Facial expression recognition in the wild using improved dense trajectories and Fisher vector encoding

Sadaf Afshar[1] Albert Ali Salah[2]

[1]Computational Science and Engineering Program, Boğaziçi University, Istanbul, Turkey
[2]Department of Computer Engineering, Boğaziçi University, Istanbul, Turkey
{sadaf.afshar, salah}@boun.edu.tr

Abstract

Improved dense trajectory features have been successfully used in video-based action recognition problems, but their application to face processing is more challenging. In this paper, we propose a novel system that deals with the problem of emotion recognition in real-world videos, using improved dense trajectory, LGBP-TOP, and geometric features. In the proposed system, we detect the face and facial landmarks from each frame of a video using a combination of two recent approaches, and register faces by means of Procrustes analysis. The improved dense trajectory and geometric features are encoded using Fisher vectors and classification is achieved by extreme learning machines. We evaluate our method on the extended Cohn-Kanade (CK+) and EmotiW 2015 Challenge databases. We obtain state-of-the-art results in both databases.

1. Introduction

Automatic video data analysis has been a growing interest in multimedia retrieval and human computer interaction. One of the most challenging parts in video analysis is the ability of evaluating human affective displays robustly. Despite intensive work on facial expression recognition, most of the previous research is done on videos collected under controlled conditions. In 2010, Lucey et al. [24] introduced an extended version of the Cohn-Kanade dataset (CK+). Around the same time, Valstar and Pantic [32] introduced the MMI facial expression dataset, which contains a large collection of facial videos encoded with the facial action coding system (FACS). These and similar databases have been crucial for evaluation of facial expression systems, but most of such resources contain posed expressions, collected under laboratory conditions. On the other hand, the EmotiW 2015 Challenge Dataset[1] contains short labeled audio-video clips depicting six universal emotions (Anger, Disgust, Fear, Happiness, Sad, and Surprise), as well as neutral faces, selected from movies with challenging illumination and pose conditions.

Evaluating human emotion in real-world videos (called “in the wild”) is still an open challenge and less addressed by researchers. Complexity of emotion recognition in real-world videos is due to many factors such as different illumination conditions, various head poses, unspecified apex of emotional expressions, scaling, occluding of faces, and complicated background. For the problem of video-based action recognition, which shares some of these challenges, improved dense trajectory features have been successfully used [36]. In this paper we propose a novel methodology, which achieves state-of-the-art results for facial expression recognition in the wild, and explores the use of improved dense trajectory features. This is our first contribution.

For designing a robust emotion recognition system in the wild, one of the most important pre-processing steps is the facial alignment, or registration. Our second contribution is an alignment system that is a combination of recent face detection, landmark localization and registration approaches. Our third contribution is the fusion of multiple feature representations that complement each other. We use feature-level fusion in our approach. In the literature, support vector machines (SVM) have been frequently used for classification of facial expressions. Instead of SVM, we use an extreme learning machine (ELM) classifier, which is faster to train, and achieves better results compared to the SVM.

The rest of this paper is organized as follows. We review the recent literature on facial expression recognition in Section 2. Our proposed approach to facial registration will be discussed in Section 3. Feature extraction will be discussed in detail in Section 4. We describe video modeling briefly in Section 5, followed by model learning in Section 6. Results and conclusions will be reported in Sections 7 and 8, respectively.

[1] https://cs.anu.edu.au/few/EmotiW2015.html

2. Related work

The facial expression recognition pipeline consists of face detection, landmark localization, alignment, feature extraction and classification steps. Alignment (or registration) is an important step, since removing rotation, scale and translation can improve the recognition system considerably.

If landmark points on faces can be located efficiently and accurately, these will guide the registration. Zhu & Ramanan [40] presented a unified tree-structured model for face detection and landmark localization in the wild, which is shown to be efficient for capturing deformations. Dibeklioglu et al. [7] offered a method for 2-D facial landmarking, which is based on the combination of a mixture model of Gabor wavelet features and a shape prior, estimated with a multivariate Gaussian mixture model. Xiong and de la Torre [37] proposed a supervised descent method for minimizing a Non-linear Least Squares (NLS) function, which achieves state-of-the-art performance in facial features detection. Asthana et al. [2] introduced the Discriminative Response Map Fitting method for landmark localization, which produces very good results.

Different alignment methods have been proposed for aligning faces, and other shapes. In the field of morphometrics, Gower [10] proposed the Generalized Procrustes Analysis (GPA), which can be used for aligning any number of shapes represented by point sets, to a reference model. This approach requires accurate landmarks to produce good results. If this condition is met, GPA will provide a very good registration for 2D or 3D faces.

Feature extraction plays a crucial role in designing a robust recognition system. For face processing resistant to local illumination effects, Local Binary Pattern features were proposed. This descriptor was later extended for modeling spatio-temporal features, as LBP-TOP [39]. Another important feature for modeling faces have been Gabor wavelets, which capture oriented edges over the facial image at different scales. The LBP-TOP features are extracted from Gabor images to get LGBP-TOP descriptors [1], which have shown promising results in the literature. Other typical descriptors are SIFT [22], HOG [4], LPQ [15].

Recently, Wang and Schmid [36] proposed the improved dense trajectories method for action recognition, which is based on extracting motion boundary histograms (MBH) [5], Oriented Histograms of Flow (HOF) [18], Histogram of Oriented Gradients (HOG) [4] and optical flow [23] trajectories. The descriptor provides a summary of changing features over the video, and it is robust to idiosyncratic variations. We postulate that this method can be useful for facial expression recognition, as we need to track changes in facial dynamics.

The last step in the recognition pipeline is classification. SVM is a very popular approach [20, 6]. Recently, extreme learning machines (ELM) introduced by [12] is shown as a viable alternative to the SVM, which is slow to train. In our study, we use ELM and show that it has good generalization performance for multi-class classification and needs shorter time for the learning phase, compared to SVM.

The pipeline of our method is illustrated in Figure 1. In the following sections, we describe its main components.

3. Facial registration

Our facial registration approach consists of two major steps of landmark detection and alignment, respectively.

3.1. Face and landmark detection

For landmark localization, we propose to combine two different methods in order to make a robust system. We use the supervised descent method (SDM) [37] in conjunction with the Discriminative Response Map Fitting (DRMF) method [2]. The Intraface (SDM) method is fast (about 0.51 second per frame) and can detect facial features precisely, but since its face detector is based on the Viola & Jones face detector [33], sometimes it fails to find faces. This is a problem in many cases of realistic videos, where different illumination, pose and complicated background conditions are present. The DRMF method, on the other hand, benefits from a tree-based face detector, which is proposed by Zhu & Ramanan [40]. Although the DRMF method works well in the wild conditions, its current implementation is very slow. It takes about 23 seconds to find the face and its corresponding landmarks in a frame. In order to preserve both accuracy and speed, we combine these two methods. The video is first processed by Intraface, using the Viola & Jones face detector and for each frame, faces and their landmarks with high confidence scores are selected. If a given video has less than three frames with detected faces, we use the combination of Zhu & Ramanan and DRMF for face and landmark localization. If this approach also fails to find faces, the video is tagged as not having any faces.

3.2. Generalized Procrustes alignment

In our proposed system, we use a single reference model and align all faces to it in order to remove translation, rotation, and scale effects. For this purpose, we use the generalized Procrustes analysis (GPA) proposed by [10]. A set of faces are represented by their landmarks and an iterative approach is employed to obtain the reference model. This procedure automatically produces the registered set of faces from the training set at the same time. Given a new facial image, GPA will find the affine transformation that aligns the face to the reference face, minimizing a distance functions that equally weights each landmark. For details, we refer the reader to [29].

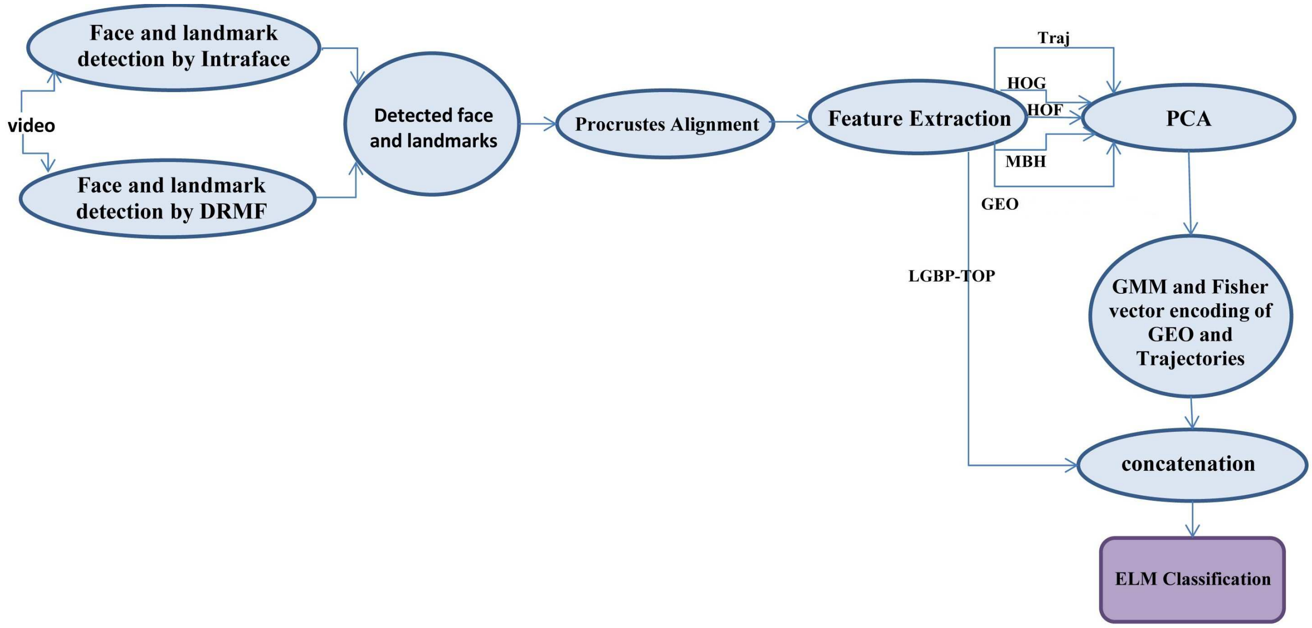

Face and landmark detection by Intraface
video
Face and landmark detection by DRMF
Detected face and landmarks
Procrustes Alignment
Feature Extraction
Traj
HOG
HOF
MBH
GEO
LGBP-TOP
PCA
GMM and Fisher vector encoding of GEO and Trajectories
concatenation
ELM Classification

Figure 1. Pipeline of the proposed system.

4. Feature extraction

After aligning faces to the reference model, we perform a histogram equalization to cope with illumination problems. A number of features and descriptors are extracted from the aligned and preprocessed faces. From each training (or test) video, we extract features including improved dense trajectories (using software provided by [36]), geometric features [16] and LGBP-TOP features. To reduce dimensionality of the descriptors, principal components analysis (PCA) is used. We then apply Fisher vector encoding on improved trajectory and geometric features for global representation. Finally, we concatenate all the features and pass this feature vector to an ELM classifier in order to predict a single emotional expression label for the entire video clip.

4.1. Improved Dense Trajectories

Improved dense trajectory features reach state-of-the-art in the action recognition problem [36]. These features are based on image descriptors (HOG, HOF and MBH descriptors) computed along tracked trajectories. We use these features to capture the changes in facial dynamics. Wang et al. illustrated that tracking densely sampled feature points from a multi-scale pyramid (built from each frame of the video) can outperform sparse sampling. Tracking is based on dense optical flow [9] and is done for a certain time window. In realistic videos, camera motion should be filtered out to prevent generating trajectories which correspond to the background. For solving this problem, Wang et al. proposed using the homography matrix of the points from continuous frames, which is extracted by the RANSAC approach. After filtering out the camera motion, 96-dimensional HOG, 108-dimensional HOF, 192-dimensional MBH and 30 dimensional trajectory features are computed to describe the appearance, shape and motion information. Figure 2 visualizes the trajectories on some videos from the EmotiW Challenge Dataset. The tracked points in the current frame are given as red dots, and the motion of each such point is indicated with a green line.

Using improved dense trajectory features has some important advantages over simple tracking of interest points. First, removing camera motion from optical flow improves HOF descriptors as discussed in [35]. Second, canceling out camera motion also removes trajectories that are produced by camera motion. Therefore, only trajectories related to face movements are kept. This specification is very important in real-world videos, where there are lots of pan and tilt camera motions.

4.2. Geometric Features

The shape of the face can be captured by its landmarks. Therefore, interpreting the movement of the landmarks during a facial expression can improve the performance of a facial expression recognition system. In this paper, we used a set of landmarks on the face detected by both Intraface and DRMF methods. The geometric features we use are mostly the same features introduced in Kaya et al. [16]. We have included three more features to enhance this set. Indices of extracted landmarks can be seen in Figure 3 and the details of the employed geometric features are tabulated in Table 1.

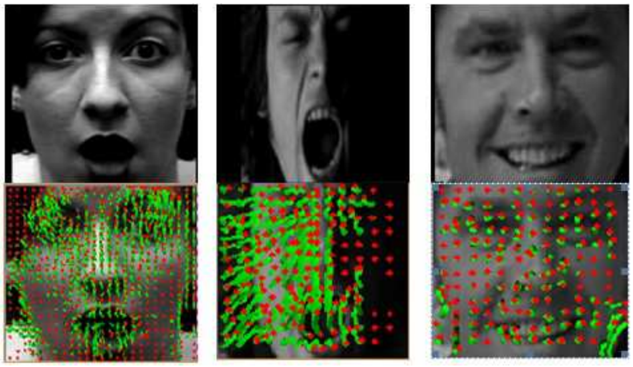


Figure 2. Original video frames (first row) and their visualized improved dense trajectories (second row). Images are selected from the CK+ and EmotiW 2015 Challenge datasets. Best viewed in color.

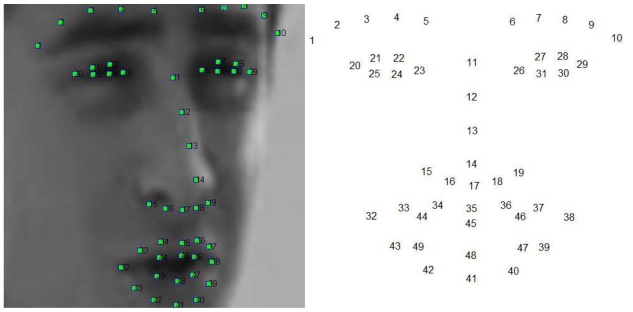


Figure 3. Landmarks extracted from the face.

4.3. Local Gabor binary patterns (LGBP)

In LGBP [1] images are convolved with a set of 2D complex Gabor filters to obtain Gabor-pictures, and then LBP-TOP is applied to each Gabor-picture. A 2D complex Gabor filter is defined as the convolution of a complex sinusoid $s(x,y)$ (carrier) with a 2-D Gaussian kernel $\omega_\tau(x,y)$ (envelope):

$$g(x,y) = s(x,y)\,\omega_\tau(x,y) \tag{1}$$

$$s(x,y) = \exp\left(j\left(2\pi\left(u_0 x + v_0 y\right) + p\right)\right), \tag{2}$$

where (u_0, v_0) stands for spatial frequency and p defines the phase of the sinusoid.

$$\omega_\tau(x,y) = K \exp\left(-\pi\left(a^2\left(x-x_0\right)_r^2 + b^2\left(y-y_0\right)_r^2\right)\right), \tag{3}$$

where a, b are scaling parameters of the Gaussian, K is amplitude and r subscript stands for a clockwise rotation around point (x_0, y_0). It should be stated that, we only use the magnitude response of the filter.

Table 1. Explanation of geometric features.

#	Features	Type and Explanation of feature
1	Eye aspect ratio	Distance, averaged over left and right parts of the face
2	Mouth aspect ratio	Distance
3	Upper lip angles	Angle, averaged over left and right parts of the face
4	Nose tip - mouth corner angles	Angle, averaged over left and right parts of the face
5	Lower lip angles	Angle, averaged over left and right parts of the face
6	Eyebrow slope	Angle, averaged over left and right parts of the face
7,8	Lower eye angles	Angle, averaged over left and right parts of the face
9	Mouth corner - mouth bottom angles	Angle
10	Upper mouth angles	Angle, averaged over left and right parts of the face
11	Curvature of lower-outer lips	Curvature, averaged over left and right parts of the face
12	Curvature of lower-inner lips	Curvature, averaged over left and right parts of the face
13	Bottom lip curvature	Curvature
14	Mouth opening	Distance
15	Mouth up/low	Distance
16	Eye - middle eyebrow distance	Distance, averaged over left and right parts of the face
17	Eye - inner eyebrow distance	Distance, averaged over left and right parts of the face
18	Inner eye - eyebrow center	Distance, averaged over left and right parts of the face
19	Inner eye - mouth top distance	Distance
20	Mouth width	Distance
21	Mouth height	Distance
22	Upper mouth height	Distance
23	Lower mouth height	Distance
24	Inner eye - mouth corner distance	Distance
25	Mouth center-left mouth corner	Distance
26	Mouth center-right mouth corner	Distance

5. Video modeling

5.1. Fisher vector encoding

The Fisher vector (FV) representation can be seen as an extension of Bag of Words (BOW). Perronnin and Dance [26] proposed the usage of GMM and Fisher kernels for producing visual vocabularies. Since then, FV has attracted a lot of attention. The Fisher vector representation benefits from characteristics of both generative statistical models (like HMM) and discriminative methods (like SVM). Unlike BOW, the Fisher vector uses both 0-order statistics (counting) and second order statistics. This specification enables the Fisher vector to find the best direction in which parameters of the GMM model are modified in order to fit to the data efficiently. In this paper, we used PCA in order to reduce the dimensionality of descriptors and for decorrelating them. Empirically, we got the best results on the training set by reducing the dimension of each trajectory to 25 and that of the other three descriptors (HOG, HOF, and MBH, respectively) to 64. The geometric features are projected to a decorrelated space by PCA, while their full dimensionality is kept.

We used GMMs with diagonal covariance matrix to produce the FV. The GMM clustering produces a visual vocabulary, where the number of clusters is a parameter of the method optimized on the training set. 20 to 60 clusters work well for each of the feature categories. In our experiments, the Fisher vectors are normalized firstly by the signed square root function, and secondly by L2 normalization. The final dimensionality of FV is $2 \times D \times K$, where D is the dimensionality of the descriptor, and $K$is the number of GMM components.

Given a set of descriptors $X = \{x_1, x_2, ..., x_N\}$ and parameters like $\lambda = \{\omega_i, \mu_i, \sigma_i\}_{|_{i=1}^{k}}$ learned from a random subset of training features, we calculate the FV as follows:

$$g^x_{\mu,i} = \frac{1}{N\sqrt{\omega_i}} \sum_{j=1}^{N} \gamma_{ij} \left(\frac{x_j - \mu_i}{\sigma_i} \right) , \qquad (4)$$

$$g^x_{\sigma,i} = \frac{1}{N\sqrt{2\omega_i}} \sum_{j=1}^{N} \gamma_{ij} \left[\left(\frac{x_j - \mu_i}{\sigma_i} \right)^2 - 1 \right] , \qquad (5)$$

where $\gamma_{i,j}$ denotes the posterior probability connecting each vector x_j with a component i in the GMM.

5.2. LGBP-TOP

The final feature we incorporate is LGBP-TOP. For this feature, LGBP histograms from three orthogonal planes (XY, XT, and YT, respectively, with X and Y representing the image plane, and T representing time) are extracted from two equal length volumes of the video, which are obtained by dividing the video over the time axis. The resulting features are concatenated in order to form the final feature vector. We take the idea of dividing the video for improving temporal modeling from Kaya et al. [16], and the Gabor pictures were obtained using an open source script [11]. Three scales and six orientations are used to prepare the Gabor filter bank, and each Gabor picture is divided into blocks. We have used 4 blocks for experiments on CK+, and 16 blocks for the EmotiW 2015 Challenge Dataset, respectively. Our approach is robust to operational parameters, and since CK+ has a smaller number of samples to train, reducing the number of blocks slightly improves generalization.

6. Classification

We have use kernel extreme learning machines (ELM) for classification [12]. ELM is actually a feedforward neural network with a single layer of hidden nodes, in which weights from the input layer to the hidden nodes are initialized randomly, and unlike neural networks, they will not be updated with backpropagation. This specification of ELM is the reason of its short training time.

If we assume that the output of the first layer of the network is represented as $H \in R^{N \times h}$ (where N and h show the number of observations and the hidden neurons, respectively), we should find the weights $\beta \in R^{h \times L}$ between H and $T \in R^{N \times L}$ to learn the classifier. Here, T is the output, and L is the number of classes. The least square solution of this linear equation starts from $H\beta = T$.

$$\beta = H^{\dagger} T, \qquad (6)$$

where $H^{\dagger}$ is the Moore-Penrose generalized inverse [27]. For a multi class classification problem, T is defined as one vs. all by below notation:

$$T_{t,L} = \begin{cases} +1 & if \quad y^t = 1 \\ -1 & if \quad y^t \neq 1 \end{cases} \qquad (7)$$

To optimize ELM, a regularization coefficient on the residual error $\parallel H\beta - T \parallel$ can be used. The idea of this optimized version of ELM is inspired by Least Square SVMs (LSSVM) via below equation:

$$\beta = H^T (\frac{I}{C} + HH^T)^{-1} T, \qquad (8)$$

where I is a $N \times N$ identity matrix and C is related to complexity parameter of LSSVM, which is used to regularize the linear kernel HH^T [31]. This formulation can be simplified as in Equation 9 given a kernel K [12, 28].

$$\beta = (\frac{I}{C} + K)^{-1} T \qquad (9)$$

7. Experiments

7.1. Datasets

We have evaluated our proposed pipeline on the Extended Cohn-Kanade (CK+) [24] and Emotiw 2015 Challenge [6] datasets. The Extended Cohn-Kanade (CK+) dataset contains 593 sequences from 123 subjects for seven facial expressions: happy, sad, surprise, anger, disgust, fear and contempt. The sequences are recorded in laboratory conditions and coded at the peak frame with the facial action coding system (FACS). All the videos start from the neutral face and end with the apex expression. Among these, only 327 samples have emotion labels, which are used in our experiments. In order to be able to compare our results with the state-of-the-art, Leave-One-Subject-Out protocol (LOSO) is used.

The EmotiW 2015 Challenge Dataset[2] consists of 723 training, 383 validation and 539 test videos. The labels of the test videos are sequestered, and the number of evaluations are limited. In the provided alignment by the organizers of the challenge, faces are detected only in 711 training and 371 validation videos. There are false positives due to challenging conditions of sequences. Our proposed alignment pipeline was able to detect 713 faces in the training set and 378 faces in the validation set, with a small amount of false positives, in a completely automatic manner.

The given alignment by the challenge organizers is good for frontal faces, but the alignment was not very efficient in the case of rotations. With our proposed method, we were able to improve the alignment. Since emotion labels of the test set are sequestered, we use cross validation on the training set to find the best parameters and then test the proposed method on the validation set. This approach is also used in previous works [20] and [34].

During our experiments, we analyze different combinations of features and also compare each block of our pipeline with other methods. We test each step of our proposed methodology to see how much each block contributes to the final result.

7.2. Descriptor types

Several experiments have been done in order to find the best combination of descriptors. We investigate the contribution of each descriptor both individually and in combination with others. We learned PCA and GMM models from each descriptor (HOG, HOF, MBH, improved trajectories, and geometric features), separately.

Results on the CK+ dataset are shown in Table 3. Concatenation of LGBP-TOP (after dimensionality reduction and power-L2 normalization) and Fisher-encoded HOG, HOF and GEO yields 94.80% accuracy on the CK+ dataset, which is among the best results obtained for this dataset so far. As expected, the combination of an appearance based feature (HOG) and a motion based feature (HOF) produces higher accuracy than combining two motion based (HOF and MBH) features. Joining only two descriptors (HOG and HOF) already gives a very promising result (93.58%) compared to baseline method [24]. Among the seven classes, "sad" is the most challenging emotion to recognize and "happy" is the easiest one. Table 2 compares several state-of-the-art approaches on the CK+ dataset, which are obtained with the same standard protocol. The confusion matrix of the final system is shown in Figure 4.

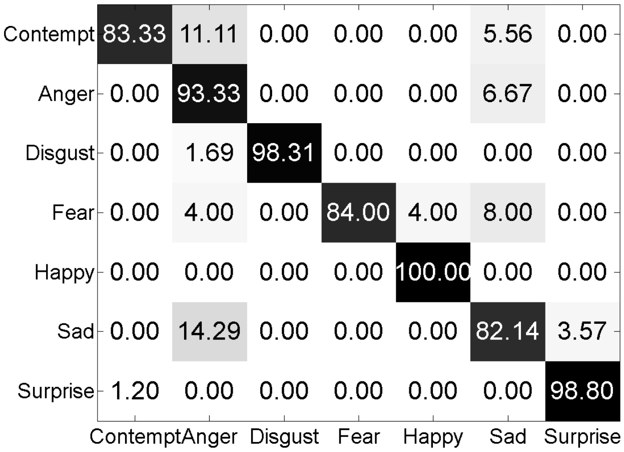
We use the same procedure on the EmotiW 2015 dataset; results are shown in Table 5. Again, the combination of HOG and HOF yields the best performance among improved trajectory features and produces higher recognition rate compared to the baseline on the validation set. Our final approach achieves 42.86% accuracy on the Emotiw 2015 validation set, which is 6.78% higher than the baseline (36.08%). The best result is obtained by a combination of Fisher encoded geometric, HOG, HOF, MBH and LGBP-TOP (after dimensionality reduction and power-L2 normal-

Table 2. State of the art results on the CK+

Algorithm	Protocol	Mean Rec. R
STPS+CAPP (baseline, Lucey et al., 2010 [24])	LOSO	88.38%
STLMBP (Huang et al., 2012 [14])	LOSO	92.62%
Cov3D (Sanin et al., 2013 [30])	LOSO	92.30%
RCC (Huang et al., 2014 [13])	LOSO	95.38%
LCRF (Walecki et al., 2015 [34])	10-fold	93.90%

Table 3. Contribution of different descriptors (CK+).

Descriptor	Dimension	Mean Rec. R
Trajectory	1250	71.56%
HOG	8192	90.52%
HOF	8192	87.77%
MBH	8192	89.91%
HOG+HOF	8192+8192	93.58%
HOG+MBH	8192+8192	92.05%
HOF+MBH	8192+8192	90.52%
HOG+HOF+MBH	8192+8192+8192	93.27%
Traj+HOG+HOF+MBH	1250+8192+8192+8192	91.13%
GEO	1352	69.42%
LGBP-TOP	75168	86.24%
GEO+HOG+HOF+ LGBP-TOP(RN)	1352+8192+8192+326	**94.80%**
GEO+HOG+HOF+ LGBP-TOP(RN) (without contempt)	1352+8192+8192+326	**95.79%**



	Contempt	Anger	Disgust	Fear	Happy	Sad	Surprise
Contempt	83.33	11.11	0.00	0.00	0.00	5.56	0.00
Anger	0.00	93.33	0.00	0.00	0.00	6.67	0.00
Disgust	0.00	1.69	98.31	0.00	0.00	0.00	0.00
Fear	0.00	4.00	0.00	84.00	4.00	8.00	0.00
Happy	0.00	0.00	0.00	0.00	100.00	0.00	0.00
Sad	0.00	14.29	0.00	0.00	0.00	82.14	3.57
Surprise	1.20	0.00	0.00	0.00	0.00	0.00	98.80

Figure 4. Confusion matrix of the final system (CK+).

[2] The Third Emotion Recognition in the Wild Challenge was held at the ACM International Conference on Multimodal Interaction, 2015, Seattle.

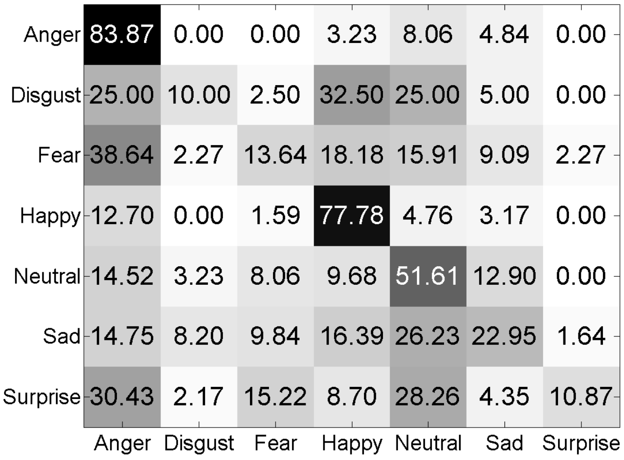
Table 4. State of the art results on the validation partition of the EmotiW 2015 Challenge Dataset.

Algorithm	Accuracy
LBP-TOP (Baseline) (Dhall et al., 2015 [6])	36.08%
LPQ+LBP-TOP+OpenSmile (Kayaoglu et al., 2015 [17])	40.70%*
AU-AFF (winner of the challenge) (Yao et al., 2015 [38])	49.09%*
RNN (Ebrahimi et al., 2015 [8])	39.60%

Table 5. Contribution of different descriptors (EmotiW 2015).

Descriptor	Dimension	Accuracy
Trajectory	1250	26.72%
HOG	8192	34.13%
HOF	8192	32.28%
MBH	8192	31.22%
HOG+HOF	8192+8192	36.77%
HOG+MBH	8192+8192	34.92%
HOF+MBH	8192+8192	29.63%
HOG+HOF+MBH	8192+8192+8192	34.92%
Traj+HOG+HOF+MBH	1250+8192+8192+8192	33.86%
GEO	1300	38.10%
LGBP-TOP(RN)	712	32.28%
GEO+HOG+HOF+ LGBP-TOP(RN)	1300+8192+8192+712	41.80%
GEO+HOG+HOF+ MBH+LGBP-TOP(RN)	1300+8192+8192+8192+712	**42.86%**

ization) features. Table 4 compares several approaches on the validation set of EmotiW 2015 dataset. The results that are marked with an asterisk are not completely comparable with the results reported here, since they do not follow the same protocol. The confusion matrix of our final system is shown in Figure 5.



	Anger	Disgust	Fear	Happy	Neutral	Sad	Surprise
Anger	83.87	0.00	0.00	3.23	8.06	4.84	0.00
Disgust	25.00	10.00	2.50	32.50	25.00	5.00	0.00
Fear	38.64	2.27	13.64	18.18	15.91	9.09	2.27
Happy	12.70	0.00	1.59	77.78	4.76	3.17	0.00
Neutral	14.52	3.23	8.06	9.68	51.61	12.90	0.00
Sad	14.75	8.20	9.84	16.39	26.23	22.95	1.64
Surprise	30.43	2.17	15.22	8.70	28.26	4.35	10.87

Figure 5. Confusion matrix of the final system (Emotiw 2015).

Table 6. ELM and SVM comparison in terms of time and performance.

Classifier	Training time	Testing time (one subject)	Accuracy(%)
ELM	0.45 s	0.0627 s	93.58
SVM	27.79 s	0.24 s	80.73

7.3. Facial alignment

In order to investigate how our registration pipeline improves the recognition performance, we apply the same procedure on the registered images that are provided by the organizers of EmotiW 2015 challenge. We were not able to extract geometric features from the provided alignment, since for a considerable number of frames, there are no landmarks, and for the rest, the number of detected landmarks is not consistent. Therefore, Fisher vector encoding of HOG, HOF, MBH concatenated with LGBP-TOP are used as feature vector. We obtain 38.54% accuracy on the validation set with the default alignment. Using the improved landmark detector and the generalized Procrustes alignment improves this result by 4.32%.

7.4. Fisher vectors vs. bag of words (BoW)

In order to compare Fisher vector and BoW on the CK+ dataset, we used 4000 cluster centers. We prepared vocabularies for each modality (i.e. HOG, HOF, MBH and improved trajectory) separately. Each BoW vector is separately normalized with L1 normalization. The concatenation of the BoW vectors is used in ELM with a linear kernel.

With BoW representation, the best result obtained on the CK+ using improved dense trajectory features (with a combination of HOG, HOF and MBH) is 88.99%, by leave one subject out protocol. FV encoding with considerable fewer number of visual words (64 words) outperforms BoW encoding (4000 words) by 4.59%.

7.5. Comparison of ELM with SVM

We contrast ELM and SVM in terms of training time and accuracy on the CK+ dataset. We have found that ELM is faster and more accurate than SVM. For this test, concatenation of Fisher vector encoding of HOF and HOG features are used. Experiments are done on a machine with an Intel (R) core i5 CPU 2.50 GHz and 6 GB of RAM. ELM reaches 93.58% accuracy, while SVM (the libsvm implementation was used [3]) achieves 80.73% accuracy. Table 6 shows the results on the CK+ dataset. We note here that it is possible to get faster computation times with more optimized SVM implementations like liblinear, but the difference remains significant.

7.6. Deep learning

Deep learning is becoming popular in the context of facial expression recognition. One of the widely used deep learning structures is the convolutional neural network (CNN). Training deep learning approaches requires very large datasets and longer training times. Here we briefly discuss the results of some deep learning methods applied on the CK+ dataset. In a recent study by Li et al. [19], 10,595 external images were used for training CNN models and 83% mean recognition rate was reported on CK+. Lv et al. [25] proposed a method based on face parsing detectors trained via deep belief networks and obtained 91.11% mean recognition rate. Liu et al. [21] proposed a new Boosted Deep Belief Network (BDBN), which yields 96.70% mean recognition rate, but it should be stated that in that work, the contempt emotion was not considered. By excluding the contempt expression, we were able to obtain 95.79% mean recognition rate. By comparing these results with the ones reported here, it can be seen that our approach is advantageous in terms of high accuracy and low complexity, as well as low training time.

8. Conclusions

In this work, we presented an approach for facial expression recognition that uses a combination of different static and dynamic features. We tested the proposed approach on the CK+ and EmotiW 2015 Challenge datasets. The results show that our method yields state-of-the-art results in both databases. The main contribution of this paper is that this is the first time that improved dense trajectory features are used for facial expression recognition in the wild. In original improved dense trajectory features human bounding box is used to remove camera motion, in case of facial expression recognition an accurate face detection can be used instead of human detection as what we have done in this paper.

In the case of the Emotiw dataset, the recall of surprise, disgust and fear classes are low, which can be due to the low number of training samples in these classes. Also, a lot of training samples can be considered to contain a mixture of two or more facial expressions (such as surprise, fear and happy), which makes the recognition more challenging. The class distribution is not balanced for the EmotiW dataset. For example, we have many more videos in happy and anger classes compared to fear, surprise and disgust. The proposed method is sensitive to small facial changes and works successfully in some of the difficult cases, which contain very small facial changes and are hard to distinguish even for a human annotator. An example is shown in Figure 6. On the other hand, the sensitivity of method to small changes and the aforementioned problem of mixing expressions in the training set, sometimes cause the failure of the method in simple cases. For instance Figure 7 illustrates such an example.

Collecting more samples for under-sampled classes and fusion with audio features, which contain significant emotional information, with visual features, may improve the performance of the system.

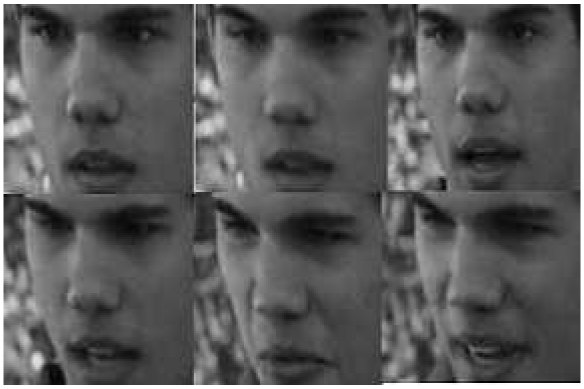


Figure 6. A correct classified sample from disgust class.

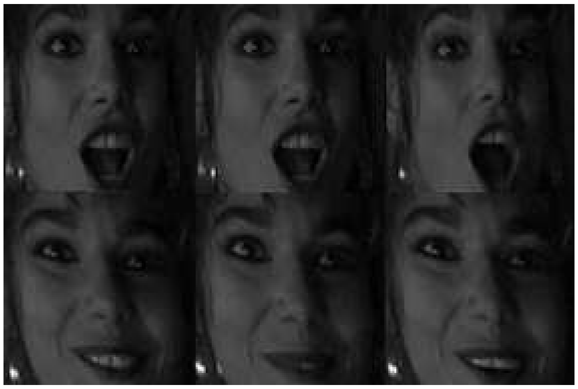


Figure 7. A misclassified sample from happy class.

References

[1] T. R. Almaev and M. F. Valstar. Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *ACII*, pages 356–361, Sept 2013. 2, 4

[2] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, June 2013. 2

[3] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM (TIST)*, 2(3):27, 2011. 7

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893 vol. 1, June 2005. 2

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, pages 428–441. Springer, 2006. 2

[6] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 423–426. ACM, 2015. 2, 5, 7

[7] H. Dibeklioılu, A. Salah, and T. Gevers. A statistical method for 2-d facial landmarking. *IEEE Transactions on Image Processing*, 21(2):844–858, Feb 2012. 2

[8] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal. Recurrent neural networks for emotion recognition in video. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 467–474. ACM, 2015. 7

[9] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image analysis*, pages 363–370. Springer, 2003. 3

[10] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. 2

[11] M. Haghighat, S. Zonouz, and M. Abdel-Mottaleb. Identification using encrypted biometrics. In *Computer analysis of images and patterns*, pages 440–448. Springer, 2013. 5

[12] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(2):513–529, 2012. 2, 5

[13] X. Huang, G. Zhao, M. Pietikainen, and W. Zheng. Robust facial expression recognition using revised canonical correlation. In *ICPR*, pages 1734–1739. IEEE, 2014. 6

[14] X. Huang, G. Zhao, W. Zheng, and M. Pietikainen. Spatiotemporal local monogenic binary patterns for facial expression recognition. *IEEE Signal Processing Letters*, 19(5):243–246, May 2012. 6

[15] B. Jiang, M. Valstar, B. Martinez, and M. Pantic. A dynamic appearance descriptor approach to facial actions temporal modeling. *Cybernetics*, 44(2):161–174, 2014. 2

[16] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah. Contrasting and combining least squares based learners for emotion recognition in the wild. In *Proceedings of the 2015 ACM*, pages 459–466. ACM, 2015. 3, 5

[17] M. Kayaoglu and C. Eroglu Erdem. Affect recognition using key frame selection based on minimum sparse reconstruction. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pages 519–524. ACM, 2015. 7

[18] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8. IEEE, 2008. 2

[19] W. Li, M. Li, Z. Su, and Z. Zhu. A deep-learning approach to facial expression recognition with candid images. In *MVA*, pages 279–282, May 2015. 8

[20] M. Liu, S. Shan, R. Wang, and X. Chen. Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition. In *CVPR*, pages 1749–1756, 2014. 2, 6

[21] Y. Liu, X. Hou, J. Chen, C. Yang, G. Su, and W. Dou. Facial expression recognition and generation using sparse autoencoder. In *SMARTCOMP*, pages 125–130. IEEE, 2014. 8

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2

[23] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981. 2

[24] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops (CVPRW)*, pages 94–101, June 2010. 1, 5, 6

[25] Y. Lv, Z. Feng, and C. Xu. Facial expression recognition via deep learning. In *SMARTCOMP*, pages 303–308. IEEE, 2014. 8

[26] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *CVPR*, pages 1–8. IEEE, 2007. 4

[27] C. Rao and S. Mitra. Generalized inverse of matrices and its applications. *Wiley series in probability and mathematical statistics Show all parts in this series*, 1971. 5

[28] R. Rifkin, G. Yeo, and T. Poggio. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences 190*, pages 131–154, 2003. 5

[29] A. A. Salah, N. Alyüz, and L. Akarun. Registration of three-dimensional face scans with average face models. *Journal of Electronic Imaging*, 17(1):011006–011006, 2008. 2

[30] A. Sanin, C. Sanderson, M. T. Harandi, and B. C. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *WACV*, pages 103–110. IEEE, 2013. 6

[31] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999. 5

[32] M. F. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the MMI facial expression database. In *The Workshop Programme*, page 65, 2010. 1

[33] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004. 2

[34] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic. Variable-state latent conditional random fields for facial expression recognition and action unit detection. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8, May 2015. 6

[35] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79, 2013. 3

[36] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, December 2013. 1, 2, 3

[37] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, June 2013. 2

[38] A. Yao, J. Shao, N. Ma, and Y. Chen. Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 451–458, New York, NY, USA, 2015. ACM. 7

[39] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):915–928, June 2007. 2

[40] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, June 2012. 2