback of this model became evident. It is too crude to accurately model the complexity of emotions people experience in everyday lives. As a response, facial action coding system (FACS) [10] was developed in order to define atomic facial muscle movements named action units (AU) spanning the whole spectrum of human facial expressions. Its aim is objectivity in the signal measurement which is separated from the final expression classification often influenced by the context. Consequentially, a group of researchers [18,22–24,49] tried to develop algorithms that recognize these simpler, intermediate categories and synthesize the final expression afterward. However, FACS annotation is a very tedious process which requires expert knowledge few people possess. Therefore, few data sets with full FACS annotations are available to the community. In this paper, we opted for the six basic expressions classification approach as it is currently the most widely used categorization in computer vision community.

As mentioned in Introduction, FER is traditionally divided into three steps: face detection, feature extraction, and classification. In most papers, face detection is not discussed in detail since the face location and size are assumed as a priori knowledge. The greatest emphasis is put on the feature selection and extraction which is often considered to be the critical part of the system while standard machine learning techniques are mostly used for the classification step. The used features can roughly be divided into appearance and geometric-based. The appearance features are extracted from facial image intensities to represent a discriminative textural pattern while the geometric ones need accurate landmark positions from which different relations can be constructed. The geometric features are, however, very sensitive to the individual face shape configuration and are therefore less consistent in person independent scenarios. It is important to note that these two types of features have recently been shown to be complementary [52]; hence, hybrid systems similar to the one we propose are gaining popularity.

An additional direction of research is to integrate temporal dimension into both appearance and geometric features when working with image sequences [19,21,32,48,60]. However, this paper focuses on single static image recognition since it is a natural first step that can be extended in future work.

2.1 Hand-crafted features

Well-known and widely successful hand-crafted features such as variations of Local Binary Patterns (LBP) [12,16,17,20,21,25,50,55,59–61] and Histogram of Oriented Gradients (HoG) [6,12,16,59], Gabor filters [17,31,41,51,55,58] and Local Phase Quantization (LPQ) [6,12] descriptors have also been considered for FER. While most approaches considered a regular grid of patches [16,17,21,47,50,55,59,60,63] or the whole face region [6,31,41] for feature extraction, there have been advances in determining common and specific salient facial regions for each expression. In [20], Happy and Routray demonstrated the importance of facial landmark detection in order to find the salient patches from which they extract features. Through the use of one-vs-one SVM classifier for each patch and each expression pair, they were able to find the most discriminative patches for each expression. A similar idea was adopted in [61]; however, a regular grid of patches was used without landmark detection which resulted in lower accuracy than in [20]. In [25], Khan et al. performed a psycho-visual experiment to track the participant's gaze and determine which regions of the face are salient for specific expression. Rivera et al. designed a novel descriptor called Local Directional Number Pattern to differentiate between bright and dark transitions which occur often in faces [47].

2.2 Feature fusion

On the other hand, a number of researchers [6,12,55,59] tried to fuse different texture encoding features in order to extract complementary information that would benefit the FER. For instance, Zhang et al. used multiple kernel learning to combine two different feature representations: HoG and LBP [59]. A different approach to feature fusion was taken in [55] where a pool of SVM classifiers was trained using either Gabor filters or LBPs as features. A genetic algorithm was then used to find the optimal ensemble of classifiers in terms of both size and accuracy. The fusion idea was tested with geometric features as well [44,51]. Wan et al. used constrained local model (CLM) to detect the facial landmarks and used their positions normalized to the mean shape as geometrical features which they concatenated to Gabor features as input to Robust Metric Learning [51]. The method was developed to recognize spontaneous expressions.

2.3 Deep learning

While all of the previously mentioned methods use hand-crafted and heuristically determined features, experiments with deep learning using convolutional neural networks (CNN) [27] on the FER problem were recently conducted as well [4,9,26,30,33,35,39,40,45,54,57]. As already mentioned in Introduction, deep learning methods have serious

over-fitting problems with small datasets that are typical for the FER. Several different approaches have recently been examined in order to cope with the mentioned problem: artificial data augmentation, data set merging, and transfer learning. For a more in-depth review of FER methods using CNN, we refer the reader to a recent survey by Pramerdorfer and Kampel [45]. Additionally, they demonstrate that modern architectural changes in deep networks reduce the over-fitting problem on a moderately large FER 2013 data set (35k images) [15].

Kim et al. used a combination of both aligned and non-aligned faces to train their ensemble of deep CNNs (DCNNs) making the method more robust to face registration problems on faces in the wild [26]. Levi et al. also used an ensemble of 20 DCNNs each having a differently preprocessed input [30]. They designed a novel transformation of image intensities to 3D spaces called mapped LBP in order to reduce the illumination variation in the training set. The mapped LBP transformations with different parameters were used as one of the inputs in the ensemble along with ordinary RGB intensities. Lopes et al. tried standard preprocessing techniques (image normalizations, synthetic samples etc.) and were able to achieve state-of-the-art results on the CK+ benchmark dataset [35]. In [39], the authors combined seven different data sets in order to have enough samples for each expression to train on, making it hard to compare to other methods which restricted their training samples to those available in the individual benchmark data sets.

Finally, transfer learning has recently emerged as the most effective approach to small data set sizes [34,42]. Ng et al. used a general object recognition pre-trained DCNN model and fine-tuned it in two stages. In the first stage, they used the large FER 2013 data set and finally the SFEW 2.0 training set. However, both Levi et al. and Zhai et al. achieved better results by using a model pre-trained on a related face recognition task with extremely large data sets (millions of images) [30,57]. State-of-the-art results on the SFEW 2.0 data set were achieved by Yu et al. using an ensemble of DCNNs, data augmentation (random affine transformations) and pre-training on the larger FER 2013 data set. An interesting approach to transfer learning was presented in [9]. The authors trained a DCNN for FER using a face recognition model's convolutional weights as regularization. Next, they appended fully connected layers and fine-tuned the network for a specific data set. Since they used a single DCNN, the authors were able to achieve an impressive run-time speed (3 ms); however, they require a high-end GPU (TitanX) which is not viable for mobile and embedded platforms.

Even though deep learning methods achieve good results, problems with over-fitting and slow run-time still remain, confirming the need for an effective and fast FER method.

3 Proposed method

The aim of the proposed method is to identify six prototype facial expressions (anger, disgust, fear, happiness, sadness, and surprise) [11] from a single static 2D image. The method uses appearance-based features due to greater robustness to face shape variations when compared to geometric-based ones [20].

As already mentioned, appearance features are extracted around facial landmarks (i.e., mouth and eyes, depicted in Fig. 1); therefore, the first step is to detect the face and its landmarks. Fortunately, face alignment has recently reached a mature state, especially in controlled laboratory conditions [28,46,62]. We exploit this fact and use a recently proposed fast method which served as an inspiration for our work [46]. It is a cascaded regression method that introduced the trainable local binary features (LBF).

3.1 Local feature learning

The key concept of this paper is the task-specific learning process for feature extraction which encodes highly discriminative texture patterns for each facial expression around the detected facial landmarks (Fig. 1). Ensembles of gentle boost decision trees [14] are trained with pixel difference features indexed to facial landmarks in order to maximize the one-vs-all posteriori probability for each expression e around each landmark l. The number of trees within an ensemble and tree depth is specified in advance.

Let E and L denote the number of basic facial expressions and landmark points, respectively. For each facial expression $e \in \{e_1, \ldots, e_E\}$, we train an ensemble of gentle boost decision trees around each landmark point $l \in \{l_1, \ldots, l_L\}$ as can be seen on the left side of Fig. 2. Let C represent the sample patches of an expression e and landmark l at the decision tree node n. Each candidate split $\theta = (p_1, p_2, t_n)$ from a random pool of generated parameters, divides the training samples in the following way:

$$C_{\text{left}}(\theta) = I(p_1) - I(p_2) \leq t_n \tag{1}$$

$$C_{\text{right}}(\theta) = C \backslash C_{\text{left}}(\theta) \tag{2}$$

where p_1 and p_2 represent the local patch positions, t_n represents the threshold, and I represents the image intensities. The positions are placed relative to corresponding landmark location as depicted on the left part of Fig. 3.

The cost function Q that is minimized consists of a Gini impurity measure:

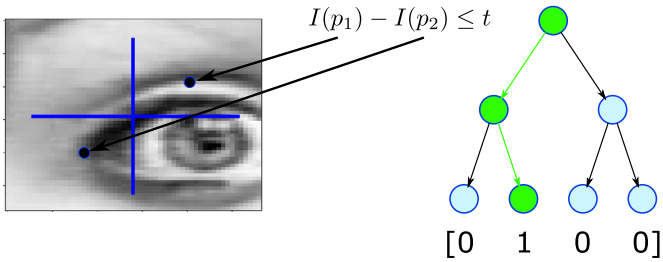

Fig. 3 The decision trees use shape-indexed Pixel Difference Features to split the training set. When encoding a sample into a local binary feature vector, a binary 1 is placed at the index of the vector corresponding to the leaf node where the sample ended up after traversing the tree

$$G(X_n) = p_n(1 - p_n) \tag{3}$$

where p_n represents the proportion of expression e observations at node n:

$$p_n = \frac{1}{N_n} \sum_{x_i \in R_n} I(y_i = e) \tag{4}$$

R_n and N_n represent the sample space and number of samples at node n, respectively. y_i and x_i represent the current ground truth label (one-vs-all binary label) and sample patch, respectively. The full cost function is a weighted sum of impurity measures for both data partitions:

$$Q(C, \theta) = \frac{n_{\text{left}}}{N_n} G(C_{\text{left}}(\theta)) + \frac{n_{\text{right}}}{N_n} G(C_{\text{right}}(\theta)) \tag{5}$$

The described decision trees are organized into ensembles with gentle boosting algorithm [14] in place. The algorithm ensures more emphasis is put on misclassified samples from the previous tree in the ensemble. In practice, each sample i has a weight w_i assigned to it which is increased or decreased depending on the output of the previous tree o_i:

$$w_i := w_i e^{-(y_i o_i)} \tag{6}$$

By doing this, each successive tree in the ensemble is forced to find even more discriminative features compared to the previous trees.

Once gentle boost ensembles for each facial expression and each landmark point are trained, local binary features are extracted as depicted in Fig. 3. Each tree of an ensemble yields in a tree vector of size equal to the number of the leaves in that tree. All elements in that tree vector are equal to 0 except the one that corresponds to the leaf in which the given sample ended up while traversing that tree. This element is equal to 1. The tree vectors are concatenated into an ensemble vector with respect to the order of the trees.

Each facial expression e gets ensemble vectors $\phi_{\mathbf{e},\mathbf{l}}$ where $l \in \{l_1, \ldots, l_L\}$.

These ensemble vectors are concatenated to acquire a global binary feature vector $\mathbf{\Phi_e}$ for each sample (Fig. 2). It represents relevant pattern information for each expression:

$$\mathbf{\Phi_e} = [\phi_{\mathbf{e},\mathbf{1}}, \ldots, \phi_{\mathbf{e},\mathbf{L}}]. \tag{7}$$

3.2 Expression classification

Feature vectors $\mathbf{\Phi_e}$ for each expression e are concatenated into a single feature vector $\mathbf{\Phi}$ which is used as an appearance-based representation of the face specifically tuned for expression differentiation in a completely automatic supervised manner:

$$\mathbf{\Phi} = [\mathbf{\Phi_1}, \ldots, \mathbf{\Phi_E}] \tag{8}$$

A shallow neural network with one hidden layer is used on the described sparse binary feature vector $\mathbf{\Phi}$. This simple network architecture (Fig. 4) has demonstrated enough capacity to model the nonlinear relationship between different expressions as shown in Sect. 4.1. The network is trained using a cross-entropy criterion which is minimized over the data set:

$$\mathbf{\Theta^N} = \arg\min_{\theta} \left(-\sum_{e=1}^{E} \log P(e) \right) \tag{9}$$

where $P(e)$ represents the probability of each expression e obtained by appending a soft-max layer at the end of the network:

$$P(e) = \frac{e^{x_e}}{\sum_{k=1}^{E} e^{x_k}} \tag{10}$$

The optimized network parameters $\mathbf{\Theta^N}$ are obtained using a quasi-Newton method for optimization called limited memory BFGS which approximates the Hessian matrix inverse

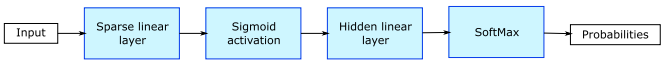

Fig. 4 The diagram of the simple neural network architecture used to predict the expression probabilities

when searching for the optimal descent direction [5]. Since all of the data sets are quite small, the whole training set is used in each iteration of the optimization. In order to improve the convergence speed, Wolfe conditions were used to modify the step length of the descent direction at each iteration [53].

4 Experiments

We evaluated our system on the four most commonly used data sets for FER: CK+ [36], MMI [43], JAFFE [37], and SFEW 2.0 [8]. Due to the small size of the data sets, all of the experiments (except SFEW 2.0 which has a defined protocol) were conducted using a 10-fold cross-validation procedure which randomly divides the data sets into 10 training and validation subsets. By doing this, every sample has both been in the training and validation set in one of the folds. The results were averaged across folds.

Furthermore, our experiments were strictly divided into person-independent (PI) and person-dependent (PD) scenarios. The PI scenario assures a strict subject division between the training and validation sets, meaning the same person cannot appear in both sets with different expressions. Naturally, the PI scenario is more complex; however, many researchers do not explicitly state their experimental procedure which makes comparisons difficult. Both six and seven class results are reported since all of the data sets include a neutral expression also.

Face detection and alignment were first applied to all samples in the data sets. Since shape-indexed local features were used, no face registration and image transformations were needed as a preprocessing step. The only operation applied to the images was a conversion to gray-scale format since only pixel intensities are relevant and sampled by the decision trees.

4.1 Experiments on CK+

The Extended Cohn-Kanade (CK+) [36] data set is a widely recognized benchmark data set for FER. It contains 593 sequences from 123 subjects posing six prototypical expressions and contempt, additionally. All sequences start with a neutral expression and end with the peak of the requested expression. The peak frames are fully FACS annotated. Unlike other data sets, each expression label was verified using the FACS manual by certified FACS coders. Using the requested labels as the ground truth proved to be unreliable by the authors, thus they added an additional validation step. After the validation, 327 of 593 sequences were determined to be of sufficient quality. Due to the comprehensiveness of the data set, we used it for the bulk of our experiments for parameter and architecture investigation.

According to the usual practice in static image FER, one neutral and three peak frames were used from each validated sequence. This amounts to the following number of samples per expression: 135 (An), 177 (Di), 75 (Fe), 207 (Ha), 84 (Sa), 249 (Su), 327 (Ne).

4.1.1 Decision tree parameters analysis

We explored the decision tree parameters (tree depth TD—and tree count in the ensembles TC) using the PI scenario on the 7-class problem from the described CK+ data set. A simple logistic regression with one-vs-all objective was used to train separate expression classifiers to set a baseline. Furthermore, the analysis using a simple logistic regression was suitable to narrow down the decision tree parameter space before analyzing the neural network architecture. The dimensionality of the final feature vector is calculated as follows:

$$D = 2^{\mathrm{TD}} * \mathrm{TC} * L * E \tag{11}$$

The tree parameters TC and TD directly affect the feature vector size, and since the dimensionality is quite high, regularization was needed to prevent over-fitting.

It is evident from Fig. 5 that TD = 2 gives the overall best results regardless of the number of trees in the ensemble. Given the large dimensionality of the feature vector and the relatively small size of the data set, it comes as no surprise that such simple trees are enough to capture relevant textural information. It is also clear from the graph that there is little or no added value in increasing the number of trees in the ensemble beyond 30. The best accuracy was achieved with TD = 2 and TC = 35, averaging 93.77%. We shall call this method LBF-LR.

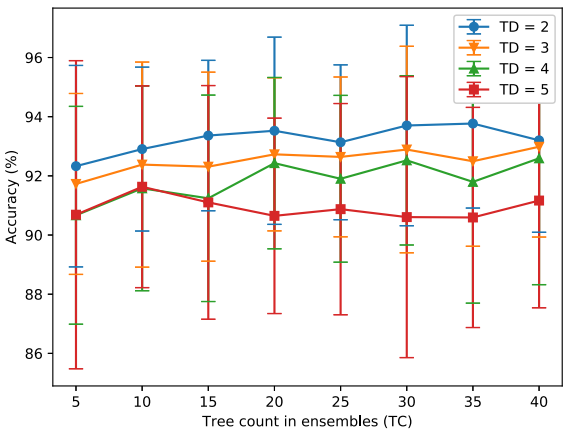

Fig. 5 The accuracies and corresponding standard deviations plotted with error bars for different tree count TC and tree depth TD parameters trained with one-vs-all logistic regression on the PI scenario with 7 classes from the CK+ data set

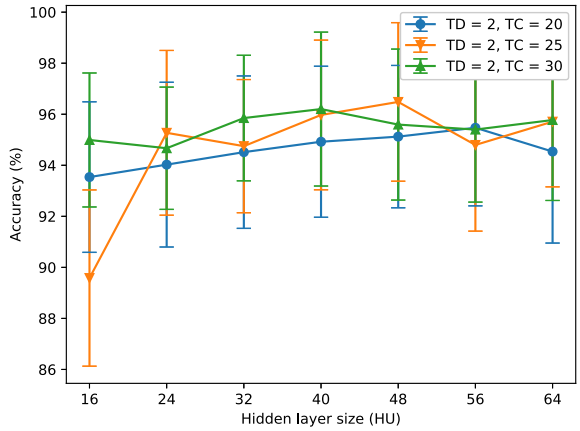

Fig. 6 The accuracies and corresponding standard deviations plotted with error bars for different hidden layer sizes with selected decision tree configurations trained with the described neural network on the PI scenario with 7 classes from the CK+ data set

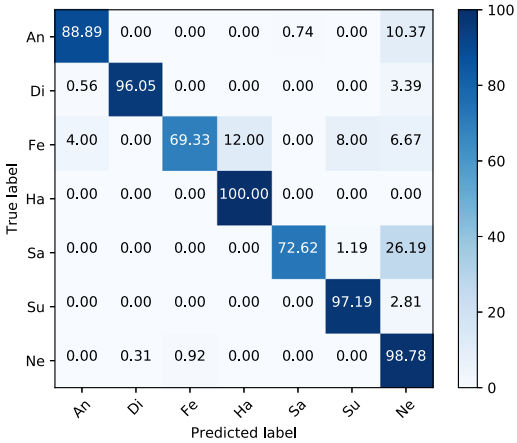

Fig. 7 The confusion matrix for LBF-LR obtained on the CK+ data set using 7 classes and the PI scenario

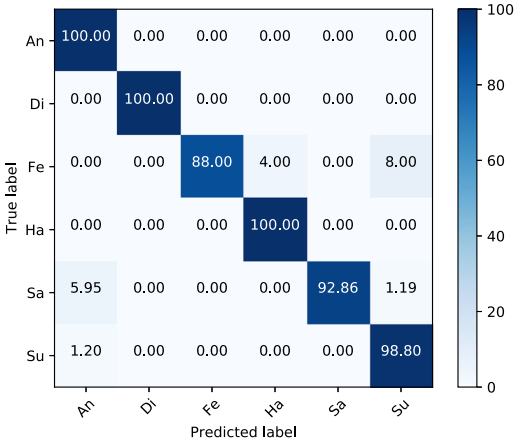

Fig. 8 The confusion matrix for the proposed LBF-NN method on the CK+ data set using 6 classes with the PI scenario

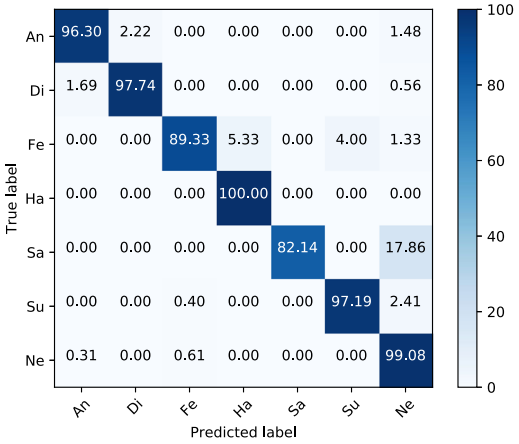

Fig. 9 The confusion matrix for the proposed LBF-NN method on the CK+ data set using 7 classes with the PI scenario

4.1.2 Neural network parameters analysis

As already described, our neural network has one hidden layer whose size needed to be determined experimentally. We used the same scenario as in the previous section. We varied the size of the hidden layer HU while keeping the decision tree parameters fixed to three configurations with the same three depth TD = 2: TC = 20, TC = 25 and TC = 30.

The results can be seen in Fig. 6 where the optimal configuration is visible for parameters TD = 2, TC = 25 and HU = 48. When compared with the separate optimization using logistic regression from Sect. 4.1.1, there is a boost in accuracy from 93.77 to 96.48% which demonstrates the need for joint optimization to recognize facial expressions. We shall call this method LBF-NN.

Upon closer examination of the confusion matrices for both LBF-LR and LBF-NN shown in Figs. 7 and 9, we can see that the most important boosts in accuracy are obvious for the most difficult expressions: fear and sadness. Incidentally, these two expressions have the least amount of samples in the data set due to the difficulty of truthfully portraying these emotions. Having a joint nonlinear optimization process, features from other expressions can prove complementary and helpful to increase the recognition rate for these difficult expressions. The recognition rate increase for fear is 20%, while for sadness is 9.52%.

4.1.3 Comparison

It is quite difficult to compare our results to previous work since there is no official protocol described for the CK+ data set. We conducted experiments on both 6 and 7 class (including neutral expression) problems with PD and PI scenario using the best configuration described in Sect. 4.1.2. The confusion matrices for the PI scenario are shown in Figs. 8 and 9. It is clear that the PD scenario is an easier task produc-

Table 1 Comparison with previous work on the CK+ data set

Method	No. of folds	No. of subjects	Scenario	No. of classes	Recognition rate (%)
Boughrara et al. [1]	10	97	PI	6	96.66
Gritti et al. [16]	10	95	Not stated	7	92.90
Gu et al. [17]	10	94	PI	7	91.51
Happy and Routray [20]	10	118	Not stated	6	94.09
Khan et al. [25]	10	Not stated	PI	6	96.70
Lee et al. [29]	118	118	PI	7 (contempt)	90.47
Zhong et al. [61]	10	96	Not stated	6	89.89
Littlewort et al. [31]	90	90	PI	7	93.30
Lopes et al. [35]	8	100	PI	6	96.76
			PI	7	95.75
Zhang et al. [59]	10	109	PI	6	95.50
			PI	7	93.60
Poursaberi et al. [44]	10	not stated	PI	6	86.10
			PD	6	90.37
Zhang and Tjondronegro [58]	10	92	PI	6	94.48
Liu et al. [33]	8	118	PI	6	96.70
Shan et al. [50]	10	96	PI	6	95.10
			PI	7	91.40
Mollahosseini et al. [39]	5	Not stated	PI	6	93.20
Zavaschi et al. [55]	10	not stated	PI	7	88.90
			PD	7	99.40
Rivera et al. [47]	10	118	PI	7 (contempt)	89.30
Burkert et al. [4]	10	210	PD	7 (contempt)	99.60
Ding et al. [9]	10	Not stated	PI	6	**98.60**
Proposed LBF-NN	10	118	PI	6	98.08
			PI	7	**96.48**
			PD	6	**99.89**
			PD	7	**99.68**

The bold results are the best (state-of-the-art) results for each experiment scenario

ing accuracies of 99.89% and 99.68% when compared to the PI scenario with accuracies of 98.08% and 96.48% for 6 and 7 class problems, respectively. It is, therefore, very important to clearly and explicitly state the protocol of the experiments when comparing to other works. Upon closer inspection of the confusion matrices, we can see that by introducing the neutral expression, overall recognition rate drops due to confusion between sadness and neutral expressions.

As we can see from Table 1, most of the previous methods differ in the number of classes, folds, and subjects used in the experiments. However, there is a positive trend of adopting the more difficult PI scenario. Our method is very competitive with other works for all experiment setups and sets a new state-of-the-art recognition rate for the CK+ data set with 96.48% for the 7 class problem. The previous best result was from Lopes et al. [35] where a CNN was used with various preprocessing methods to artificially increase the training set size and prevent over-fitting. The nature of our simpler LBF features makes it easier to train on smaller data sets and proves to be a viable alternative to heavy-weight convolutional features. Similarly, current state-of-the-art method for the 6 class problem uses a trained face recognition network to regularize and prevent over-fitting of the expression DCNN [9].

4.2 Results on MMI

MMI [43] data set contains more than 2900 videos and images of 75 subjects. It is an ongoing work to provide large volumes of data of facial expressions to the research community. Along with 6 basic emotions, it also contains single FACS Action Unit activation samples and naturalistic expressions. All of the videos include the starting neutral expression along with the onset, apex and offset phases. A major problem is that the apex frames are not indexed; therefore, it is hard

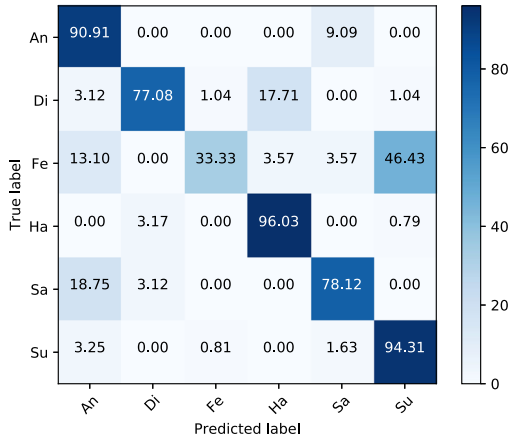

Fig. 10 The confusion matrix for the proposed LBF-NN method on the MMI data set using 6 classes with the PI scenario

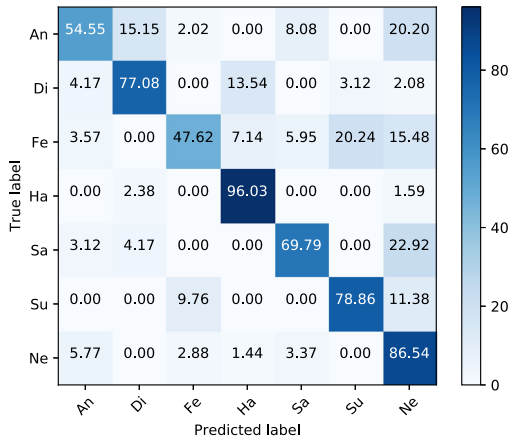

Fig. 11 The confusion matrix for the proposed LBF-NN method on the MMI data set using 7 classes with the PI scenario

to compare since researchers manually choose the frames to include into the training and validation sets.

We filtered the data set to frontal view and 7 basic expressions (including neutral) which resulted in 208 sequences (one sequence was corrupted) and 31 subjects. One neutral frame and three manually selected apex frames were used totaling with the following number of samples per expression: 99 (An), 96 (Di), 84 (Fe), 126 (Ha), 96 (Sa), 123 (Su), 208 (Ne). Again, no preprocessing was applied to the images except for the gray-scale conversion and the face detection/alignment to find the facial landmarks used in our method.

Four experiments were conducted similarly to the CK+ experiments, including 6 and 7 class recognition in both PI and PD scenarios. The confusion matrices for the PI scenario are presented in Figs. 10 and 11. Once again, the PD scenario was easily solved with 99.84% and 99.88% recognition rates for 6 and 7 class problems, respectively. However, PI scenario proved to be much more difficult with recognition rates of

Table 2 Optimal parameters for the PI scenario on the MMI, JAFFE and SFEW 2.0 data sets

Data set	No. of classes	TD	TC	HU	L_2
MMI	6	2	20	16	0
MMI	7	2	25	24	0.0001
JAFFE	6	2	25	48	0
JAFFE	7	2	25	24	0
SFEW 2.0	7	2	30	24	0.0001

78.88% and 73.73% with optimal parameters presented in Table 2. A small L_2 regularization coefficient was used on the 7-class problem in the PI scenario that helped prevent over-fitting.

There are a number of reasons for these results. First of all, the MMI data set is much more challenging than the CK+ data set due to a large age span between subjects (19–62 years) and the fact that many subjects wore accessories like glasses and hats. Secondly, the sequences were not filtered by expert annotators; therefore, there is no guarantee that challenging expressions such as fear and sadness were acted out correctly and consistently across subjects. It is evident from the confusion matrices that it is very difficult to discern, i.e., fear from surprise and sadness from disgust. Thirdly, the results are very dependent on the peak frames used in the data set which needed to be manually selected since the sequences are of varying length and different expression dynamics.

We compared ourselves with previous work in Table 3. Again, comparison on this data set is even harder since the data acquisition is an ongoing process. Also, as can be seen from Table 3, there is a large variation in the number of subjects and sequences used for training and testing. Some of the authors manually discarded sequences with poorly acted expressions. The method from Zhang et al. [59] uses an almost identical set in their experiments and achieve state-of-the-art recognition rate. However, they use hand-crafted features (fusion of LBPH and HOG) coupled with a multi-kernel SVM. Due to the hand-crafted features making their model less complex, it is also less prone to over-fitting on small data sets. Another important point to note is that they fine-tuned their hyper-parameters on each fold in the cross-validation tests making the models highly specialized for combinations of specific fold training and test sets. Our tests were done with hyper-parameters optimized using the average accuracy across folds, not at the fold level. Furthermore, no cross-database experiments were conducted by the authors to test the generalization ability of their models. Another hand-crafted features method from Rivera et al. [47] achieves the state-of-the-art performance in the PI scenario with 7 classes. The problem compared to this method is that only 168 sequences are available now from the 238 sessions they used.

Table 3 Comparison with previous work on the MMI data set

Method	No. of folds	No. of sequences/subjects	Scenario	No. of classes	Recognition rate (%)
Lee et al. [29]	20	150/21	PD	6	93.81
Zhong et al. [61]	10	205/Not stated	Not stated	6	77.39
Fang et al. [13]	10	203/Not stated	Not stated	6	75.96
Zhang et al. [59]	10	209/Not stated	PI	6	93.60
			PI	7	**92.80**
Poursaberi et al. [44]	10	Not stated	PI	6	86.10
			PD	6	90.37
Shan et al. [50]	10	96/20	PI	7	86.90
Mollahoseini et al. [39]	5	Not stated/not stated	PI	6	77.60
Rivera et al. [47]	10	238/28	PI	6	**95.80**
Burkert et al. [4]	10	187/?	PD	6	98.63
Proposed LBF-NN	10	208/31	PI	6	78.88
			PI	7	73.73
			PD	6	**99.84**
			PD	7	**99.88**

The bold results are the best (state-of-the-art) results for each experiment scenario

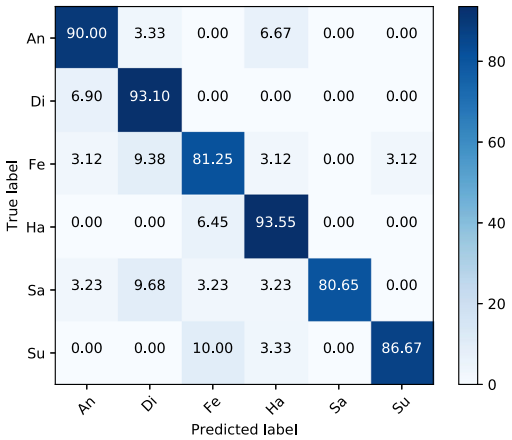

Fig. 12 The confusion matrix for the proposed LBF-NN method on the JAFFE data set using 6 classes with the PI scenario

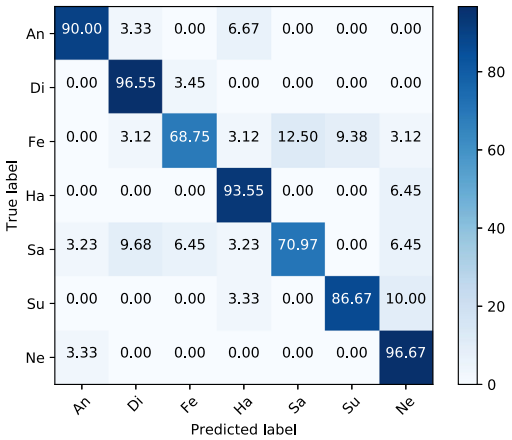

Fig. 13 The confusion matrix for the proposed LBF-NN method on the JAFFE data set using 7 classes with the PI scenario

4.3 Results on JAFFE

The Japanese Female Facial Expression database (JAFFE) [37] contains images of 10 Japanese female models posing 7 basic emotions. The total number of images is 213 which makes it the smallest data set by far on which we tested our method. An additional problem is that the data set obviously lacks diversity with respect to gender, age, and race.

The same experiments were conducted as with the other two data sets, and similarly, the PD scenario recognition rates were extremely high above 98% for both 6 and 7 class problems. However, as can be seen from the confusion matrices in Figs. 12 and 13, in the PI scenario our method struggled again to discern difficult and similar expressions such as fear and sadness. This can again be explained by the difficulty of sincerely portraying such emotions on demand. Nevertheless, in the easier 6 class task, our method achieves recognition rates above 80% for each expression.

Table 4 compares our method to previous work on this data set. We achieve the state-of-the-art results in the PD scenario due to the high flexibility of our method to adapt its feature extraction process. In the PI scenario, we achieve competitive recognition rates of 87.22% and 85.88% for the 6 and 7 class problems, respectively.

4.4 Results on SFEW 2.0

The Static Facial Expressions in the Wild (SFEW) [7] data set aims to benchmark the performance of FER methods in realistic conditions with unconstrained lighting, head poses,

Table 4 Comparison with the previous work on the JAFFE data set

Method	No. of folds	No. of images	Scenario	No. of classes	Recognition rate (%)
Gu et al. [17]	10	213	PI	7	89.67
Happy and Routray [20]	10	183	Not stated	6	91.80
Lee et al. [13]	20	213	PD	6	94.70
Lopes et al. [35]	10	213	PI	6	53.44
			PI	7	53.57
Poursaberi et al. [44]	10	213	PI	7	91.12
			PD	7	95.04
Zhang and Tjondronegoro [58]	10	213	PI	6	92.93
Liu et al. [33]	10	213	PI	7	**91.80**
Shan et al. [50]	10	213	PI	7	81.00
Owusu et al. [41]	10	213	PD	6	96.83
Zavaschi et al. [55]	10	213	PI	7	70.00
			PD	7	96.20
Rivera et al. [47]	10	213	PI	6	**93.40**
			PI	7	90.60
Proposed LBF-NN	10	213	PI	6	87.22
			PI	7	85.88
			PD	6	**98.33**
			PD	7	**98.10**

The bold results are the best (state-of-the-art) results for each experiment scenario

and occlusions. The second version of the data set that we used in our experiments was released as part of the EmotiW 2015 challenge [8]. The images were extracted and annotated semi-automatically from movies and, even though the emotions are acted, the data set can be considered as a spontaneous one since professional actors were involved.

The data set has a well-defined protocol with a strict division of training (958 images), validation (436 images), and test (372 images) sets. Since we could not obtain the labels for the test set, we report the results on the validation set only. The division of the data set is strictly person independent, and it contains 7 basic expressions with the following number of samples (training and validation sets combined): 255 (An), 89 (Di), 145 (Fe), 271 (Ha), 236 (Ne), 245 (Sa), 153 (Su).

Due to the unconstrained nature of the data set, we needed to modify the preprocessing pipeline to some extent. First, the face detector could not detect all of the faces so we manually annotated 8 images. Next, we used a more powerful face alignment method [3] that was trained on unconstrained head poses and can accurately align profile faces as well. Furthermore, we utilized the 2D landmark positions to remove the in-plane rotations of the faces which reduced the variation of the relevant expression patterns around landmarks. Finally, we used horizontal mirroring to double the size of the training set. Even though this preprocessing step did not improve the results on other data sets, it proved to be beneficial here due to the asymmetry caused by large variations in head pose, illumination, and occlusions.

It is clear from the baseline results of the EmotiW 2015 challenge [8] (35.93% and 39.13% accuracy on validation and test sets, respectively) that this is a very challenging benchmark. The optimal parameters for this data set are shown in Table 2 and the confusion matrix in Fig. 14. It is evident once again that happiness is the easiest expression to recognize even in the unconstrained environment (80.82%); however, neutral and anger achieve respectable recognition rates as well (69.77% and 53.25%, respectively). Disgust and fear are traditionally very difficult to identify.

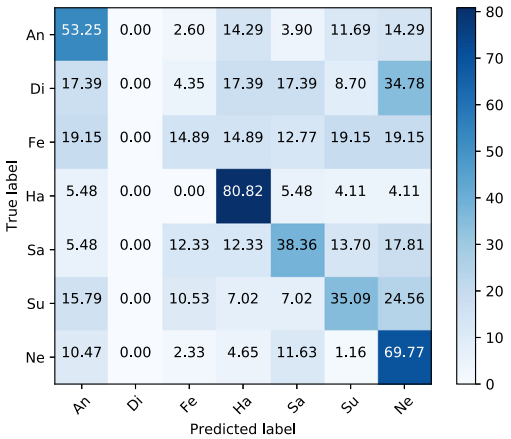

Fig. 14 The confusion matrix for the proposed LBF-NN method on the SFEW 2.0 validation set

Table 5 Comparison with the previous work on the SFEW 2.0 data set

Method	No. of images			Recognition rate (%)		External data
	Train	Val	Test	Val	Test	
Zong et al. [63]	958	436	372	38.07	50.00	Yes
Mollahosseini et al. [39]	332	331	–	47.70	–	Yes
Ng et al. [40]	958	436	372	48.50	55.60	Yes
Zhai et al. [57]	958	436	–	48.51	–	Yes
Levi and Hassner [30]	891	431	372	51.75	54.56	Yes
Ding et al. [9]	891	431	–	55.15	–	Yes
Yu and Zhang [54]	958	436	371	**55.96**	**61.29**	Yes
Ding et al. [9]	891	431	–	48.19	–	No
Proposed LBF-NN	958	436	–	**49.31**	–	No

The bold results are the best (state-of-the-art) results for each experiment scenario

The proposed method achieves an average recognition rate of 49.31% without using any additional training data which is the state-of-the-art result in such conditions. However, the best results are achieved by leveraging transfer learning with large related data sets (usually face recognition sets) and large ensembles of DCNNs [30,54]. As it can be seen from Table 5, all of the deep learning methods need auxiliary data sets and, even then, our method is very competitive. The displayed results demonstrate good robustness of the proposed method to unconstrained conditions. Furthermore, the method has once again shown an excellent ability to learn relevant information from very limited amount of data.

4.5 Cross-database results

In order to test the generalization ability of our method, we conducted cross-database experiments with 7 classes. We trained our method on CK+ and tested on the MMI data set and vice versa. We chose these two databases because they have a similar number of samples and they are at the opposite ends of the difficulty spectrum. The achieved results actually confirm these presumptions. When trained on the consistent and constrained CK+ data set and tested on the more challenging MMI set, we achieve the average recognition rate of 62.74%. When the situation is reversed, an impressive recognition rate of 78.79% is achieved. In fact, both results show a great generalization capacity of the proposed method since results in cross-database experiments are generally much worse than within database experiments.

It is interesting to observe here that the within database results for the MMI data set are *worse* (73.73%) than in the cross-database experiment with CK+ as the test set. This confirms the theory that the MMI data set is not consistently annotated and is quite difficult to train on. In Table 6, we compared our cross-database results with previous work which provided similar experiments. Our method achieves the state-of-the-art result when generalizing from MMI to CK+ data set with an improvement of 14.22% from the previous best result.

Table 6 Comparison of cross-database recognition rates with 7 classes

Method	Train	Test	Recognition rate (%)
Zhang et al. [59]	CK+	MMI	**66.9**
	MMI	CK+	61.2
Shan et al. [50]	CK	MMI	51.1
Lee et al. [29]	MMI	CK+	64.57
Proposed LBF-NN	CK+	MMI	62.74
	MMI	CK+	**78.79**

The bold results are the best (state-of-the-art) results for each experiment scenario

4.6 Computational performance analysis

We tested the recognition run-time of our method on a PC with an Intel Core i7-7500U CPU operating at 2.70 GHz frequency. The method is not parallelized and uses a single CPU core. The average computing time of our method on the JAFFE data set is approximately 1 ms which makes it ideal for mobile and embedded applications. Due to its simple pixel difference features coupled with shallow decision tree ensembles and 2-layer neural network, the online recognition phase is extremely efficient. The first neural network layer weight matrix is the largest one, and the multiplication with the large input feature vector would be the bottleneck of the system; however, due to the sparse binary nature of the feature vector, it can be computed with a simple series of memory lookups and additions. The run-time is written in C++ which contributes to fast execution.

We compared our method to the previous work which stated their execution time in Table 7. It is clear that our method achieves an order of magnitude improvement over all previous works. Ding et al. [9] achieve a real-time performance of 3 ms; however, they use a high-level GPU optimized code which is impractical for mobile and embedded systems.

Table 7 Comparison of computation time in milliseconds

Method	CPU	Feature extraction	Classification	Total
Happy and Routray [20]	Intel i5 3.2 GHz	?	?	295.5
Khan et al. [25]	?	10	?	?
Lee et al. [29]	Pentium 3.50 GHz	110	40	150
Lopes et al. [35]	?	–	–	10
Zhang et al. [59]	Intel i5 2.66 GHz	?	30	?
Zhang and Tjondronegoro [58]	Core Duo 1.66 GHz	?	?	125.8
Liu et al. [33]	6-core 2.4 GHz	?	?	210
Shan et al. [50]	?	30	?	?
Owusu et al. [41]	?	?	?	14.5
Levi et al. [30]	Amazon GPU g2.8xlarge instance	?	?	500
Ding et al. [9]	Titan X GPU	?	?	3
Proposed LBF-NN	Intel i7-7500U 2.70 GHz	?	?	**1**

The bold results are the best (state-of-the-art) results for each experiment scenario

5 Conclusion

We presented a fast facial expression recognition method based on a trainable feature extraction process using ensembles of decision trees producing sparse binary feature vectors (LBF) and a shallow neural network. The 2-layer neural network is capable of modeling the nonlinear relationship between expressions as demonstrated in Sect. 4.1.2 which boosted the recognition rates of especially difficult expressions such as fear and sadness. The method uses static images and achieves state-of-the-art results on the most widely used CK+ database and demonstrates great generalization abilities in the cross-database experiments and robustness on the "in the wild" SFEW 2.0 data set. The great accuracy results are accompanied by an extremely fast computation time of 1 ms on a single CPU which is an order of magnitude improvement in speed compared to recent work. The accuracy and speed of the method make it ideal for FER in environments with limited resources such as embedded and mobile platforms. It is a viable alternative to end-to-end CNNs in scenarios with limited data sets and run-time resources.

A number of factors contributed to the success of the proposed method. Unlike layers of trainable convolutional kernels used in deep learning methods, ensembles of decision trees have demonstrated great generalization ability deduced from small data sets due to their simplistic nature. By limiting the possible feature space to local regions around prominent facial landmarks, their expressive power is further boosted which resulted with highly discriminative and specialized features. Furthermore, joint classification using a shallow neural network utilized inter-class information which contributed to correct classification of ambiguous expressions.

As future work, the method could be extended to incorporate temporal information through the use of increasingly popular variants of recurrent neural networks such as long short-term memory (LSTM) networks. It would be natural since expressions are dynamic in nature and their intensity changes over time. Another course of action would be to integrate occlusion and head pose information to make it more robust on "in the wild" images and videos.

Compliance with ethical standards

Conflict of interest Authors I. Gogić and M. Manhart have received grants from the company Visage Technologies. Authors I. S. Pandžić and J. Ahlberg own stock in and are members of the board of directors of the company Visage Technologies.

References

1. Boughrara, H., Chtourou, M., Amar, C.B., Chen, L.: Facial expression recognition based on a MLP neural network using constructive training algorithm. Multimed. Tools Appl. **75**(2), 709–731 (2016)
2. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and regression trees. CRC Press, Boca Raton (1984)
3. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2D and 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In: International Conference on Computer Vision, vol. 1, p. 4 (2017)
4. Burkert, P., Trier, F., Afzal, M.Z., Dengel, A., Liwicki, M.: Dexpression: Deep convolutional neural network for expression recognition. arXiv preprint arXiv:1509.05371 (2015)
5. Byrd, R.H., Lu, P., Nocedal, J., Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. **16**(5), 1190–1208 (1995)
6. Dhall, A., Asthana, A., Goecke, R., Gedeon, T.: Emotion recognition using PHOG and LPQ features. In: 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops (FG 2011), pp. 878–883. IEEE (2011)
7. Dhall, A., Goecke, R., Lucey, S., Gedeon, T.: Static facial expression analysis in tough conditions: data, evaluation protocol and benchmark. In: 2011 IEEE International Conference on Com-

puter Vision Workshops (ICCV Workshops), pp. 2106–2112. IEEE (2011)
8. Dhall, A., Ramana Murthy, O., Goecke, R., Joshi, J., Gedeon, T.: Video and image based emotion recognition challenges in the wild: Emotiw 2015. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 423–426. ACM (2015)
9. Ding, H., Zhou, S.K., Chellappa, R.: Facenet2expnet: Regularizing a deep face recognition net for expression recognition. In: 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), pp. 118–126. IEEE (2017)
10. Ekman, P., Friesen, W.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists, San Francisco (1978)
11. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
12. Eleftheriadis, S., Rudovic, O., Pantic, M.: Discriminative shared gaussian processes for multiview and view-invariant facial expression recognition. IEEE Trans. Image Process. **24**(1), 189–204 (2015)
13. Fang, H., Mac Parthaláin, N., Aubrey, A.J., Tam, G.K., Borgo, R., Rosin, P.L., Grant, P.W., Marshall, D., Chen, M.: Facial expression recognition in dynamic sequences: an integrated approach. Pattern Recogn. **47**(3), 1271–1281 (2014)
14. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. **28**(2), 337–407 (2000)
15. Goodfellow, I.J., Erhan, D., Carrier, P.L., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., et al.: Challenges in representation learning: a report on three machine learning contests. In: International Conference on Neural Information Processing, pp. 117–124. Springer, Berlin (2013)
16. Gritti, T., Shan, C., Jeanne, V., Braspenning, R.: Local features based facial expression recognition with face registration errors. In: 8th IEEE International Conference on Automatic Face and Gesture Recognition, 2008. FG'08, pp. 1–8. IEEE (2008)
17. Gu, W., Xiang, C., Venkatesh, Y., Huang, D., Lin, H.: Facial expression recognition using radial encoding of local gabor features and classifier synthesis. Pattern Recogn. **45**(1), 80–91 (2012)
18. Gudi, A., Tasli, H.E., den Uyl, T.M., Maroulis, A.: Deep learning based facs action unit occurrence and intensity estimation. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, pp. 1–5. IEEE (2015)
19. Guo, Y., Zhao, G., Pietikäinen, M.: Dynamic facial expression recognition with atlas construction and sparse representation. IEEE Trans. Image Process. **25**(5), 1977–1992 (2016)
20. Happy, S., Routray, A.: Automatic facial expression recognition using features of salient facial patches. IEEE Trans. Affect. Comput. **6**(1), 1–12 (2015)
21. Huang, X., Zhao, G., Pietikäinen, M., Zheng, W.: Robust facial expression recognition using revised canonical correlation. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 1734–1739. IEEE (2014)
22. Jaiswal, S., Martinez, B., Valstar, M.F.: Learning to combine local models for facial action unit detection. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), vol. 6, pp. 1–6. IEEE (2015)
23. Jiang, B., Martinez, B., Valstar, M.F., Pantic, M.: Decision level fusion of domain specific regions for facial action recognition. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 1776–1781. IEEE (2014)
24. Jiang, B., Valstar, M., Martinez, B., Pantic, M.: A dynamic appearance descriptor approach to facial actions temporal modeling. IEEE Trans. Cybern. **44**(2), 161–174 (2014)
25. Khan, R.A., Meyer, A., Konik, H., Bouakaz, S.: Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recogn. Lett. **34**(10), 1159–1168 (2013)
26. Kim, B.K., Dong, S.Y., Roh, J., Kim, G., Lee, S.Y.: Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 48–57 (2016)
27. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
28. Lee, D., Park, H., Yoo, C.D.: Face alignment using cascade Gaussian process regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4204–4212 (2015)
29. Lee, S.H., Plataniotis, K.N.K., Ro, Y.M.: Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. IEEE Trans. Affect. Comput. **5**(3), 340–351 (2014)
30. Levi, G., Hassner, T.: 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, pp. 503–510. ACM (2015)
31. Littlewort, G., Bartlett, M.S., Fasel, I., Susskind, J., Movellan, J.: Dynamics of facial expression extracted automatically from video. Image Vis. Comput. **24**(6), 615–625 (2006)
32. Liu, M., Shan, S., Wang, R., Chen, X.: Learning expression lets on spatio-temporal manifold for dynamic facial expression recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1749–1756 (2014)
33. Liu, P., Han, S., Meng, Z., Tong, Y.: Facial expression recognition via a boosted deep belief network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1805–1812 (2014)
34. Long, M., Cao, Y., Wang, J., Jordan, M.I.: Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791 (2015)
35. Lopes, A.T., de Aguiar, E., De Souza, A.F., Oliveira-Santos, T.: Facial expression recognition with convolutional neural networks: coping with few data and the training sample order. Pattern Recogn. **61**, 610–628 (2017)
36. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (ck+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101. IEEE (2010)
37. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Proceedings of Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200–205. IEEE (1998)
38. Mehrabian, A.: Silent Messages, vol. 8. Wadsworth, Belmont (1971)
39. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–10. IEEE (2016)
40. Ng, H.W., Nguyen, V.D., Vonikakis, V., Winkler, S.: Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 443–449. ACM (2015)
41. Owusu, E., Zhan, Y., Mao, Q.R.: A neural-adaboost based facial expression recognition system. Expert Syst. Appl. **41**(7), 3383–3390 (2014)
42. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
43. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: IEEE International Conference on Multimedia and Expo, 2005. ICME 2005, pp. 5. IEEE (2005)

44. Poursaberi, A., Noubari, H.A., Gavrilova, M., Yanushkevich, S.N.: Gauss-laguerre wavelet textural feature fusion with geometrical information for facial expression identification. EURASIP J. Image Video Process. **2012**(1), 17 (2012)
45. Pramerdorfer, C., Kampel, M.: Facial expression recognition using convolutional neural networks: state of the art. arXiv preprint arXiv:1612.02903 (2016)
46. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment via regressing local binary features. IEEE Trans. Image Process. **25**(3), 1233–1245 (2016)
47. Rivera, A.R., Castillo, J.R., Chae, O.O.: Local directional number pattern for face analysis: face and expression recognition. IEEE Trans. Image Process. **22**(5), 1740–1752 (2013)
48. Rudovic, O., Pavlovic, V., Pantic, M.: Multi-output Laplacian dynamic ordinal regression for facial expression recognition and intensity estimation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2634–2641. IEEE (2012)
49. Sandbach, G., Zafeiriou, S., Pantic, M.: Markov random field structures for facial action unit intensity estimation. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 738–745 (2013)
50. Shan, C., Gong, S., McOwan, P.W.: Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis. Comput. **27**(6), 803–816 (2009)
51. Wan, S., Aggarwal, J.: Spontaneous facial expression recognition: a robust metric learning approach. Pattern Recogn. **47**(5), 1859–1868 (2014)
52. Whitehill, J., Bartlett, M.S., Movellan, J.R.: Automatic facial expression recognition. Soc. Emot. Nat. Artifact **88**, 58 (2013)
53. Wolfe, P.: Convergence conditions for ascent methods. SIAM Rev. **11**(2), 226–235 (1969)
54. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, pp. 435–442. ACM (2015)
55. Zavaschi, T.H., Britto, A.S., Oliveira, L.E., Koerich, A.L.: Fusion of feature sets and classifiers for facial expression recognition. Expert Syst. Appl. **40**(2), 646–655 (2013)
56. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans. Pattern Anal. Mach. Intell. **31**(1), 39–58 (2009)
57. Zhai, Y., Liu, J., Zeng, J., Piuri, V., Scotti, F., Ying, Z., Xu, Y., Gan, J.: Deep convolutional neural network for facial expression recognition. In: International Conference on Image and Graphics, pp. 211–223. Springer, Berlin (2017)
58. Zhang, L., Tjondronegoro, D.: Facial expression recognition using facial movement features. IEEE Trans. Affect. Comput. **2**(4), 219–229 (2011)
59. Zhang, X., Mahoor, M.H., Mavadati, S.M.: Facial expression recognition using $\{l\}_{\{p\}}$-norm MKL multiclass-SVM. Mach. Vis. Appl. **26**(4), 467–483 (2015)
60. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
61. Zhong, L., Liu, Q., Yang, P., Huang, J., Metaxas, D.N.: Learning multiscale active facial patches for expression analysis. IEEE Trans. Cybern. **45**(8), 1499–1510 (2015)
62. Zhu, S., Li, C., Change Loy, C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4998–5006 (2015)
63. Zong, Y., Zheng, W., Huang, X., Yan, K., Yan, J., Zhang, T.: Emotion recognition in the wild via sparse transductive transfer linear discriminant analysis. J. Multimodal User Interfaces **10**(2), 163–172 (2016)

Ivan Gogić is a research associate at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, since 2015. He graduated at the same institution in July 2009 with a thesis in the area of face detection and tracking. Currently, he is in charge of face tracking and analysis research in cooperation with Visage Technologies d.o.o.

Martina Manhart is a research associate at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, since 2015. Between 2007 and 2014, she was a research fellow at the Department of Mathematics, Faculty of Science, University of Zagreb, Croatia, where she had received her B.Sc. and M.Sc. degrees in December 2006 with thesis Accurate singular values and eigenvalues of matrices with acyclic graphs. Her fields of interest are numerical mathematics, especially (indefinite) matrix factorization algorithms and machine learning algorithms in face analysis. Currently, she is in charge of Face Analysis development, coordination between Visage Technologies d.o.o. and HOTLab and coordination of HOTLab students.

Igor S. Pandžić is a Professor at the Department of Telecommunications, Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. He leads the Human-Oriented Technologies Laboratory (HOTLab). He teaches undergraduate and postgraduate courses in the fields of virtual environments and communications. His main research interests are in the field of computer graphics and, more recently, computer vision, with particular interest in face analysis and animation and strong focus on applications of these technologies. Igor also worked on networked collaborative virtual environments, computer-generated film production and parallel computing. He published five books and around 100 papers on these topics.

Jörgen Ahlberg received his M.Sc degree in Computer Science and Engineering in 1996 and his Ph.D. in Electrical Engineering in 2002, both from Linköping University, Sweden. He then held positions as scientist and research leader at FOI, the Swedish Defence Research Agency for nine years. He is currently an Adjunct Senior Lecturer at Linköping University and runs R&D projects at Visage Technologies, Termisk Systemteknik, and Glana Sensors, companies specializing in visual, thermal and hyperspectral computer vision, respectively. Research interests, in the general area of image analysis and vision, include tracking and analysis of facial images as well as automatic detection, recognition, and tracking in thermal and hyperspectral systems. He has published more than 50 scientific papers, of which four are award-winning.