

Combining feature-level and decision-level fusion in a hierarchical classifier for emotion recognition in the wild

Bo Sun¹ · Liandong Li¹ · Xuewen Wu¹ · Tian Zuo¹ · Ying Chen¹ ·
Guoyan Zhou¹ · Jun He¹ · Xiaoming Zhu¹

Received: 28 January 2015 / Accepted: 12 October 2015 / Published online: 18 November 2015
© OpenInterface Association 2015

Abstract Emotion recognition in the wild is a very challenging task. In this paper, we investigate a variety of different multimodal features (acoustic and visual) from video clips to evaluate their discriminative abilities in human emotion analysis. For each clip, we extract MSDF BoW, LBP-TOP, PHOG, LPQ-TOP and Audio features. We train different classifiers for every type of feature on the AFEW dataset from the ICMI 2014 EmotiW Challenge, and we propose a novel hierarchical classification framework, which combines the feature-level and decision-level fusion strategy for all of the extracted multimodal features. The final achievement we gain on the AFEW test set is 47.17%, which is considerably better than the best baseline recognition rate of 33.7%. Among all of the teams participating in the ICMI 2014 EmotiW challenge, our recognition performance won the first runner-up award. Furthermore, we test our method on

FERA and CK datasets, the experimental results also show good performance.

Keywords Emotion recognition · Multimodal features · Feature-level fusion · Decision-level fusion · Multiple kernel learning · Hierarchical classifier

1 Introduction

Psychologists believe that facial expressions and verbal messages are sometimes used as the main channel of human communication [1]. In recent years, automatic emotion recognition has received great attention. The development of technologies in emotion recognition is surprisingly fast, and many of them have already been used in real life. At the early stage, researchers focused mostly on emotion analysis from single static facial images under constrained circumstances [2]. The recognition in the real world is certainly quite different. As human emotions are actually dynamic streaming, the research is turning into recognition through video or image sequences. There are multiple methods for emotion recognition, which contain different types of feature descriptors, such as Gabor [3], local binary pattern (LBP) [4], histogram of oriented gradients (HOG) [5], and active appearance models [6] etc. Using a single feature descriptor to recognize emotion is easy to implement, but the result can hardly be satisfactory.

The Emotion Recognition in the Wild Challenge and Workshop (EmotiW) Grand Challenge in the ICMI 2014 is an audio–video based emotion classification challenge, which mimics real-world conditions [8]. The dataset is part of the Acted Facial Expression in the Wild (AFEW) database [11], which consists of short video clips extracted from popular Hollywood movies. For the challenge, the database

✉ Xuewen Wu
xuewenwu@hotmail.com

Bo Sun
tosunbo@bnu.edu.cn

Liandong Li
bnulee@hotmail.com

Tian Zuo
tian.zuo21@gmail.com

Ying Chen
ying.ch@mail.bnu.edu.cn

Guoyan Zhou
zgy@mail.bnu.edu.cn

Jun He
hejun@bnu.edu.cn

Xiaoming Zhu
zhuxm@elec.bnu.edu.cn

¹ College of Information Science and Technology, Beijing Normal University, Beijing 100875, China



Fig. 1 AFEW dataset

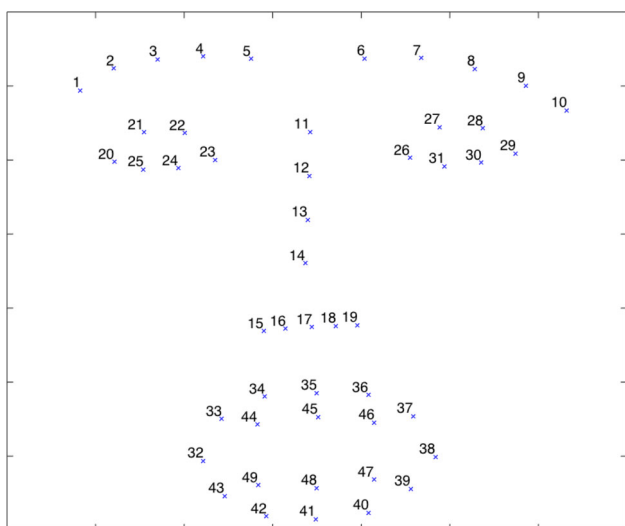


Fig. 2 49 facial landmarks tracked by Intraface

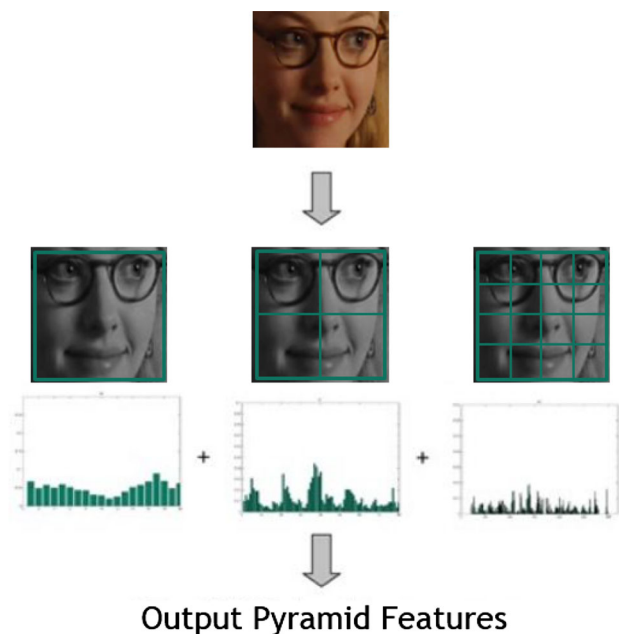


Fig. 3 Spatial pyramid matching

is divided into three subsets: Train set (578 samples), Validation set (383 samples) and Test set (407 samples). The task of the challenge is to classify a sample (audio–video clip) into one of the seven basic emotion categories, namely Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise. The Train set and the Val set are labeled and can be used for training or validating models, while the final classification accuracy on the Test set is taken into account for the overall Grand Challenge results. In Fig. 1a–g, we listed seven typical clips of those categories. We also conduct our experiment on other two facial expression databases: the Cohn Kanade [45] (CK) database and the Facial Expression Recognition and Analysis (FERA) [44] database.

From the clips in Fig. 1, we can see that emotion recognition in the wild may face the challenge of face detection, face alignment and facial feature extraction in the real world. The facial expression and speaking voice are common contributed

to the emotion recognition, so we choose to extract the features from them. For data preprocessing, openly available tools such as MoPS [9] and Intraface [10] are used for face detection and alignment. For facial expression, we extract the features of shapes, texture and visual bags. Then, we propose a novel hierarchical classification fusion framework, which is a decision-level fusion method for improving the result of emotion recognition. Pipeline of our method is shown in Fig. 4. Our classifier fuses different features and gains a promising recognition performance. In the additional experiments, we also compare the result of it with that of other multiple kernel methods.

In this paper, based on our work and research in the EmotiW 2014 challenge, we introduce the novel Hierarchical

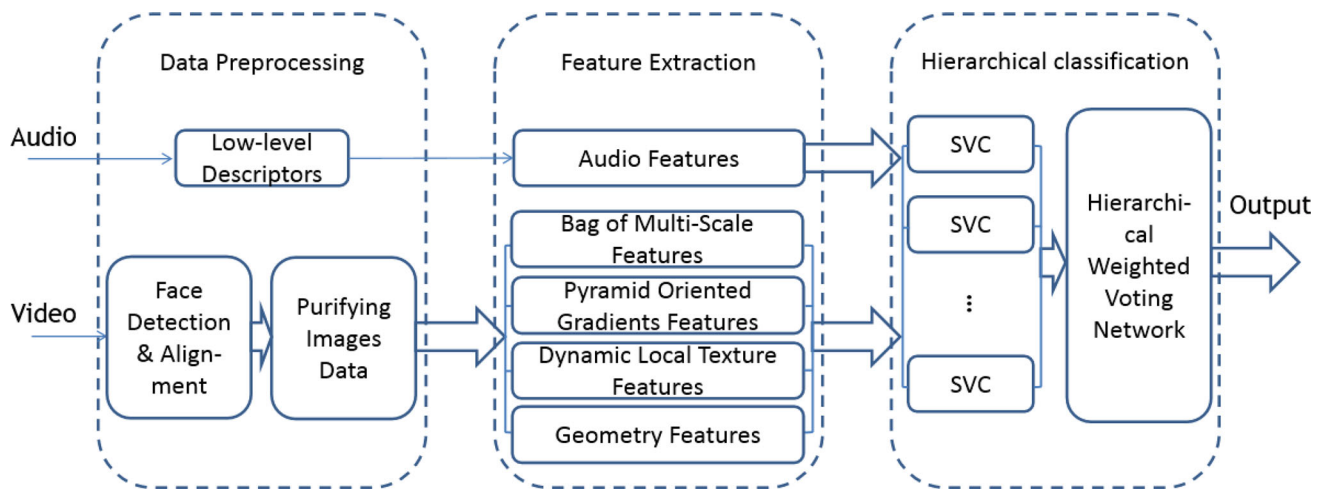


Fig. 4 Our proposed method

Classification Framework. The paper is organized as follows: In Sect. 2 we review the related works. Section 3 lists the data preprocessing steps. Section 4 gives the 4 groups of visual features extracted in the method, whose detailed ideas will be introduced. Section 5 introduces the classifiers used in our work, including support vector machine (SVM), MKL and our proposed classifier. Section 6 gives the entire experiments performed on the study, especially the recognition rates of our experiments both on the validation set and test set of the three datasets. Finally, the conclusion is given in Sect. 7.

2 Related works

For most automatic emotion recognition schemes, the first step is to locate and extract the position of a face in the whole scene. There are many methods to detect faces in the real world which were surveyed by Zafeiriou et al. in [29]. With aligned faces, there are many ways to recognize emotion through facial expression [38]. There are also researches focusing on voices or speech based emotion recognition [37]. When testing on real world data [8, 46], it seems that the previous works got lower results than that in lab controlled recognition. Since no feature descriptor can handle the problem alone, fusion method can be used to combine the discriminative abilities of multimodal features.

Previous works on fusion strategies can be broadly categorized into feature level fusion and decision level fusion. Feature level fusion aims to directly combine feature vectors by concatenation [8, 48] or kernel methods [33, 36, 39]. In recent years, multiple kernel learning (MKL) has been a popular fusion method at feature level. MKL aims at weighted combining the feature kernels for SVM. Gönen and Alpaydm

[7] surveyed quite a few kinds of multiple kernels methods for common classification problems. Bucak et al. reviewed the state-of-the-art for MKL in [41], including different formulations and algorithms for solving the related optimization problems, with the focus on their applications to object recognition. Sikka et al. [33] explored the use of general MKL for emotion recognition in the wild. Chen et al. [36] used a SimpleMKL method to combine visual and acoustic features. Huang et al. [39] used UFO-MKL for feature fusion. As combined feature is different from the original features, the discriminative ability can be further improved by a decision level fusion.

Decision level fusion combines the prediction scores of each single classifier. The advantage of decision level fusion is that it can combine different types of classifiers like logistic regression and SVM [34, 35]. Previous works usually conducted it by a single layer averaging [49] or weighted voting [28]. Kahou et al. proposed a voting matrix in [34] and used random search to tune weight parameters. Results showed that the voting matrix has lower results on testing set compared with that on validation set. There are also boosting methods for feature fusion: multiple kernel boosting (MKB) [40] and boosted trees [47], which may have high computation costs for retraining the basic classifiers.

3 Data preprocessing

3.1 Face alignment

We follow the face extraction and tracking method of Sikka et al. [33] and Dhall et al. [8]. First a mixture of tree structured part model [9] face detector is used to detect the position of face in the first frame of a video. Then the

Intraface toolkit used supervised descent method [10] to track facial landmarks of the rest frames in a Parameterized Appearance Model. The 49 landmark points can be used to align faces and generate geometric features for expression classification.

Through experiments, we choose the position of eyes and mouth as base landmarks: the first base point is the middle of two eyes; the second one is the central point of mouth. All frames are aligned to this base face through affine transformation and cut to 128×128 pixels.

3.2 Image purification

Due to the complex environment background of the AFEW datasets, the face alignment result is not very well. Then, we try to obtain a better set of face images by data purifying. We tried three methods: head frontalization, low-rank face decomposition, and removing bad images judged by principal component analysis projection.

Head frontalization method aims to get front view face images through the estimated head positions generated by Intraface. By rotating a face in 3D space and padding the side one, Hassner et al. [31] generated the front view for faces in the wild. The low-rank decomposition [30] method aims to remove the high-rank part of the face images which is considered as occlusion. By other means, Liu et al. [28] used PCA to remove the images that differed from others in a sequence by reconstruction error. By our experiments, using PCA test to purify images has some effect.

3.3 Audio feature extraction

The audio feature is computed by extracting features using the OpenSmile toolkit. OpenSmile, Speech & Music Interpretation by Large Space Extraction [24], is a fast and real-time audio feature extraction utility for automatic speech, music and paralinguistic recognition research. It is capable of extracting low-level descriptors (LLD) [25] and applying various filters, functions, and transformations to these descriptors. Delta regression coefficients can be computed from the low-level descriptors, and a moving average filter can be applied to smooth the feature contours. Next, functions (statistical, polynomial regression coefficients and transformations) can be applied to the low-level features.

The audio feature set we used consists of 34 energy and spectral related $LLD \times 21$ functions, 4 voicing related $LLD \times 19$ functions, 34 delta coefficients of energy and spectral $LLD \times 21$ functions, 4 delta coefficients of the voicing related $LLD \times 19$ functions and 2 voiced/unvoiced durational features [8]. The dimension of the audio feature is $34 \times 21 + 4 \times 19 + 34 \times 21 + 4 \times 19 + 2 = 1582$.

4 Multimodal visual features

4.1 Geometry feature

The geometric feature is based on the theory of emotion action unit defined by Ekman [32]. With it, Valstar et al. successfully detected AU detection and then achieved better result than some complex descriptors [42]. Similarly, in our method, the aligned 49 landmarks (Fig. 2) tracked by Intraface are spilt into three regions: left eye (6–10, 26–31), right eye (1–5, 20–25) and mouth (32–49). For every region, we compute the angles between three points and distances between two points. Then the positions of 49 points of this frame and the frame before are concatenated into the vector. At last, the distance of 49 landmarks to the mean central facial position is added to the geometric feature. The feature vector has a length of $71 + 98 \times 2 + 49 = 316$ at last. The geometry feature of a video is by taking the mean feature values of every frame.

4.2 Pyramid oriented gradients features

The pyramid of histograms of orientation gradients (PHOG) [19] consists of a histogram of orientation gradients over each image subregion at each resolution level, which is used for object detection. PHOG features are frequently used to describe the local spatial image shape and are used by Dhall et al. [21] for emotion recognition. The PHOG descriptor is based on two sources: (1) the HOG and (2) Spatial Pyramid Matching (SPM). We also use the PHOG features to obtain the statistical local shape information of a face.

Let G_h and G_v represent the gradients along the X and Y directions. Then, the gradient intensity and orientation are defined as M and θ :

$$M(x, y) = \sqrt{G_h(x, y)^2 + G_v(x, y)^2} \quad (1)$$

$$\theta(x, y) = \arctan(G_h(x, y)/G_v(x, y))$$

The HOG descriptor is then implemented by dividing the image window into small spatial cells. For each cell we accumulate a local 1-D histogram of gradient directions θ over the pixels of the cell. All histograms concatenate to be HOG feature.

The spatial layout follows the scheme proposed by Lazebnik et al. [16], which is based on spatial pyramid matching [20]. Each image is divided into a sequence of increasingly finer spatial grids by repeatedly doubling the number of divisions in each axis direction as shown in Fig. 3. Through this type of pyramid form, the overall HOG feature of the image has more representativeness.

The number of pyramids was set to four, and the bin number was eight. We divided the face into four levels: $1 \times 1, 2 \times 2,$

4×4 , and 8×8 and the final PHOG vector is a weighted concatenation of histograms for all levels [19]. The dimension is $(1^2 + 2^2 + 4^2 + 8^2) \times 8 = 680$.

4.3 Dynamic local texture features

4.3.1 LBP-TOP

LBP from three orthogonal planes (LBP-TOP) [22] are efficient representations of the dynamic-image texture, and has been successfully applied to facial expression recognition. LBP-TOP is widely used in ordinary texture analysis, which is calculated with Eq. (2). While, $I(O, N)$ means the bool comparison between a pixel O_p and its neighboring pixels N . The binary labels form a LBP. Then LBP-TOP feature is generated by concatenating LBPs on three orthogonal planes XY, XT and YT. The XY plane provides the spatial texture information, while the XT and YT planes provide information regarding the space-time transitions.

$$d = \sum_{O}^{XY, XT, YT} \sum_p \sum_{i=1}^k 2^{i-1} I(O_p, N_i) \tag{2}$$

The face images are divided into 4×4 blocks from whose LBP features are extracted and concatenated into an enhanced feature vector. In [4], Ojala et al. found that the vast majority of the LBP patterns in a local neighborhood are 59 “uniform patterns”. All features extracted from each block volume are connected into one vector to represent the appearance and motion of the facial expression sequence, whose dimension is $59 \times 3 \times 16 = 2832$.

4.3.2 LPQ-TOP

Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) [23] is a variant of LBP-TOP which is robust to blur. It bases on binary encoding of the phase information of the local Fourier transform at low frequency points and is an extension to the LPQ operator for spatial texture analysis.

The local Fourier transform is computed efficiently using 1-D convolutions for each dimension in a 3-D volume. The achieved features are reduced to a smaller dimension through PCA before a scalar quantization procedure. Finally, a histogram of all binary codewords from dynamic texture is formed. Similar to the LBP-TOP features, we first divide the dynamic volume into 16 blocks, then compute the LPQ-TOP features on each block and finally concatenated them together to form a LPQ-TOP feature, which is of $256 \times 3 \times 16 = 12,288$ dimension.

4.3.3 LGBP-TOP

Local Gabor Binary Patterns is a kind of descriptor that is robust to illumination changes and misalignment. It first takes a video frame convolved with a number of Gabor filters. Almaev and Valstar [43] proposed LGBP-TOP descriptor for AU detection. It is followed by the LBP feature extraction through the set of Gabor magnitude response images. The resulting binary patterns are histogrammed and concatenated into a single feature histogram. Like other volume local texture features, in our sense, a video is blockwisely to 4×4 to from XY plane. The LGBP-TOP feature with those three spatial frequencies and six Gaussian orientations is then extracted, and a length of $18 \times 4 \times 4 \times 59 \times 3 = 50,976$ feature is available.

4.4 Bag of multi-scale dense SIFT features

The Bag of Words (BoW) model based on multi-scale dense SIFT features (MSDF) [12] is used to extract visual features, which has shown remarkable performance on object recognition [13] and facial expression recognition [14]. The superiority of this approach lies in the combination of dense SIFT feature extraction, feature encoding using locality-constrained linear coding (LLC) [15] and spatial information fusion by multi-scale pooling over spatial pyramid [16]. LLC is known to be more robust to local spatial translations and captures more salient properties of visual patterns compared to the original simple histogram spatial encoding. Most importantly, a linear kernel SVM is sufficient to achieve good performance with LLC encoding, thus avoiding the computational expense of applying non-linear kernels [18] as in the case with spatial histogram based encoding [13, 16].

First, we extracted multi-scale dense SIFT features with a stride of two pixels. To retain more spatial information at different scales, we employed four different scales in MSDF extraction, defined by setting the width of the SIFT spatial bins to 4, 8, 12 and 16 pixels. The vocabulary for BoW is constructed by using approximate K-means clustering, which is based on calculating data-to-cluster distances using the approximate nearest neighbor (ANN) algorithm. For clustering, we used a feature subset extracted from 100 randomly selected images in the train samples. The size of the dictionary was set to 800 based on empirical experiments.

After the codebook of MSDF cluster centers was generated, LLC was used to encode the features and pool them together. LLC projects each descriptor to a local linear subspace spanned by some codewords by solving an optimization problem. Then, max pooling [17] was used to construct one single code for each region based on the maximization principle on each code dimension.

The traditional bag of words model is robust to spatial translations but sacrifices spatial layout information during

the histogram computing process. SPM is also combined to address this limitation. In our method, by experiments, each image was partitioned into $2^l \times 2^l$ segments at multiple scales $l = 1, 2, 4, 8, 16$. The BoW representation was then computed within each of these segments using LLC, and all of these LLC codes were concatenated into a single feature vector of $(1^2 + 2^2 + 4^2 + 8^2 + 16^2) \times 800 = 272,800$ dimensions.

5 Fusion feature classification

5.1 Support vector classifier

SVM is a supervised learning model with associated learning algorithms that analyze data and recognize patterns. It is used for classification and regression analysis [26]. SVM constructs a hyperplane in a high or infinite dimensional space, which can be used for classification. Given a training set of N data points $\{x_k, y_k\}$, $k = 1, \dots, N$, where $x_k \in \mathbb{R}^n$ is the k th input pattern and $y_k \in \mathbb{R}$ is the k th output pattern, the support vector method aims at constructing a classifier of the following form:

$$y(x) = \text{sign} \left[\sum_{k=1}^N a_k y_k k \psi(x, x_k) + b \right] \quad (3)$$

In multi-class classification, binary SVM classifiers (SVC) are built to distinguish between (i) one of the labels and the rest (OVR) or (ii) between every pair of classes (OVO). Classification of new instances for the one-versus-all case is performed by a winner-takes-all strategy in which the classifier with the highest output function assigns the class. For the one-versus-one approach, classification is performed by a max-wins voting strategy in which every classifier assigns the instance to one of the two classes. Then, the vote for the assigned class is increased by one vote. Finally, the class with the most votes determines the instance classification.

We use the LIBSVM [26] and LIBLINEAR [27], which are both popular open source machine learning libraries. For each feature listed above, we perform the SVM classifier followed by the OVO and OVR multi-class classifier strategy. For audio and MSDF BoW features, the linear kernel is used for classification. For the other features, we use a grid search to find the suitable cost parameter C and Gamma for the RBF kernels. The accuracies of those methods are shown in Tables 3 and 10.

5.2 Feature-level fusion multiple kernel classifier

Kernel methods have proven to be efficient for solving learning problems such as classification or regression. Feature-Level Fusion aims at directly combining the discriminative

ability of multiple feature kernel of SVM. It can be done in the ways of combined kernel or voted results. Recent applications have shown that using multiple kernels instead of a single kernel can enhance the interpretability of the decision function. Multiple kernel learning (MKL) [7] is a popular type of method for its ability of simultaneously learning kernels and can be used for feature-level fusion for different features. Multiple kernel boost [40] uses the idea of boosting to combine the classification result of single kernel SVMs, but it has a high computation cost for retraining a classifier.

For the MKL problem, a convenient approach is to consider that $K(x, x')$ is actually a convex combination of basis kernels:

$$K(x, x') = \sum_{m=1}^M d_m K_m(x, x'), \quad d_m \geq 0, \quad \sum_{m=1}^M d_m = 1$$

Here, M is the total number of kernels. Then, the form of classifier changes to:

$$y(x) = \text{sign} \left[\sum_{k=1}^N a_k y_k \sum_{m=1}^M \beta_m K_m(x, x_i) + b \right] \quad (4)$$

In our experiment, the RBF kernel we used is in the form of $K = \exp(-r|x - x_i|^2)$, x and x_i is the training data and testing data.

Different multiple kernel methods differ at combining method, functional form, target function, training method and base learners. We will compare those common-used nine multiple kernel methods in Sect. 6.3 on their fusion ability of combining multimodal features for emotion recognition in the wild.

5.3 Decision-level fusion hierarchical classifier

Single feature kernel SVM classifier has different results on each specific emotion. To make use of multiple features, we propose a hierarchical classifier to combine SVCs together. The proposed classifier is shown in Fig. 5 and consists of three levels: the feature layer, fusion layer and decision layer. Whole pipeline is showed in Fig. 4. For the fusion layer, the weight of each feature is represented as ρ , and for the decision layer, the weight of each fusion classifier is λ . Suppose I is the number of features, and J is the number for the fusion method in the second layer. Then, the proposed hierarchical classifier model is:

$$Y(x) = \text{sign} \left[\sum_{j=1}^J \lambda_j \sum_{i=1}^I \rho_i y_i(x) \right] \quad (5)$$

First all of the SVM classifiers with each feature make up the feature layer classification. For the testing set of N

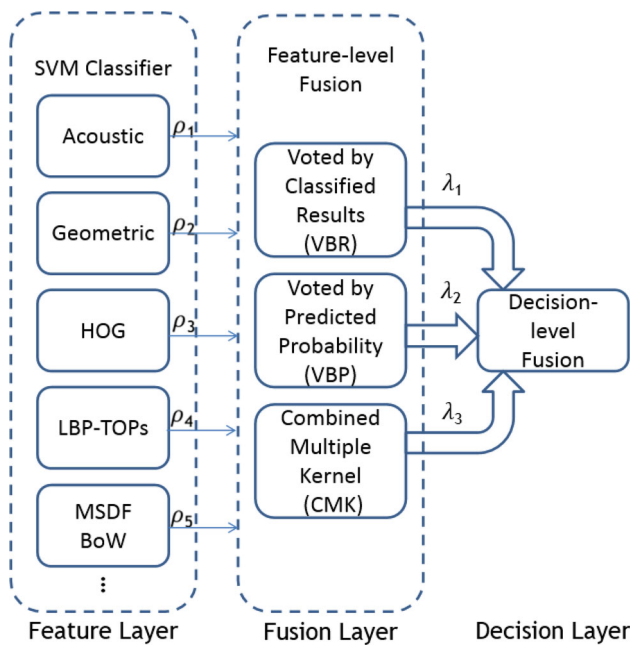


Fig. 5 Layers of hierarchical classifier

data points $x_k \in \mathbb{R}^n$, the i th SVM outputs, $\{x_k, y_k, p_k\}$, $k = 1..N$, where $y_k \in \{1, 2, \dots, 7\}$ is the predicted class label and p_k is the estimated prediction probability of the predicted label y_k .

On the fusion layer, we use three different feature-level fusion methods for voting classification. (1) The first one is voting based on classification results (VBR). We simply use parameter ρ_i as the voting weights of each feature-layer SVM classifier. Each SVC classifies the data $x_k \in \mathbb{R}^n$ to one emotion categories $y_k \in \{1, 2, \dots, 7\}$, and the highest voted emotion is the result of this one. Then, we tune the voting weights ρ to obtain the best results for the validation set. (2) The second fusion classifier is the same as the VBR classifier except that it is voted on by the prediction probability (VBP) estimated by the SVM classifier. Each feature layer classifier votes with the estimated probability p_k as its weight to class y_k . (3) The last fusion classifier is the SVM classifier with the combined multiple kernel (CMK). In our experiment, all kernels can be combined with the same weights after normalization as HCF2 in Sect. 6.3 or just with the best single kernel as HCF1. These second layer classifiers represent feature-level fusion.

At decision layer, λ_j is defined as the weight of the second layer fusion classifiers j and used to vote for each data $x_k \in \mathbb{R}^n$. We tune the parameters to find the best results by using the grid search strategy. The hierarchical structure allows us to tune parameters layer by layers. Through the three-layer fusion network, we can achieve the best recognition result.

In conclusion, the proposed classification framework combines both feature-level fusion and decision-level fusion.

Its advantage is that it combines the result of different fusion methods, which guarantees the robust accuracy of overall recognition. In the next section, we give the detailed experimental analysis of our method.

6 Experiments

6.1 Data preprocessing result

We first use IntraFace toolkit to get the positions of 49 facial landmarks for every video frame. Then, LBP-TOP feature and RBF kernel SVC are used to test the result of face alignment and image purification. Parameters of RBF kernel are tuned by grid search and five-fold cross validation.

The result of face alignment with different resolution (width \times height) and base points are listed in Table 1. Data of the FERA set [44] and the CK set [45] are also tested. Details are given in Sect. 6.4.

We set the threshold to 0.27 and remove the non-face or badly aligned face images on the EmotiW2014 training set similar to the ones in Fig. 6b. Trained with the purified data set, the accuracy of the LBP-TOP classifier is improved. But the Frontalization Method and RASL actually leads to a worse result as showed in Table 2. The samples are shown in Fig. 6.

6.2 Single feature classification

We extract the features listed in Sect. 4 and apply the SVM classification. Because each frame produces a Geometric, MSDF BoW and HOG feature vector, information from all frames of the video were combined using pooling [33], which was accomplished by taking the maximum or mean value of the MSDF BoW or PHOG feature vectors over all frames. By experiment, max pooling has better results for MSDF BoW and HOG, while the Geometric feature fits well with mean pooling. The classification accuracies and parameters are listed in Table 3. Because the character of LLC coding, we choose linear kernel SVM classifier for MSDF BoW feature. Actually RBF kernel is too complex with so large

Table 1 Face alignment results

Dataset	Resolution	Base points	SVC result (%)
AFEW	128 \times 128	L and R eyes	38.27
AFEW	128 \times 128	Eye and mouth	39.15
AFEW	128 \times 160	Eye and mouth	38.72
FERA	200 \times 200	L and R eyes	87.10
FERA	128 \times 128	Eye and mouth	87.10
FERA	128 \times 160	Eye and mouth	85.81
CK	60 \times 80	L and R eyes	87.16

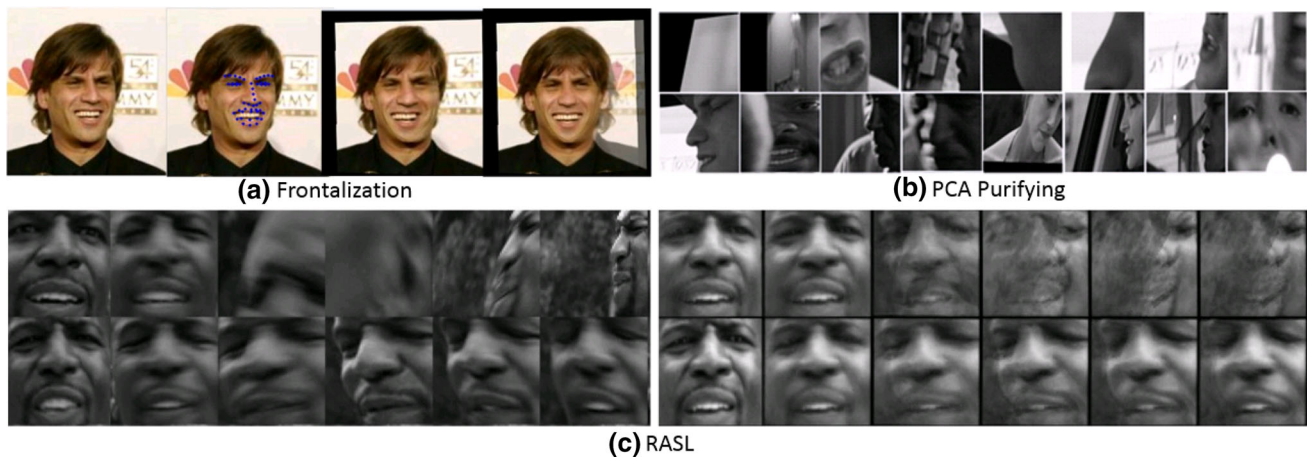


Fig. 6 **a** Head frontalization progress. **b** Samples with largest mean reconstruction error. **c** Sample images before (*left*) and after (*right*) RASL alignment

Table 2 Results of image purification on AFEW

Original	RASL	Frontalization	PCA
39.15 %	35.94 %	20.11 %	40.50 %

Table 3 Our single feature result in EmotiW challenge

Feature	OVR accuracy (%)	OVO accuracy	Kernels and parameters
MSDF-BoW	42.32	N/A	Linear, $c = 0.9$
PHOG	31.27	30.73 %	RBF, $c = 128, g = 8$
LBP-TOP	36.12	33.15 %	RBF, $c = 1.2, g = 2.8$
LPQ-TOP	19.41	19.68 %	RBF, $c = 2, g = 8$
Audio	26.68	16.98 %	Linear, $c = 2$

dimensions. For the other features except audio feature, RBF kernels achieve better results while consuming more training time.

The results on the validation set are listed in Fig. 7. Here, we observe that Angry, Happy and Neutral are relatively easy to distinguish. In addition, those features' abilities to classify specific emotion is quite different.

Obviously, the bag of multi-scale dense SIFT features has the highest validation result. Take note that we do not have the labels of the AFEW test set. All of the training progress is performed on the training set, and we tune the parameters through 5-fold cross validation and test the trained model on validation set.

6.3 Results of fusion classification

Observing the classification accuracy of each single feature SVM classifier in Fig. 7, we find that those features show

different results on each specific emotion [34]. To make use of multiple features, we propose a hierarchical classification framework (HCF) to combine the results of each SVM classifier.

Then our proposed HCF is performed on AFEW validation set. Mean weights of each feature are listed in Tables 4 and 5. The difference between HCF1 and HCF2 is the weight in combined multiple kernel (CMK). HCF1 only uses MSDF BoW kernel while HCF2 use the mean of all kernels. After testing HCF1 on the validation set, we obtained an accuracy of 45.55 % with decision level weights λ_1 (VBR) = 0.3, λ_2 (VBP) = 0.3 and λ_3 (CMK) = 0.4. Detailed weight parameters of HCF1 and HCF2 are listed in Tables 4 and 5. Note that although the Audio feature did not perform so well in SVM classification, it has a large scale of weight. This shows that acoustic feature has some information that is complementary to the homogeneous visual features. With the proposed method, in August 2014, we submitted our HCF1 prediction of test set and got the final result of 47.17 %, which make us the first runner-up of EmotiW 2014 Challenge.

We also compare our HCFs to eight representatives multiple kernel methods by using Gönen and Alpaydın's implement in [7]. We changed their code to suit multi-class classification with One-VS-Rest strategy. The methods we compare are: MSDF Single Kernel SVM, alignment-based ABMKL, group Lasso-based GLMKL, generalized GMKL, localized LMKL, original MKL, nonlinear NLMKL, rule-based RBMKL and SimpleMKL. For comparison, we all use LIBSVM for their SVM solver. The kernels for each features are the same as in Sect. 6.2. Grid searching is used to get best training cost parameters of each method. The comparing results are listed in Table 6.

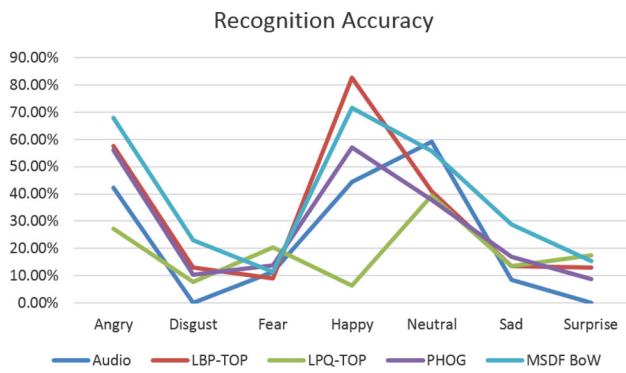


Fig. 7 Single feature classification accuracies of each emotions in AFEW validation set

6.4 Results of proposed method on different datasets

We also test our HCF method on other datasets: the Cohn Kanade [45] (CK) database and the facial expression recognition and analysis (FERA) [44] database. CK set consist of 327 image sequences of seven kind of emotion. We use half

of the sequences for training, 1/4 for validation and the rest for testing. FERA set has 155 training videos. We also divide the training set by half for training and validation. Though they are not specific dataset for emotion recognition in the wild, we can still use them to test the generalized ability of proposed method. FERA set and AFEW set are tracked and aligned in the same way, while CK set is aligned by manually aligning eyes for first image of a sequence.

As our LPQ-TOP feature has not work well in EmotiW 2014 set, we bring in LGBP-TOP and Geometric features to

Table 7 HCF Result on different datasets

Dataset	Baseline	Best SVC (%)	Validation result (%)	Test result (%)
AFEW	33.2 %	42.32	45.82	47.2
FERA	55.6 %	83.12	90.91	N/A
CK	N/A	82.93	89.02	85.19

Table 4 Weights ρ_i of features in our hierarchical classification framework (HCF1) on AFEW

λ_j	Audio	MSDF BoW	LBP-TOP	LPQ-TOP	PHOG
VBR, $\lambda_1 = 0.3$	0.375	0.375	0.042	0.083	0.125
VBP, $\lambda_2 = 0.3$	0.467	0.400	0.033	0.016	0.083
CMK, $\lambda_3 = 0.4$	0	1	0	0	0
Mean weight	0.2526	0.6325	0.0260	0.0315	0.0574

Table 5 Weights ρ_i of features in HCF2 on AFEW

λ_j	Audio	MSDF BoW	LBP-TOP	LPQ-TOP	PHOG
VBR, $\lambda_1 = 0.4$	0.375	0.375	0.042	0.083	0.125
VB1, $\lambda_2 = 0.2$	0.467	0.400	0.033	0.016	0.083
CMK, $\lambda_3 = 0.4$	0.2	0.2	0.2	0.2	0.2
Mean weight	0.3234	0.31	0.1035	0.1165	0.0574

Table 6 Comparing between different multiple kernel methods

Method	Mean kernel weights					Training cost C	Accuracy (%)
	Audio	MSDF BoW	LBP-TOP	LPQ-TOP	PHOG		
Single SVM	0	1	0	0	0	0.9	42.32
ABMKL	0.0532	0.3083	0.0975	0.4435	0.0975	2	38.27
GLMKL	0.0527	0.2344	0.3349	0.0430	0.3349	16	43.94
GMKL	0.0595	0.1983	0.3427	0.0128	0.3867	4	45.55
LMKL	N/A	N/A	N/A	N/A	N/A	0.25	34.50
MKL	0.0424	0.1884	0.3781	0.0130	0.3781	2	43.94
NLMKL	0.2041	0.2028	0.1996	0.1940	0.1996	0.125	26.95
RBMKL	0.2	0.2	0.2	0.2	0.2	2	42.32
SimpleMKL	0.0459	0.1981	0.3456	0.0132	0.3971	4	45.01
HCF1	0.2526	0.6325	0.0260	0.0315	0.0574	N/A	45.55
HCF2	0.3234	0.31	0.1035	0.1165	0.1466	N/A	45.82

Table 8 Weights ρ_i of features in HCF on CK dataset

λ_j	MSDF BoW	LBP-TOP	LGBP-TOP	HOG	PHOG
VBR, $\lambda_1 = 0.3$	0.2	0.3	0.3	0	0.2
VBP, $\lambda_2 = 0.4$	0.1	0.6	0.2	0.1	0
CMK, $\lambda_3 = 0.3$	0.2	0.2	0.2	0.2	0.2
Mean weight	0.16	0.39	0.23	0.10	0.12

Table 9 Weights ρ_i of features in HCF on FERA dataset

λ_j	MSDF BoW	LBP-TOP	LGBP-TOP	HOG	PHOG	Geometry
VBR, $\lambda_1 = 0.3$	0.2	0.2	0.1	0.1	0.2	0.2
VBP, $\lambda_2 = 0.4$	0.1	0.2	0.2	0.1	0.1	0.3
CMK, $\lambda_3 = 0.3$	0	0	0	0	0	1
Mean weight	0.10	0.14	0.11	0.07	0.10	0.48

Table 10 Cross validation result of training sets

Feature	AFEW		FERA		CK		Multi-class strategy
	Accuracy (%)	Kernels and parameters	Accuracy	Kernels and parameters	Accuracy	Kernels and parameters	
Audio	32.68	Linear, $c = 2$	N/A	N/A	N/A	N/A	OVR
Geometry	42.04	RBF, $c = 131072$, $g = 1.86e-9$	80.00 %	RBF, $c = 2048$, $g = 4.77e-7$	N/A	N/A	OVO
HOG	37.74	RBF, $c = 2$, $g = 0.5$	78.71 %	RBF, $c = 8$, $g = 2$	70.64 %	RBF, $c = 8$, $g = 0.125$	OVO
PHOG	35.27	RBF, $c = 120$, $g = 8$	76.12 %	RBF, $c = 2048$, $g = 0.5$	71.25 %	RBF, $c = 128$, $g = 8$	OVO
LBP-TOP	39.15	RBF, $c = 8$, $g = 0.125$	87.1 %	RBF, $c = 2048$, $g = 0.031$	87.16 %	RBF, $c = 128$, $g = 0.031$	OVO
LGBP-TOP	37.04	RBF, $c = 8$, $g = 0.031$	74.84 %	RBF, $c = 32$, $g = 0.031$	81.35 %	RBF, $c = 2048$, $g = 1.22e-4$	OVO
MSDF BoW	46.96	Linear, $c = 1$	90.67 %	Linear, $c = 4$	83.44 %	Linear, $c = 16$	OVR

replace. Note that CK set is consisting of grey images; the Intraface video tracking method is unsuitable.

We list the results in Table 7. We can see our fusion method achieve improvements on both sets, comparing to the best single kernel SVC. Parameters on those sets are shown in Tables 8 and 9.

6.5 Discussion

Result of data preprocessing show that adding image pixels to gain more information or using complex methods to purify images do not have much improvement on classification. The large size of images may add more noise to disturb the feature extraction progress on the contrary. Data purifying by takeout bad images have some effect, but could not solve the misalignment problem thoroughly, which is still one of the most challenging part for emotion recognition.

Single feature classification shows that no single feature SVC can achieve convincing result for emotion recognition

Table 11 Accuracies on validation set of datasets

Features	AFEW (%)	FERA (%)	CK (%)
Audio	26.68	N/A	N/A
Geometry	29.30	76.62	N/A
HOG	29.38	80.52	64.63
PHOG	31.27	74.03	63.41
LBP-TOP	36.12	81.82	82.93
LGBP-TOP	28.57	79.22	80.49
MSDF BoW	42.32	83.12	78.05

in the wild. MSDF BoW and LBP-TOP are most robust features for all three datasets. On CK set, LBP-TOP beats MSDF BoW; this is caused by the low image size that limits the amount of Dense SIFT descriptors. Another interesting phenomenon is that PHOG is no better than HOG on some situations in Tables 10 and 11. The concatenation way of SPM feature may need to be replaced by multi kernel methods.

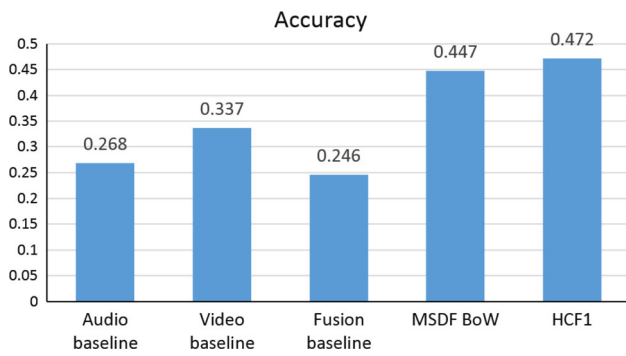


Fig. 8 Submission results on test set

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.74	0.00	0.00	0.03	0.12	0.09	0.02
Di	0.12	0.19	0.04	0.27	0.35	0.04	0.00
Fe	0.24	0.02	0.17	0.07	0.24	0.24	0.02
Ha	0.12	0.04	0.02	0.59	0.11	0.11	0.00
Ne	0.12	0.03	0.03	0.04	0.65	0.13	0.00
Sa	0.11	0.15	0.02	0.21	0.28	0.23	0.00
Su	0.15	0.00	0.12	0.12	0.38	0.23	0.00

Fig. 9 Confusion matrix of our final result on AFEW test set

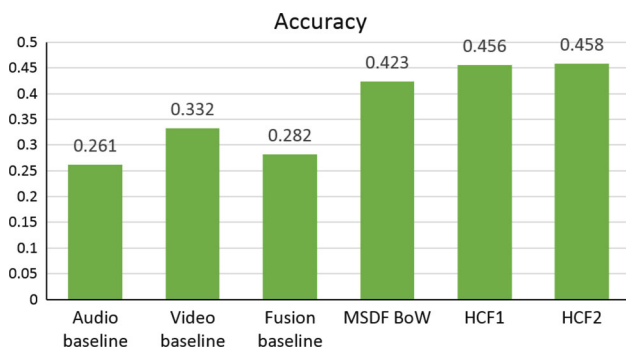


Fig. 10 Final experimental results on validation set

Due to the different discriminate ability of those features, we propose the hierarchical fusion framework to combine them. Comparison results in Table 6 show that our HCF achieves the highest result among the multiple kernel methods we have reviewed. Actually, the MKL algorithms can be used to replace our simply combined kernel SVC in feature level. Our fusion method shows robust result on all three datasets. Results and confusion matrices of them on AFEW validation and test sets are showed in Figs. 8, 9, 10, 11. HCF2 only got result on validation set. As acoustic feature has large weight in many fusion methods, dig more information through audio may be a way to improve the whole recognition result. Another indication is our HCF improves more when single feature has better recognition rates. For better recognition result, we should continue to improve the result of single feature SVC.

	An	Di	Fe	Ha	Ne	Sa	Su
An	0.81	0.03	0.03	0.03	0.05	0.02	0.02
Di	0.31	0.13	0.03	0.26	0.18	0.08	0.03
Fe	0.45	0.05	0.09	0.18	0.09	0.11	0.02
Ha	0.08	0.00	0.02	0.84	0.05	0.02	0.00
Ne	0.10	0.03	0.02	0.25	0.52	0.05	0.03
Sa	0.24	0.08	0.02	0.24	0.20	0.19	0.03
Su	0.30	0.09	0.24	0.13	0.15	0.04	0.04

Fig. 11 Confusion matrix of our final result on AFEW validation set

7 Conclusions

In this paper, we present several image sequence feature extraction methods and investigate different features and classifiers for multimodal human emotion recognition in the real world. For each feature, we trained individual classifiers on the train data. The experimental results show that different features have different discriminative abilities for emotion classification. We propose the feature-level fusion strategy to combine different features, which significantly improves the emotion recognition performance. Then, we propose a novel decision-level fusion hierarchical classifier framework, which combines these single feature classifiers and combined multiple kernel classifier results using adaptive weighted parameters to further improve the multimodal recognition performance. The method is evaluated on several datasets and gains very promising achievements on the validation set and test set.

In the future, we will try to conduct more experiments on more public benchmark databases to obtain more detailed performance analysis on our work. We will do more research on feature level to get better single kernel result.

Acknowledgments This work is supported by the Fundamental Research Funds for the Central Universities of China (2014KJJC15, 2012YBXS10) and the National Education Science Twelfth Five-Year Plan Key Issues of the Ministry of Education (DCA140229).

References

- Knapp M, Hall J, Horgan T (2013) Nonverbal communication in human interaction. Cengage Learning, Oklahoma
- Pantic M, Rothkrantz LJM (2000) Automatic analysis of facial expressions: the state of the art. Pattern Anal Mach Intell IEEE Trans 22(12):1424–1445
- Wu T, Bartlett MS, Movellan JR (2010) Facial expression recognition using Gabor motion energy filters. In: Computer vision and pattern recognition workshops (CVPRW), 2010 IEEE computer society conference on IEEE, pp 42–47
- Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Anal Mach Intell IEEE Trans 24(7):971–987
- Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: Computer vision and pattern recognition.

- CVPR 2005. IEEE computer society conference, vol 1. IEEE, pp 886–893
6. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. *IEEE Trans Pattern Anal Mach Intell* 23(6):681–685
 7. Gönen M, Alpaydın E (2011) Multiple kernel learning algorithms. *J Mach Learn Res* 12:2211–2268
 8. Dhall A, Goecke R, Joshi J, Sikka K, Gedeon T (2014) Emotion recognition in the wild challenge 2014: Baseline, data and protocol. In: *Proceedings of the 16th international conference on multimodal interaction*. ACM, pp 461–466
 9. Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on IEEE*, pp 2879–2886
 10. Xiong X, De la Torre F (2013) Supervised descent method and its applications to face alignment. In: *Computer vision and pattern recognition (CVPR), IEEE conference on IEEE*, pp 532–539
 11. Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimed* 19(3):34–41
 12. Vedaldi A, Fulkerson B (2010) VLFeat: an open and portable library of computer vision algorithms. In: *Proceedings of the international conference on multimedia*. ACM, pp 1469–1472
 13. Zhang J, Marszałek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: a comprehensive study. *Int J Comput Vis* 73(2):213–238
 14. Sikka K, Wu T, Susskind J, Bartlett M (2012) Exploring bag of words architectures in the facial expression domain. In: *Computer vision-ECCV 2012. Workshops and demonstrations*. Springer, Berlin, pp 250–259
 15. Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y (2010) Locality-constrained linear coding for image classification. In: *Computer Vision and pattern recognition (CVPR), 2010 IEEE conference on IEEE*, pp 3360–3367
 16. Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *Computer vision and pattern recognition, IEEE computer society conference on IEEE*, vol. 2, pp 2169–2178
 17. Yang J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: *Computer vision and pattern recognition, CVPR 2009. IEEE conference on IEEE*, pp 1794–1801
 18. Chatfield K, Lempitsky V, Vedaldi A, Zisserman A (2011) The devil is in the details: an evaluation of recent feature encoding methods. *BMVC* 2(4):239–259
 19. Bosch A, Zisserman A, Munoz X (2007) Representing shape with a spatial pyramid kernel. In: *Proceedings of the 6th ACM international conference on image and video retrieval*. ACM, pp 401–408
 20. Grauman K, Darrell T (2005) The pyramid match kernel: discriminative classification with sets of image features. In: *Computer vision, ICCV 2005. Tenth IEEE international conference on IEEE*, vol. 2, pp 1458–1465
 21. Dhall A, Asthana A, Goecke R, Gedeon T (2011) Emotion recognition using PHOG and LPQ features. In: *Automatic face & gesture recognition and workshops (FG 2011), IEEE international conference on IEEE*, pp 878–883
 22. Zhao G, Pietikainen M (2007) Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Anal Mach Intell IEEE Trans* 29(6):915–928
 23. Päivärinta J, Rahtu E, Heikkilä J (2011) Volume local phase quantization for blur-insensitive dynamic texture classification. In: *Image analysis*. Springer, Berlin, pp 360–369
 24. Eyben F, Wöllmer M, Schuller B (2010) Opensmile: the Munich versatile and fast open-source audio feature extractor. In: *Proceedings of the international conference on multimedia*. ACM, pp 1459–1462
 25. Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu X, Woodland P (2006) The HTK book (for HTK version 3.4). *Camb Univ Eng Dep* 2(2):2–3
 26. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol (TIST)* 2(3):27
 27. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 9:1871–1874
 28. Liu M, Wang R, Huang Z, Shan S, Chen X (2013) Partial least squares regression on grassmannian manifold for emotion recognition. In: *Proceedings of the 15th ACM on international conference on multimodal interaction*. ACM, pp 525–530
 29. Zafeiriou S, Zhang C, Zhang Z (2015) A survey on face detection in the wild: past, present and future. *Comput Vis Image Underst* 138:1–24
 30. Peng Y, Ganesh A, Wright J, Xu W, Ma Y (2012) RASL: robust alignment by sparse and low-rank decomposition for linearly correlated images. *Pattern Anal Mach Intell IEEE Trans* 34(11):2233–2246
 31. Hassner T, Harel S, Paz E, Enbar R (2014) Effective face frontalization in unconstrained images. Preprint [arXiv:1411.7964](https://arxiv.org/abs/1411.7964)
 32. Ekman P, Friesen WV (1977) Facial action coding system. In: Blacking J (ed) *Anthropology of the body*. Academic Press, New York
 33. Sikka K, Dykstra K, Sathyanarayana S, Littlewort G, Bartlett M (2013) Multiple kernel learning for emotion recognition in the wild. In: *Proceedings of the 15th ACM on international conference on multimodal interaction*. ACM, pp 517–524
 34. Kahou SE, Pal C, Bouthillier X, Froumenty P, Gülçehre Ç, Memisevic R, Wu Z (2013) Combining modality specific deep neural networks for emotion recognition in video. In: *Proceedings of the 15th ACM on international conference on multimodal interaction*. ACM, pp 543–550
 35. Liu M, Wang R, Li S, Shan S, Huang Z, Chen X (2014) Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In: *Proceedings of the 16th international conference on multimodal interaction*. ACM, pp 494–501
 36. Chen J, Chen Z, Chi Z, Fu H (2014) Emotion recognition in the wild with feature fusion and multiple kernel learning. In: *Proceedings of the 16th international conference on multimodal interaction*. ACM, pp 508–513
 37. El Ayadi M, Kamel MS, Karray F (2011) Survey on speech emotion recognition: features, classification schemes, and databases. *Pattern Recognit* 44(3):572–587
 38. De la Torre F, Cohn JF (2011) Facial expression analysis. In: Moeslund TB, Hilton A, Krüger V, Sigal L (eds) *Visual analysis of humans*. Springer, London, pp 377–409
 39. Huang X, He Q, Hong X, Zhao G, Pietikainen M (2014) Improved spatiotemporal local monogenic binary pattern for emotion recognition in the wild. In: *Proceedings of the 16th international conference on multimodal interaction*. ACM, pp 514–520
 40. Xia H, Hoi SC (2013) Mkboost: a framework of multiple kernel boosting. *Knowl Data Eng IEEE Trans* 25(7):1574–1586
 41. Bucak SS, Jin R, Jain AK (2014) Multiple kernel learning for visual object recognition: a review. *Pattern Anal Mach Intell IEEE Trans* 36(7):1354–1369
 42. Valstar M, Girard J, Almaev T, McKeown G, Mehu M, Yin L, Cohn J (2015) Fera 2015-second facial expression recognition and analysis challenge. *Proceeding of the IEEE ICFG*
 43. Almaev TR, Valstar MF (2013) Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In: *Affective computing and intelligent interaction (ACII), humaine association conference on IEEE*, pp 356–361
 44. Valstar MF, Jiang B, Mehu M, Pantic M, Scherer K (2011) The first facial expression recognition and analysis challenge. In: *Automatic*

- face & gesture recognition and workshops (FG 2011), IEEE international conference on IEEE, pp 921–926
45. Tian YL, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. *Pattern Anal Mach Intell IEEE Trans* 23(2):97–115
 46. Sun B, Li L, Zuo T, Chen Y, Zhou G, Wu X (2014) Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild. In: *Proceedings of the 16th international conference on multimodal interaction*. ACM, pp 481–486
 47. Day M (2013) Emotion recognition with boosted tree classifiers. In: *Proceedings of the 15th ACM on international conference on multimodal interaction*. ACM, pp 531–534
 48. Tariq U, Lin KH, Li Z, Zhou X, Wang Z, Le V, Han TX (2011) Emotion recognition from an ensemble of features. In: *Automatic face & gesture recognition and workshops (FG 2011), IEEE international conference on IEEE*, pp 872–877
 49. Meudt S, Zharkov D, Kächele M, Schwenker F (2013) Multi classifier systems and forward backward feature selection algorithms to classify emotional coloured speech. In: *Proceedings of the 15th ACM on international conference on multimodal interaction*. ACM, pp 551–556