The University of Passau Open Emotion Recognition System for the Multimodal Emotion Challenge

Jun Deng[1], Nicholas Cummins[1], Jing Han[1], Xinzhou Xu[1,2], Zhao Ren[3],
Vedhas Pandit[1], Zixing Zhang[1], and Björn Schuller[1]

[1] Chair of Complex & Intelligent Systems, University of Passau, Passau, Germany
[2] Technische Universität München, Munich, Germany
[3] Northwestern Polytechnical University, Xi'an, PR China
jun.deng@uni-passau.de

Abstract. This paper presents the University of Passau's approaches for the Multimodal Emotion Recognition Challenge 2016. For audio signals, we exploit Bag-of-Audio-Words techniques combining Extreme Learning Machines and Hierarchical Extreme Learning Machines. For video signals, we use not only the information from the cropped face of a video frame, but also the broader contextual information from the entire frame. This information is extracted via two Convolutional Neural Networks pre-trained for face detection and object classification. Moreover, we extract facial action units, which reflect facial muscle movements and are known to be important for emotion recognition. Long Short-Term Memory Recurrent Neural Networks are deployed to exploit temporal information in the video representation. Average late fusion of audio and video systems is applied to make prediction for multimodal emotion recognition. Experimental results on the challenge database demonstrate the effectiveness of our proposed systems when compared to the baseline.

Keywords: Multimodal Emotion Recognition, Bag-of-Audio-Words, Transfer Learning, Long Short-Term Memory, Convolutional Neural Networks

1 Introduction

Emotion recognition 'in the wild' is attracting growing interest due to its practical importance in many real-world applications, such as human-computer interaction (HCI), e-learning, and health care. Despite a large number of existing research efforts to collect and analyse spontaneous or in the wild emotion databases in English, French or German [5], there have only been a small number of similar investigations undertaken on Chinese databases [2]. To advance spontaneous emotion recognition in the Chinese context, the *Multimodal Emotion Recognition Challenge* 2016 (MEC) provides a common benchmark database consisting of audiovisual clips taken from Chinese movies and TV programs [16]. In this paper, we present our approaches to the audio, video, and multimodal emotion recognition tasks introduced in this challenge [16].

Given the variety of acoustic events that can potentially occur within the selected audiovisual clips present in the challenge data, *Bag-of-Audio-Words* (BoAW) represents a potentially robust audio representation of the signal. This technique has been successfully used in similar emotion recognition tasks [20, 22, 23]. Inspired by this success, we create BoAW features based on five different *Low-Level-Descriptors* (LLDs) sets available from our open source toolkit openSMILE [7] and investigate their suitability for in-the-wild emotion detection.

Neural Networks based classification is widely used in audio-based emotion detection systems [3, 8, 17]. *Extreme Learning Machines* (ELMs) are a feedforward neural networks with a single hidden layer, currently gaining considerable interest in the machine learning community. This is due in part to their fast training and ease of implementation [11]. ELMs have shown competitive performance compared to *Support Vector Machines* (SVMs) and *Deep Neural Networks* (DNNs) in similar tasks [8, 14] and are a key component in our audio-based system.

Many of the latest video recognition approaches are based on the features extracted from deep *Convolutional Neural Networks* (CNNs). However, the major challenge of applying this technique in the emotion recognition field is the lack of adequate training data. We overcome this challenge by using CNN models pre-trained on a large scale of data. The basic idea is to leverage the pre-trained model as a feature extractor for the new dataset at hand. We first make use of a pre-trained CNN model, referred to as *VGGFace* [19], to extract features relevant to the facial expressions present in a video frame. In addition to the features from the face in a video frame, we leverage another pre-trained CNN model, referred to as *VGG* [28], to extract features relevant to the broader contextual information. Further, recent work [33] has verified the significance of Facial *Action Units* (AUs) as features for emotion recognition. Thus, we also incorporate AU features into our video-based system.

The remainder of this paper is organised as follows. First, the Sections 2 to 4 present each of the models used for different modalities. Next, Section 5 briefly introduces the MEC 2016 data and presents the results on the data. Finally, in Section 6 we conclude this paper and highlight future work directions.

2 Audio Systems

2.1 Feature Representations

BoAW audio feature representations are gaining popularity in many paralinguistic classification tasks [20,22,23]. BoAW involves generating a fixed length audio representation of each clip by first identifying a set of audio words, and then quantising (bagging) the original feature space, with respect to the generated codebook, to form a histogram representation of each data. The final BoAW representation represents the frequency of each previously identified audio word in a given instance [23].

Our basic framework to extract BoAW representations is depicted in Figure 1. During training, acoustic LLDs are extracted from the audio files and are stan-

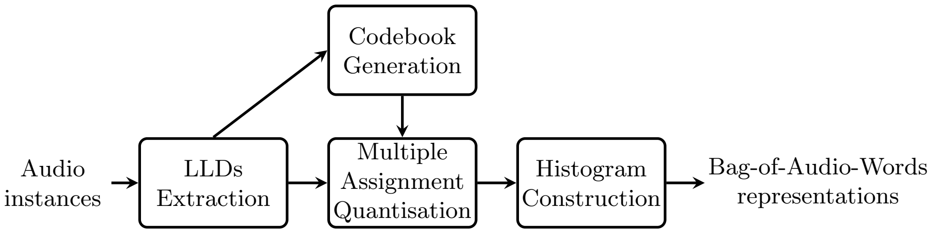


Fig. 1: Overview of the *Bag-of-Audio-Words* (BoAW) generation framework used in our audio systems. Figure adapted from [18].

dardised to zero mean and unit standard deviation. The next step is to generate the codebook. Work presented in [22] indicates that the codebooks generated via random sampling of the extracted LLDs offer similar emotion recognition performance to codebooks generated using k-means clustering. Given this result, we also employ a random sampling of the extracted LLDs to generate our codebooks.

After the codebook is built, a multi-assignment quantisation technique is applied to map LLDs from each frame to the first d closest audio-words, as measured by Euclidean distance. Preliminary experiments in [22] show that, for speech emotion recognition, the multi-assignment ($d > 1$) outperforms the uni-assignment ($d = 1$); therefore this paradigm is employed for all our extracted BoAW representations. A histogram is generated by calculating the counts of occurrence of each audio-word in all frames of one audio file. Finally, to generate a BoAW representation of a file, its corresponding relative counts are normalised to sum to one. This final step is undertaken to help minimise effects relating to disparities caused by various lengths of the files present in the MEC dataset. Note that, for the test data the LLDs of each frame are mapped to the audio words from the codebook pre-generated during the training phase.

2.2 Classifiers

Classification is performed in our audio systems either using a SVM or ELM back-end. SVMs are used due to their proven ability to handle a small dataset, relative lack of computational expense and established software implementations. Further, SVMs can also be regarded as a de-facto classifier for audio-based emotion detection systems [25].

As previously mentioned, ELMs have demonstrated competitive performance in similar audio based classification tasks [14]. The ELM is a single-hidden-layer feedforward neural network which is exceptionally fast to train as the weights and biases corresponding to the hidden layer are randomly assigned [11] and never tuned. The basic theory behind ELMs, as introduced in [12], is that the first layer (the inputs weights and biases) can be regarded as carrying an unsupervised feature mapping and the only learning being performed is between the output

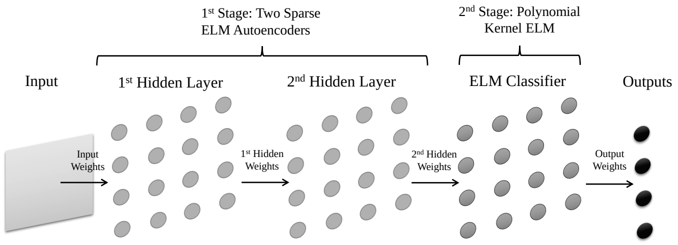


Fig. 2: The *Hierarchical Extreme Learning Machine* (HELM) framework consists of two phases: firstly two ELM based auto-encoders are used to form a sparse representation of the input feature space follwed by a *Extreme Learning Machine* (ELM) classifier. Figure adapted from [30].

of the second layer (the output weights) and the label matrix. This is achieved efficiently using a least-squares solution. The generalised output function $f(\mathbf{x})$ of an ELM is given by:

$$f\left(\mathbf{x}\right) = \sum_{i=1}^{L} \beta_i h_i\left(\mathbf{x}\right) = \mathbf{H}\beta, \tag{1}$$

where $\mathbf{x}$ is the input feature space, $\mathbf{H}$ a non-linear feature mapping (hidden layer) and $\beta = [\beta_1, \ldots, \beta_L]$ the (learnt) output weight vector. $h_i\left(\mathbf{x}\right)$, the output of the i^{th} hidden node can be obtained using a range of different activation functions, i. e., Sigmoid, Hyperbolic tangent, Gaussian, etc. As mentioned, β is found by minimising the least squared error:

$$\min_{\beta} \left\|\mathbf{H}\beta - \mathbf{T}\right\|^2, \tag{2}$$

where $\mathbf{T}$ is the training target matrix. For further details on the ELM, the reader is referred to [10, 11].

A wide variety of variants to the basic ELM structure have been proposed which include: *Regularised ELMs* (RELM), *Kernel ELMs* (KELM), and *Hierarchical*-ELM (HELM) which can be regarded as the ELM analogue to deep learning [11, 30]. Given the effectiveness of both DNNs and ELMs for performing speech-based emotion classification [8], an aim of this paper is to explore the suitability of the HELM framework for in-the-wild audio based emotion classification.

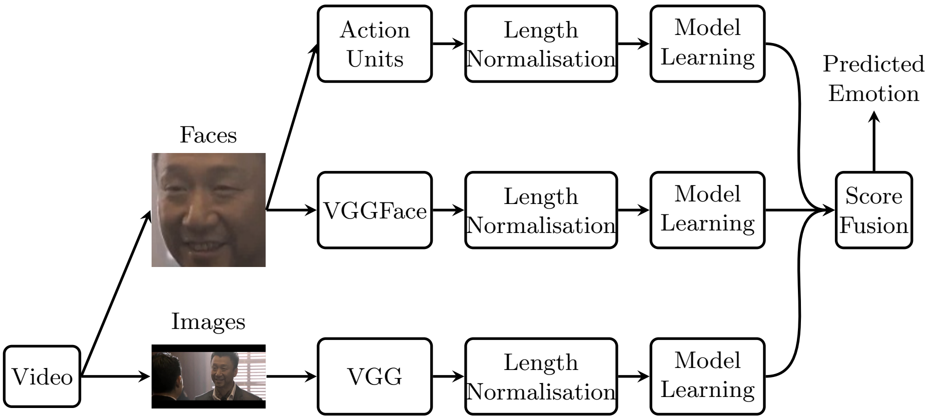


Fig. 3: Video Emotion Recognition System Architecture.

The HELM framework is an extension of the original ELM that allows ELMs to operate in a similar manner to a multilayer perceptron [30]. As seen in Figure 2, the HELM framework consists of three layers; two ELM based autoencoders and an ELM classifier. The role of the two ELM-based autoencoder layers is to form a sparse hierarchical representation of the original input feature space; the aim of this step is to exploit hidden information present in the training data. Sparsity is achieved by enforcing an ℓ_1 penalty during ELM training:

$$\min_{\beta}\left\{\|\mathbf{H}\beta-\mathbf{X}\|^2+\|\beta\|_{\ell_1}\right\}. \tag{3}$$

Each hidden layer in the HELM is an independent module and **H** is a randomly initialized output which does not require fine tuning.

In the original HELM framework, standard ELM classifiers was proposed to perform the classification step; however, as per the hidden layers, this module is independent and can be replaced with any ELM variant. Initial experiments (results not given) indicated that the use of a Polynomial Kernel ELM improved overall system performance. The closed-form solution for β is given by:

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

where **I** is the identity matrix, C is the regularization coefficient, and **H** is the given polynomial transformation of the input feature space. For further detail on the KELM and HELM, the reader is referred to [10] and [30] respectively.

3 Video Systems

Given a video sequence, Figure 3 illustrates the architecture of our proposed video emotion recognition system. Like many other recent approaches for video analysis, CNNs can be directly applied to learn salient features from the input image. Our video system makes use of two CNN models (i. e *VGG* [28] and *VGGFace* [19]) pre-trained for object classification and face recognition on a large scale of data. Here, the idea is to leverage the pre-trained models as feature extractors to provide features from each frame image and cropped face. In addition, we exploit facial action units as complementary features to enhance the video system. *Length normalisation* techniques such as *max* or *temporal k-max pooling* on frames are employed in order to give a fixed length input to a following classifier (i. e., SVMs or *Long Short-Term Memory Recurrent Neural Networks* (LSTM RNNs)), making it easy to perform model learning. Finally, an average decision rule is used to aggregate the scores predicated by the different models.

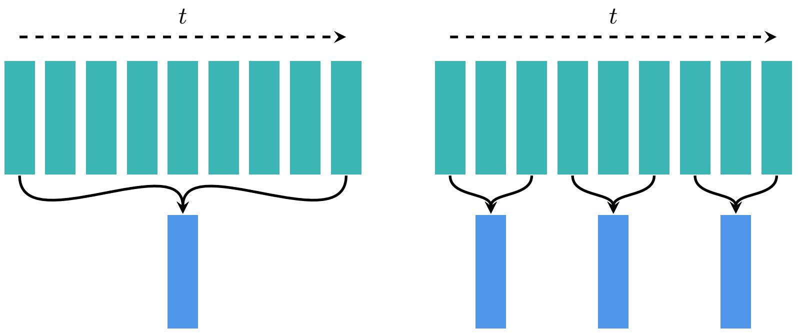
3.1 CNN Features

Deep CNNs are currently the most dominant approach in both video action recognition [32] and video emotion recognition [33]; this is due to their overwhelming accuracy. Deep CNNs trained on natural images exhibit an interesting phenomenon: the features learned from bottom to top layers are from *general* to *specific*. On the one hand, the first layer learns the features that are similar to Gabor filters and colour blobs. On the other hand, the higher-level layers are usually well trained for specific datasets and tasks. Consequently, the outputs of higher-level layers are widely chosen for recognition tasks because they combine all the general features into a rich image representation [4]. Thus, a deep CNN pre-trained on a large scale of image data can be used as a feature extractor for a task of interest.

A number of deep CNN architectures, which were originally proposed for image classification tasks, are popularly applied to directly extract deep CNN features from input pixels in a variety of computer vision tasks. These architectures include VGG-16, VGG-19 [28], AlexNet [15], and GooLeNet-22 [29]. In this challenge, we select the representative VGG-16 network (*VGG*), which consists of 13 convolutional layers, and 3 fully connected layers. Specifically, we use the 'FC7' features (i. e., the last feature layer, 4096 dimensions); FC7 is the most widely used deep feature extraction method for other computer vision tasks.

Facial-based features are known to be well suited for emotion recognition. Therefore, in addition to the pre-trained VGG model, we use VGGFace [19] to extract visual face descriptors of each frame. The VGGFace network has the same network architecture as VGG, but was trained on a very large-scale face data (2.6M images, 2.6k people). Hence, VGGFace tends to yield visual features with a more specific focus on faces than VGG.

Further, for CNN features in the video system, we want not only use the broad contextual information from the entire frame, but also the more specific information from the face. To this end, as illustrated in Figure 3, the feature extraction of our proposed system achieves both an image-based video representation and a face-based video representation by using VGG and VGGFace, respectively.



(a) Max Pooling (b) Temporal k-max Pooling

Fig. 4: Illustration of max pooling and temporal k-max pooling, which are used for length normalisation in order to provide fixed length feature vectors for a later emotion classifier.

3.2 Action Units

Facial action units (AUs), which reflect facial muscle movements, are of importance in nonverbal behaviour and emotion recognition systems [21]. For example, [33] recently presented the use of AUs, resulting in a superior multimodal emotion recognition system in the wild; this result encouraged us to look into AUs in this challenge. The aim of AUs is to provide features that are complementary to the CNN features. We estimate action units for each video frame by using the OpenFace toolkit [1], resulting in 14 AU intensity factors and 6 AU occurrence factors extracted for each frame.

3.3 Length Normalisation

Length normalisation methods encode video sequence data into a fixed-length vector video representation by pooling all the descriptors from all the frames. Max pooling over video frames, as shown in Figure 4, is typically used in video emotion recognition and is considered as the default length normalisation method tested in this paper. Such a pooling method results in one identical video representation, which simplifies model learning.

Max pooling, however, ignores all temporal information within the video. This information has been found important for distinguishing between different emotions. Therefore, we also test temporal k-max pooling in order to preserve valuable temporal information. The temporal k-max pooling, shown in Figure 4, is applied to the frame-level features (e. g., CNN features) where the whole frame sequence is divided into k sub-sequences in a temporal manner and a max pooling step is used over frames in each sub-sequence. Note that, temporal k-max pooling corresponds to max pooling when k is equal to 1.

3.4 Temporal Modelling with LSTM RNNs

Since temporal k-max pooling captures valuable time varying information within the video sequences, it enables us to perform temporal modelling. A large number of previous works suggest that LSTM RNNs are good at exploiting temporal information [31, 34, 35]. Therefore, in addition to SVMs used in model learning, LSTM RNNs are also used to leverage temporal information.

The LSTM RNN model uses one or multiple LSTM blocks [9, 35]. Every memory block consists of self-connected linear memory cells $\mathbf{c}$ and three multiplicative gate units: an input gate $\mathbf{i}$, a forget gate $\mathbf{f}$, and an output gate $\mathbf{o}$. Given an input $\mathbf{x}_t$ at the time step t, the activations of the input gate $\mathbf{i}_t$, the forget gate $\mathbf{f}_t$, the output gate $\mathbf{o}_t$, the candidate state value $\mathbf{g}_t$, the memory cell state $\mathbf{c}_t$ are separately computed by the following equations:

$$\mathbf{i}_t = \mathrm{sigm}(\mathbf{W}_{ix}\mathbf{x}_t + \mathbf{W}_{ih}\mathbf{h}_{t-1} + \mathbf{b}_i), \tag{5}$$

$$\mathbf{f}_t = \mathrm{sigm}(\mathbf{W}_{fx}\mathbf{x}_t + \mathbf{W}_{fh}\mathbf{h}_{t-1} + \mathbf{b}_f), \tag{6}$$

$$\mathbf{o}_t = \mathrm{sigm}(\mathbf{W}_{ox}\mathbf{x}_t + \mathbf{W}_{oh}\mathbf{h}_{t-1} + \mathbf{b}_o), \tag{7}$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_{gx}\mathbf{x}_t + \mathbf{W}_{gh}\mathbf{h}_{t-1} + \mathbf{b}_g), \tag{8}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, \tag{9}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{10}$$

where $\mathbf{W}$ is a weight matrix of the mutual connections; $\mathbf{h}_t$ represents the output of the hidden block; $\mathbf{b}$ indicates the block bias, $\odot$ indicates the convolution operation.

4 Multimodal Systems

Our multimodal systems are based on the aforementioned audio and video systems (see Sections 2 and 3). Late fusion or decision level fusion are adopted simply because it has been constantly proven efficient in multimodal emotion recognition tasks [33]. Specifically, we select a simple yet powerful, and widely used *average rule* to fusion scores from different models.

5 Results

5.1 The MEC 2016 Data and Evaluation Metrics

The MEC 2016 data are a subset of the Chinese Natural Emotional Audio-Visual Database (CHEAVD) that consists of video clips from a variety of Chinese movies and TV programs. This subset was chosen with the aim to provide natural emotion data close to real-world environments [2, 16]. The challenge data contain samples labelled in eight emotional states: *angry*, *anxious*, *disgust*, *happy*, *neutral*, *sad*, *surprise* and *worried*. In total, there are 2 852 examples, which are partitioned into a *Training* (Tr.) set (1 981), a *Validation* (Val.) set (243), and a *Test* set (628). The full details of the challenge data can be found in [16]. As the dataset is unbalanced, *Macro Average Precision* (MAP) is used as the primary metric in this challenge.

Table 1: Audio results (in %) on the MEC 2016 test data. The last 3 runs combined the validation (Val.) set and the training (Tr.) set to form a larger training set.

Methods	MAP (Accuracy)
Audio baseline	24.02 (24.36)
Run 1: BoAW-SVM (Tr.)	19.04 (25.16)
Run 2: BoAW-HELM (Tr.)	30.61 (25.16)
Run 3: BoAW-SVM (Tr. + Val.)	23.95 (25.80)
Run 4: BoAW-SVM (Tr. + Val., excluded talk-show data)	23.54 (25.32)
Run 5: BoAW-SVM + Up-sampling (Tr. + Val.)	**36.11** (32.17)

Table 2: Confusion matrix of the best audio system on the test set in the 8-way (i. e., Angry (Ang), Anxious (Anx), Disgust (Di), Happy (Ha), Neutral (Ne), Sad (Sa), Surprise (Su), Worried (Wo)) emotion recognition.

		Predicted Labels							
		Ang	**Anx**	**Di**	**Ha**	**Ne**	**Sa**	**Su**	**Wo**
True Labels	**Ang**	33	1	5	11	8	15	0	2
	Anx	0	26	1	4	3	3	3	2
	Di	24	3	3	18	1	34	3	0
	Ha	9	2	3	13	4	18	4	3
	Ne	14	5	2	21	12	40	8	1
	Sa	32	4	3	26	5	67	2	3
	Su	13	0	4	10	14	15	16	1
	Wo	3	1	1	8	2	4	0	32

5.2 Audio Results

A wide range of preliminary experiments were performed to establish a suitable BoAW and SVM combination; the aim of this testing was to establish the suitability of BoAW for in-the-wild emotion classification. In preliminary experiments, we tested five distinct LLD sets: the 18 LLDs of the GeMAPS feature set [6], the 23 LLDs of eGeMAPS [6], the 16 LLDs and their corresponding delta regression coefficients of the INTERSPEECH 2009 Emotion Challenge feature set (IS09-Emotion) [26], the 38 LLDs and their deltas of the INTERSPEECH 2010 Paralinguistics Challenge set (IS10-Paraling) [24], and the 65 LLDs and their deltas of the ComParE feature set [27]. To optimise the codebook dimension, we chose three different sizes, i. e., 500, 1 000, 2 000. To optimise the number of assignments, LLDs of each frame were mapped to the $d = [25, 50, 100]$ closest words.

These experiments (results not given) revealed that the best combination was found to be BoAW formed formed from the IS10-Paraling feature set (Codebook size 500, $d = 25$), in combination with a polynomial SVM (Degree: 1, Cost: 1). This set-up gave a validation MAP of 32.62 %. Despite this relatively strong performance on the validation set, this audio system did not perform well on the test set, achieving a MAP of 19.04 %.

As with the BoAW-SVM system, a wide range set of preliminary experiments was performed to establish a robust set-up for the BoAW-HELM system. Again, the IS10-Paraling feature set was identified as the most suitable for forming the BoAW representation (Codebook size 500, $d = 100$). Further initial testing also revealed the benefits of applying *Canonical Correlation Analysis* (CCA) feature selection [13], before HELM training. It was observed during these initial tests that it was easy for the ELM systems to overfit to their training set. Therefore to establish a robust set-up, system configurations were chosen which performed well under both validation (train on training set, tested on validation set) and pseudo-test conditions (trained on training and validation sets, tested the extra set of labelled data released by the challenge organisers). This testing set-up revealed a single layer polynomial kernel ELM (Degree: 5, C: 0.2) set-up was able to achieve a MAP of 33.78 %, and our BoAW-HELM system (Degree: 9, C: 500) was able to achieve a MAP of 37.88 % representing a 12 % relative improvement over the single layer ELM system. However, as with the BoAW-SVM system, the strong performance of the HELM system in validation did not generalise onto the test set where it achieved a MAP of 30.61 %; which is a relative increase of 27 % over the challenge test set baseline.

To help minimise the effects related to potential overfitting in the BoAW-SVM and the BoAW-HELM audio systems, we re-trialled the BoAW-SVM system combining the training set and the validation set to form a larger training set with 2 473 instances. We then performed a series of ten-fold cross validation preliminary experiments on the combined training set and found that the BoAW representation (Codebook size 2000, $d = 25$), of IS09-Emotion in conjunction with polynomial SVM (Degree: 1, Cost: 0.04) produced the best ten-fold cross validation performance; this set-up was used for the rest audio systems. This system achieved a MAP of 45.99 % under the cross validation condition; however, disappointingly it achieved a MAP of 23.95 % on the test set.

Because the MEC dataset includes both movie clips and talk-show data, the training set potentially exhibits an unwanted source of noise. Therefore, to investigate this potential effect, our fourth audio system excludes the talk-show data from the whole training data and uses the cleaned training data for modelling. This system achieved a MAP of 74.66 % under the cross validation condition. However, this system obtained a MAP of only 23.54 % on the test set.

Inspired by the baseline systems using sampling techniques to balance the training data with talk-show data, we up-sampled the full training set before the model learning for the fifth audio system. The final submission system yields a MAP of 54.66 % under the cross validation condition and a test set MAP of 36.11 %. This is a relative increase of 50 % over the challenge baseline. Table 1 summarises the performance of our five audio systems on the test set and the confusion matrix of the fifth audio system is presented in Table 2.

Table 3: Video results (in %) on the MEC 2016 test data.

Methods	MAP (Accuracy)
Video baseline	34.28 (19.59)
Run 1: VGG (max), VGGFace (k=1,3), AU (Average late fusion)	42.04 (35.89)
Run 2: VGG (max), VGGFace (k=1,3,5,7), AU (Majority voting)	45.21 (35.03)
Run 3: VGGFace (k = 3) (LSTM)	**53.43** (32.17)
Run 4: VGG (max), VGGFace (k=1,3,5), AU (Average late fusion)	43.34 (35.67)
Run 5: VGG (max), VGGFace (k=1,3,5,7), AU (Average late rule)	44.36 (34.13)

Table 4: Confusion matrix of the best video system on the test set.

		Predicted Labels							
		Ang	**Anx**	**Di**	**Ha**	**Ne**	**Sa**	**Su**	**Wo**
True Labels	**Ang**	16	0	0	3	0	44	12	0
	Anx	3	22	0	1	0	8	7	1
	Di	9	0	1	3	0	38	35	0
	Ha	3	0	0	26	0	18	9	0
	Ne	13	0	0	5	1	56	28	0
	Sa	14	0	2	7	0	93	25	1
	Su	9	0	0	3	0	40	20	1
	Wo	2	0	2	2	0	15	7	23

5.3 Video Results

As mentioned, our videos systems make full use of the VGG video representation, the VGGFace video representation, as well as the AU video representation in conjunction with SVMs and LSTM. A large number of preliminary experiments were performed to identify for the best set-up on the validation set. Consequently, we found that the VGG-based SVM system, the VGGFace-based SVM system, and the AU-based SVM system obtain a validation MAP of 17.17 %, 33.31 %, and 30.16 %, respectively, which are all higher than the video baseline. It is worth noting that the VGG-based SVM system is surprisingly competitive with the baseline video system; this indicates that CNN video representations derived from whole frames are as informative as hand-crafted video representation derived from faces. Average pooling was also tested as a length normalisation method. However, it resulted in worse performance than max pooling, and hence, was not used in our emotion recognition systems.

Encouraged by the preliminary experiments above, we next investigated the combination of the VGGFace video representation, the temporal k-max pooling, and LSTM on the given training and validation sets. We trained the LSTM network by feeding the k-max pooling video representations from the VGGFace descriptors to the network and using the last sequence to produce the class prediction. We found that the VGGFace-based LSTM system obtains the best validation macro precision of 47.17 % when $k = 3$ for k-max pooling.

Table 5: Multimodal results (in %) on the MEC 2016 test data. Each multimodal system was found by performing average late fusion of the equivalent audio and video systems.

Methods	MAP (Accuracy)
Multimodal baseline	30.63 (21.18)
Average late fusion of ‘Run 1 from Table 1’ and ‘Run 1 of Table 3’	**49.43** (34.87)
Average late fusion of ‘Run 2 from Table 1’ and ‘Run 2 of Table 3’	44.78 (34.39)
Average late fusion of ‘Run 3 from Table 1’ and ‘Run 3 of Table 3’	36.70(28.34)
Average late fusion of ‘Run 4 from Table 1’ and ‘Run 4 of Table 3’	43.55 (35.35)
Average late fusion of ‘Run 5 from Table 1’ and ‘Run 5 of Table 3’	43.82 (34.24)

Table 6: Confusion matrix of the best multimodal system on the test set.

		Predicted Labels							
		Ang	**Anx**	**Di**	**Ha**	**Ne**	**Sa**	**Su**	**Wo**
True Labels	**Ang**	8	1	0	4	0	54	8	0
	Anx	1	22	0	1	0	11	7	0
	Di	5	0	1	5	0	54	20	1
	Ha	1	0	0	31	0	21	3	0
	Ne	3	0	0	8	0	70	21	1
	Sa	9	0	0	12	0	106	10	5
	Su	2	1	0	4	1	45	20	0
	Wo	0	0	0	5	0	13	2	31

Furthermore, we realised that the strong performance of a video system in validation may not generalise to the test set. Hence, we decided to use five-fold cross validation on the training set plus the validation set to select a robust video model. As with the audio, we also excluded the talk-show data in an attempt to clean the training data for model learning. Using this set-up, we obtain a cross validation macro precision of 37.47 %, 52.05 %, and 23.55 % for the VGG-based, VGGFace-based, and AU-based systems, respectively.

Using the set-up established in our preliminary experiments, all our five submission systems achieve very notable performance on the test data. Table 3 shows their results on the MEC 2016 test data. Four of our submissions used a Radial Basis Function SVM (Gamma: 2^{-12}, Cost: 10) as this set-up obtained consistently good cross-validation performance across different video features. Our other system (Run 3, Table 3) used LSTM and temporal k-max pooling, where the network has one hidden layer with 256 hidden units; $k = 3$ for k-max pooling achieved our best video-system MAP of 53.43 %, which is a relative increase of 56 % over the challenge baseline. Table 4 presents the confusion matrix for this system.

5.4 Multimodal Results

Table 5 shows the performance of each of our submitted runs for the multimodal task. Each result has been found by performing average late fusion of the equivalent audio and video systems, e. g., Run 1 in Table 5 is the fusion of Audio System Run 1 (from Table 1) and Video System Run 1 (from Table 3). It can been seen from Table 5 that all of our systems outperformed the baseline multimodal system. Table 6 depicts the confusion matrix of the best multimodal system which yielded a MAP of 49.43 %; a relative increase of 61 % over the challenge baseline.

6 Conclusions

This paper presented the University of Passau's audio, video and multimodal systems for submission to the Multimodal Emotion Recognition Challenge 2016. For our audio systems, we investigated the effectiveness of a BoAW representation in a combination with ELMs, Hierarchical ELMs, and SVMs. Disappointingly, the strong performance of the audio systems during system development did not generalise to testing. For our video-based systems we leveraged LSTM RNNs based on visual features extracted from deep CNNs and facial action units. All of our videos systems outperformed the challenge baseline, indicating the benefits of using video features extracted from CNNs pre-trained for object recognition or face recognition, for emotion classification. Our multimodal results highlight the benefit of using late fusion. As with our video system, all of our multimodal approaches outperformed the challenge baseline.

Future audio work will explore the benefits of using different techniques to form BoAW codebook and the advantages offered by different Neural Network based classifiers. To further improve the performance of our video systems, we will investigate the use of Bidirectional Long Short-Term Memory Recurrent Neural Networks. We will also consider sampling methods for audiovisual emotion recognition as a way to balance the training data.

Acknowledgments. This work has been partially supported by the BMBF IKT2020-Grant under grant agreement No. 16SV7213 (EmotAsS), the European Communitys Seventh Framework Programme through the ERC Starting Grant No. 338164 (iHEARu), and the EU's Horizon 2020 Programme agreement No. 688835 (DE-ENIGMA), and the European Union's Horizon 2020 Programme through the Innovative Action No. 645094 (SEWA). It was further partially supported by research grants from the China Scholarship Council (CSC) awarded to Xinzhou Xu.

References

1. Baltrušaitis, T., Robinson, P., Morency, L.P.: OpenFace: an open source facial behavior analysis toolkit. In: Proc. WACV. pp. 1–10. Lace Placid, USA (2016)

2. Bao, W., Li, Y., Gu, M., Yang, M., Li, H., Chao, L., Tao, J.: Building a Chinese natural emotional audio-visual database. In: Proc. ICSP. pp. 583–587. IEEE, Hangzhou, China (2014)
3. Deng, J., Zhang, Z., Marchi, E., Schuller, B.: Sparse autoencoder-based feature transfer learning for speech emotion recognition. In: Proc. ACII. pp. 511–516. IEEE (2013)
4. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A deep convolutional activation feature for generic visual recognition. In: Proc. ICML. pp. 647–655. Beijing, China (2014)
5. El Ayadi, M., Kamel, M.S., Karray, F.: Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44(3), 572–587 (2011)
6. Eyben, F., Scherer, K., Schuller, B., Sundberg, J., André, E., Busso, C., Devillers, L., Epps, J., Laukka, P., Narayanan, S., Truong, K.: The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing 7(2), 190–202 (2016)
7. Eyben, F., Weninger, F., Groß, F., Schuller, B.: Recent developments in openSMILE, the Munich open-source multimedia feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia. pp. 835–838. MM '13, ACM (2013)
8. Han, K., Yu, D., Tashev, I.: Speech emotion recognition using deep neural network and extreme learning machine. In: Proc. INTERSPEECH. pp. 223–227. Singapore (2014)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
10. Huang, G., Huang, G.B., Song, S., You, K.: Trends in extreme learning machines: A review. Neural Networks 61, 32–48 (2015)
11. Huang, G.B.: What are extreme learning machines? Filling the gap between Frank Rosenblatt's Dream and John von Neumann's Puzzle. Cognitive Computation 7(3), 263–278 (2015)
12. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: Theory and applications. Neurocomputing 70(1-3), 489–501 (2006)
13. Kaya, H., Eyben, F., Salah, A.A.: CCA based feature selection with application to continuous depression recognition from acoustic speech features. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3757–3761. IEEE, Florence, Italy (2014)
14. Kaya, H., Salah, A.A.: Combining modality-specific extreme learning machines for emotion recognition in the wild. In: Proceedings of the 16th International Conference on Multimodal Interaction. pp. 487–493. ICMI '14, ACM (2014)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proc. NIPS. pp. 1106–1114. Lake Tahoe, USA (2012)
16. Li, Y., Tao, J., Schuller, B., Shan, S., Jiang, D., Jia, J.: MEC 2016: The multimodal emotion recognition challenge of CCPR 2016. In: Proc. CCPR. Chengdu, China (2016), 11 pages
17. Mao, Q., Dong, M., Huang, Z., Zhan, Y.: Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia 16(8), 2203–2213 (2014)
18. Pancoast, S., Akbacak, M.: Bag-of-audio-words approach for multimedia event classification. In: Proc. INTERSPEECH. pp. 2105–2108. Portland, USA (2012)

19. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proc. BMVC. pp. 41.1–41.12. Swansea, UK (2015)
20. Pokorny, F., Graf, F., Pernkopf, F., Schuller, B.: Detection of negative emotions in speech signals using bags-of-audio-words. In: Proc. ACII. pp. 879–884. Xi'an, China (2015)
21. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: A survey of registration, representation, and recognition. IEEE Trans. Pattern Anal. Mach. Intell. 37(6), 1113–1133 (2015)
22. Schmitt, M., Ringeval, F., Schuller, B.: At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech. In: Proc. INTERSPEECH. San Francsico, USA (2016), 5 pages
23. Schmitt, M., Schuller, B.W.: openxbow-introducing the passau open-source crossmodal bag-of-words toolkit. CoRR abs/1605.06778 (2016)
24. Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S.S.: The INTERSPEECH 2010 Paralinguistic Challenge. Proc. INTERSPEECH pp. 2794–2797 (2010)
25. Schuller, B., Batliner, A.: Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. Wiley Publishing, Chichester, United Kingdom (2013)
26. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 emotion challenge. In: Proc. INTERSPEECH. pp. 312–315. Brighton, UK (2009)
27. 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.: The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In: Proc. INTERSPEECH. pp. 148–152. Lyon, France (2013)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR (2014), `http://arxiv.org/abs/1409.1556`
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proc. CVPR. pp. 1–9. Boston, USA (2015)
30. Tang, J., Deng, C., Guang, G.B.: Extreme learning machine for multilayer perceptron. IEEE Transactions on Neural Networks and Learning Systems 27(4), 809–821 (2015)
31. Wöllmer, M., Kaiser, M., Eyben, F., Schuller, B., Rigoll, G.: LSTM-modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing, Special Issue on Affect Analysis in Continuous Input 31(2), 153–163 (2013)
32. Xu, Z., Yang, Y., Hauptmann, A.G.: A discriminative CNN video representation for event detection. In: Proc. CVPR. pp. 1798–1807. Boston, USA (2015)
33. Yao, A., Shao, J., Ma, N., Chen, Y.: Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In: Proc. ICMI. pp. 451–458. ACM, Seattle, USA (2015)
34. Zhang, Z., Pinto, J., Plahl, C., Schuller, B., Willett, D.: Channel mapping using bidirectional long short-term memory for dereverberation in hands-free voice controlled devices. IEEE Transactions on Consumer Electronics 60(3), 525–533 (2014)
35. Zhang, Z., Ringeval, F., Han, J., Deng, J., Marchi, E., Schuller, B.: Facing realism in spontaneous emotion recognition from speech: Feature enhancement by autoencoder with LSTM neural networks. In: Proc. INTERSPEECH. San Francsico, CA (2016), 5 pages