ABSTRACT

This paper presents our work on ACM MM Audio Visual Emotion Corpus 2014 (AVEC 2014) using the baseline features in accordance with the challenge protocol. For prediction, we use Canonical Correlation Analysis (CCA) in affect sub-challenge (ASC) and Moore-Penrose generalized inverse (MPGI) in depression sub-challenge (DSC). The video baseline provides histograms of Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) features. Based on our preliminary experiments on AVEC 2013 challenge data, we focus on the inner facial regions that correspond to eyes and mouth area. We obtain an ensemble of regional linear regressors via CCA and MPGI. We also enrich the 2014 baseline regressor with Local Phase Quantization (LPQ) features extracted using Intraface toolkit detected/tracked faces. Combining both representations in a CCA ensemble approach, on the challenge test set we reach an average Pearson’s Correlation Coefficient (PCC) of 0.3932, outperforming the ASC test set baseline PCC of 0.1966. On the DSC, combining modality specific MPGI based ensemble systems, we reach 9.61 Root Mean Square Error (RMSE).

Categories and Subject Descriptors

G.3 [Mathematics of Computing]: Probability and Statistics–Correlation and regression analysis; I.4.7 [Image Processing and Computer Vision]: Feature Measurement; I.5.4 [Computing Methodologies]: Pattern Recognition–Signal processing

General Terms

Human-Computer Interaction

Keywords

canonical correlation analysis; audio-visual emotion corpus; ensemble learning; emotion prediction; depression prediction

1. INTRODUCTION

A fundamental aspect of human communication is non-verbal. Analyzing non-verbal cues from audio and video for the purpose of human behavior understanding is attracting increasing attention. Such an endeavor has manifold applications ranging from intelligent tutoring systems to affect sensitive robots. A proposal is to start the study for a variety of affective states (e.g. stances, attitudes, personality, emotion) from analysis of emotion [8], since it is a central notion inter-related with other short and long term states.

In order to boost the study on emotion and related concepts and to bridge the gap between state-of-the-art research on signal processing/machine learning and the studies in the field, challenges in audio [30, 32, 33, 31] and audio-visual [34, 39, 38] modalities are being organized. These challenges provide a unique opportunity for comparability and transparency of the work in the field, while bringing together experts from different disciplines such as psychology, audio-visual signal processing, and machine learning.

The most recent ACM MM Audio-Visual Emotion Corpus and Challenge (AVEC 2014) focuses on prediction of self-reported level of depression and multiple rater human annotated 3 affective dimensions (arousal, valence and dominance) [38]. The providers organized a baseline video and acoustic features along with the video clips partitioned into training, developments and test sets.

In this paper, we use an ensemble of Canonical Correlation Analyzers (CCA) to predict affective dimensions from video modality. We also utilize Moore-Penrose Generalized Inverse (MPGI) and Extreme Learning Machines (ELM) for audio-visual depression recognition. CCA is a statistical method that finds linear projections for two views (representations) of a semantic object to maximize the mutual correlation [13, 12]. CCA has a variety of applications in pattern recognition ranging from multi-modal fusion [12, 26] to feature selection [27, 19]. CCA also has successful applications in affective computing [36, 35]. In [35] Shan et al. use CCA for feature level fusion of body gestures and facial expressions on gesture and emotion recognition tasks. In [36] the authors utilize CCA introducing Matrix CCA (MCCA) for facial action recognition and facial parts synthesis.

A previous study on AVEC 2013 ASC [29] applies CCA between MFCC based low level descriptors and the affective dimensions. Another study on AVEC 2013 benefits from CCA as an acoustic feature selector for depression, where the authors utilize the projection vector to rank the features [19].

Here, we use CCA for feature extraction and regression in
video modality. In a recent work on the same corpus, CCA is used for extraction of audio-visual depression covariates [20].

While the competition measure in ASC is correlation, in DSC the measure is RMSE. Therefore, a different base learner than CCA, which maximizes correlation, should be used in this case. In the latter sub-challenge, we use Moore-Penrose generalized inverse [25] motivated from output layer learning rule of Extreme Learning Machines (ELMs) [17]. ELMs provide a unified framework for regression and multi-classification for a generalized Single Layer Feedforward Networks (SLFNs) including Support Vector Machines (SVMs) [15]. For both sub-challenges our base learners are related to ELMs. However, we use Principal Component Analysis (PCA) for the input layer instead of random projections used in basic ELM [17]. Unlike ELMs, we do not utilize non-linear activation functions at hidden layer. For the output layer, the learning rule in DSC is the same with basic ELM that provides a least squares solution. As mentioned, for ASC CCA is employed to learn output layer weights.

To allow comparability with AVEC 2013 baseline set [39] and to enrich the 2014 baseline video feature set, we extract Local Phase Quantization (LPQ) features. In both feature representations we focus on regions corresponding to eyes (including eyebrows) and mouth area, as these are found to be the most informative in our preliminary experiments with AVEC 2015 corpus. We attribute the better performance of eyes and mouth to information they carry about the target tasks (emotion, depression) while the remaining regions are thought to show stronger tendency to represent the speaker identity.

The remainder of this paper is organized as follows. In the next section we provide background on CCA and ELM. Then in Section 3 we briefly introduce corpus and baseline feature sets. In Section 4 we give the experimental results. Finally, Section 5 concludes with future directions.

2. BACKGROUND

In this section, background information regarding CCA and ELMs that are used extensively in the proposed systems are provided.

2.1 Canonical Correlation Analysis

Proposed early in 1936 by Hotelling [13], CCA seeks to maximize the mutual correlation between two sets of variables by finding linear projections for each set. Mathematically, CCA seeks to maximize the mutual correlation between two views of the same semantic phenomenon (e.g. audio and video of a speech) denoted $X \in \mathbb{R}^{n \times d}$ and $Y \in \mathbb{R}^{n \times p}$, where $n$ denotes the number of paired samples, via:

$$\rho(X, Y) = \sup_{w, v} \text{corr}(Xw, Yv),$$  

(1)

where “corr” corresponds to Pearson’s correlation, $w$ and $v$ correspond to the projection vectors of $X$ and $Y$, respectively. Let $C_{XY}$ denote the cross-set covariance between the sets $X$ and $Y$, and similarly let $C_{XX}$ denote within set covariance for $X$. The problem given in eq. (1) can be reformulated as:

$$\rho(X, Y) = \sup_{w, v} \frac{w^T C_{XY} v}{\sqrt{w^T C_{XX} w} \sqrt{v^T C_{YY} v}}.$$  

(2)

The formulation in Eq. (2) can be converted into a generalized eigenproblem for both projections (i.e. $w$ and $v$), the solution can be shown [12] to have the form of:

$$C_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{YX} = \lambda w,$$  

(3)

where the correlation appears to be the square root of eigenvalue:

$$\rho(X, Y) = \sqrt{\lambda}.$$  

(4)

To attain maximal correlation, the eigenvector corresponding to the largest eigenvalue in Eq. (3) should be selected. Similarly, by restricting the new vectors to be uncorrelated with the previous ones, it can be shown that the projection matrices for each set are spanned by the $k$ eigenvectors corresponding to the $k$ largest eigenvalues. Non-linearity can be incorporated into CCA using the kernel trick [12] or deep neural networks [3]. CCA is related to two extensively used statistical methods: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [11]. When CCA is applied between feature matrix and identity matrix, the result is PCA. Barlett showed that when LDA is a special case of CCA, and can be obtained when CCA is applied between feature matrix and 1-of-K coded label matrix [4]. Another recent method called Covariance Operator Inverse Regression (COIR) [22] is proved to give identical central space with Kernel CCA.

2.2 CCA for Regression

CCA is commonly used for several tasks ranging from co-variate extraction, to modality/representation fusion [26]. It has also applications in ranking feature selection [27, 19]. However, though it can be used for non-linear regression, this use is not common to the best of our knowledge. Among few studies in this vein, Nicolaou et al. introduce Correlated-Spaces Regression (CSR) inspired from CCA and the high inter-correlation of emotion dimensions [23].

To employ CCA as a regressor using eq. (1), we note that as the correlation is 1, the covariates of the two views $X$ and $Y$ (i.e. in the projected space) become identical:

$$\rho(X, Y) = \sup_{w, v} \text{corr}(Xw, Yv) = 1 \iff Xw = Yv,$$  

(5)

where we assume both $X$ and $Y$ are mean-normalized. When correlation is close to 1, we can use one representation to reconstruct the other. Let column vectors $\mu_X$ and $\mu_Y$ denote respective training set means used to normalize the sets, eq. (5) can be rewritten as:

$$(Y - 1 \mu_Y^T) v \simeq (X - 1 \mu_X^T) w,$$  

(6)

where $1$ is a column vector of ones having length equal to respective dataset cardinality. It is straightforward to convert eq. (6) to reconstruct one side (here $Y$):

$$Y \simeq (X - 1 \mu_X^T) w^T v^T + 1 \mu_Y^T,$$  

(7)

where $^T$ denotes a generalized inverse. Note that in the case of regression $v$ is a scalar. This approach can be used in regression where one view represents the target variables. This study, this CCA based regression scheme is used on regional video features versus affective dimensions.

2.3 Extreme Learning Machines

Extreme Learning Machine (ELM) was first introduced a decade earlier [16] 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. [17], 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; and is related to generalized SLFN class including Least Square SVM (LSSVM) [37, 15]. Due to fast and accurate results obtained via ELMs, the method is applied in many real life tasks ranging from gesture recognition to representational learning [14, 7]. In this section, we provide a brief introduction to the paradigm.

The argument of basic ELM introduced by Huang et al. 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; the second layer (output weights) can be effectively and efficiently solved via least squares [17]. It can be thought that the input layer carried out unsupervised feature mapping, then the activation function outputs (the output matrix) is subjected to a supervised learning procedure. Let $x \in \mathbb{R}^d$ denote an input sample, $h(x) \in \mathbb{R}^d$ denote the hidden node output. Similarly, let $X \in \mathbb{R}^{n \times d}$ denote the dataset and $H \in \mathbb{R}^{n \times p}$ denote the hidden node output matrix. The hidden node activation via randomly generated mapping matrix $W$ and bias vector $b$ is defined as in regular SLFN:

$$H(l, t) = h_i(x^l) = g(x^l, w_l, b_l), l = 1, \ldots, L, t = 1, \ldots, N,$$

where nonlinear activation function $g()$ can be any infinitely differentiable bounded function [17]. A common choice for $g()$ is sigmoid function:

$$g(x, a, b) = \frac{1}{1 + \exp(-a \cdot x + b)}.$$

(9)

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

$$T_{i,m} = \begin{cases} +1 & \text{if } y_i = m, \\ -1 & \text{if } y_i \neq m. \end{cases}$$

(10)

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

$$\beta = H^T \beta = H^T,$$

(11)

where $H^T$ is the Moore-Penrose generalized inverse [25] that gives not only minimum $L_2$ norm solution to $\|H\beta - T\|$, but also minimizes the norm of projection $\|\beta\|$. The use of this special generalized inverse is motivated by Barlett’s theory stating that for networks approximating an arbitrarily small training error, the smaller the norm of weights is, the better the generalization capability of the network [5]. The universal approximation and classification capability of ELMs have been rigorously discussed in the literature (cf. [15]), and are beyond the scope of this paper. However, it is important to mention that ELM is related to Least Square SVMs via the following output weight learning formulation:

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

(12)

where $I$ is $N \times N$ identity matrix, and $C$ used to regularize the linear kernel $HH^T$ is indeed the complexity parameter of LSSVM [37]. The approach is extended to use any valid kernel. A popular choice for kernel function is Gaussian (RBF):

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

(13)

In both (basic and kernel) approaches, the prediction of $x$ is given via $\hat{y} = h(x)\beta$. In case of multi-class classification the class with maximum score in $\hat{y}$ is selected. In our study we utilize both basic and kernel version 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.

We next introduce the AVEC 2014 data used for experimental validation.

### 3. THE CORPUS AND FEATURES

AVEC 2013 and 2014 [39, 38] use a subset of the audio-visual depressive language corpus (AVDLC), which includes 340 video clips of subjects performing a Human-Computer Interaction task while being recorded by a webcam and a microphone. In AVDLC, the total number of subjects is 292 and only one person appears per clip, i.e. some subjects feature in more than one clip. The clip duration ranges from 20 to 50 minutes, with a total duration of 240 hours and an average of 25 minutes. The age of subjects ranges from 25 to 63. The target variable (BDI-II score) range is 0-45 [39]. Recorded behavior includes speaking out loud while solving a task, counting from one to ten, reading excerpts of a novel and a fable, singing, free talk: telling the best event and a sad event from childhood. The depression levels were labeled per clip using Beck Depression Inventory-II (BDI-II) [6], a subjective self-reported 21 item multiple-choice inventory.

For the AVEC 2014 challenge, only two of the 12 tasks from AVDLC are used. These are referred as Freeform and Northwind tasks. In the Freeform task the subject is asked to recount a good or bad memory from the past using her own words. In the Northwind task a German fable about a competition between Sun and the Northwind is read by the subject. This story has markedly depressing undertones, with the main character facing failure and giving up after experiencing helplessness. For both tasks, the recordings are split into three partitions: training, development, and test sets of 50 recordings each, respectively.

For audio modality, a set of 2268-length openSMILE [10] features, which were introduced in AVEC 2013 [39], are provided to participants. The acoustic feature sets are arranged in three segmentation settings: short (3 s. overlapping frames with 1 s. shifts), long (20 s. overlapping frames with 1 s. shifts), voice-activity detected (VAD) segments. In VAD segmentation a voice activity detector [9] is used to split the clip when there is a pause for more than 200 ms. The statistical functionals (e.g. moments, extremes) are then applied on the Low Level Descriptor (LLD) contours (e.g. MFCC 1-16, F0, jitter) of segments. The reader is referred to the paper on the challenge [38] for further details of acoustic features. In AVEC 2013 challenge paper, features of short and long segmentations are reported to work well on affect and depression, respectively [39].
3.1 Video Feature Sets

The baseline video features of AVEC 2014 consist of Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [1]. LGBP-TOP combines the power of Gabor wavelet representation of image with TOP extension. Gabor wavelet is obtained from convolution of a Gaussian and a sinusoid. With varying rotation and phase angles, a set of Gabor pictures are obtained for each video frame then Local Binary Patterns (LBP) are computed for three orthogonal planes (i.e. XY, XZ and YZ). The patterns are finally represented as a histogram. Since this histogram representation does not keep the structural information of facial features, the face is divided into $4 \times 4 = 16$ regions and an LGBP-TOP histogram is computed per region. In AVEC 2014 baseline set, 18 Gabor pictures with 59 dimensional uniform pattern LBP [24] are computed on XY plane resulting in 1062-dimensional histograms per region.

To accompany the baseline feature set provided by the challenge organizers, we extract LPQ features that served as baseline in AVEC 2013 and are previously shown to be useful for facial action detection [18]. We detect and track the faces via a freely available facial landmark detector tool developed by Xiong and De La Torre[40]. Using the 49 landmark points provided by the detector we determine the facial area, align it using the eye centers, then crop and scale it such that the provided by the detector we determine the facial area, align it using the eye centers, then crop and scale it such that the inter-ocular distance is 120 pixels and the right eye center is located at (65, 80) coordinates.

The LPQ features are computed by taking 2-D Discrete Fourier Transform (DFT) of M-by-M neighborhood of each pixel in the gray scale image. In our implementation we use $M = 9$. The 2D-DFT is computed at four frequencies $\{[a, 0]^T, [0, a]^T, [a, a]^T, [a, -a]^T\}$ with $a = 1/M$, which correspond to four of eight neighboring frequency bins centered at the pixel of interest. The real and imaginary parts of resulting four complex numbers are separately quantized using a threshold of zero, that gives an eight bit string. The eight bit string is then converted into an integer value in the range 0-255. The pixel based values are finally converted into a histogram of 256 bins. Since this histogram representation does not keep the structural information of facial features, the face is divided into $5 \times 5 = 25$ regions and an LPQ histogram is computed per region. Here we focus on only 5 regions: two containing eyes with brows and three covering the mouth area. We augment the raw LPQ descriptors with first and order second deltas.

Before we proceed with the experiments, we provide canonical correlation analysis of video baseline features in the next section.

3.2 Analysis of Features

As a preliminary probe into the descriptive/predictive power of video features, we carried out analysis on the Freeform training set video features using CCA. The summary of Canonical Correlation (CC) of video features against the affective labels are given in Table 1. From the table we observe that even using a single region, the video features are highly informative, moreover CCA can be used as a predictor for the development/test set. We further analyzed the same sets of video features for a multi-task learning. In this scheme we applied CCA between the video features and all three affective dimensions. Though in this work we focus on predicting individual affective dimensions using CCA, a probe into the multi-task learning performance may be useful for further studies. The canonical correlations obtained via multi-task CCA are given in Table 2. Similar to individual projections, we see that both the statistical method and the features deserve further analysis for a multi-task study. Note that here the projection vectors do not correspond to affective dimensions, however they can be used for feature reduction to be subsequently used in all three affective dimensions.

<table>
<thead>
<tr>
<th>Region</th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right eye</td>
<td>0.9092</td>
<td>0.9189</td>
<td>0.9284</td>
</tr>
<tr>
<td>Left eye</td>
<td>0.9100</td>
<td>0.9129</td>
<td>0.9294</td>
</tr>
<tr>
<td>Right part of mouth</td>
<td>0.9058</td>
<td>0.9166</td>
<td>0.9265</td>
</tr>
<tr>
<td>Left part of mouth</td>
<td>0.9082</td>
<td>0.9161</td>
<td>0.9278</td>
</tr>
<tr>
<td>All regions combined</td>
<td>0.9563</td>
<td>0.9558</td>
<td>0.9639</td>
</tr>
</tbody>
</table>

Table 1: Canonical correlations of video features vs. individual affect labels on the Freeform training set. Regions 6 and 7 correspond to right and left eyes including brows, while 10 and 11 together correspond to mouth area.

<table>
<thead>
<tr>
<th>Region</th>
<th>AVD-v1</th>
<th>AVD-v2</th>
<th>AVD-v3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right eye</td>
<td>0.9365</td>
<td>0.8635</td>
<td>0.8624</td>
</tr>
<tr>
<td>Left eye</td>
<td>0.9357</td>
<td>0.8679</td>
<td>0.8527</td>
</tr>
<tr>
<td>Right part of mouth</td>
<td>0.9334</td>
<td>0.8679</td>
<td>0.8587</td>
</tr>
<tr>
<td>Left part of mouth</td>
<td>0.9358</td>
<td>0.8630</td>
<td>0.8541</td>
</tr>
<tr>
<td>All regions combined</td>
<td>0.9683</td>
<td>0.9336</td>
<td>0.9230</td>
</tr>
</tbody>
</table>

Table 2: Canonical correlations of video features vs. all affect labels on the Freeform training set. Since we have 3 linearly independent target variables (AVD) CCA outputs three projection vectors. The canonical correlations associated with each covariate are given in descending order.

4. EXPERIMENTAL RESULTS

In our experiments we adhere to the challenge protocol. We use the training set and optimize our model hyper parameters on the development set. It is important to note that all our experiments were carried out via training and predicting on the same task (e.g. train and predict on the Freeform task). No cross task learning is tested. When a viable method and hyper-parameter set is obtained, we retrain a model using the training and the development set to predict the independent test set, the labels of which are sequestered. The sub-challenge measures for DSC and ASC are Root Mean Squared Error (RMSE) and Pearson’s correlation, respectively.

4.1 Visual Emotion Prediction

In the ASC, the target variables are continuous and the challenge requires casting a prediction at 30Hz, which is the frame rate of video. Therefore, the video baseline features are found more suited for the task 1. The video modality baseline scores for the ASC are given in Table 3. Note that compared to AVEC 2013 [39], the baseline scores for the

1We have also experimented on baseline acoustic features for emotion prediction, however they were not found to contribute to video system.
ASC are dramatically higher. This is partly due to employing multiple annotators and averaging the scores, and partly due to LGBP-TOP descriptors.

Table 3: AVEC 2014 Challenge Video Modality Baselines (Pearson’s Correlation Coefficient Computed over All Sequences)

<table>
<thead>
<tr>
<th>Partition</th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Devel.</td>
<td>0.412</td>
<td>0.355</td>
<td>0.319</td>
<td>0.362</td>
</tr>
<tr>
<td>Test</td>
<td>0.2062</td>
<td>0.1879</td>
<td>0.1959</td>
<td>0.1966</td>
</tr>
</tbody>
</table>

Since the ASC requires the prediction of continuous affective labels per video frame, to overcome the severe negative effects of undetected faces we carry out smoothing over the prediction contours after setting the training set mean of relevant affective dimension as rough prediction for the undetected frame. Mean smoothing is carried out on predictions of each clip over a window of 2K frames with respect to frame of interest.

As our motivation is to divide and conquer the data using ensemble CCA, we first show that ensemble averaging provides better results than using the whole set of features. This is intuitive as the error resulting from variance of predictors in known to decrease via combining multiple learners [2]. The feature level fusion performance of four regions of interest on the development set is given in Table 4. The results are given in Pearson’s Correlation Coefficient (PCC) that is the challenge measure for ASC. We observe that the smoothing has a major effect on performance, however the results only reach the development set baseline even when smoothed with K=120. The development set performance of CCA ensembling can be seen in Table 5. We see that when projected separately, even a simple mean combination outperforms the baseline. Combined with smoothing, it is possible to reach an average PCC of 0.4066 using CCA regression and 0.4073 when 1-Nearest Neighbor regressor is used on the extracted CCA covariate. Since the performance difference of regressors in smoothed predictions is negligible and without smoothing the former performs better, we report further results with only CCA based regression.

We next add the LPQ features into the loop and carry out tests on the development set. We use all the features from 5 aforementioned regions in LPQ, since individually they were not found to yield good performance. Although LPQ features do not provide as good individual performance as LGBP-TOP, they contribute to overall performance in ensemble setting (see Table 6). We see both the non-smoothed and smoothed performance increase yielding a maximum PCC performance of 0.449. The computation of PCC over all sequences regardless of the task gives a result of 0.449. We see that the results change dramatically when the two tasks are handled separately. Using K=150 for smoothing on LPQ plus LGBP-TOP features, the PCC result in the Northwind task is 0.538, while in the Freeform the same setting gives a PCC performance of 0.423. The dramatic difference stems from the nature of the Freeform task: since this is a more in-the-wild task there is a large variation in terms of spoken content and type of reactions. Moreover, difference between the average of two task-dependent correlations (0.481) and the PCC computed over all sequences (0.449) is attributed to difference of variances in two tasks.

Figures 1 and 2 show the effect of smoothing on PCC performance of the Freeform and Northwind tasks, respectively. Both tasks use CCA ensemble in the same setting given in Table 6. The figures also show that the effect of smoothing is mostly salient with smaller values of K, where the slope of contribution of smoothing decreases with increasing K. When we wish to build real-time emotion recognition systems, a small value of K (e.g. in range 10-30), would be appropriate considering the performance-efficiency dilemma.

Table 6: Using CCA regression ensemble of LPQ features (1 covariate) and LGBP-TOP features (4 regional covariates). K=150 is used for smoothing.

<table>
<thead>
<tr>
<th></th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>No smooth</td>
<td>0.4636</td>
<td>0.3992</td>
<td>0.3256</td>
<td>0.3961</td>
</tr>
<tr>
<td>Smoothing</td>
<td>0.5161</td>
<td>0.4424</td>
<td>0.3871</td>
<td>0.449</td>
</tr>
</tbody>
</table>

4.1.1 Enhanced Visual System for Test Set

Motivated by the performance of feature partitioning based CCA ensembles with simple averaging as fusion rule, we extend our test set system in three ways. First, for compu-
Table 4: Performance of CCA correlate of four inner regions of the baseline video features. Here, all the regions are combined at feature level. K=120 is used for smoothing.

<table>
<thead>
<tr>
<th></th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlate</td>
<td>0.3085</td>
<td>0.3626</td>
<td>0.2855</td>
<td>0.3189</td>
</tr>
<tr>
<td>Smoothed Correlate</td>
<td>0.3503</td>
<td>0.4187</td>
<td>0.3375</td>
<td>0.3688</td>
</tr>
<tr>
<td>1-NN Pred. On Correlate</td>
<td>0.2841</td>
<td>0.3282</td>
<td>0.2515</td>
<td>0.2879</td>
</tr>
<tr>
<td>Smoothed 1-NN Pred. On Correlate</td>
<td>0.3525</td>
<td>0.4207</td>
<td>0.3356</td>
<td>0.3696</td>
</tr>
</tbody>
</table>

Table 5: Performance of CCA correlates of four inner regions projected separately. K=120 is used for smoothing.

<table>
<thead>
<tr>
<th></th>
<th>Arousal</th>
<th>Valence</th>
<th>Dominance</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Correlate</td>
<td>0.4569</td>
<td>0.3916</td>
<td>0.2748</td>
<td>0.3744</td>
</tr>
<tr>
<td>Smoothed Mean Correlate</td>
<td>0.4984</td>
<td>0.4201</td>
<td>0.3012</td>
<td>0.4066</td>
</tr>
<tr>
<td>1-NN Pred. On Mean Correlate</td>
<td>0.4012</td>
<td>0.3444</td>
<td>0.2385</td>
<td>0.3280</td>
</tr>
<tr>
<td>Smoothed 1-NN Pred. On Mean Correlate</td>
<td>0.4987</td>
<td>0.4206</td>
<td>0.3025</td>
<td>0.4073</td>
</tr>
</tbody>
</table>

4.2 Audio-visual Depression Prediction

For the DSC video sub-system, we carry out a slight modification on the ASC pipeline. We keep frame level learning as is taking into account only features from face detected frames. We replace first level CCA by MPGI and second level CCA by simple averaging. In our preliminary experiments, we observe that MPGI yields better RMSE performance on the development set than CCA. To our surprise, simple averaging is also found to give better RMSE performance compared to CCA based fusion for final stage. Different from ASC, the predictions from both the Northwind and Freeform tasks can be combined to give a final score for the clip. We therefore train PCA + MPGI based ensemble ELM systems per task, then evaluate their individual and combined (averaged) performance.

For audio sub-system, we utilize long segmented baseline acoustic feature sets, since this segmentation is shown to suit depression sub-challenge [39, 19]. We used ELMs with Linear Kernel\(^2\) and optimized the complexity parameter in the range \(2^{(-15, -14, \ldots, 0)}\). Here also, we trained separate models on the training and development sets in a 2-Fold cross validation manner. Prior to model learning acoustic features are normalized such that they are in the range [-1,1]. We observe that models learned from the Freeform task does not yield better performance than challenge baseline. So,

\(^2\)Available from [www.extreme-learning-machines.org](http://www.extreme-learning-machines.org)
we used only predictions from the Northwind task in audio sub-system. Averaging models learned from the training and development sets, we obtain our acoustic system’s final output. For audio-visual fusion, we combine the Northwind audio and video subsystems for consistency of the video task. The sequenced test set scores of five systems are listed in Table 8. The results indicate that audio only modality generalizes better than video only modality, that was found to give lower than 10 RMSE on the development set. The best results are obtained with audio-visual decision fusion.

Table 8: Challenge test set scores of five systems for predicting depression severity level

<table>
<thead>
<tr>
<th>Task</th>
<th>Modality</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freeform</td>
<td>Video</td>
<td>8.284</td>
<td>10.519</td>
</tr>
<tr>
<td>Northwind</td>
<td>Video</td>
<td>8.254</td>
<td>10.315</td>
</tr>
<tr>
<td>Northwind+Freeform</td>
<td>Video</td>
<td>8.202</td>
<td>10.269</td>
</tr>
<tr>
<td>Northwind</td>
<td>Audio</td>
<td>7.962</td>
<td>9.785</td>
</tr>
<tr>
<td>Northwind</td>
<td>Audio-visual</td>
<td>7.693</td>
<td>9.611</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS AND OUTLOOK

In this study we utilize CCA to extract affective covariates in visual modality as well as audio-visual fusion of ELM based systems for predicting depression severity level. We further employ CCA as a regressor and combine regional facial features in ensemble setting. The ensemble is obtained by training the most informative facial regions (corresponding to eyes and mouth area) separately. We show that CCA ensembling improves over mere feature fusion, and the predictions can be dramatically improved via mean smoothing. We also introduce AVEC 2013 baseline features to accompany the AVEC 2014 video baseline set. We observe that LPQ features are not individually sufficient but improve prediction performance collectively. For the ASC test set predictions, ensemble CCA is extended to instance space partitioning in addition to feature space partitioning, reaching an overall PCC score of 0.3932, which outperforms the challenge baseline PCC of 0.1966, dramatically. The visual ensemble system is also used in audio-visual depression recognition, where CCA is replaced with Moore-Penrose generalized inverse to find a least squares solution with minimum $L_2$ norm projection. For audio based depression prediction, we utilize ELMs with Linear kernel. Combining audio and video sub-systems learned from the Northwind task, we reach a test set RMSE score of 9.611 improving the challenge baseline RMSE score of 10.859.

In the study, we did not utilize any feature selection method, which we think might be useful especially for audio systems. In our future studies, we are planning to apply variants of a recently introduced multi-view feature selection approach [21] utilizing domain knowledge to partition the feature set. Combining the modalities at feature level for depression recognition also deserves further attention.

6. ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for beneficial suggestions. The first author is a member of Faculty Member Training Program (ÖYP) supervised by the Turkish Higher Education Council (YÖK). This study is partly supported by Bogazici University research project BAP 6531.

7. REFERENCES


