EYES WHISPER DEPRESSION: A CCA BASED MULTIMODAL APPROACH

Heysem Kaya, Albert Ali Salah
Department of Computer Engineering, Boğaziçi University, İstanbul, Turkey
{heysem, salah}@boun.edu.tr

Introduction

- Depression: «a state of low mood and aversion to activity»
- **In this study, we focus on extracting covariates of the target variable as highly informative features**
- We also use CCA for audio-visual feature level fusion
- Use mode and range functionals to summarize low level descriptors
- Analysis of regional audio-visual features' regregression performance

Background: Canonical Correlation Analysis

- CCA seeks to maximize the mutual correlation between two sets of variables by finding linear projections for each set. Let *A* and *B* denote two representations data, *C* denote covariance

$$\rho(A,B) = \sup_{w,v} corr(w^T A, v^T B) = \sup_{w,v} \frac{w^T C_{AB} v}{\sqrt{w^T C_{AA} w \cdot v^T C_{BB} v}}.$$

- The optimization problem can be converted into a Lagrangian, which has the eigenform of

$$C_{AA}^{-1} C_{AB} C_{BB}^{-1} C_{BA} w = \lambda w, \quad \rho(A,B) = \sqrt{\lambda}.$$

- To maximize correlation, we select the eigenvectors corresponding to the largest eigenvalues.

Corpus and Features

- We experiment on a AVEC 2013 (Valstar et al.,2013), which uses a subset of Audio-visual Depression Language Corpus (AVDLC). We adhere to the challenge protocol, use baseline acoustic (openSMILE) and video (Local Phase Quantization) features.

Table 1. Statistics of the AVDLC

Property	Statistic
# of Clips	340
# of Subjects	292
Range of Clip Length	20-50 min.
Mean Clip Length	25 min.
Total Duration	240 hours
Age Range of Subjects	18-63 years
Mean±Std of Age of Subjects	31.5±12.3 years
BDI-II Score Range	0-45

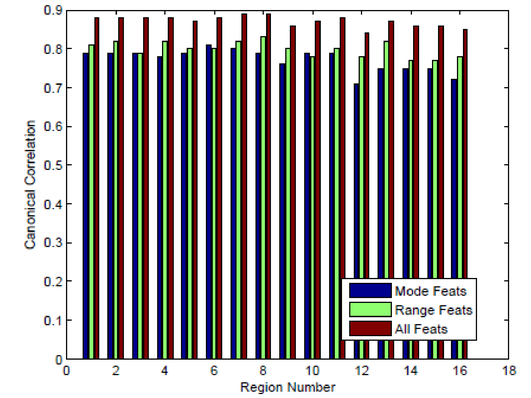
Table 2: State-of-the-art test set results on AVEC 2013 corpus

Work	Modality	RMSE
Kaya et al. [8]	Audio	9.78
Cummins et al. [4]	Audio	10.17
Meng et al. [9]	Audio-visual	10.96
Cummins et al. [3]	Audio	11.37

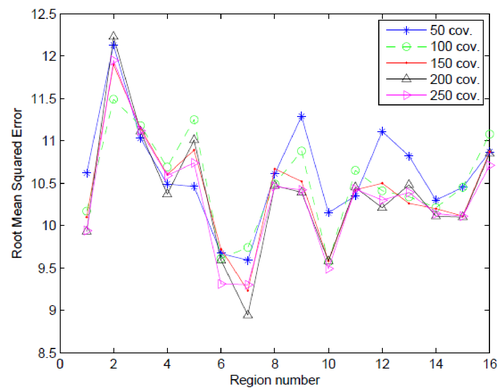
Analysis of Features

Table 3: CCA analysis of training set features vs depression labels

Set	Dim	ρ
All	8208	0.959
Mode (LPQ+Geo)	4104	0.893
Range (LPQ+Geo)	4104	0.952
LPQ	8192	0.948
Geo	16	0.301



Results



System (on Development Set)	MAE	RMSE
Audio baseline	8.66	10.75
Video baseline	8.74	10.72
Regional dep. cov. (16 dim)	6.90	8.61
Decision fusion of regional reg.	7.61	9.16
6 best regional regressors	**7.07**	**8.56**

System (on Test Set)	Mod.	MAE	RMSE
S1: 16 regional covariates	Video	7.97	9.94
S2: Best 3 performing regions	Video	7.86	9.72
S3: CCA based audio-visual fusion	AV	8.79	10.81
S4: Fusion of S2 with audio system from Kaya et al.*	AV	**7.68**	**9.44**

Conclusions and Outlook

- In this study we utilized CCA for feature extraction, and audio visual fusion
- We combined our visual system with audio system that use CCA for acoustic feature selection > best result
- The best development set performance is obtained with inner facial regions (eyes and mouth area) : higher action information and more robust to registration errors

* Kaya et al. CCA Based Feature Selection with Application to Continuous Depression Recognition from Acoustic Speech Features. (ICASSP 2014)