Semantic Feature Projection for Continuous Emotion Analysis

Prasanth Lade, Troy McDaniel, Sethuraman Panchanathan
Center for Cognitive Ubiquitous Computing
Arizona State University
prasanthl@asu.edu, troy.mcdaniel@asu.edu, panch@asu.edu

ABSTRACT

Affective computing researchers have recently been focusing on continuous emotion dimensions like arousal and valence. This dual coordinate affect space can explain many of the discrete emotions like sadness, anger, joy, etc. In the area of continuous emotion recognition, Principal Component Analysis (PCA) models are generally used to enhance the performance of various image and audio features by projecting them to a new space where the new features are less correlated. We instead, propose that quantizing and projecting the features to a latent topic space performs better than PCA. Specifically we extract these topic features using Latent Dirichlet Allocation (LDA) models. We show that topic models project the original features to a latent feature space that is more coherent and useful for continuous emotion recognition than PCA. Unlike PCA where no semantics can be attributed to the new features, topic features can have a visual and semantic interpretation which can be used in personalized HCI applications and Assistive technologies. Our hypothesis in this work has been validated using the AVEC 2012 continuous emotion challenge dataset.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Feature Measurement—*Feature Representation*

Keywords

Topic models , Continuous Affect Recognition, Feature Comparison

1. INTRODUCTION

Medical diagnosis of patients, federal investigations, human-computer interactions and many other applications need an understanding of human emotions. This research work addresses the application of social interaction assistance for people who are blind. When an individual who is blind interacts with a person, it is necessary to perceive his or her counterpart's emotions so that there is a positive influence and feedback during the conversation. This application requires a two layered approach where in the lower layer the blind user is provided with actual facial movements that are correlated to the person's facial emotions and the upper layer provides the continuous or discrete emotions to the user. This needs pattern recognition algorithms that can extract discernible features from image and audio data, learn their relationship with emotions, predict the appropriate emotional state and deliver the prediction to the visually impaired. Happy, sad, disgust, surprise, contempt and anger are most widely accepted states and will be called discrete emotions throughout this document. In recent years new dimensions of emotions are gaining popularity viz. arousal (energy) and valence (positivity). Unlike discrete emotions, each of these dimensions can be assigned a real number within a given range and thus are called continuous dimensions and arousal and valence can be used to define most of the discrete emotions.

To address these problems, algorithms that can extract meaningful features and can learn an association between features and emotions are needed. In recent years, probabilistic topic models have made significant contribution to both feature learning as well as supervised learning. Features extracted from video and audio frames are projected to a new latent topic space and these topics are used to predict emotions. This latent topic space is richer than the original feature space as it considers correlations and co-occurrences within features. In this work we show that these topic features are semantically richer and perform better than the most popularly used lower dimension projection techniques such as Principal Component Analysis (PCA).

MM '14 November 03 - 07 2014, Orlando, FL, USA

http://dx.doi.org/10.1145/2647868.2655042.

2. RELATED WORK

Feature extraction and analysis is a primary area of research in emotion recognition. Facial landmarks extracted using active shape models (ASM) and active appearance models (AAM) are used as geometric features to predict emotions in [3]. Mean appearance models, linear binary patterns (LBP) [5], local phase quantizations (LPQs), histogram of gradients (HOG), scale invariant feature transform (SIFT) and Gabor features are few examples of appearance features used for emotion recognition. Once features are extracted, their dimensionality is reduced using either Principal Component Analysis [4] or Independent Component Analysis [8].

In general, the image and audio feature space is projected to a lower dimensional space which is then mapped to the

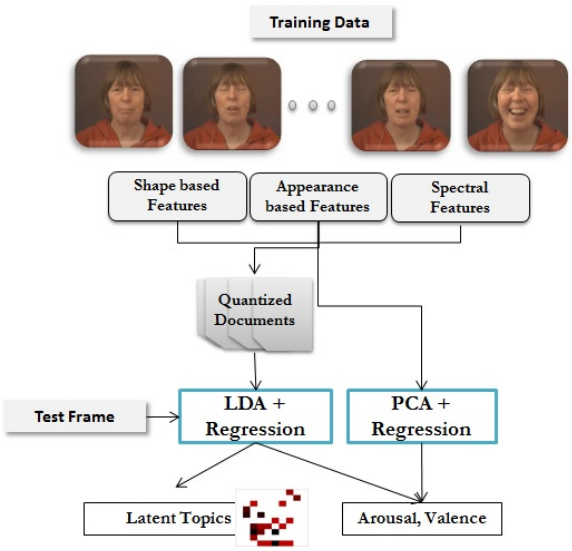

Figure 1: Experimental Methodology used in this work

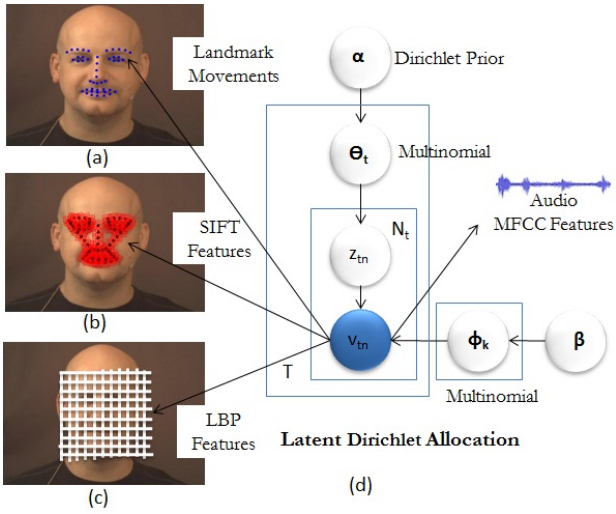

Figure 2: The LDA graphical model and the four features used in this work

emotion space. Existing dimension reduction techniques try to reduce the correlation between features but lose semantic information in the process. In text mining, probabilistic topic models are used to project the word features to a latent topic simplex and the new features are used in unsupervised and supervised settings. Blei et al [1] have proposed the Latent Dirichlet Allocation (LDA) model which is a more generalized topic model with explicit modeling of all the variables using multinomial-dirichlet conjugate distributions. In recent years topic models have been gaining popularity in image and video analysis and few works like [6], [7] and [2] have used LDA on audio and images for affect analysis. In this work we propose the use of topic models to project geometric, appearance and audio features from videos onto the topic space, analyze their interpretation and performance on continuous emotion recognition and compare its performance with PCA.

3. PROPOSED METHODOLOGY

The experimental methodology used in this work is shown in Figure 1, where a set of video frames are used to train the models. Video frames are converted to *video documents* by quantizing various image and audio features. These training documents are supplied to Latent Dirichlet Allocation Model to learn the latent topic space. The new topic features generated using LDA are used to train the regression algorithms with arousal and valence emotions as the dependent variables. The features of an unseen test frame are projected to the new topic space and the new features are used to predict emotions. PCA can only be used to predict emotions whereas it can be seen that LDA outputs extra topic projections that can be visualized or provided to users through Haptic interfaces. Topic models propose that every document contains latent concepts called topics. Probabilistic topic models are mixture models that model each document as a multinomial mixture of topics and each topic as a multinomial distribution over words. In the context of facial video processing, we can consider either an entire video or an individual video frame as a document. There is a lot of activity within each facial video or facial image document which can be possibly explained using latent topics. We hypothesize that the affective state of a person can be explained by identifying these latent topics. Similar to the definition of a textual topic which is a group of related words, video topics are defined as collections of co-occurring image/video features. This co-occurrence can be either spatial (in the case of an individual frame) or temporal (in the case of a video) and can be exploited to recognize different emotional states.

3.1 Topic Extraction from Base Features

Figure 2(d) shows the Latent Dirichlet Allocation graphical model, where the outer plate represents a document and the inner plate corresponds to a word. $\boldsymbol{\theta_t}$ is a multinomial distribution of topics in a document t and $\boldsymbol{\phi_k}$ is a multinomial distribution of words in a topic k. Given a set of K topics, each word v_{tn} in a document t is generated by assigning a topic z_{tn} from the distribution $\boldsymbol{\theta_t}$ and then generating a word from the topic distribution $\boldsymbol{\phi_{z_{tn}}}$. In LDA model the only observed variable is $\boldsymbol{v}$ and all other variables are latent variables. Now, given a set of documents, distributions $\boldsymbol{\theta_t}$ and $\boldsymbol{\phi_k}$ are inferred using Gibbs Sampling or Variational EM. In this work, we have used the variational Bayes approach to estimate the topics and the two parameters α and β have been selected using cross validation.

In order to use the LDA model the features have to be quantized to words and the details follow. As shown in Figure 2, four different *base* features viz. facial landmark movements (LM), SIFT features, Linear binary patterns from image frames and MFCC spectral features from audio frames have been considered. We have used the API provided by Xuehan et al. [9] to track 49 landmarks from facial videos. The landmarks from all videos are rotated, scaled and aligned with respect to the neutral frame (used the one shown in Figure 2) and the changes in the movements are quantized to $R-\theta$ space with four orientations for θ and 20 bins for the R. The 49 tracked landmarks are interpolated to 71 points as shown in Figure 2 (b) and SIFT features have been extracted at a scale of $S = 2$ and the features at the

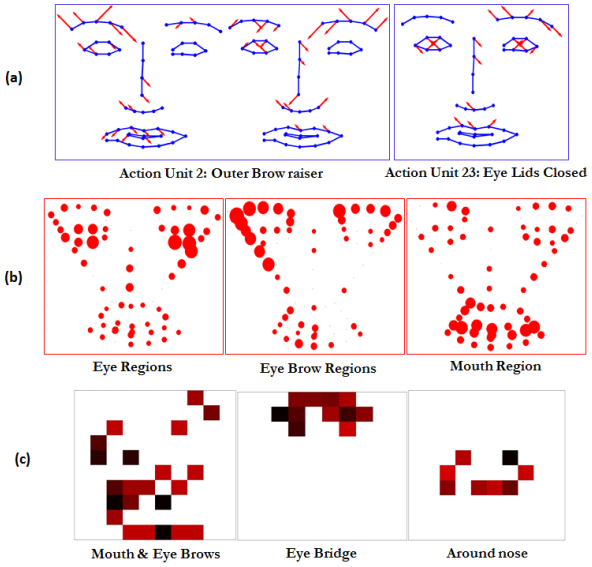

Figure 3: Plots of topics extracted from (a) Landmark (b) SIFT and (c) LBP features

optimal orientations have been considered. KMeans with cosine, KL-divergence and euclidean distances have been used to quantize SIFT features at each landmark. Figure2 (c) shows LBP features extracted from 100 image blocks and KMeans has been used to vector quantize these features. The final set of features are the spectral coefficients (1242) from audio at each frame. Each feature vector is quantized using Kmeans and the nearest centroid is assigned. The total vocabulary sizes of quantized words used for LDA are 196, 3550, 5900 and 62100 for LM, SIFT, LBP and Audio features respectively.

4. VISUAL INTERPRETATION

We hypothesize that latent topic features have a semantic interpretation and thus can be used to convey lower level information about changes in facial texture to a visually impaired person. We have used videos provided for the AVEC2012 challenge [5] to extract topics and perform our analysis. We have used LDA model with the optimal parameters (see Section 5 for details) to extract topics from different features. The topics extracted using geometric features, correspond to patterns of movements on the human face and the appearance features based topics provide information about changes in textures of the face. Facial plots of various topics can be seen in Figure 3, where the topics extracted from landmark, SIFT and LBP features are displayed in (a), (b) and (c) respectively. The angle and length of arrows in Figure 3(a) correspond to the direction and amount of landmark movements. It can be observed that sets of topics in fact have a correspondence with Facial Action units. The sizes of blobs plotted in Figure 3(b) correspond to the amount of influence (probability) the SIFT features at that landmark have on that topic. We find that each topic is modeling spatially correlated landmarks (that belong to a region) as they have similar textures e.g. the topics plotted model eyes, eyebrows and mouth regions respectively. The topics from LBP are plotted in 10 x 10 blocks as each word corresponds to a block. Figure 3(c) shows three

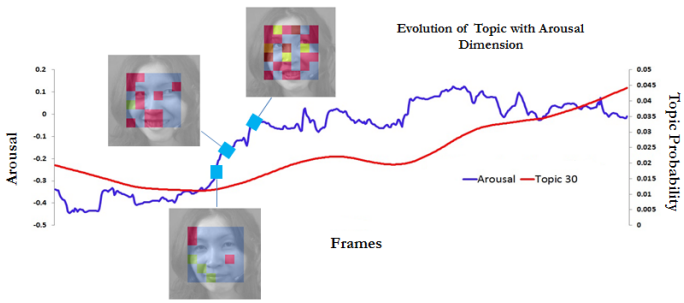

Figure 4: Evolution of LBP based topics with Arousal dimensions.

such topics active in the mouth-eyebrow, nose/eye bridge and around nose-cheek regions. We observe that each of the features produce different topics whose influence is felt in different ways and regions.

Along with a semantic connotation associated with topics, our experiments have also shown that these topics can be displayed in real time on a human face. The best topics corresponding to a region or a landmark can be delivered through haptic or visual interfaces which provide overall changes in face textures and movements. Another interesting observation is that topics are correlated with arousal and valence emotions. Figure 4 shows a sample plot of a video where the changes in arousal have been plotted along with the probability of Topic 30 extracted from block wise LBP features. There is a direct correlation between the topic probability and the emotion and the topic plots at different points of time are overlaid on the plot.

5. PERFORMANCE EVALUATION

To test the performance of projected features on continuous emotion recognition, we used 31 training videos (7 subjects) from AVEC2102 dataset and used the 32 development videos (8 subjects) for testing. We evaluated the mean cross correlations between the actual and predicted response labels averaged across all development videos. The parameters corresponding to the LDA model are the α, β and the number of topics K. We used a grid search approach to select the best parameters using cross validation on training videos. Video frames have been sampled at 30 ams and audio frames that do not have audio have been filtered. We have used 50 K-Means clusters for quantizing Audio and SIFT features. The optimal number topics for LM was 50 whereas 30 topics were selected for Audio, LBP and SIFT features. We used topic distributions extracted from four features viz. landmarks (LM), SIFT, LBP and Audio features which we address as Base features, as new LDA features. Once the dimensionality of the base features is reduced using PCA and LDA we have used Support Vector Regressors with to RBF kernels to predict arousal and valence separately. The results of arousal and valence prediction using different features are shown in Table 1. The mean cross correlations between predicted and actual emotions over all development videos using each of the features are shown on the left side. Since the size (number of frames) of each development video is different, we also calculated the weighted cross correlation weighted by the video size so that longer videos get more weightage. This is useful

	Arousal						Valence					
	Mean			Weighted Mean			Mean			Weighted Mean		
	Base	PCA	LDA	Base	PCA	LDA	Base	PCA	LDA	Base	PCA	LDA
LM	0.037	0.023	**0.177**	0.009	0.003	**0.196**	0.011	0.011	**0.127**	0.010	0.010	**0.190**
LBP	0.190	0.090	**0.191**	0.150	0.138	**0.195**	0.207	0.240	**0.264**	0.280	0.310	0.308
SIFT	0.190	**0.242**	0.207	0.170	**0.230**	0.190	0.07	0.022	**0.128**	0.050	0.030	**0.078**
Audio	0.0609	0.069	**0.277**	0.032	0.023	**0.294**	0.001	0.001	**0.113**	0.013	0.130	**0.195**

Table 1: Cross-correlation values on AVEC development videos

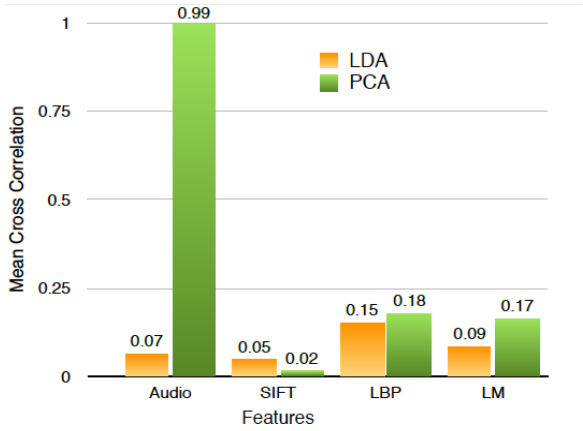

Figure 5: Mean cross correlations between features extracted using LDA and PCA (the lower the better)

because longer videos have more changes in emotions and better performance on longer videos implies the algorithm is able to model these changes. The Base algorithm uses the base features without any projection and dimensionality reduction. We have compared the results from two dimensionality reduction techniques, PCA and LDA. We can observe from the results that almost across all features on board, LDA features have performed better than the Base and PCA features. SIFT features are an exception to this and the probable reason could be that the quantization technique may not be sufficiently capturing all the histograms due to high dimensionality (128 dimensions).

Table 1 also gives useful insights about the effect of each feature on arousal and valence. Audio based topic features have performed better in arousal prediction than valence. Landmark features in contrast performed equally well in predicting both arousal and valence. LBP and SIFT features counter interact in modeling valence and arousal respectively. In order to show the contrast between PCA and LDA we also calculated correlations within LDA and PCA based features. Figure 5 shows the mean cross correlations within features extracted using PCA and LDA. It can be seen that LDA features are less correlated to each other in comparison to PCA except for the SIFT features which may explain the lower performance of LDA.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have used Latent Dirichilet Allocation models to project image and audio features to a latent topic space. These topic features have meaning and can be used to provide lower level facial information to individuals who are blind. We also showed that this projection technique is comparable to PCA and evaluated its performance on arousal and valence prediction. Also the LDA features are less correlated than the PCA based projections. During our analysis we observed that certain topics that are highly correlated to arousal or valence are given low probability which effects the prediction capability of features. To handle this, in future, we want to use supervised LDA models where the emotions can influence the topics that are extracted. We also plan to use continuous topic models where the features need not be quantized and a continuous Gaussian distribution will be used instead of a discrete multinomial distribution.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation (NSF) under Grant No. 1116360. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

8. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. M. L. R.*, Mar. 2003.

[2] P. Lade, V. Balasubramanian, H. Venkateswara, and S. Panchanathan. Detection of changes in human affect dimensions using an adaptive temporal topic model. In *IEEE ICME*, 2013.

[3] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 2010.

[4] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani. Robust continuous prediction of human emotions using multiscale dynamic cues. In *ACM ICMI*, 2012.

[5] B. Schuller, M. Valstar, R. Cowie, and M. Pantic. Avec 2012: The continuous audio/visual emotion challenge - an introduction. In *ACM ICMI*, 2012.

[6] M. Shah, L. Miao, C. Chakrabarti, and A. Spanias. A speech emotion recognition framework based on latent dirichlet allocation: Algorithm and fpga implementation. In *ICASSP*, 2013.

[7] L. Shang and K.-P. Chan. A temporal latent topic model for facial expression recognition. In *ACCV*, 2010.

[8] Y.-s. Shin. Recognizing facial expressions with pca and ica onto dimension of the emotion. pages 916–922. Springer Berlin Heidelberg, 2006.

[9] Xuehan-Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *IEEE CVPR*, 2013.