SIViP (2015) 9 (Suppl 1):S245–S253
DOI 10.1007/s11760-015-0810-4

ORIGINAL PAPER

Automated facial expression recognition based on histograms of oriented gradient feature vector differences

Uroš Mlakar[1] · Božidar Potočnik[1]

Received: 8 January 2015 / Revised: 23 June 2015 / Accepted: 28 July 2015 / Published online: 13 August 2015

Abstract This article proposes an efficient automated method for facial expression recognition based on the histogram of oriented gradient (HOG) descriptor. This subject-independent method was designed for recognizing six prototyping emotions. It recognizes emotions by calculating differences on a level of feature descriptors between a neutral expression and a peak expression of an observed person. The parameters for the HOG descriptor were determined by using a genetic algorithm. Support vector machines (SVM) were applied during the recognition phase, whereat one SVM classifier was trained for one emotion. Each classifier was trained using difference vectors obtained by subtraction of HOG feature vectors calculated for the neutral and apex emotion subjects image. The proposed method was tested by using a leave-one-subject-out validation strategy for 106 subjects on 1232 images from the Cohn Kanade, and for 10 subjects on 192 images from the JAFFE database. A mean recognition rate of 95.64 % was obtained using the Cohn Kanade database, which is higher than the recognition rates for almost all other single-image- or video-based methods for facial emotion recognition.

Keywords Facial expression recognition · Histogram of oriented gradients · Support vector machines · Histogram differences · Genetic algorithm

✉ Uroš Mlakar
uros.mlakar@um.si

[1] Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia

1 Introduction

Facial expressions are facial changes that represent a person's emotional states (i.e., human emotion), intentions, or social communication [30]. Researchers focus on recognizing the following prototypical emotional states, namely 'Disgust,' 'Anger,' 'Surprise,' 'Fear,' 'Happiness,' 'Sadness,' and sometimes even 'Neutral' emotions. Generally, a computer system for automatic emotion recognition consists of three modules [31]: a face acquisition module, facial data extraction accompanied by a feature selection module, and a facial expression recognition module.

Generally, emotion recognition methods first describe a face either by geometrically based (e.g., active appearance models—AAM [8]) or appearance-based features (e.g., local binary patterns—LBP [18], local ternary patterns—LTP, Gabor filters [14], binary features [15]). Afterward, the facial expression is recognized by using the constructed feature vectors either indirectly as a collection of facial action units (see FACS system [9]) or directly as one of the prototypical emotions [30], whereat very diverse classifiers were used ranging from k-nearest neighbor (kNN) [34], family of Bayes classifiers [4], support vector machines (SVM), hidden Markov model (HMM) [32], etc., combined by principal component analysis (PCA), independent component analysis (ICA) [2], and linear discriminant analysis (LDA). It should be stressed that recognition approaches may also be classified as frame-based or video sequence-based, depending on whether temporal information is used [30]. The more important methods for facial expression recognition are summarized in Table 7. They are accompanied by their key features, some information about the validation procedure, and recognition accuracies typically calculated on the Cohn Kanade (CK) database.

This article proposes a method for recognizing facial expressions from two images (a neutral image and image with peak facial expression) based on the texture information. The method observes changes or gradient information between these two images. Recently, a histogram of oriented gradient (HOG) descriptor has attracted the attention of the facial expression recognition research community due to its invariance to geometric (except object orientation) and photometric transformations. This texture descriptor has been applied in several methods for human emotion recognition from a single 2D facial image, e.g., in [7,14,23,34]. The highest recognition rate amongst them was obtained in [34] by combining the HOG descriptor and the Weber local descriptor (WLD), whereas kNN was used as classifier. Despite recognition rates of above 95 % (see [7] and [34]), it is believed that algorithms' robustness (and eventually recognition rates) would increase if additional information were to be provided during the recognition process. This is especially apparent by recognizing spontaneous facial expressions that are usually much harder to recognize compared with acted expressions. Additional information could be brought into the recognition process by inspecting two or more facial images (e.g., video) of a particular person expressing emotion.

Several so-called video sequence-based methods, which exploit temporal information, have been reported in the literature. Michel et al. [22] developed a system based on feature displacements between neutral and peak expressions. A set of important facial landmarks are tracked in the input video, which are hand-labeled in the first frame. Valstar et al. [32] tracked a set of 20 fiducial points to model the temporal activations of different AUs in an input video. They reported 95.3 % recognition accuracy for 22 AUs for posed facial expressions (72 % for spontaneous). Fang et al. [12] developed a dynamic framework which observes salient information extracted from successive frames in a video. The best recognition rate at 71.57 % was obtained on a dataset comprised of spontaneous facial expressions with the nearest neighbor classifier based on fuzzy sets. Siddiqi et al. [29] extracted facial movements using optical flow in combination with a stepwise LDA. Using the HMM classifier, they reported a 99.33 % recognition rate, but the experiments were done for ten subjects only.

It can be seen that many video sequence-based methods rely on feature point-tracking mechanisms (or on fitting a facial model, to a specific frame—this is not the focus of this article), which can mean a lot of extra processing time. The motivation for this work is to show that by observing changes (gradients) between two selected frames and with minimal additional processing time with respect to single-image-based methods, state-of-the-art recognition accuracy can be obtained. A computationally more efficient method than video-based methods is proposed in this article, since information is extracted from just two facial images of an observed person, while at the same time preserving all the advantages of texture-single-image-based methods, whereas comparable recognition accuracy as state-of-the-art methods is obtained. This subject-independent method, based on HOG descriptors, recognizes the emotions by calculating differences on a level of feature descriptors between the neutral expression and peak expression (apex) of an observed person. Our method indeed follows the idea of motion-based methods (see Table 7), but here the displacements between feature vectors are measured and not between geometric features or intensity images. This method is designed for recognizing six prototypical emotions, whereas one SVM classifier is trained for one emotion. The key novelties of our approach are in (1) automated emotion recognition based on comparing the HOG descriptor's feature vectors for two facial images, where one contains the subjects neutral expression and one his/her peak expression (in contrast, the method in [22] observes the displacements of geometric features between neutral and peak images), (2) a genetic algorithm for HOG descriptor parameter selection, and (3) the method as a whole.

2 Proposed algorithm

Our proposed method follows the typical three-modular structure of computer systems for emotion recognition.

2.1 Face acquisition module

The Viola-Jones detector [33] is used to roughly locate a face in the input grayscale image (color images should first be transformed into grayscale). Next, a possible in-plane rotation of the face is eliminated. This elimination is based on accurate locations of eyes and nose. Since the Viola-Jones detector proved to be insufficiently accurate, an AAM from [26] was applied instead of the bounding box with the face. Afterward, the coordinates of the eyes and nose obtained from the AAM were used to remove in-plane facial rotation. Finally, the extracted image was scaled to a predefined size of 130 × 150 pixels. Figure 1 depicts this intermediate result on the sample image.

It can be seen from Fig. 1 that the extracted facial image still contains several unwanted features such as hair and ears, which might significantly influence the recognition procedure. That is why the face should be accurately clipped out of the image. Our clipping procedure is based on the fact that the human face resembles an ellipse. The center of the ellipse was calculated during our research by using eyes and nose locations (see Fig. 2a). The longer ellipse axis was defined as the distance from the ellipse center to the bottom of the image. On the other hand, the shorter axis was determined by searching the local minima of the intensity function on the

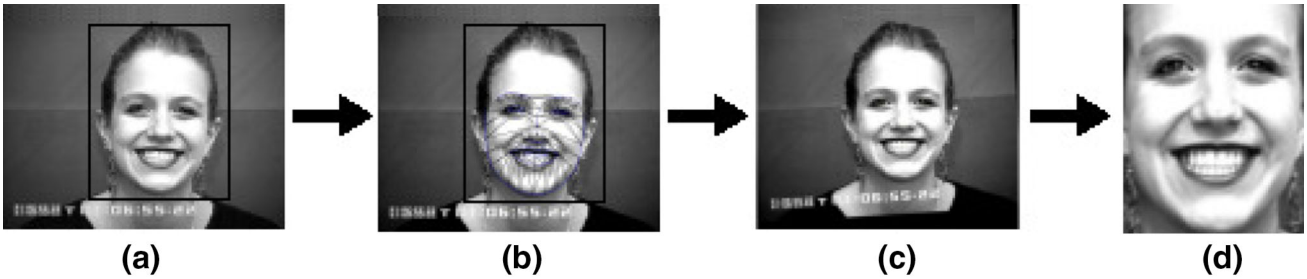

Fig. 1 Face acquisition module: **a** face detection, **b** AAM model fitting, **c** registration, and **d** cropped face

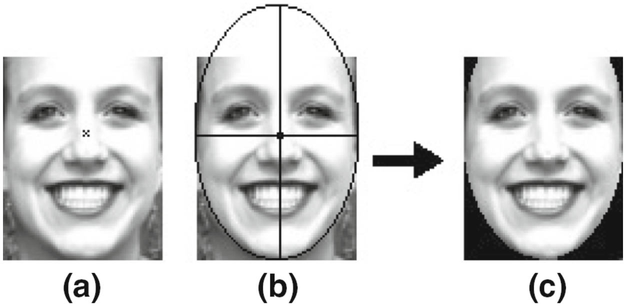

Fig. 2 Face approximation using an ellipse: **a** cropped face, with the marked ellipse center, **b** ellipse and both axes, **c** final approximated facial image

left and right sides of the ellipse center in the y direction. If no such minima was found, the smaller axis was set to half of the image width (see Fig. 2b). Finally, the calculated ellipse was used for accurately extracting the face from the image (see Fig. 2c).

Let us emphasize that our input images were not additionally preprocessed during this research. In the case of the presence of noise in the images, such interference would of course need to be removed.

2.2 Facial regions description using the HOG descriptor

Although the AAM was used for face extraction, no information regarding the geometric features obtained from the AAM was used for building the feature vector for emotion recognition. It should be stressed that the used AAM model had a small number of key points, which means that the model quickly fitted to the image, but on the other hand, it was insufficiently detailed for emotion recognition. Besides, the AAM models are not the topic of this research. Consequently, the extracted facial region is described by the HOG descriptor.

Recently, this descriptor has played an important role within the field of facial expression recognition (see Table 7 and the previous section). The strength of this descriptor is in its invariance to geometric and photometric transformations (except object rotation), which is particularly important for unevenly illuminated scenes. The HOG features are extracted by dividing the image into small regions called 'cells,' then computing a histogram of gradient directions over pixels in the cell. Finally, the facial image is represented by concatenation of these histograms. For improved performance, the local histograms are normalized to local contrast by calculating a measure of the intensity over larger spatial regions called 'blocks,' where this calculated measure is used to normalize all cells within the block. The normalization results are more invariant to changes in illumination and shadowing [14] (for details about HOG, see [6]).

2.3 Facial expression recognition

The feature vectors obtained in the previous module are not fed directly into the classifier, but rather an idea is followed, as described in this section. If the neutral image (i.e., an image in which a subject does not express any emotion) and the so-called emotion image (i.e., the image in which a subject expresses emotion) are compared, then those differences that emerge due to expressing emotion can be easily captured and advantageously used within the recognition process. Two variants were experimented on with, namely (1) comparison on a pixel level and (2) comparison on a feature vector level. In the first variant, the difference between both images was calculated (only the extracted face region was considered) and afterward this difference image was described by the HOG descriptor and fed into the classifier. However, a subtraction of images introduces noise at the border of the ellipse used for face approximation. The HOG descriptor gives greater importance to these phantom-strong gradients, thus leading to poor classification results. This approach was thus rejected.

Therefore, the second variant was developed within a functioning system. Firstly, the difference Δ between the feature vectors calculated for the emotional image (i.e., feature vector E) and the neutral image (i.e., feature vector N) was calculated in this approach as $\Delta = E - N$. Figure 3 demonstrates the concept of difference vector calculation.

These so-called difference vectors were then directly fed into a classifier. SVMs were selected as classifiers during this research due to their huge nonlinear classifying capabilities. Six SVMs were constructed in this research, one for each pro-

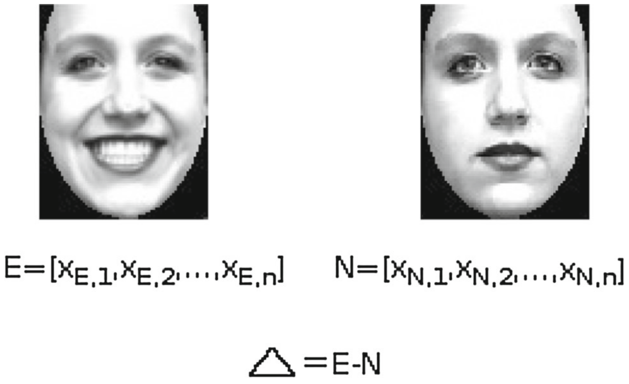
E=[x_{E,1},x_{E,2},...,x_{E,n}]
N=[x_{N,1},x_{N,2},...,x_{N,n}]
△=E-N

Fig. 3 Calculation of difference vector Δ

totyped emotion. During the training phase, SVM was trained by a subset of feature vectors for this particular emotion (i.e., positive examples) and by a subset of feature vectors of all the remaining emotions (i.e., negative examples). During the recognition phase, the subjects' facial images (i.e., neutral and emotion images) were first described by feature vectors; then, the difference vector was calculated, and afterward, this vector was fed into each trained SVM. The SVM returning the highest probability was the winner, and consequently, the subject was marked as expressing such an emotion for which the winning SVM was trained. It should be emphasized that our SVM classifiers do not need to be trained by testing subjects' images. Actually, our approach generalizes to novel subjects very well (i.e., subjects not within the training set), which will be demonstrated in the results section.

2.4 Parameter tuning

Our method has several parameters, where the setting of a HOG appearance descriptor is crucial for better recognition accuracy. The real-value-coded genetic algorithm (GA) [5,35] was used during this research for the selecting of HOG descriptor parameters. At the beginning, a population (i.e., solutions to the problem) is randomly initialized and afterward evolved into final solutions by using genetic operators.

A population of size 20, a one-point crossover, and roulette wheel selection were applied during this research. Within each generation of the GA algorithm, a small subset of subjects (in our case 10 %) was chosen from the validation database. This subset was used to evaluate the current population of HOG descriptor parameters. The evaluation was carried out as follows. Firstly, the subset was divided into learning and testing sets. Afterward, our algorithm was trained on the learning set, followed by testing on the testing set. The obtained evaluation results on the testing set were then used as the fitness function within the GA algorithm. The better individuals found by the GA algorithm determined the HOG descriptor parameters. It should be stressed that each individual defined one setting of the HOG descriptor parameters.

3 Results

Our method being designed for recognizing six prototyping emotions (i.e., 'Disgust,' 'Anger,' 'Surprise,' 'Fear,' 'Happiness,' and 'Sadness') was validated on the CK [16,20], and the japanese female facial expression (JAFFE) [21] database.

The CK database is well established and more often cited within the emotion recognition research field consisting of 593 image sequences from 123 subjects. All sequences are fully FACS-coded, while some are also labeled with the prototypical emotion. Only those sequences where the emotion label was available were selected for our experiments. In this way, 106 subjects were included during our study. Image sequences for each subject could have annotations ranging from one to six prototypical emotions. These image sequences were then sampled as follows. Three peak frames (i.e., frames around the apex) and one neutral image were selected from each sequence. Such sampling resulted in 1232 images being available for the evaluation procedure. Table 1 presents the number of images per emotion used during our evaluation. On the other hand, the JAFFE database consists of 213 images from 10 subjects. All images are labeled with the prototypic emotion (i.e., six prototypic emotions and neutral state). All emotion images and one neutral image from each subject were used during the evaluation procedure.

Our recognition algorithm is based on a set of parameters, the setting of the HOG descriptor's parameters being particularly important. The GA algorithm was used for tuning these parameters, as described in Sect. 2.4. The better ten HOG descriptor parameters' settings, ranked from the best to the worst, are gathered in Table 2 for the CK database, while the best setting of the HOG parameters for the JAFFE database is in Table 3 (other settings and results are omitted due to limited article size). The operation of our recognition algorithm is also significantly affected by SVM machines. The LibSVM toolbox [3] was used to implement SVM machines during this research. A SVM with a radial basis function kernel was selected due to the encouraging results for the facial expression recognition problem (see [28] and [22]).

Table 1 Number of images per emotion from CK and JAFFE databases used in our experiments

Database	Neutral	Disgust	Anger	Surprise	Fear	Happiness	Sadness	Σ
CK	308	177	135	187	75	204	84	1232
JAFFE	10	29	30	29	32	31	31	192

Table 2 The better ten settings of HOG descriptor parameters as found by the GA algorithm for the CK database

No.	Bins	Cell size (px)	Block size	Orientation*	Clip value
1.	11	15	2	1	0.20
2.	11	14	2	1	0.56
3.	9	16	2	1	0.20
4.	8	16	2	1	0.16
5.	9	16	4	1	0.93
6.	9	16	2	1	0.50
7.	8	13	4	1	0.33
8.	8	13	4	1	0.52
9.	16	13	4	1	0.70
10.	9	16	2	0	0.20

Settings are ranked with respect to recognition accuracy in GA from the best to the worst
* Orientation: 1—signed (0°–360°), 0—unsigned (0°–180°)

Table 3 The best setting of HOG descriptor parameters as found by the GA algorithm for the JAFFE database

No.	Bins	Cell size (px)	Block size	Orientation*	Clip value
1.	11	13	4	1	0.25

* Orientation: 1—signed (0°–360°), 0—unsigned (0°–180°)

The established metric ERR (i.e., emotion recognition rate) was used from the literature for evaluating our algorithm recognition accuracy. This metric ERR, which in the literature can also be referred to without abbreviation, is defined as:

$$\mathrm{ERR}_i = \frac{\sum_{i,\mathrm{corr}}}{\sum_i}, \qquad (1)$$

where $\sum_{i,\mathrm{corr}}$ is the number of correctly classified images, and $\sum_i$ is the total number of all images for the selected emotion i.

Our evaluation was carried out by the so-called leave-one-subject-out strategy. Let's detail this strategy. Firstly, all images of the selected subject are left out during the current evaluation iteration. The remaining images are then used for training SVM machines. Finally, all the subjects' images are used for assessing the algorithms recognition accuracy. As described above is carried out for all subjects, and finally, the results of all iterations are simply combined. Table 4 presents the calculated emotion recognition rates for the six prototyping emotions for our algorithm for the CK database, where each column presents the recognition rates for each particular emotion. The results are presented for the better ten settings of HOG descriptor parameters (see also Table 2 for the explanation), where each row presents the calculated emotion recognition rates for each particular setting. The mean emotion recognition rates and their corresponding standard deviations are calculated both with respect to the particular emotion and the HOG descriptor setting. The better results within each category are marked as bold. On the other hand, Table 5 presents the recognition results for JAFFE database for the best setting of HOG parameters.

4 Discussion and conclusion

The results are analyzed and compared to the state-of-the-art methods in this section. Firstly, an experiment was conducted that confirmed that our method using neutral and one emotional (apex) images performs better than the same procedure using just one single emotion image (i.e., single-image version). The latter means that the same algorithm was applied where the feature difference vector Δ was sub-

Table 4 Emotion recognition rates for our algorithm for the six prototypic emotions (columns) and the better ten settings of HOG descriptor (rows) for CK database

No.	Disgust	Anger	Surprise	Fear	Happiness	Sadness	$\overline{x}$	σ
1.	**98.31**	94.07	**98.80**	82.67	**100.00**	**100.00**	**95.64**	6.13
2.	96.61	93.33	98.80	**84.00**	100.00	98.81	95.26	5.48
3.	96.61	**94.81**	98.80	82.6	100.00	96.43	94.89	5.72
4.	96.61	94.07	98,80	78.67	100.00	91.67	93.30	7.11
5.	97.18	89.63	98.39	77.33	100.00	94.05	92.76	7.67
6.	96.61	89.63	97.19	77.33	100.00	95.24	92.66	7.54
7.	97.74	91.85	98.80	70.67	100.00	96.43	92.58	10.13
8.	97.74	91.85	98.39	70.67	100.00	96.43	92.51	10.09
9.	96.05	85.93	98.39	74.66	100.00	97.62	92.11	9.04
10.	96.05	90.37	97.59	68.00	100.00	88.10	90.02	10.66
$\overline{x}$	96.95	91.55	98.40	76.67	100.00	95.48		
σ	0.72	2.59	0.54	5.31	0.00	3.31		

Mean ($\overline{x}$) and standard deviation (σ) values are calculated for the particular emotion and HOG descriptor setting. The higher rates are marked as bold

Table 5 Emotion recognition rates for our algorithm for the six prototypic emotions (columns) and the best setting of HOG descriptor for JAFFE database

No.	Disgust	Anger	Surprise	Fear	Happiness	Sadness	$\overline{x}$	σ
1.	75.86	86.66	93.10	90.62	96.77	83.87	**87.82**	6.78

Mean ($\overline{x}$) and standard deviation (σ) values are calculated

Table 6 Emotion recognition rates for the proposed algorithm without feature vector differences (i.e., single-image version) for the six prototypic emotions (columns) and the best setting of HOG descriptor for JAFFE and CK databases

Database	Disgust	Anger	Surprise	Fear	Happiness	Sadness	$\overline{x}$	σ
CK	96.61	88.89	98.39	78.67	99.51	86.90	91.49	8.13
JAFFE	58.62	83.33	68.97	56.25	77.41	48.39	65.50	13.40

stituted during the recognition phase by the HOG descriptor vector calculated on the emotional image. The results of this experiment for the best setting of HOG descriptor parameters are collated in Table 6 for both databases. When comparing Tables 4 and 6, and Tables 5 and 6, it can be seen that our proposed method outperformed the single-image version of our algorithm on both databases. The mean recognition rate improvement for CK was around 4 %, while for JAFFE it was roughly 20 %.

Let us analyze the recognition accuracy of our proposed method for facial expression recognition based on HOGs and the feature vector differences. The mean recognition rate of our method for the best setting of the HOG descriptor (see row 1 in Table 4) was 95.64 % for the CK database. When using this setting, the emotions ‘Happiness’ and ‘Sadness’ were recognized with 100 % accuracy, with slightly deviated recognition rates for emotions ‘Surprise’ and ‘Disgust’ at 98.8 and 98.31 %, respectively, followed by an acceptable 94.07 % recognition rate for ‘Anger,’ while the emotion ‘Fear’ was recognized with just 82.67 % accuracy. It should be stressed that if the accuracy of our method were to be calculated as a ratio between the number of correctly classified images (897) and the total number of testing images (924), then the recognition rate would be 97.10 %. One of reasons for a slightly lower recognition accuracy for the emotion ‘Fear’ could be found in the smaller number of images available within the CK database for the learning and testing of this emotion (see also Table 1). Some psychological experiments have demonstrated that similar muscle activities are noticed when expressing emotions ‘Fear,’ ‘Disgust,’ and ‘Anger.’ Detailed analysis of our classifier pointed out that the emotion ‘Fear’ was regularly misclassified either as the emotions ‘Disgust’ or ‘Anger.’ A similar trend of recognition results was also noticed for JAFFE database. Otherwise, the results are slightly lower than for CK database, which is a consequence of the not as expressive emotions on images from the JAFFE database.

Let us also assess the time complexity of our proposed recognition procedure based on HOG descriptors. Training SVM machines is certainly very time-consuming and can take several hours. However, it should be stressed that this learning phase is carried out off-line within real-world applications. On the other hand, once the classifier is trained, the recognition phase executes practically in the real time. It should be stated that facial expression recognition by using our procedure implemented in C++ requires approximately 65 ms of CPU time per image on today’s typical computer system with an Intel Core i5-4570 processor having a 3.2 GHz system clock and 16 GB of DDR3 SDRAM.

Our recognition method was also compared to the facial expression recognition methods from the literature. Table 7 presents the recognition accuracies of the compared methods accompanied by key features of the methods, used classifiers, and some information about the validation procedure. The methods are ordered according to their mean recognition rates. Our proposed method (marked bold) has also been added to this table for easier comparison. It can be seen that our approach with respect to recognition accuracy surpassed practically all methods except three, and also that the best results were obtained on the CK database. The method in [29] resulted in 99.33 % recognition accuracy, but it should be stressed that videos from just ten subjects were used (unlike 106 subjects used in our research). The method in [19] is a pure video-based facial expression recognition approach. This means that this method exploits (and requires) more information than ours, a difference is also in the different validation procedure where the method in [19] randomly selects 60 subjects for learning and 33 subjects for testing over 10 runs. The second method in [34] outperforms our proposed approach with respect to recognition accuracy by just 0.22 %. Both methods are based on HOG descriptors, while using different classifiers. The method in [34] is single-image-based, while ours is based on a comparison between two images. It should be stressed that both methods were developed independently of each other. The main difference is in the validation procedures. Only 32 subjects with 6 images per emotion were applied in [19], in contrast to a bigger population of 106 subjects with just 3 images per emotion used in

Table 7 Comparison between the recognition accuracy of the proposed method with state-of-the-art methods

Method	Facial features	Classifier*	Performance				
			Input**	Emotions	Database	Samples	Recognition rate (%)
[29]	SWLDA + optical flow	HMM	VI	6	CK	10 Subjects videos	99.33
[19]	Spatial filters + ICA	SVM	VI	6	CK	317 Videos	97.80
[34]	HOG + WLD	kNN	SI	7	CK	1344 Images	95.86
Our method	HOG difference vector	SVM	FI (2)	6	CK	1232 Images	95.64
[32]	Geometric features	Gentle SVM + HMM	VI	22 AUs	MMI	244 Videos	95.3
[7]	HOG of important facial features	SVM	SI	6	CK	N/A	95.00
[13]	Gabor wavelets + LDA	DCA	VI	10	N/A	N/A	94.13
[37]	LBP-TOP	SVM + Adaboost	VI	6	CK	374 Videos	93.85
[27]	Motion units	kNN	SI	4	CK	212+ Images	93
[14]	LBP	SVM	SI	7	CK	1240 Images	92.9
[15]	HDBF + local FDA	Convolutional NN	SI	7	CK	327 Images	91.3
[36]	Pixel intensity of face	kNN + HMM	VI	6	CK	488 Videos	90.7
[17]	Gabor features	SVM	SI	6	JAFFE	N/A	88.1
Our method	HOG difference vector	SVM	FI (2)	6	JAFFE	192 Images	87.82
[11]	Log-Gabor filters	RHF	VI	6	CK	344 Videos	87.10
[24]	Face projections on eigenspace	NN	SI	7	CK	97 Images	86
[1]	Spatiotemporal vectors	Weighted kNN	VI	4	CK	25 Videos	85.0
[23]	HOG features of log-likelihood maps	SVM and DT	SI	5	CK	300 Images	83.3
[25]	Frontal and profile facial points	Rule-based	SI	9	MMI	196 Images	83
[10]	24 Facial points	DBN	VI	6	Mind reading DVD	164 Videos	77.4
[38]	34 Points converted to a labeled graph	Nonlinear canonical correlation analysis	SI	6	JAFFE	183	77
[4]	Motion units	TAN	VI	6	CK	53 Videos	73
[22]	Feature displacements (neutral-peak)	SVM	VI	6	CK	75 Videos	71.8

*Classifier: *TAN* tree augmented naive Bayes, *DT* decision trees, *RHF* randomized hough forest, *WLD* Webber local descriptor, *DBN* dynamic Bayesian network, *DCA* dynamic cellular automata, *SWLDA* stepwise linear discriminant analysis, *HDBF* high dimensional binary features

**Input: *FI(N)* N images from video, *SI* single image, *VI* video our validation. In addition, the authors in [19] used half of the subjects' emotions (randomly selected) for training and the remaining emotions for testing over four runs. In this way, the classifier dealt with images of the same subject during the learning and testing phases. Indeed, images of particular subject used for training were from those emotions not used when testing the remaining images of this subject, but such an evaluation might still influence recognition accuracy.

Our proposed method differs from the single-image-based methods in a way that utilizes information from the additional image (i.e., neutral image). This idea significantly improves facial expression recognition accuracy (see Table 7), despite the fact that our method does not apply any advanced preprocessing routines, no special feature weighting, no advanced recognition approaches or supplements (e.g., Adaboost, LDA, ICA, ...). It applies just sophisticated HOG parameter selection by using a real-coded genetic algorithm. All the previously mentioned can of course be used as future work directions. A recognition rate higher than the recognition rates for practically all the other compared single-image- or video-based methods was obtained by simple subtraction of the HOG feature vectors calculated for the neutral and apex emotional images for the CK database. The same is true, if the JAFFE database is considered.

It should also be emphasized that our proposed method is fully automatic and needs no manual input. This method is person independent and generalizes very well when recognizing facial expressions for those new persons not used during the training period. The main research direction foreseen for future work is namely that our method could be upgraded by an intelligent sub-feature selection procedure (i.e., a procedure for selecting the more important blocks in the HOG descriptor).

References

1. Bourel, F., Chibelushi, C.C., Low, A.A.: Recognition of facial expressions in the presence of occlusion. In: BMVC, pp. 1–10. Citeseer (2001)
2. Buciu, I., Kotropoulos, C., Pitas, I.: Comparison of ica approaches for facial expression recognition. Signal Image Video Process. **3**(4), 345–361 (2009)
3. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**:27:1–27:27 (2011). Software available at url http://www.csie.ntu.edu.tw/cjlin/libsvm
4. Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.S.: Facial expression recognition from video sequences: temporal and static modeling. Comput. Vis. Image Underst. **91**(1), 160–187 (2003)
5. Corcoran, A.L., Sen, S.: Using real-valued genetic algorithms to evolve rule sets for classification. In: Proceedings of the First IEEE Conference on Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence, pp. 120–124. IEEE (1994)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Schmid, C., Soatto, S., Tomasi, C., (eds.) International Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 886–893. INRIA Rhône-Alpes, ZIRST-655, av. de l'Europe, Montbonnot-38334 (2005)
7. Donia, M., Youssif, A., Hashad, A.: Spontaneous facial expression recognition based on histogram of oriented gradients descriptor. Comput. Inf. Sci. **7**(3), 31–37 (2014)
8. Edwards, G.J., Cootes, T.F., Taylor, C.J.: Face recognition using active appearance models. In: Computer Vision ECCV98, pp. 581–595. Springer, Berlin (1998)
9. Ekman, P., Friesen, W.V.: Measuring facial movement. Environ. Psychol. Nonverbal Behav. **1**(1), 56–75 (1976)
10. El Kaliouby, R., Robinson, P.: Real-time inference of complex mental states from facial expressions and head gestures. In: Real-Time Vision for Human-Computer Interaction, pp. 181–200. Springer, US (2005)
11. Fanelli, G., Yao, A., Noel, P.-L., Gall, J., Van Gool, L.: Hough forest-based facial expression recognition from video sequences. In: Trends and Topics in Computer Vision, pp. 195–206. Springer (2012)
12. Fang, H., Mac Parthaláin, N.M., Aubrey, A.J., Tam, G., Borgo, R., Rosin, P., Grant, P., David, M., Chen, M.: Facial expression recognition in dynamic sequences: an integrated approach. Pattern Recogn. **47**(3), 1271–1281 (2014)
13. Geetha, P., Narayanan, V.: Evolutionary computational method of facial expression analysis for content-based video retrieval using 2-dimensional cellular automata. (2010). arXiv preprint arXiv:1009.1983
14. Gritti, T., Shan, C., Jeanne, V., Braspenning, R.: Local features based facial expression recognition with face registration errors. In: FG'08. 8th IEEE International Conference on Automatic Face and Gesture Recognition, 2008, pp. 1–8. IEEE (2008)
15. Kahou, S.E., Froumenty, P., Pal, C.: Facial expression analysis based on high dimensional binary features. In: Computer Vision-ECCV 2014 Workshops, pp. 135–147. Springer, Switzerland (2015)
16. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000, pp. 46–53. IEEE (2000)
17. Kotsia, I., Buciu, I., Pitas, I.: An analysis of facial expression recognition under partial facial image occlusion. Image Vis. Comput. **26**(7), 1052–1067 (2008)
18. Lajevardi, S., Hussain, Z.: Automatic facial expression recognition: feature extraction and selection. Signal Image Video Process. **6**(1), 159–169 (2012)
19. Long, F., Wu, T., Movellan, J.R., Bartlett, M.S., Littlewort, G.: Learning spatiotemporal features by using independent component analysis with application to facial expression recognition. Neurocomputing **93**, 126–132 (2012)
20. 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)
21. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J.: Coding facial expressions with gabor wavelets. In: Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, 1998, pp. 200–205. IEEE (1998)
22. Michel, P., Kaliouby, R.E.: Real time facial expression recognition in video using support vector machines. In: Proceedings of the 5th International Conference on Multimodal Interfaces, pp. 258–264. ACM (2003)
23. Orrite, C., Gañán, A., Rogez, G.: Hog-based decision tree for facial expression classification. In: Pattern Recognition and Image Analysis, pp. 176–183. Springer, Berlin (2009)
24. Padgett, C., Cottrell, G.W.: Representing face images for emotion classification. Adv. Neural Inf. Process. Syst. **9**, 894– 900 (1997)

25. Pantic, M., Rothkrantz, L.: Case-based reasoning for user-profiled recognition of emotions from face images. In: ICME'04. 2004 IEEE International Conference on Multimedia and Expo, vol. 1, pp. 391–394. IEEE (2004)
26. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 1034–1041. IEEE (2009)
27. Sebe, N., Lew, M.S., Sun, Y., Cohen, I., Gevers, T., Huang, T.S.: Authentic facial expression analysis. Image Vis. Comput. **25**(12), 1856–1863 (2007)
28. Shan, C., Gong, S., McOwan, P.W.: Robust facial expression recognition using local binary patterns. In: ICIP 2005. IEEE International Conference on Image Processing, 2005, vol. 2, pp. II–370. IEEE (2005)
29. Siddiqi, M., Ali, R., Khan, A., Kim, E., Kim, G., Lee, S.: Facial expression recognition using active contour-based face detection, facial movement-based feature extraction, and non-linear feature selection. Multimed. Syst. (2014). doi:10.1007/s00530-014-0400-2
30. Tian, Y., Kanade, T., Cohn, J.F.: Facial expression recognition. In: Handbook of Face Recognition, pp. 487–519. Springer, London (2011)
31. Tian, Y.-L., Kanade, T., Cohn, J.F.: Facial Expression Analysis. Springer, Berlin (2005)
32. Valstar, M., Pantic, M.: Fully automatic recognition of the temporal phases of facial actions. IEEE Trans Syst. Man Cybern. Part B Cybern. **42**(1), 28–43 (2012)
33. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001, vol. 1, pp. I–511. IEEE (2001)
34. Wang, X., Jin, C., Liu, W., Hu, M., Xu, L., Ren, F.: Feature fusion of hog and wld for facial expression recognition. In: 2013 IEEE/SICE International Symposium on System Integration (SII), pp. 227–232 (2013)
35. Wu, C.-H., Tzeng, G.-H., Goo, Y.-J., Fang, W.-C.: A real-valued genetic algorithm to optimize the parameters of support vector machine for predicting bankruptcy. Expert Syst. Appl. **32**(2), 397–408 (2007)
36. Yeasin, M., Bullot, B., Sharma, R.: Recognition of facial expressions and measurement of levels of interest from video. IEEE Trans. Multimed. **8**(3), 500–508 (2006)
37. Zhao, G., Pietikäinen, M.: Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. Pattern Recognit. Lett. **30**(12), 1117–1127 (2009)
38. Zheng, W., Zhou, X., Zou, C., Zhao, L.: Facial expression recognition using kernel canonical correlation analysis (kcca). IEEE Trans. Neural Netw. **17**(1), 233–238 (2006)