AN AUDIO-VISUAL APPROACH TO LEARNING SALIENT BEHAVIORS IN COUPLES' PROBLEM SOLVING DISCUSSIONS

James Gibson, Bo Xiao, Panayiotis G. Georgiou, and Shrikanth Narayanan
Signal Analysis & Interpretation Lab, University of Southern California, CA, USA

ABSTRACT

We present a method for characterizing salient behavioral events from audio-visual data of dyadic human interactions. This behavioral signal processing work is aimed at supporting observational analysis of domain experts such as psychologists and clinicians. We extract prosodic and spectral speech features as well as visual motion vector features on overlapping windows from a multimodal corpus. We then apply a technique called multiple instance learning to detect salient audio and visual instances for predicting human expert annotated behavior ratings. We demonstrate the performance gains achieved through multimodal fusion in characterizing complex behavior patterns of interest such as blame and acceptance in recordings of couples' problem solving discussions during marital therapy.

Index Terms— audio-visual signal processing, multiple instance learning, behavioral signal processing

1. INTRODUCTION

Modeling human behavior is an inherently complex and multimodal task. Speech, language, and physical gestures are all used to communicate the affective state of a person. People are trained (socially or professionally as in the psychological sciences) to recognize and react to one another's behavioral and emotional displays using these multimodal cues, each carrying useful information. Computational models of human behavior, therefore, should also rely on such multimodal signal information. Evangelopoulos et al. discuss modeling audio-visual signal saliency for movie summarization in [1]. In this work, we fuse audio and visual information for the task of estimating salient behavioral events for classifying observed human behaviors. Black et al. introduced methodology to use audio processing and analysis for automatically coding behaviors in the couples' therapy dataset in [2] and discuss the extraction of automatically derived lexical features and their fusion with audio features in [3].

Estimating the temporal saliency of observed feature streams using multiple instance learning (MIL) for couples data, as will be further examined in this paper, was first explored using audio features in [4] and using lexical features in [5]. These papers demonstrated the promise of MIL for modeling saliency in this type of data. The multiple instance learning framework used in these papers makes a *bag-of-instances* assumption on the data observations. We adopt the same *bag-of-instances* assumption in this work. More explicitly, we assume that each session (*bag*) is comprised of many behavioral expressions (*instances*). These instances can contain features derived from the observed audio and visual streams independently or jointly. In this work, we examine prototypical example representations learned in the multiple instance framework for observed human behaviors using audio, visual, and audio-visual features. We then use these features for automatically predicting psychologist annotated behavioral ratings. Ali and Shah used a similar type of analysis, i.e., multiple instance learning applied to human motion features, for human action recognition in [6]. The major differences in this work and [6] are the visual features, fusion with audio features, and we attempt to predict abstract behaviors versus predicting clearly defined human actions.

2. COUPLES THERAPY CORPUS

The Couples Therapy Corpus consists of 574 audio-visual recordings of real married couples having dyadic conversations. Each conversation lasted ten minutes and focused on a problem in the couple's relationship chosen by one of the spouses. This was then repeated for a topic chosen by other spouse. This was part of a longitudinal couples therapy study conducted by psychology researchers at the University of California, Los Angeles and the University of Washington [7]. Each pair of conversations for a particular couple were recorded at three different times during their ongoing marital counseling. The first was before therapy, the next was 26 weeks into therapy, and the last was 2 years after therapy had ended.

The data were manually annotated by psychologists using standardized, domain-relevant behavioral observation techniques. In each video, both spouses were observed and given a session-level rating, y_i, on a $1-9$ scale for presence of behaviors defined in the Couples Interaction Rating System 2 and Social Support Interaction Rating System manuals (1 corresponds to low presence of a behavior and 9 corresponds to high presence) [8, 9]. In this study we focus on the task of automatically classifying *acceptance* and *blame*[1].

This work was funded by the USC Annenberg Fellowship Program and the NSF.

[1]These terms are italicized to emphasize that they are terms of art used to represent very specific behaviors in the couples psychology community.

The quality of these clinical recordings are varied (their original purpose was human analysis), and in a pre-processing step, data were rejected based on audio and visual signal quality. The audio signals were recorded at rate of 16 kHz and encoded with 16 bit linear pcm. A 5 dB signal-to-noise ratio threshold was used to determine audio signals suitability for automatic processing. The video format is 704×480 pixels, at a rate of 30 frames-per-second (fps). Face-detection was performed using OpenCV [10] at a rate of one fps. Data were then rejected based on whether a face was detected in at least 70% of the frames. After data exclusion, 213 total speaker/session videos were deemed suitable for both audio and visual feature extraction.

3. METHODOLOGY

An overview of the experimental methodology of this work is as follows: first, we extract features from the audio and visual signals independently over short-time windows; next, we use a particular formulation of multiple instance learning known as diverse density to estimate saliency with respect to 2 second time windows (overlapped by one second) for both the audio, visual streams independently, and a concatenation of the audio-visual feature vectors; finally, we use the saliency estimates to represent each session by a small number of salient prototypes and use these features for classification.

3.1. Feature Extraction

3.1.1. Audio Features

The audio stream corresponding to each subject is represented by three types of low-level audio features: log of fundamental frequency ($f0$), intensity, and 13 mel-frequency cepstral coefficients (MFCCs). All the low-level features are mean normalized by speaker. Additionally, we use voice activity detection (VAD) to determine if there is speech represented by each particular low-level audio feature vector. These features are extracted from 25 ms windows incremented at a 10 ms rate. We then compute functionals of these low-level features over two second windows (200 low-level feature windows) with one second overlap. We use the six basic functionals described in [11]. They are: mean, median, standard deviation, 99^{th} percentile (robust max), 1^{st} percentile (robust min), and $99^{\text{th}}-1^{\text{st}}$ percentile (robust range). We compute functionals using only the low-level feature vectors with voice activity. A 90 dimensional feature vector is produced by the six functionals computed for each low-level feature (6 functionals × 15 low-level features). We will now refer to these feature representations as the *audio instances*. Each approximately ten minute session is comprised of approximately 600 audio instances. However, we use only audio instances with at least one second of voice activity for subsequent modeling.

3.1.2. Visual Features

We employ visual features based upon those used in [12], that represent subjects' head motion. Face-tracking is the first step in the visual feature extraction. The face is tracked over a two second window with one second overlap (as in the audio features). In the detected face region, we extract motion vectors which represent the motion of pixels corresponding to the face. We take the mean of the motion vectors magnitude with respect to the x (horizontal) and y (vertical) directions. We denote the motion in each direction as $M_x(t)$ and $M_y(t)$ for a window centered at time t. Similar to voice activity detection for audio features, we use kinesis activity detection (KAD) to discriminate between windows that contain motion and those that do not. We compute kinesics activity detection using the magnitude of the mean motion vector stream, $M(t) = \sqrt{M_x^2(t) + M_y^2(t)}$. Motion versus non-motion is modeled with a 2-mixture GMM of $M(t)$ and the transitions between motion and non-motion states are modeled using a 2 state hidden Markov model (HMM). We perform principal component analysis on the mean motion vectors to compensate for tilting of the head. We then compute ten linear prediction coefficients with respect to each axis over each two second window to model the dynamics of the head motion within that window. However, we use only instances with at least one second of kinesics activity for modeling. Figure 1 shows an illustration of the visual feature extraction process. Xiao et al. give a more detailed description of the visual feature extraction process in [12].

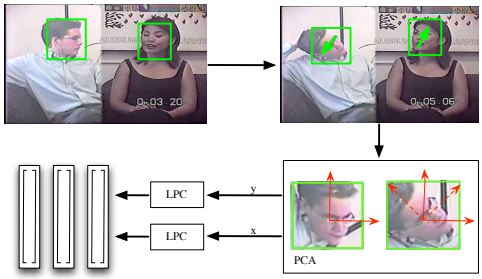

Fig. 1. diagram of visual instance feature extraction

3.2. Machine Learning

3.2.1. Multiple Instance Learning of Audio-Visual Instances

As stated earlier, we take each rated session to be a *bag-of-instances*, denoted B_i, $\forall i = 1, .., M$, where M is the total number of sessions. Each bag is comprised of several features vectors, which we refer to as *instances*. The instances of the i^{th} bag are denoted by B_{ij}, $\forall j = 1, ..., N_i$, where N_i is the number of instances in the i^{th} bag. Note that N_i varies from bag-to-bag based on the amount of speech/motion detected in the session. Each bag has a single label, l_i, which corresponds

to the behavior being highly present ($l_i = 1$) or not highly present ($l_i = 0$).

We estimate saliency of instances from each bag using a particular approach of the MIL problem known as the Diverse Density (DD) algorithm [13]. The goal of the Diverse Density is to learn concepts in the feature space that discriminate positively labeled bags from negatively labeled bags. The intuition is that points in feature space that are in high density areas surrounded by many different bags of the same label are *concepts* that distinguish bags of that particular label from bags of a different label. The diverse density value of a particular point, $\mathbf{x} = [x_1, x_2, ..., x_d]$, is given by:

$$DD(\mathbf{x}, \mathbf{w}) = \prod_{i=1}^{M} \left[1 - \left| l_i - \max_{1 \leq j \leq N_i} \left(e^{-\sum_{k=1}^{d} w_k |B_{ijk} - x_k|^2} \right) \right| \right], \quad (1)$$

where d is the dimension of the feature vector and $\mathbf{w}$ is a vector of weights for each feature that is learned in the maximization of the diverse density. We use the Expectation-Maximization Diverse Density algorithm (EMDD) to learn discriminative concepts in the instance feature space [14]. The instances from the bags of the male and female participants with the highest rating, y_i, to the behavior of interest are used to initialize the EMDD optimization. EMDD then performs expectation-maximization from these initialization points using gradient descent at each iteration to maximize the diverse density of the point with respect to $\mathbf{x}$ and $\mathbf{w}$.

After EMDD is performed about all the initialization points, we have a set of candidate concepts to choose as bases for modeling the rest of the bags. We sort the candidate concepts according to their diverse density values and keep candidates with diverse densities equal to or above the 95$^{\text{th}}$ percentile of diverse density values for all the candidates. We now refer to these candidate concepts with the highest diverse density values as the *salient concept prototypes*, $c^{(l)} = \{\hat{x}^{(l)}, \hat{w}^{(l)}\}, \forall l = 1, ..., N_c$, where N_c is the number of candidates with diverse density in the top 95$^{\text{th}}$ percentile, which we will use to discriminate between bags of high and low labels.

We use a single dimensional feature to represent each bag based on the salient concept prototypes. For a particular bag, B_i, we compute the distance of all its instances to each of the salient concept prototypes. Then we take the median of these distances to be the bag feature $\phi(B_i)$, i.e.,

$$\phi(B_i) = \underset{1 \leq l \leq N_c}{\text{median}} \left[\min_{1 \leq j \leq N_i} \left(\sum_{k=1}^{d} \hat{w}_k^{(l)} |B_{ijk} - \hat{x}_k^{(l)}|^2 \right) \right]. \quad (2)$$

These bag features are then used to predict the labels, l_i, with a linear support vector machine (SVM) classifier.

3.2.2. Audio-Visual Fusion

We take two approaches to fusing information derived from the observed audio and visual signals:

- The first, is an early fusion technique where we simply concatenate the instance feature vectors in each bag. A drawback of this approach is that instances are now excluded based on the voice activity and kinesics activity thresholds. These instances are then used for multiple instance learning as described in the previous section and classification is performed in the same manner.
- The second approach is late fusion. We use three separate classifiers' decisions to determine the final session label. The first two classifiers are the SVMs with linear kernels used for single modality classification. The third classifier is a SVM with a polynomial kernel trained on the concatenation of the audio and visual bag features (giving a two dimensional feature vector). The polynomial kernel allows for interaction terms between the two modalities to influence the classification decision. The mode of the three classifiers' decisions is taken as the late fusion label prediction.

4. EXPERIMENTS AND RESULTS

Our goal is to establish the usefulness of multimodal data and of saliency in extracting domain-relevant information. Such information is vital for domain experts.

4.1. Behavioral Observation Rating Classification

The labels are determined from ratings given by trained evaluators as discussed in section 2. At least three evaluators rate each session and the mean across evaluators is taken as the behavioral rating, $y_i \in [1, 9]$, for each bag with respect to the behavior of interest. To determine the label l_i, a threshold of the top and bottom quartiles of $\mathbf{y} = y_1, y_2, ...y_M$ and sessions with ratings above the top quartile threshold, $y_{75\%}$, are labeled as 'high' ($l_i = 1$) and session with rating below the bottom quartile, $y_{25\%}$, are labeled as 'low' ($l_i = 0$). We do not include the sessions with ratings in the middle 50% in this analysis.

We report accuracies determined from using a leave-one-couple-out cross validation scheme. That is, all the sessions from each husband and wife pair are used as the test data for a single fold with the sessions from all other couples used as the training data of that fold. We report total percentage of sessions that were correctly classified across all cross validation folds for audio, visual, and and audio-visual fusion techniques in table 1.

Early fusion marginally improves upon audio-only classification accuracy in the case of *blame* but degrades performance for classifying *acceptance*. This may indicate that saliency in the two modes are more congruent for *blame* than for *acceptance*. Another possible explanation for the lower

Table 1. *Classification accuracy (%) with audio, visual and audio-visual fusion. Chance accuracy is 50%.*

behavior	audio	visual	fusion	
			early	late
acceptance	70.5	62.5	64.3	**72.3**
blame	69.4	57.4	70.4	**71.3**

accuracy for *acceptance* is that we have less available data for training in early fusion because we require that each instance has both voice and motion activity.

Late fusion gives an improvement over audio-only and visual-only classification for both target behaviors. The polynomial fusion classifier allows interaction terms to influence the model, which places more emphasis on sessions that have high median saliency values with respect to both modalities rather than those that are only salient with respect to one. Taking the mode of the fusion classifier and the audio and visual-only classifiers helps prevent over fitting which often results from higher dimensional kernels. We also attempted late fusion by learning importance weights for the modalities and/or the samples. However, this method did not perform competitively with taking the mode.

We found that the audio-only and visual-only classifiers agree for 54.5% of the sessions and when they agree they correctly classify 80.3% of the sessions (for classifying *acceptance*). When they disagree, audio-only and visual-only classifiers give 58.8% and 41.2% accuracy, respectively. If we knew which modality to trust whenever the two disagree we would achieve 89.9% overall accuracy. So clearly, there is a large margin for improvement through with a fusion scheme that is able to determine which is the more reliable classifier at each sample. We are currently investigating such a scheme.

5. FUTURE WORK

In the future, we plan to continue developing novel human behavior representations using automatically signal derived features. We are especially interested in testing the usefulness of the salient instance representation of couples' interactions, that are presented in this work, for studying how each spouse reacts and influences the other's behaviors during the problem solving discussions. We also plan to study further methodologies for learning salient concepts in multimodal, time-varying signals and for exploiting non-congruent saliency.

6. REFERENCES

[1] G. Evangelopoulos, K. Rapantzikos, A. Potamianos, P. Maragos, A. Zlatintsi, and Y. Avrithis, "Movie summarization based on audiovisual saliency detection," in *Proc. ICIP*, 2008.

[2] M. Black, A. Katsamanis, B. Baucom, C. Lee, A. Lammert, A. Christensen, P. Georgiou, and S. Narayanan, "Toward automating a human behavioral coding system for married couples interactions using speech acoustic features," *Speech Communication*, 2011.

[3] M. Black, P. G. Georgiou, A. Katsamanis, B. Baucom, and S. Narayanan, "You made me do it: Classification of blame in married couples interactions by fusing automatically derived speech and language information," in *Proc. Interspeech*, 2011.

[4] J. Gibson, A. Katsamanis, M. Black, and S. Narayanan, "Automatic identification of salient acoustic instances in couples' behavioral interactions using diverse density support vector machines," in *Proc. Interspeech*, 2011.

[5] A. Katsamanis, J. Gibson, M. P. Black, and S. S. Narayanan, "Multiple instance learning for classification of human behavior observations," in *Proc. Affective Computing and Intelligent Interaction (ACII)*, 2011.

[6] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 288–303, 2010.

[7] A. Christensen, D. Atkins, S. Berns, J. Wheeler, D. Baucom, and L. Simpson, "Traditional versus integrative behavioral couple therapy for significantly and chronically distressed married couples," *Journal of Consulting and Clinical Psychology*, vol. 72, no. 2, pp. 176–191, 2004.

[8] C. Heavey, D. Gill, and A. Christensen, *Couples interaction rating system 2 (CIRS2)*, University of California, Los Angeles, 2002.

[9] J. Jones and A. Christensen, "Couples interaction study: Social support interaction rating system," *University of California, Los Angeles*, 1998.

[10] G. Bradski, "The opencv library," *Dr. Dobb's Journal of Software Tools*, 2000.

[11] M. Black, A. Katsamanis, C. Lee, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *Proc. Interspeech*, 2010.

[12] B. Xiao, P. G. Georgiou, B. Baucom, and S. S. Narayanan, "Data driven modeling of head motion towards behavioral analysis in couples' interactions," *accepted to ICASSP*, 2012.

[13] O. Maron, "Learning from ambiguity," Ph.D. dissertation, Massachusetts Institute of Technology, 1998.

[14] Q. Zhang and S. Goldman, "Em-dd: An improved multiple-instance learning technique," *Advances in neural information processing systems*, vol. 2, pp. 1073–1080, 2002.