Eyes Whisper Depression
A CCA based Multimodal Approach

Heysem Kaya
Department of Computer Engineering
Boğaziçi University
34342, İstanbul, Turkey
heysem@boun.edu.tr

Albert Ali Salah
Department of Computer Engineering
Boğaziçi University
34342, İstanbul, Turkey
salah@boun.edu.tr

ABSTRACT
This paper presents our work on ACM MM Audio Visual Emotion Corpus 2013 (AVEC 2013) depression recognition sub-challenge using the baseline features in accordance with the challenge protocol. We use Canonical Correlation Analysis for audio-visual fusion as well as covariate extraction for the target task. The video baseline provides histograms of local phase quantization features extracted from 4x16 regions of the detected face. We summarize the video features over segments of length 20 seconds using mode and range functionals. We observe that features of range functional that measure the variance tendency provides statistically significantly higher canonical correlation than mode functional features that measure the mean tendency. Moreover, when audio-visual features are used with varying number of covariates per region, the regions that were consistently found the best are the ones corresponding to two eyes and the right part of the mouth. We reach 9.44 Root Mean Square Error on the challenge test set using audio-visual decision fusion, improving the video baseline 30% relative.

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; audio-visual fusion; feature extraction; depression recognition

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

1. INTRODUCTION
A fundamental aspect of human communication is non-verbal. In order for computing machinery to advance, not only the commands but also the traits (e.g. cultural background, personality) and the states (e.g. sleepiness, emotions) of the interacting individual should be taken into account. Moreover, the computer based systems can also be employed to detect and monitor the level of a long term illness, such as Alzheimer’s disease or depression. This way, it will be possible for the patients to self-monitor their progress anytime, anywhere. The study on the field can help reduce the economic cost of diagnosis at a global scale and allow the patients/medical doctors to get timely feedback.

ACM MM Audio-Visual Emotion Corpus and Challenge 2013 (AVEC 2013) focused on prediction of self-reported level of depression [13]. Depression can be defined as a state of low mood and aversion to activity that can affect a person’s thoughts, feelings, behaviors, and sense of well-being. The participating individuals are recorded via a web cam while they are guided by a presentation to do various tasks such as singing, reading and counting. The level of depression is measured by Beck-Depression Index II (BDI-II), a 21 item multiple-choice inventory [2]. The organizers provided baseline video and acoustic features along with the video clips partitioned into training, developments and test sets.

In this paper, we focus on extracting covariates of the target variable as highly informative features. We employ Canonical Correlation Analysis (CCA) for this purpose. We also use CCA for audio-visual feature level fusion. CCA is a statistical method that aims find linear projections for two views/representations of a semantic object that maximizes the mutual correlation [7, 6]. It is shown that when two representations of the same object (no matter whether they are from the same modality or not) are fused using CCA, the resulting features are more robust than the individual views [10]. CCA has a variety of applications in pattern recognition, ranging from multi-modal fusion [6, 10] to feature selection [8].

A recent study on AVEC 2013 corpus benefits from CCA as an acoustic feature selector, where the authors utilize the projection vector to rank the features [8]. Here instead, we use CCA for feature extraction in video and audio-visual modalities. We fuse the regional video features with the selected acoustic features from [8]. To focus on the discriminative power of the extracted covariate, we use 1-Nearest Neighbor (1-NN) as a smoother, and then average the predictions over the clip to get a final score. When individ-
nal regression performances of fused features are compared, the best results are consistently observed with regions corresponding to right eye, left eye and right half of the mouth. The stillness of the eyes, measured with the low variance of features from these regions, is found to be correlated with depression. Therefore, the preliminary studies show that eyes indicate depressive mood.

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

2. BACKGROUND: CANONICAL CORRELATION ANALYSIS

Proposed early in 1936 by Hotelling [7], 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 representations 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 $n$ denote 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:

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

(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 [6] to have the form of:

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

(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 [6] or deep neural networks (DNN) [1]. Non-linear feature reduction using CCA can be benefited in classification exploiting non-linear feature interactions such as SVM.

We next introduce the AVEC 2013 data used for experimental validation.

3. THE CORPUS AND FEATURES

AVEC 2013 [13] uses 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. The age of subjects ranges from 25 to 63. The target variable (BDI-II score) range is 0-45. Recorded behavior includes speaking out loud while solving a task, counting from 1 to 10, 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) [2], a subjective self-reported 21 item multiple-choice inventory.

For the AVEC 2013 challenge, the recordings were split into three partitions: training, development, and test sets of 50 recordings each, respectively.

3.1 Baseline Feature Sets

The AVEC 2013 audio baseline feature set, which provides an extended set of features with respect to AVEC 2012 [12], consists of 2268 features extracted using TUM’s open source feature extractor openSMILE [5]. Due to space limitations, the reader is referred to the challenge paper [13] for the details of the Low Level Descriptors (LLDs - e.g. F0, MFCC bands 0-12) and functionals (e.g. mean, std). The audio features are obtained using functionals operating over LLD contours in four segmentations: short (3 sec. with 2 sec. shifts), long (20 sec. with 2 sec. shifts) Voice Activity Detected segmentation (VAD-Seg) and per clip. Since the challenge paper reports best results on Depression Sub-Challenge (DSC) with long segmented acoustic feature set, we focus on this segmentation and synchronize the video features accordingly. The distribution of instances is given in Table 1.

<table>
<thead>
<tr>
<th>#</th>
<th>Train</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Per Clip</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>VAD Seg</td>
<td>6015</td>
<td>5763</td>
<td>5946</td>
</tr>
<tr>
<td>Short Seg</td>
<td>23863</td>
<td>23513</td>
<td>23824</td>
</tr>
<tr>
<td>Long Seg</td>
<td>23439</td>
<td>23087</td>
<td>23399</td>
</tr>
</tbody>
</table>

The baseline video features consist of geometric features (e.g. head pose and coordinates) and Local Phase Quantization features (LPQ). The geometric features (Geo) are 8 dimensional. 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. The 2D-DFT is computed at four frequencies $\{a, 0\}^T$, $\{0, a\}^T$, $\{a, a\}^T$, $\{a, -a\}^T$, $a = 1/M$, which correspond to 4 of 8 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 0, that gives an 8 bit string. The 8 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 $4 \times 4 = 16$ regions and an LPQ histogram is computed per region. On the overall, the baseline LPQ feature set provides $16 \times 256 = 4096$ features for each face detected frame. Before we proceed with the experiments, we provide canonical correlation analysis of long segmented baseline features in the next section.
3.2 Analysis of Features

As mentioned earlier, the video features are synchronized with the long segmented audio features. For this purpose, the frame-wise video features are summarized over windows of 20 seconds with a shift of 2 seconds. Two summarizing functionals are used. We use mode functional as a robust alternative measuring the mean tendency and range functional as a robust alternative for variance. The reason we prefer these alternatives is that data contains 0 values when face is not detected. For mode computation, all the features are independently discretized into histograms of 1000 bins in min-max feature value range. Then after mode computation, they were transformed back using the mean value of the corresponding bin. The summary of Canonical Correlation ($\rho$) of synchronized video features against the depression labels are given in Table 2.

Table 2: Canonical correlations of video features vs. depression labels in the training set

<table>
<thead>
<tr>
<th>Set</th>
<th>Dim</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>8208</td>
<td>0.959</td>
</tr>
<tr>
<td>Mode (LPQ+Geo)</td>
<td>4104</td>
<td>0.893</td>
</tr>
<tr>
<td>Range (LPQ+Geo)</td>
<td>4104</td>
<td>0.952</td>
</tr>
<tr>
<td>LPQ</td>
<td>8192</td>
<td>0.948</td>
</tr>
<tr>
<td>Geo</td>
<td>16</td>
<td>0.301</td>
</tr>
</tbody>
</table>

It is important to note that the long segmented audio features provide a canonical correlation of 0.794 against the labels. From the table we observe that features using range functional give more information about the target variable than the mode features. To investigate this effect in statistical manner, we computed the $\rho$ of video features against the target variable in 16 regions, separately. The mean±std $\rho$ statistics computed over regions are 0.77 ± 0.03, 0.80 ± 0.02, and 0.87 ± 0.01 for mode features, range features and combination of them, respectively. Administering a paired two tailed t-test on the canonical correlation of regional features, we found that range functional provides statistically higher canonical correlation ($p < 0.001$) than mode features. Moreover, the combined set provides much better performance compared to range only features ($p < 10^{-12}$). The reason for high canonical correlation of mode features and better performance of range functional can be accounted by the tendency of mean features to encode the identity of the individual rather than the depression level.

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. When a viable method and hyper parameter set is obtained, we re-train a model using training plus development set to predict the independent test set, the labels of which are unknown. Our aim is to surpass the challenge test set baseline provided in the paper. The sub-challenge measure for depression is Root Mean Squared Error (RMSE), while to ease the comparison Mean Absolute Error (MAE) is also provided. The test set baseline RMSE scores for DSC are 14.12 and 13.61 for audio and video modalities, respectively.

The best test set RMSE results reported on AVEC 2013 Corpus/DSC are listed in Table 3. We see that the best two scores are obtained using audio only modality. In this paper, we are interested in utilizing the visual cues and the robustness of audio-visual fusion.

Table 3: Best test set results reported on AVEC 2013/DSC

<table>
<thead>
<tr>
<th>Work</th>
<th>Modality</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kaya et al. [8]</td>
<td>Audio</td>
<td>9.78</td>
</tr>
<tr>
<td>Cummins et al. [4]</td>
<td>Audio</td>
<td>10.17</td>
</tr>
<tr>
<td>Meng et al. [9]</td>
<td>Audio-visual</td>
<td>10.96</td>
</tr>
</tbody>
</table>

4.1 Experiments with Video Features

In the previous section $\rho$ of regional features were given. Here we utilize the canonical covariates obtained from regional video features. For this, we remove the training set mean of each regional LPQ features (mode plus range) and apply the projection to development set. We then apply 1-NN regressor on the resulting 16-dimensional covariate space and obtain 6.90 MAE, 8.61 RMSE on the development set. If we combine the decisions of regional regressors instead, we get 7.61 MAE, 9.16 RMSE. When only the best 6 regional regressors are combined, the performance barely reaches the feature level combination: 7.07 MAE, 8.56 RMSE.

4.2 Audio-Visual Fusion

Here we repeat the feature selection experiments on a recent CCA based study on the same corpus and challenge [8]. We fuse the selected 387 acoustic features with each of regional visual features using CCA. Since CCA provides projections from either views, i.e. video and audio features, we utilize the video projections. The number of covariates in CCA is limited to the minimum of matrix ranks of two views. Since the selected audio features are smaller in number (387 as opposed to 512 video features) and they are linearly independent, the maximum number of covariates is 387. We apply a second level CCA between $p = \{50, 100, 150, 200, 250\}$ covariates from each region and the target labels to obtain a single depression covariate for regression. The RMSE performance of regional regressors with varying number of covariates are given in Fig. 1. We observe that the best three regional regressors are always number 6, 7 and 10, which correspond to left eye, right eye and the right part of the mouth, respectively. This is intuitive as the most potent parts of the face for action recognition are eyes and mouth area. This preliminary result helps reduce the computations to one quarter by focusing on the inner 2-by-2 square in the 4-by-4 partitioning of facial image. Moreover, the RMSE is observed to decrease up to 200 covariates but does not improve further. We finally concatenate all the local audio-visual covariates into a feature vector and apply CCA against the target labels to obtain a single depression covariate. The best RMSE results were obtained when 100 regional audio-visual covariates are combined (i.e. 1600 features) for second stage CCA in a similar fashion to Ensemble CCA [11]. An interesting finding is that the development set correlation of audio-visual covariates is higher than the training set. With 100 local covariates, the training set $\rho$ is found as 0.930 and development set $\rho$ is 0.952. For video modality $\rho$ reduces from 0.948 to 0.636, whereas in audio modality the decrease is more dramatic (from 0.794 to 0.306).
5. CONCLUSIONS AND OUTLOOK

In this study we utilize Canonical Correlation Analysis to extract depression covariates in visual and audio-visual modalities. We see facial regions that correspond to eyes and mouth area provide the best regression scores. Moreover, when a functional over the frame level features is used, range functional measuring the variance tendency is observed to be significantly better than the mode functional for mean tendency. We also observe that multi-modal fusion yields more robust covariates in the development set, which may be important for affect recognition. The challenge test set scores indicate the efficacy of our simple approach. The best test set results are obtained with audio-visual decision fusion. In our future work, we will extend the study to affect recognition and focus more on inner facial regions.

Table 4: Challenge test set results

<table>
<thead>
<tr>
<th>Model</th>
<th>Modality</th>
<th>MAE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>Video</td>
<td>7.97</td>
<td>9.94</td>
</tr>
<tr>
<td>M2</td>
<td>Video</td>
<td>7.86</td>
<td>9.72</td>
</tr>
<tr>
<td>M3</td>
<td>Audio-visual (feature level)</td>
<td>8.79</td>
<td>10.81</td>
</tr>
<tr>
<td>M4</td>
<td>Audio-visual (decision fusion)</td>
<td>7.68</td>
<td>9.44</td>
</tr>
</tbody>
</table>

Considering scores given in Table 3, our best test set score obtained using depression covariates as features and 1-NN regressor ranks the second after the state-of-the-art. The regressor and its hyper parameter were not optimized in this study, in order to sharpen the effect of the feature transformation. The study can be improved by employing more potent visual descriptors and more sophisticated regressors.

6. REFERENCES


