J Multimodal User Interfaces (2016) 10:139–149
DOI 10.1007/s12193-015-0175-6

ORIGINAL PAPER

Combining modality-specific extreme learning machines for emotion recognition in the wild

Heysem Kaya[1] · **Albert Ali Salah**[1]

Received: 30 January 2015 / Accepted: 21 April 2015 / Published online: 1 May 2015

Abstract This paper proposes extreme learning machines (ELM) for modeling audio and video features for emotion recognition under uncontrolled conditions. The ELM paradigm is a fast and accurate learning alternative for single layer Feedforward networks. We experiment on the acted facial expressions in the wild corpus, which features seven discrete emotions, and adhere to the EmotiW 2014 challenge protocols. In our study, better results for both modalities are obtained with kernel ELM compared to basic ELM. We contrast several fusion approaches and reach a test set accuracy of 50.12 % (over a video-only baseline of 33.70 %) on the seven-class (i.e. six basic emotions plus neutral) EmotiW 2014 Challenge, by combining one audio and three video sub-systems. We also compare ELM with partial least squares regression based classification that is used in the top performing system of EmotiW 2014, and discuss the advantages of both approaches.

Keywords Audio-visual emotion corpus · Audio-visual fusion · Feature extraction · Emotion recognition in the wild · Extreme learning machines

1 Introduction

Emotion recognition from video and audio is gaining increasing attention, especially because its outputs can be used in many related domains [5]. In the last decade, a considerable amount of research efforts spent in the field was on controlled, laboratory-condition data. In some of such corpora (e.g. Berlin emotional speech database [4]) it was possible to obtain classification scores even better than human perception [34]. Now the field is moving on to less controlled conditions, including noisy audio-visual background, large variance in facial appearance and spoken content.

Audio-visual emotion related challenges have been instrumental in improving the state-of-the-art in this field. The challenges provide a great opportunity for the researchers in the field and help advance the state-of-the-art by bringing together experts from different disciplines, such as signal processing and psychology. One such challenge series is emotion recognition in the wild (EmotiW) [7,8] that provides out of laboratory data—acted facial expression wild (AFEW)—collected from videos that mimic real life [6].

In this paper, we apply a powerful classification paradigm, extreme learning machines (ELM), to audio-visual emotion recognition. We investigate feature/group selection in both modalities to enhance generalization of learned models. We further extract audio features using the most recent INTERSPEECH Computational Paralinguistic Challenge baseline set [37] with the freely available openSMILE tool [10] and augment the AFEW dataset with four other publicly available emotional corpora: Berlin EMODB [4], Danish emotion database (DES) [9], eNTERFACE Database [29], and the Turkish emotional database (BUEMODB) [23], respectively.

Further contributions of this paper are as follows. In addition to the baseline feature sets, we use new visual feature types and compare ELMs with a Partial Least Squares based classifier, which yields the best performance in the state-of-the-art system on the EmotiW 2014 Challenge [25]. We extract dense SIFT features from images, representing the videos (image sets) using a linear subspace obtained via singular value decomposition, the data covariance matrix and

✉ Heysem Kaya
heysem@boun.edu.tr

Albert Ali Salah
salah@boun.edu.tr

[1] Department of Computer Engineering, Boğaziçi University, 34342, Bebek, Istanbul, Turkey

the distribution statistics (assuming a normal distribution), all of which lie on Riemannian manifolds. In line with [25], Riemannian kernels are used in classifiers. We also extract video features using local Gabor binary patterns from three orthogonal planes (LGBP-TOP), which is shown to be less sensitive to registration errors compared to LGBP and LBP-TOP [1]. We finally use a weighted score fusion strategy, searching for the optimal weights in a pool of randomly generated fusion matrices. This combination of modality-specific models boosts the accuracy of individual models.

The remainder of this paper is organized as follows. In the next section we provide background on ELM. Then in Sect. 3 we overview the corpora and describe our proposed approach. In Sect. 4 we give experimental results, and conclude in Sect. 5.

2 Background: extreme learning machines

The extreme learning machine (ELM) classifier was first introduced in [14] as a fast alternative training method for single layer Feedforward networks (SLFNs). The rigorous theory of the ELM paradigm is presented in 2006 by Huang et al. [15], where the authors compare the performance of ELM, SVM, and back propagation (BP) learning based SLFN in terms of training time and accuracy. The basic ELM paradigm has matured over the years to provide a unified framework for regression and classification, related to generalized SLFN class including least square SVM (LSSVM) [16,39].

Despite the speed and accuracy of ELMs, they were only recently employed in affective computing exhibiting outstanding performance with typically undersampled, high dimensional datasets [12,22]. In one of the recent studies, Han et al. [12] use deep neural networks (DNN) for extraction of higher level features (class distribution) from segment based acoustic descriptors, then summarize these features over the utterances using simple statistical functionals (e.g. mean, max). The suprasegmental features were stacked as input to ELMs. They show that ELM based systems outperform both SVM and HMM based systems.

The argument of the basic ELM introduced by Huang et al. [15] is that the first layer (input layer) weights and biases of a neural network classifier do not depend on data and can be randomly generated, whereas the second layer (output weights) can be effectively and efficiently solved via least squares. The input layer can be considered as carrying out unsupervised feature mapping, and the activation function outputs (the output matrix) are subjected to a supervised learning procedure. Let $(\mathbf{W}, \mathbf{b}, \mathbf{H}, \beta)$ denote an SLFN, where the output with respect to input $\mathbf{x} \in \mathbb{R}^d$ is given as $\hat{\mathbf{y}} = \mathbf{h}(\mathbf{W}\mathbf{x} + \mathbf{b})\beta$. Here, $\mathbf{W}$ and $\mathbf{b}$ denote the randomly generated mapping matrix, and the bias vector, respectively. $\mathbf{h}(\mathbf{x}) \in \mathbb{R}^p$ denotes the hidden node output and $\mathbf{H} \in \mathbb{R}^{N \times p}$ denotes the hidden node output matrix. β is the analytically learned second layer weight matrix.

The nonlinear activation function $h()$ can be any infinitely differentiable bounded function. A common choice for $h()$ is the sigmoid function:

$$h(\mathbf{a}) = \frac{1}{1 + exp\,(-\mathbf{a})}. \tag{1}$$

ELM proposes unsupervised, even random generation of the hidden node output matrix $\mathbf{H}$. The actual learning takes place in the second layer between $\mathbf{H}$ and the label matrix $\mathbf{T}$. $\mathbf{T}$ is composed of continuous annotations in case of regression, therefore is a vector. In the case of K-class classification, $\mathbf{T}$ is represented in one vs. all coding:

$$\mathbf{T}_{t,k} = \begin{cases} +1 & \text{if } \; y^t = k, \\ -1 & \text{if } \; y^t \neq k. \end{cases} \tag{2}$$

The second level weights β are learned by least squares solution to a set of linear equations $\mathbf{H}\beta = \mathbf{T}$. Proving first that random projections and nonlinear mapping with $L \leq N$ result in a full rank $\mathbf{H}$, the output weights can be learned via:

$$\beta = \mathbf{H}^{\dagger}\mathbf{T}, \tag{3}$$

where $\mathbf{H}^{\dagger}$ is the Moore–Penrose generalized inverse [33] that gives the minimum L_2 norm solution to $\|\mathbf{H}\beta - \mathbf{T}\|$, simultaneously minimizing the norm of $\|\beta\|$. It is important to mention that ELM is related to Least Square SVMs via the following output weight learning formulation:

$$\beta = \mathbf{H}^T \left(\frac{\mathbf{I}}{C} + \mathbf{H}\mathbf{H}^T \right)^{-1} \mathbf{T}, \tag{4}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix, and C, which is used to regularize the linear kernel $\mathbf{H}\mathbf{H}^T$, is indeed the complexity parameter of LSSVM [39]. The approach is extended to use any valid kernel. A popular choice for kernel function is Gaussian (RBF):

$$K(\mathbf{x_k}, \mathbf{x_l}) = \phi(\mathbf{x_k}) \cdot \phi(\mathbf{x_l}) = exp\left(-\frac{\|\mathbf{x_k} - \mathbf{x_l}\|}{\sigma^2} \right) \tag{5}$$

In both (basic and kernel) approaches, the prediction of $\mathbf{x}$ is given via $\hat{\mathbf{y}} = \mathbf{h}(\mathbf{x})\beta$. In case of multi-class classification, the class with maximum score in $\hat{\mathbf{y}}$ is selected. In our study, we utilize both basic and kernel versions of ELM. Inspired by the success of SLFN based auto-encoders for feature enhancement and the relationship between principal component analysis (PCA) and SLFNs [2], in this study, we further consider the use of PCA instead of random generation of input weights.

3 The corpus and features

3.1 The AFEW corpus

The EmotiW 2014 Challenge [7] uses the AFEW 4.0 database, which is an extended version of AFEW 3.0 used in the EmotiW 2013 Challenge [8]. The corpus contains videos that are clipped from movies with the guidance of emotion related keywords in the movie script for the visually impaired [6]. The 2014 challenge provides a total of 1368 video clips collected from movies, representing close-to-real-world conditions [6]. The challenge is a seven-class classification problem, where the video is assigned to one of A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), and SU(rprise) classes. The corpus is partitioned into training, development and test sets. The challenge participants are expected to develop their system using the first two sets and send their predictions for the test set, whose labels are sequestered.

The nature of the data collection poses challenging conditions, e.g. in terms of background noise, head pose and illumination. In most of the clips there is a single actor. However, in some cases there are multiple actors on the scene. Speech is generally accompanied by the background music and noise. While the challenge annotations come with a standard face detector result, the difficult conditions cause problems even in the early stages of the processing pipeline. Some example aligned images illustrating this problem are shown in Fig. 1. We observe that in addition to precisely aligned frontal faces, there are misaligned or occluded faces, or images that do not contain faces.

3.2 Baseline feature sets

The baseline video features consist of local binary patterns from three orthogonal planes (LBP-TOP), compacted via uniform LBP [31] extracted from the detected and aligned faces in the videos. After face alignment and conversion to gray scale, LBP computation amounts to finding the sign of difference with respect to a central pixel in a neighborhood, transforming the binary pattern into an integer and finally converting the patterns into a histogram. Uniform LBP maps the patterns into 59 bins, and takes into account occurrence statistics of common patterns [31]. To add structural information to the histogram representation, the face is divided into non-overlapping $4 \times 4 = 16$ regions and an LBP histogram is computed per region. The TOP extension applies the relevant descriptor on XY, XT and YT planes (T represents time) independently and concatenates the resulting histograms. In total, we have $59 \times 3 = 177$ dimensional visual descriptors per region.

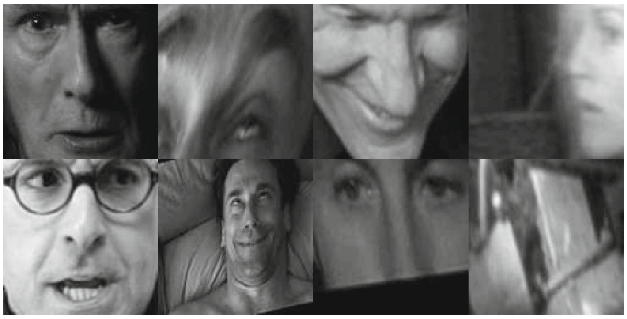


Fig. 1 Illustration of aligned images with varying conditions

The baseline audio features are extracted via freely available openSMILE tool [10] using INTERSPEECH 2010 Paralinguistic challenge baseline set [36]. The 1582 dimensional feature set covers a range of popular low level descriptors such as fundamental frequency (F0), Mel-frequency cepstral coefficients (MFCC) [0–14], line spectral Pairs frequency (LSF) [0–7] mapped to a fixed length feature vector by means of functionals such as arithmetic mean and extrema.

MFCC features correspond to Inverse Fourier Transform or preferably the discrete cosine transform (DCT) of the log of the Mel-scaled Fourier transform of the speech signal. Mel-scale mimics the human hearing capabilities in the way that it allows discriminating lower frequencies better than the higher frequencies.

LSF feature representation of speech is proposed by Itakura [17] and provides efficient and robust estimation of formant frequencies. Formant frequencies, especially the first two, are known to carry affect related information.

3.3 Exracted features

In our experiments, we use the aligned faces provided by the challenge organizers for visual signal processing. The images are first resized to 64×64 pixels. In the preprocessing step, we use PCA based data purification as shown to be effective in [24,38]. The idea is to measure the mean reconstruction error per image $x_i \in \mathbb{R}^D$ with $Err_i = \frac{1}{D}\|(x_i - \mu) - W_{pca}^T W_{pca}(x_i - \mu)\|$, where $\mu \in \mathbb{R}^D$ is the training set mean vector, and W_{pca} is the reduced PCA projection coefficient matrix. We discard the frames with a high reconstruction error, as these are probably poorly detected or aligned images. In our study, we use the L_1 norm and remove the videos that have less than three valid images from training and validation sets. In our preliminary studies on AFEW 4.0, we observe a considerable accuracy increase due to purification.

We first extract dense SIFT features from images, due to their popularity in compact representation of local appearance [27]. The dimensionality of image features are reduced via PCA, whose coefficients are learned from the training set, prior to video modeling. We use the same parameters as in [25] to extract SIFT features: typical 128 dimensional SIFT descriptors are extracted from 16 × 16 pixel patches with steps of 8 pixels that gives 7 × 7 = 49 overlapping blocks. Therefore, the dimensionality of the concatenated SIFT feature vector is 49 × 128 = 6272. As mentioned earlier, the feature dimensionality is reduced via PCA, preserving 90 % of the total variability.

In addition to dense SIFT based video representation, which will be detailed below, we also implemented LGBP-TOP feature representation [1]. The basic idea here is to combine the power of static LGBP descriptor and the dynamic LBP-TOP. The work of Almaev and Valstar [1] has shown that LGBP-TOP descriptor outperforms LGBP and LBP-TOP for Facial Action Unit (FAU) recognition and it is less susceptible to rotation errors compared to these methods.

3.3.1 Video representation for dense SIFT descriptor

After extraction of image features, the image sets are represented via four alternatives from which kernels are obtained. The first and simplest approach is using statistical functionals to provide a baseline. Here we use mean and range of image features over frames. Let $X_v \in \mathcal{R}^{d \times F_v}$ be the matrix representing d dimensional features of video v having F_v frames. Using mean and range functionals results in a $2 \times d$ dimensional video feature vector.

The second approach is taking singular value decomposition (SVD) of the video feature matrix X. Let r be the rank of matrix X. SVD gives an orthonormal decomposition in the form:

$$X = U \Lambda V^T, \tag{6}$$

where columns of U are normalized eigenvectors of XX^T, rows of V^T are normalized eigenvectors of $X^T X$, and first r diagonal elements of Λ are square the roots of corresponding sorted eigenvalues. Representing the video with the first $l \leq r$ columns of U leads to a matrix $L \in \mathbb{R}^{d \times l}$. This linear subspace representation is known to lie in a Grassmanian manifold $G(l, d)$, which is a special case of Riemannian manifold [11].

Our third approach represents the image set $X_v \in \mathbb{R}^{d \times F_v}$ with its $d \times d$ covariance matrix Σ. The fourth extends this by introducing the mean statistic μ of the features, thus obtaining a multivariate Gaussian. To embed the Gaussian in a Riemannian manifold, it is represented as a symmetric positive definite (SPD) matrix [26]:

$$\mathcal{N}(\mu, \Sigma) \sim M = |\Sigma|^{-\frac{1}{d+1}} \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix} \tag{7}$$

3.3.2 Riemannian kernels

In functional based video representation, the video feature vectors are used as if they reside in a regular vector space when kernels are computed. For the other three video representations, however, we use Riemannian kernels to compute video similarity. For the SVD-based linear subspace, the similarity of the video matrices L_i and L_j is computed via Mercer kernels that map the points in a Grassmanian manifold to Euclidean space [11,25]. Linear projection kernel $\mathcal{K}_{i,j}^{Proj.-Lin.}$ is defined as:

$$\mathcal{K}_{i,j}^{Proj.-Lin.} = \|L_i^T L_j\|_F^2, \tag{8}$$

where $\|\cdot\|_F^2$ is the Frobenius norm. The RBF kernel is defined over the mapping $\Phi_{Proj.} = L_i L_i^T$ [40]:

$$\mathcal{K}_{i,j}^{Proj.-RBF.} = exp\left(-\gamma \|L_i L_i^T - L_j L_j^T\|_F^2\right) \tag{9}$$

The covariance matrix and the Gaussian matrix representation of videos are both symmetric positive definite (SPD). A popular distance measure for SPD matrices is the log-Euclidean distance (LED), which is based on a matrix logarithm operator [3]. The proposed Linear and RBF kernels between SPD matrices S_i and S_j are formulated as [40,41]:

$$\mathcal{K}_{i,j}^{LED-Lin.} = trace\left[log(S_i) log(S_j)\right] \tag{10}$$

$$\mathcal{K}_{i,j}^{LED-RBF.} = exp\left(-\gamma trace\left[log(S_i) - log(S_j)\right]\right) \tag{11}$$

While the obtained kernels can be given to kernel machines as input, they can also suitably be used in other learners, where similarity to training instances is considered as a new feature representation. In this study, we train models using partial least squares (PLS) and ELM on the obtained kernels.

3.3.3 LGBP-TOP

In LGBP-TOP, the images are convolved with a set of 2D complex Gabor filters to obtain Gabor-videos, then LBP-TOP is applied to image blocks from each Gabor-video. A 2D complex Gabor filter is the convolution of a 2D sinusoid (carrier) having phase P, spatial frequencies u_0 and v_0 with a 2D Gaussian kernel (envelope) having amplitude K, orientation θ, and spatial scales a and b. In line with [1], for simplicity we take $a = b = \sigma$, $u_0 = v_0 = \phi$ and $K = 1$ to obtain:

$$\begin{aligned} G(x, y) = \exp&\left(-\pi\sigma^2((x - x_0)_r^2 + (y - y_0)_r^2)\right) \\ &\times exp\left(j(2\pi\phi(x + y) + P)\right), \end{aligned} \tag{12}$$

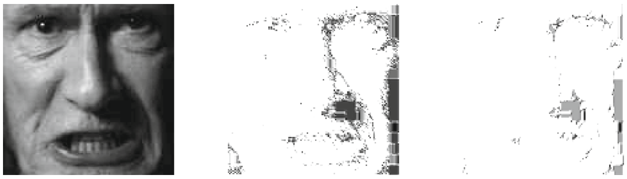


Fig. 2 A face image and two Gabor magnitude responses

where the subscript r stands for a clockwise rotation operation around reference point (x_0, y_0) such that:

$$(x - x_0)_r = (x - x_0)cos\theta + (y - y_0)sin\theta$$
$$(y - y_0)_r = -(x - x_0)sin\theta + (y - y_0)cos\theta \qquad (13)$$

Note that the effect of the phase is canceled out, since only the magnitude response of the filter is used for the descriptor. A sample video image with Gabor magnitude response images is given in Fig. 2.

When 2D complex Gabor filters are formed, all video frames are convolved and separate Gabor-videos are stacked to LBP-TOP operation. For LBP-TOP computation, we use non-overlapping blocks of 4 frames and divide all planes (i.e. XY, XT and YT) into 16 non-overlapping, equal-size regions. Also in our implementation, we divide the video into two equal length volumes over the time axis and extract LGBP-TOP features from each volume to further enhance temporal modeling. Using three spacial frequencies ($\phi = \{\pi/2, \pi/4, \pi/8\}$) and six orientations ($\theta = k\pi/6, k \in \{0..5\}$), we form a set of 18 Gabor filters. The dimensionality of the feature vector is therefore 2 × 18 × 16 × 58 × 3 = 100,224.

4 Experimental results

In our experiments we test the suitability of basic and kernel ELM for the problem of audiovisual emotion recognition. To probe the individual performance, we handle the video and audio separately and then combine the decisions of best performing modality-specific ELMs. Fast ELM training gives us the ability to simulate a wide range of hypotheses with moderate system requirements.[1]

4.1 Comparison of basic and kernel ELM

As a preliminary system development step, we carried out tests using the full set of baseline features in both modalities to see which ELM type is better suited to the problem at hand. For basic ELM, we chose the sigmoid activation function, since it provided the best among alternatives (sine, hard limit and Gaussian functions). Considering the dimensionality of video and audio features, we tested basic ELM with a different numbers of hidden units ($h \in 2^{\{1,2,\ldots,9\}}$) both using random projection and PCA (ranked eigenvectors) for input weights. For kernel ELM, we experimented with both Linear and RBF kernels, and tested different regularization parameter values $\tau \in 10^{\{-5,\ldots,5\}}$. The same set of hyper parameters was tested for the scatter parameter σ of the RBF kernel.

Table 1 Validation set classification accuracy of modality-specific ELMs with different kernels and input weight arrangements

	Basic ELM		Kernel ELM	
	Random (%)	PCA (%)	Linear (%)	RBF (%)
Audio	27.94	32.64	35.77	35.77
Video	25.07	37.74	38.81	39.35

The best results from the preliminary experiments with EmotiW 2014 baseline features (without feature selection) are given in Table 1. The challenge validation set baselines are 26.2 and 34.4 % for audio and video, respectively. We see that the using PCA improves the input layer modeling on this dataset. Moreover, the video modality features seem to benefit more from PCA compared to audio features. Although basic ELM with PCA modeling of input weights is found to outperform the challenge baselines, we use kernel ELM for the rest of our experiments, as it provides better performance in both modalities. The best accuracy obtained in both kernel types are either the same (for audio) or very close.

4.2 Experiments with baseline visual features

For visual classification, we assess the contribution of different facial regions. Apart from reduced dimensionality for better modeling, the reason of focusing on a small number of facial regions are (1) partial peripheral occlusion of face, (2) cluttered background in case the face is tilted and (3) robustness of different regions to alignment issues. Some example facial group configurations are given in Fig. 3.

Using the same experimental setting described in the previous section for kernel hyper parameters, we issue ELM tests with Linear and RFB kernels on various facial region combinations. The best results for each region and kernel type are provided in Table 2. Even using features of a single facial region (e.g. region 11), it is possible to outperform the challenge baseline score. Moreover, using 2-by-2 inner facial regions or 2-by-4 midface vertical regions, we can obtain better performance than the full set of 16 regions. Our results indicate that for emotion related tasks with difficult registration conditions, focusing on the inner face reduces the feature dimensionality, while preserving discriminative information, and results in better classification rates. Note that we carry

[1] http://extreme-learning-machines.org/.

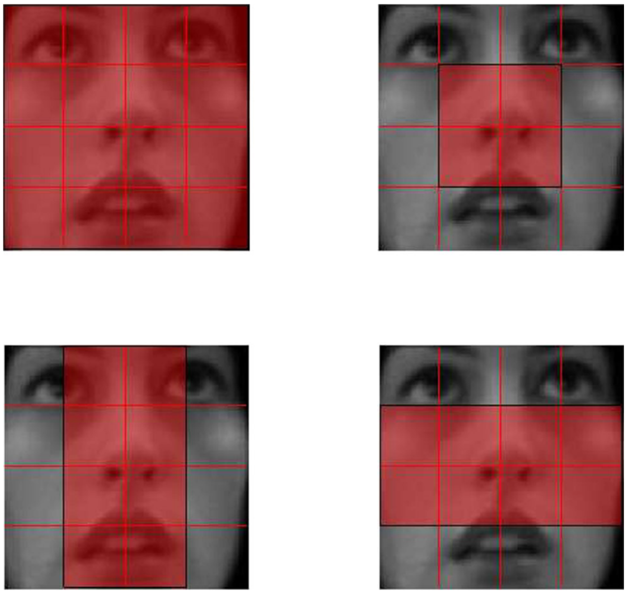


Fig. 3 Illustration of facial regions tested in the study. *Top-left* full face, *top-right* inner face (regions 6, 7, 10 and 11), *bottom-left* midface vertical, *bottom-right* midface horizontal

Table 2 Validation set accuracy comparison of facial regions using Linear and RBF

Facial regions (#)	RBF (%)	Linear (%)
Whole face (16)	39.35	38.81
Midface horizontal (8)	37.47	39.08
Midface vertical (8)	38.89	40.16
Inner face (4)	39.89	39.08
Regions 6 and 7 (2)	36.39	35.31
Regions 10 and 11 (2)	37.47	34.77
Regions 6 and 11 (2)	37.74	35.31
Regions 6 and 10 (2)	36.12	36.12
Regions 7 and 10 (2)	36.12	36.66
Regions 7 and 11 (2)	37.74	37.74

Region numbering starts from the top-left corner of the face and proceeds in reading order

out group-wise feature selection, since individual selection of histogram features are not meaningful or sufficient for pattern recognition.

To keep the table uncluttered, the parameters yielding the reported results are not included. The best RBF kernel results using at least four regions are obtained with scatter parameter $\sigma = 10^3$ and regularization parameter $\tau = 10$. On the other hand, best results with linear kernel are obtained with $\tau = 10^{-2}$. Both in video and audio modalities, we record the hyper parameters giving the best results to be later used in test set predictions.

Table 3 Validation set performance comparison of acoustic feature selection methods

	RBF (%)	Linear (%)
All baseline feats	35.77	35.77
mRMR ranked feats	33.16	34.46
mRMR-LLD (k = 10)	33.68	32.38
SLCCA-LLD (k = 10)	35.51	35.25

4.3 Experiments with baseline acoustic features

After the first probe into the full set of baseline acoustic features, we applied several feature selection methods. This was followed by extraction of a more recent and larger feature set that was used in INTERSPEECH 2013 via openSMILE tool.

We first used the iterative minimum Redundancy Maximum Relevance (mRMR) filter [32] for feature selection. mRMR adds features to a set of selected features one by one. At the kth step, mRMR maximizes the difference between relevance and redundancy terms to add the kth feature to the set [32]:

$$\max_{x_j \in X - S_{k-1}} \left[MI(x_j, t) - \frac{1}{k-1} \sum_{x_i \in S_{k-1}} MI(x_j, x_i) \right], \qquad (14)$$

where $MI(x, y)$ is mutual information between random variables x and y. As suggested by Peng et al. [32], we discretized the continuous acoustic features into seven bins based on z-scores.[2]

In addition to mRMR, we use a multi-view feature filter based on canonical correlation analysis (CCA). CCA is a statistical method that seeks to maximize the mutual correlation between two sets of variables by finding linear projections for each set [13]. We apply samples versus labels CCA (SLCCA) Filter [20] to Low Level Descriptor (LLD) based feature groups and then concatenate the ranked k features from each group as in [21]. When all features are subjected to CCA against the labels, the absolute value of the projection matrix $\mathbf{V}$ can be used to rank the features [19]. We extend the LLD based approach using mRMR and combine top $k = \{5, 10, 15, 20\}$ ranking features from each LLD group. For mRMR, the first 200 features are tested with steps of 10, each with the set of ELM hyper parameters discussed in previous sections. Together with regular mRMR, we test three feature selection approaches on baseline acoustic features.

The best validation set results of feature selection approaches utilized in the study are given in Table 3. We observe that the best LLD-based SLCCA-Filter outperforms

[2] The z-score ranges are $\{(-\infty, -2.5], (-2.5, -1.5], (-1.5, -0.5], (-0.5, 0.5], (0.5, 1.5], (1.5, 2.5], (2.5, \infty)\}$.

the best performance obtained with LLD-based mRMR. However, no feature selection method performs better than the full set of features. The superior performance of the full set can be attributed to the ELM learning rule, which minimizes the norm of the projection, thus making use of all features without over-fitting. In a regular Neural Network where the weights are learned via gradient descend based back-propagation, feature selection would help avoid over-fitting, therefore may yield better results than the full set. Another reason that a selected subset does not perform better than the full set is the fact that the paralinguistic information can be distributed over a wide range of features. This is the reason why state-of-the-art results in the field are obtained with very high dimensional supra-segmental acoustic features [34].

We further included four other publicly available emotional corpora to test whether additional corpora would improve training or not. These are Berlin Emotional Speech Database (EMODB) [4], Danish Emotion Database (DES) [9], eNTERFACE Database [29], and the Turkish Emotional Database (BUEMODB) [23]. Note that all corpora are acted, though two are recorded in studio. Here, we use only the instances belonging to the seven classes of AFEW.

Cross-corpus evaluation results are given in Table 4. Class distribution and some basic information about the corpora are given in Table 5. All corpora are individually normalized to range [−1, 1]. Without corpus-wise normalization, the corpora are found to impair the generalization of the learner. This finding is in accordance with cross-corpus work of Schuller et al. [35]. We see a performance decrease with respect to given baseline features. eNTERFACE and EMODB provide some performance increase with respect to INTERSPEECH 2013 baseline set, whereas BUEMODB and DES do not contribute at all. Even when all additional corpora are included, the performance is below the accuracy obtained via only EmotiW 2014 Challenge features.

Table 4 Best validation set performance of multi corpus training

Corpora	Accuracy (%)
AFEW 4.0 IS13 Features	33.42
+eNTERFACE	34.20
+EMODB	34.20
+BUEMODB	33.42
+DES	33.42
+ENTERFACE + EMODB	33.42
+All corpora	34.20

4.4 Comparison of ELM with PLS classifier on baseline and extracted features

We compare the performance of kernel ELM with the partial least squares regression based classifier used in [25], which reports state-of-the-art results for EmotiW 2014 Challenge. Similar to ELM, PLS enjoys the capabilities of fast learning and accurate prediction, however, is not as popular as SVM. We do not include SVM in our comparison, since it is already compared to PLS on this dataset and was shown to give inferior results [25]. For a full description of PLS regression, the reader is referred to [42]. In [25], PLS regression is applied to classification in one-versus-all setting. The class that gives the highest regression score is taken as prediction.

We compare the two classifiers first on challenge baseline features. The best validation set results of two methods, and corresponding test set results of ELM are given in Table 6. Note that the audio-only results given here with PLS are higher than those reported in [25]. This is because we use kernels, while in [25] the acoustic features were used directly. Analyzing the scores on the baseline feature sets, we observe an overall better performance with ELM, and the margin increases with modality-fusion. To show that the validation set scores are highly indicative of the test set performance, we give the test set results of ELM on the right most columns of Table 6. As discussed earlier, inner facial regions generalize better than the whole face due to reduced sensitivity to occlusions and registration errors.

For further comparison on extracted dense SIFT features, we experiment on statistical functional based video representations. Using only mean and range statistics gives 39.84 and 41.19 % validation set accuracy for PLS and ELM, respectively. Note that ELM performance here is higher compared to the best video-only result on baseline features. This might be partly due to data purification. Finally, we compare the two methods using the six Riemannian Kernels described in Sect. 3.3.2. The best validation set performances are listed in Table 7. In comparative experiments, we use the same kernels for two methods, optimizing their hyper-parameters

Table 5 Class distribution of additional emotional corpora. Classes correspond to A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), SU(rprise)

Corpus	Content	A	D	F	H	N	SA	SU	#All	Type
EMODB	German	127	38	55	64	78	53	–	415	studio
DES	Danish	85	–	–	86	85	84	79	419	non-studio
eNTERFACE	English	200	189	189	205	–	195	182	1170	non-studio
BUEMODB	Turkish	121	–	–	121	121	121	–	484	studio

All corpora are acted and the spoken content is the same for all subjects/emotion classes

Table 6 Comparison of PLS and ELM performance on EmotiW 2014 baseline feature sets

Acc (%)	Validation				Test
Kernel	Linear		RBF		RBF
Classifier	PLS	ELM	PLS	ELM	ELM
Video (WF)	39.08	39.89	38.27	39.35	36.11
Video (IF)	37.74	39.08	39.08	39.89	39.07
Audio	35.25	35.77	34.46	35.77	37.84
Fusion (WF)	41.51	43.13	39.62	42.86	43.00
Fusion (IF)	40.16	42.32	40.43	44.20	44.23

IF inner face, *WF* whole face

Table 7 Comparison of validation set accuracies of PLS and ELM on Riemannian Kernels for video representation

	SVD		Covariance		Gaussian	
	Linear	RBF	Linear	RBF	Linear	RBF
PLS	41.46	40.92	38.21	40.65	39.84	37.94
ELM	43.63	43.09	39.84	41.46	39.30	39.02

on the validation set. The PLS performance using dense SIFT is slightly lower compared to those reported in [25], which may be attributed to the number of PCA eigenvectors prior to video representation. Similar to experiments on baseline features, we observe better overall performance with ELM, giving higher than 43 % accuracy on Grassmanian kernels (SVD). When we probe the test set performance of the best models (SVD representation with Linear Kernel), we get accuracies of 40.29 % for PLS and 43.23 % for ELM, respectively. This difference is not found to be statistically significant with McNemar's test [30]. While the results confirm the good performance of PLS as classifier, it is also clear that the state-of-the-art performance of [25] is largely due to using an ensemble of 24 visual systems, which complement each other.

Lastly, we compare the performance of the two classifiers on extracted LGBP-TOP features. We optimize the σ parameter of the Gabor kernel by observing its effect on the Gabor pictures, and set it empirically to 0.5. On the overall, no optimization is done for other parameters of the Gabor filters. Considering the massive dimensionality, filter and feature selection have a high potential of improving generalization. This is left for future work. Focusing on the inner facial regions in LGBP-TOP did not provide a performance increase as in baseline LBP-TOP features. We attribute this to the added data purification step, which eliminates partially occluded or badly aligned faces. For linear kernels, the best validation set performances are 42.05 and 39.35 % for ELM and PLS, respectively. With RBF kernel, the validation performances become 41.78 % for ELM and 41.51 % for PLS.

All our results are obtained on powerful feature sets with good preprocessing. Subsequently, while the ELM classifier usually reaches higher accuracies than the PLS classifier, these differences were not significant. We have recently contrasted these classifiers on a new emotional speech corpus (EmoChildRU), which is collected from 3 to 7 years old Russian children in naturalistic conditions [28]. The data are annotated for three valence related affective classes: comfort, discomfort and neutral. Our results indicate that PLS is highly sensitive to preprocessing and to feature representation, whereas ELM consistently gives (in most cases significantly) better results.

4.5 Multimodal fusion and test set results

We test the best performing modality-specific systems using flat averaging (FA) and weighted fusion (WF) schemes. In FA, we average the class-wise predictions to get a fused score, whereas in WF, class-wise weights are used for each sub-system. Using ELM with the baseline features, the best performing single modality systems give 37.84 % (audio full set) and 39.07 % (video innerface) accuracy on the validation set. Using the extracted features from purified images, we observe 42.05 % accuracy with LGBP-TOP and 43.63 % accuracy with dense SIFT (Linear Grassmanian kernel).

We first analyze FA fusion on the modality-specific ELMs learned on the baseline features. Then we combine the best modality-specific ELMs using WF, where the optimal fusion weights are searched over a random pool of fusion matrices. This approach is inspired from the success of the top performing work in EmotiW 2013 [18]. For this, we randomly generate 50,000 fusion matrices for each alternative combination, and normalize each matrix over the models. To avoid overfitting on the validation set, fusion weights are rounded to three decimal digits. Since RBF kernels are observed to give better performance in decision fusion, all combination experiments are carried out with scores obtained from RBF ELMs.

Validation and test set performances of multimodal decision fusion of ELMs learned on baseline feature sets are provided in Table 8. We observe that using inner facial regions in audiovisual fusion provides better generalization than the full set of features. While multimodal fusion with midface features provides the highest validation set accuracy, it does not yield a high score on the test set. This might be attributed to over-fitting to the validation set, however the hyper-parameters are not specifically optimized for this system. The most probable reason of the high discrepancy between validation and test set performances of systems is partial occlusions and registration errors. The higher generalization performance of inner facial features is not only due to the relevant information they contain, but also due to resilience to occlusion and environmental noise.

Table 8 Validation and test set accuracies (%) for decision fusion of modality-specific kernel ELMs trained on baseline feature sets

System	Val	Test
LBP-TOP (midface vertical) and audio SLCCA-LLD (k = 10)	47.17	38.33
LBP-TOP (midface vertical) and Audio all	44.47	38.08
LBP-TOP (inner face) and audio all	44.20	44.23
LBP-TOP (inner face) and audio SLCCA-LLD (k = 10)	43.13	43.98
LBP-TOP (wholeface) and audio all	42.86	43.00

Flat averaging scheme is used for score fusion

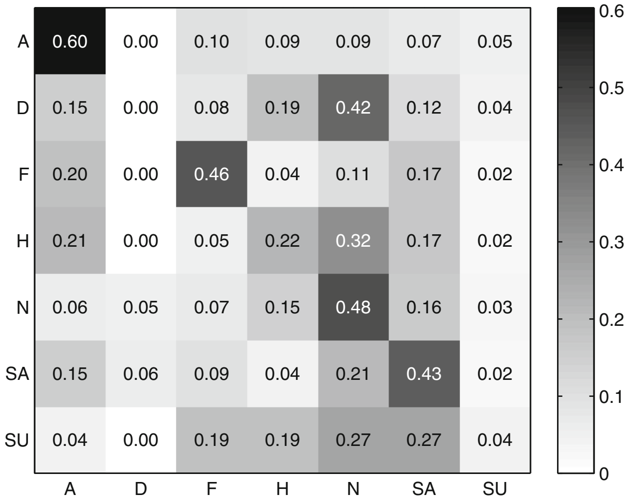


Fig. 4 Test set confusion matrix of audio modality system. Classes correspond to A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), SU(rprise)

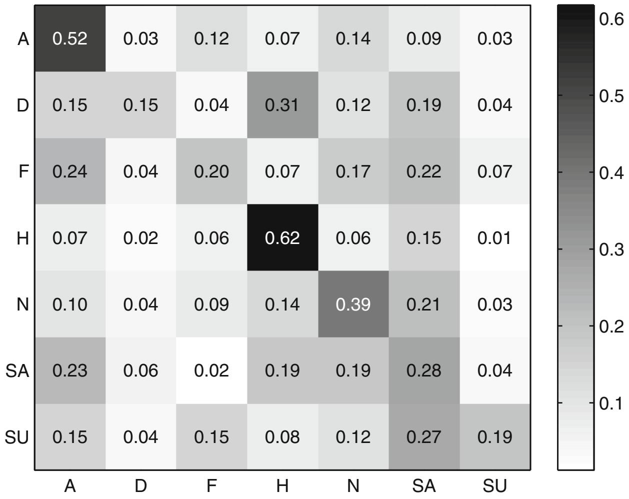


Fig. 5 Test set confusion matrix of video modality system. Classes correspond to A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), SU(rprise)

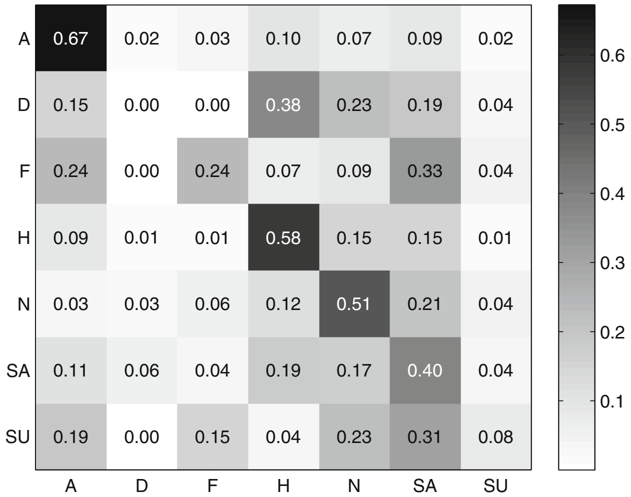


Fig. 6 Test set confusion matrix of multimodal score fusion system. Classes correspond to A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), SU(rprise)

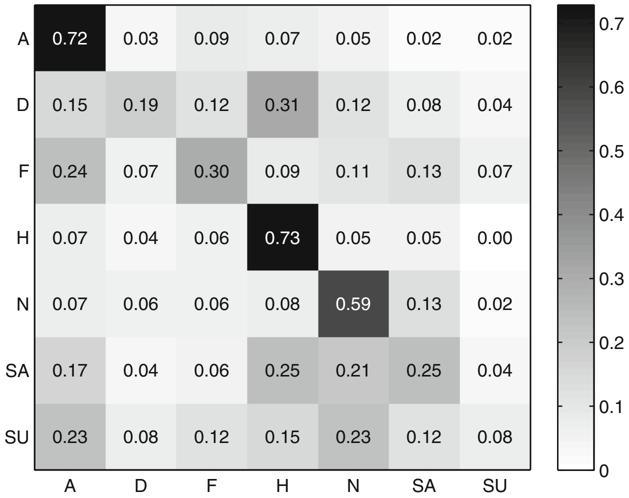
The test set confusion matrices for systems obtained with audio baseline features, video baseline features and their fusion are given in Figs. 4, 5 and 6, respectively. The diagonal elements indicate the recall of the corresponding classes. On the overall, we observe that fusion boosts the performance of single modality systems. However, since the audio-based system does not recognize Disgust, Happiness and Surprise classes well (if at all), the fusion system shows a lower recall in these classes compared to the video-based system. On the other hand, recall performance of the audio-based system outperforms the video-based system in the remaining four classes. These results imply that a confidence based fusion of modality-specific systems can advance the overall recognition. Therefore, we use a weighted fusion scheme in further experiments, where we employed combinations sub-systems.

Weighted score fusion of LBP-TOP (inner face), SIFT SVD and Audio All gave a validation set accuracy of 49.04 % and a test set accuracy of 46.44 %. Inclusion of LGBP-TOP in this scheme led to accuracies of 51.49 and 50.12 %, on the validation and the test set, respectively. It is worthy to note that using a weighted fusion of only four sub-systems, we reach the state-of-the-art test set performance obtained by [25] that combines 25 sub-systems. We attribute the attained state-of-the-art performance to three factors. First is complementarity of base feature types and modalities. We observed that including two new visual sub-systems improved the accuracy. Second is the generalization power

Table 9 Fusion weights for the best performing system

	SIFT SVD	LGBP-TOP	LBP-TOP	Audio
A	0.006	0.254	0.377	0.363
D	0.277	0.028	0.656	0.039
F	0.095	0.285	0.097	0.523
H	0.052	0.458	0.238	0.252
N	0.307	0.237	0.131	0.325
SA	0.398	0.303	0.215	0.084
SU	0.446	0.034	0.306	0.214

Classes correspond to A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), SU(rprise)



	A	D	F	H	N	SA	SU
A	0.72	0.03	0.09	0.07	0.05	0.02	0.02
D	0.15	0.19	0.12	0.31	0.12	0.08	0.04
F	0.24	0.07	0.30	0.09	0.11	0.13	0.07
H	0.07	0.04	0.06	0.73	0.05	0.05	0.00
N	0.07	0.06	0.06	0.08	0.59	0.13	0.02
SA	0.17	0.04	0.06	0.25	0.21	0.25	0.04
SU	0.23	0.08	0.12	0.15	0.23	0.12	0.08

Fig. 7 Test set confusion matrix of the best performing weighted score fusion system. Classes correspond to A(nger), D(isgust), F(ear), H(appiness), N(eutral), SA(dness), SU(rprise)

of ELM. On the overall, we observe very close validation and test set performance of models trained with ELM. The last is the weighted fusion strategy. Since the base models differ in their confusion matrices, both class and model level weighted fusion outperforms FA. Moreover, using random weights instead of a meta classifier (e.g. a second level ELM) reduces the risk of over-fitting. Fusion weights used in the best system and corresponding confusion matrix can be found in Table 9 and Fig. 7, respectively.

5 Conclusions and outlook

In this study, we introduce ELMs for audiovisual emotion recognition in the wild. ELMs provide accurate results with several orders of magnitude faster training compared to SVMs and SLFNs. Typically, this leads to more time for parameter search and optimization. We test facial feature group selection, as well as recently proposed acoustic feature selection approaches for this problem.

We compared ELM with a PLS based classifier that is used in the top system of EmotiW 2014, and obtained better results with ELM. We achieve the best validation and test set results with decision fusion of modality-specific ELM models. While our results verify the importance of multimodal fusion and combination of diverse classifiers, they also highlight the importance of the fusion strategy.

The tested systems performed very poorly on some of the classes. In particular, it was very difficult to classify happiness and surprise from audio, whereas fear and sadness are difficult to classify from video. Disgust is difficult for both modalities. This result shows that in-the-wild emotions are much more difficult to recognize compared to controlled conditions typically used in the literature.

Our tests with additional speech corpora to augment training did not contribute to accuracy. One possible cause for the lack of improvement is the difference in the acquisition conditions of the corpora. Furthermore, acoustic feature selection was not found to improve the performance. On the other hand, in video modality using a semantically meaningful subset of facial regions, it was possible to obtain better recognition results than the full set, both in the development and the test set.

References

1. Almaev TR, Valstar MF (2013) Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In: 2013 humaine association conference on affective computing and intelligent interaction (ACII), IEEE, pp 356–361
2. Alpaydin E (2010) Introduction to machine learning, 2nd edn. The MIT Press, Cambridge
3. Arsigny V, Fillard P, Pennec X, Ayache N (2007) Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J Matrix Anal Appl 29(1):328–347
4. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A database of German emotional speech. In: Proc. of INTERSPEECH 2005, pp 1517–1520
5. Cowie R, Sussman N, Ben-Ze'ev A (2011) Emotion: concepts and definitions. In: Petta P, Pelechaud C, Cowie R (eds) Emotion-oriented systems: the humaine handbook. Springer, Berlin, pp 9–32
6. Dhall A, Goecke R, Lucey S, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. IEEE Multimed 19(3):34–41
7. 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, ICMI '14, pp 461–466
8. Dhall A, Goecke R, Joshi J, Wagner M, Gedeon T (2013) Emotion recognition in the wild challenge 2013. In: Proc. of the 15th ACM Intl. conf. on multimodal interaction (ICMI 2013), ACM, pp 509–516
9. Engberg I, Hansen A (1996) Documentation of the Danish emotional speech database (DES). Internal AAU Report, Center for Person Kommunikation, Denmark
10. Eyben F, Wöllmer M, Schuller B (2010) OpenSMILE: the Munich versatile and fast open-source audio feature extractor. In: Proc. of the intl. conf. on multimedia, ACM, pp 1459–1462

11. Hamm J, Lee DD (2008) Grassmann discriminant analysis: a unifying view on subspace-based learning. In: Proceedings of the 25th international conference on machine learning, pp 376–383
12. Han K, Yu D, Tashev I (2014) Speech emotion recognition using deep neural network and extreme learning machine. In: Proceedings of INTERSPEECH, ISCA, Singapore, pp 223–227
13. Hardoon DR, Szedmak S, Shawe-Taylor J (2004) Canonical correlation analysis: an overview with application to learning methods. Neural Comput 16(12):2639–2664
14. Huang GB, Zhu QY, Siew CK (2004) Extreme learning machine: a new learning scheme of feedforward neural networks. Proc IEEE Int Joint Conf Neural Netw 2:985–990
15. Huang GB, Zhu QY, Siew CK (2006) Extreme learning machine: theory and applications. Neurocomputing 70(1):489–501
16. Huang GB, Zhou H, Ding X, Zhang R (2012) Extreme learning machine for regression and multiclass classification. IEEE Trans Syst Man Cybern Part B Cybern 42(2):513–529
17. Itakura F (1975) Line spectrum representation of linear predictor coefficients of speech signals. J Acoust Soc Am 57(S1):S35
18. Kahou SE, Pal C, Bouthillier X, Froumenty P, Gülçehre c, Memisevic R, Vincent P, Courville A, Bengio Y, Ferrari RC, Mirza M, Jean S, Carrier PL, Dauphin Y, Boulanger-Lewandowski N, Aggarwal A, Zumer J, Lamblin P, Raymond JP, Desjardins G, Pascanu R, Warde-Farley D, Torabi A, Sharma A, Bengio E, Côté M, Konda KR, 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, ICMI '13, pp 543–550
19. Kaya H, Özkaptan T, Salah AA, Gürgen F (2015) Random discriminative projection based feature selection with application to conflict recognition. IEEE Signal Process Lett 22(6):671–675. doi:10.1109/LSP.2014.2365393
20. Kaya H, Eyben F, Salah AA, Schuller BW (2014) CCA Based feature selection with application to continuous depression recognition from acoustic speech features. In: Proceedings of IEEE International conference on acoustics, speech, and signal processing (ICASSP 2014), pp 3757–3761
21. Kaya H, Özkaptan T, Salah AA, Gürgen F (2014) Canonical Correlation analysis and local fisher discriminant analysis based multi-view acoustic feature reduction for physical load prediction. In: Proceedings of INTERSPEECH, ISCA, Singapore, pp 442–446
22. Kaya H, Salah AA (2014) Combining modality-specific extreme learning machines for emotion recognition in the wild. In: Proceedings of the 16th international conference on multimodal interaction, ACM, ICMI '14, pp 487–493
23. Kaya H, Salah AA, Gurgen SF, Ekenel H (2014) Protocol and easeline for experiments on Bogazici university Turkish emotional speech corpus. In: IEEE Signal processing and communications applications conf. (SIU), 2014, pp 1698–1701
24. 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, ICMI '13, pp 525–530
25. 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, New York, NY, USA, ICMI '14, pp 494–501
26. Lovrić M, Min-Oo M, Ruh EA (2000) Multivariate normal distributions parametrized as a Riemannian symmetric space. J Multivar Anal 74(1):36–48
27. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
28. Lyakso E, Frolova O, Dmitrieva E, Grigorev A, Kaya H, Karpov AA (2015) EmoChildRu: emotional child russian speech corpus. INTERSPEECH (**submitted**)
29. Martin O, Kotsia I, Macq B, Pitas I (2006) The eNTERFACE '05 audio-visual emotion database. In: Proceedings of IEEE workshop on multimedia database management
30. McNemar Q (1947) Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12(2):153–157. doi:10.1007/BF02295996
31. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
32. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
33. Rao CR, Mitra SK (1971) Gen Inverse Matrices Appl, vol 7. Wiley, New York
34. Schuller B (2011) Voice and speech analysis in search of states and traits. In: Salah AA, Gevers T (eds) Computer analysis of human behavior. Springer, Berlin, pp 227–253
35. Schuller B, Vlasenko B, Eyben F, Wollmer M, Stuhlsatz A, Wendemuth A, Rigoll G (2010) Cross-corpus acoustic emotion recognition: variances and strategies. IEEE Trans Affect Comput 1(2):119–131
36. Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller CA, Narayanan SS (2010) The INTERSPEECH 2010 paralinguistic challenge. In: Proceedings of INTERSPEECH, pp 2794–2797
37. Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, Chetouani M, Weninger F, Eyben F, Marchi E, Mortillaro M, Salamin H, Polychroniou A, Valente F, Kim S (2013) The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In: Proceedings of INTERSPEECH, ISCA, ISCA, Lyon, France, pp 148–152
38. 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, New York, NY, USA, ICMI '14, pp 481–486
39. Suykens JA, Vandewalle J (1999) Least squares support vector machine classifiers. Neural Process Lett 9(3):293–300
40. Vemulapalli R, Pillai JK, Chellappa R (2013) Kernel learning for extrinsic classification of manifold features. In: IEEE conference on computer vision and pattern recognition (CVPR 2013), pp 1782–1789
41. Wang R, Guo H, Davis LS, Dai Q (2012) Covariance discriminative learning: a natural and efficient approach to image set classification. In: IEEE conference on computer vision and pattern recognition (CVPR 2012), pp 2496–2503
42. Wold H (1985) Partial least squares. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences. Wiley, New York, pp 581–491