Multiple Instance Learning for Behavioral Coding

James Gibson, Athanasios Katsamanis, *Member, IEEE*, Francisco Romero,
Bo Xiao, *Student Member, IEEE*, Panayiotis Georgiou, and Shrikanth Narayanan

Abstract—We propose a computational methodology for automatically estimating human behavioral patterns using the multiple instance learning (MIL) paradigm. We describe the incremental diverse density algorithm, a particular formulation of multiple instance learning, and discuss its suitability for behavioral coding. We use a rich multi-modal corpus comprised of chronically distressed married couples having problem-solving discussions as a case study to experimentally evaluate our approach. In the multiple instance learning framework, we treat each discussion as a collection of short-term behavioral expressions which are manifested in the acoustic, lexical, and visual channels. We experimentally demonstrate that this approach successfully learns representations that carry relevant information about the behavioral coding task. Furthermore, we employ this methodology to gain novel insights into human behavioral data, such as the local versus global nature of behavioral constructs as well as the level of ambiguity in the expression of behaviors through each respective modality. Finally, we assess the success of each modality for behavioral classification and compare schemes for multimodal fusion within the proposed framework.

Index Terms—Behavioral coding, behavioral signal processing, couple therapy, multi-modal signal processing, multiple instance learning

1 Introduction

Human behavior is inherently multimodal and complex, characterized by heterogeneity and variability in its patterning. This presents unique challenges and opportunities for signal processing and machine learning researchers to contribute to the behavioral sciences. These contributions can most readily be evaluated by relating them to measures that are already established and relevant to a given application domain, e.g., study of distressed relationships. A common method for evaluating human behavior is manual *behavioral coding*, which seeks to create standardized measures for characterizing observed behaviors along dimensions of interest, e.g., affect, engagement, withdrawal [1]. These measures are often applied in a holistic, summative fashion. That is, expert annotators will observe subjects in situations that elicit expressions of particular behaviors of interest and then provide their judgements on the degree that these behavioral constructs are exhibited in the overall session of observation. While this method can provide valuable insights into how these behaviors relate to outcomes, e.g., relationship success, it does not provide insight into which particular expressions contribute the most to the assigned behavioral codes. We present a method based on Multiple Instance Learning (MIL) that seeks to discover the most prominent segments of these sessions that contribute to specific behavioral characterizations. This can in turn provide insights at a detailed level into how particular manifestations of behaviors of interest impact coding of behavioral observation data.

The focus on behaviorally-salient segments within an interaction has been a feature in various behavioral analyses, most notably in psychological theory of thin slicing [2]. Thin slicing posits that judgements of certain behaviors can be made based upon brief but informative observations. This theory has been evaluated in several domains including: predicting teacher evaluations [3], judgements of personality and intelligence [4], and detecting psychopathy [5]. These studies demonstrate that short observations of expressive behaviors do not significantly differ from much longer observations for many behaviors of interest to the human behavioral research community. What is considered a "short" observation varies between studies. These studies use thin slices that range from 300 seconds to as little as 2 seconds. As one might suspect, the predictive accuracy of thin slices varies with respect to the behavior being observed, that is, some judgements require longer observation than others. Another key finding is that in some situations too much information was actually confusing to the raters [2], [5]. The thin slices in these studies are chosen at random from the longer behavioral observations. A key general question is whether selecting the "right" information will lead to higher predictive value of the thin slices. These findings motivate and set the stage for the current work.

We posit three main hypotheses:

1) *Not all observational segments are created equal*: by utilizing signal processing and machine learning techniques we can identify segments which lead to better prediction of presence of the judged behavior. Furthermore certain behaviors will be better predicted

- *J. Gibson, F. Romero, B. Xiao, P. Georgiou, and S. Narayanan are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089. E-mail: {jjgibson, faromero, boxiao}@usc.edu, {georgiou, shri}@sipi.usc.edu.*
- *A. Katsamanis is with the School of Electrical and Computer Engineering, National Technical University of Athens, Greece. E-mail: nkatsam@cs.ntua.gr.*

Manuscript received 9 Feb. 2015; revised 11 Oct. 2015; accepted 14 Dec. 2015. Date of publication 21 Dec. 2015; date of current version 24 Feb. 2017. Recommended for acceptance by G. N. Yannakakis.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2015.2510625

by thin slices due to their local nature versus those that occur more globally.

2) *Different behaviors will vary in the nature of their expression*: the number of ways that a particular behavior can be expressed will vary indicating the level of ambiguity in the expression of the behavior.
3) The predictive power of short observations will vary with respect to the modality with which they are evaluated and multiple modalities will provide complementary information for behaviors for which multiple channels are utilized for expression.

We evaluate these hypotheses using married couples conflict discussion data as a case study for the approach. First, we discuss behavioral coding and its relation to the emerging field of behavioral signal processing (BSP). Next, we describe multiple instance learning and one particular algorithmic approach, the diverse density (DD) algorithm. We then describe the application of its progeny, the incremental diverse density (IDD) algorithm, to the couples therapy data and discuss how this algorithm estimates *behavioral concepts*. These concepts are then used to predict extreme instances of behaviors relevant to the couples domain. Next, we analyze and give examples of the concepts learned and the insights they can provide into the domain. Lastly, we evaluate how our hypotheses fare with respect to our experimental results and discuss future work.

1.1 Behavioral Coding Domains

Behavioral coding is a methodology for assigning ratings for specifically defined behaviors of interest in behavioral observation studies. The goal of this practice is to assign clear and broadly applicable behavioral codes that characterize target behavioral constructs that relate to objectives of the study, e.g., approach-avoidance patterns. Behavioral coding is common in many behavioral health fields including: diagnosing autism [6]; psychotherapy [7]; family studies [1]; and evaluating marital therapy [8]. A major challenge of behavioral coding is training reliable coders who are responsible for assigning the behavioral ratings to the observation data in a consistent manner. Behavioral codes are often related to outcomes (e.g., of diagnosis or specific intervention) and their ability to predict these outcomes highlights their importance as a building block for understanding behavior and informing interventions.

1.2 Behavioral Signal Processing

This work is part of the emerging field of behavioral signal processing [9]. BSP is the development and application of signal processing tools for aiding behavioral sciences research and translation including notably in mental and behavioral health domains. Engineering approaches have the potential to offer fine data-centric insights which are otherwise inaccessible to clinicians and researchers working in the behavioral sciences. A common approach in BSP is to develop signal-derived representations and use these representations with appropriate pattern recognition methods to correlate with or predict desired behavioral codes [10].

Affective computing is an exemplary domain in behavior analysis facilitated by the use of signal processing and machine learning [11]. There have been numerous studies focusing on automatically deriving multimodal representations and performing classification experiments with emotional data [12], [13]. More recently, it has been of interest to identify prototypical expressions of emotions so as to deal with inherent ambiguity that arises from the large variability of expressions between (and within) subjects [14], [15]. With the advances in computational research on human emotions we are able to draw much inspiration toward computational approaches to understand higher level human behaviors.

In this work, we focus on evaluating the proposed behavioral signal processing approaches in the couples therapy domain. Couples marital therapy interactions represent a rich domain in which many high level human behaviors are elicited from the subjects and are used to help guide the course and evaluate the effectiveness of therapy. This domain has been the subject of several recent BSP studies, including studies seeking to develop representations of the acoustic, lexical, and visual signals as they relate to target behavioral codes [10], [16], [17], [18], as well as studies that focus on the interaction between subjects within the sessions [19], [20].

1.3 Why MIL for Behavioral Coding?

Multiple instance learning is a popular paradigm for learning problems with ambiguous or summative global labels. This is often the case with observational data of human behaviors. The ambiguity in behavioral observation data arises from variability in behavior expressions over the course of an observation. For example, if someone is rated as having negative affect, it does not imply all their expressions during the entire session used for coding were negative. In some cases the rating could be summative, meaning the observer takes every expression of the subject into account. In other cases a rating could result from an isolated expression, meaning it only required a few instantiations of a particular behavior to receive a specific high behavioral rating. Hence, a subject receiving a high rating for one of the behavior codes indicates that behavior was strongly expressed at some point during the session. It does not, however, give insight into where in the session or in what manner the behavior was expressed. By treating the session as a bag of behavioral expressions from the participants we can compare these short-term expressions in order to ascertain their contribution to the raters' judgements.

2 Multiple Instance Learning

Multiple instance learning is a machine learning framework in which labeled *bags* contain many *instances*. Each instance is represented by a feature vector and thus bags are collections of feature vectors that share a single label. The task then is to determine the label to assign to the bag without having specific information as to how the instances of that bag correspond to its assigned label. MIL was introduced by Dietterich et al. for drug activity prediction [21]. The paradigm has since been applied to several machine learning tasks including: natural scene classification [22]; image categorization [23], [24]; and text classification [25].

While the MIL framework has been more often applied to object recognition tasks for images, more recently it has been applied to human generated signals, such as speech, gestural, and linguistic data. In these data the objects being recognized are prototypical displays of a labeled action or behavior of interest. Ali and Shah applied the MIL framework to human action recognition [26]. Their target labels were clearly defined physical actions such as bending or hand waving. The MIL framework has also been applied to more abstract human expressions such as affect and behavior. Schuller and Rigoll proposed using a bag of speech frames framework for recognizing speakers' level of interest [27]. We proposed the application of MIL to couples problem-solving discussions data using acoustic features [28], lexical features [29], and audio-visual fusion [30]. This paper integrates components of these works as well as extending the framework through recent developments and analysis.

For this work we use a particular formulation of the MIL paradigm called the Diverse Density algorithm. This algorithm seeks to identify *concepts*, which are instances that occur frequently (*density*) in different (*diverse*) bags of the same label but do not occur in bags of the opposite label. The DD algorithm has been the subject of many advancements including: an expectation maximization formulation [31]; methods that seek to speed and boost learning [32]; integration with support vector machines (SVMs) [23], [24]; and an incremental version for learning multiple disjunct concepts [33].

2.1 Prominent Instance Selection Using the Diverse Density Algorithm

The Diverse Density algorithm was introduced by Maron and Lozano-Pérez [34] as an approach for handling the labeling ambiguity of multiple instance problems. It was first applied to the problem of drug activity prediction. The diverse density algorithm seeks to find concepts that are representative of a particular class of interest. Concept points lie in the feature space in areas of overlap between the instances multiple bags of the same label and far from any instances from bags of the opposite label. To perform the search for concept points we use the expectation maximization form of the diverse density called EM-DD because it is much faster than an exhaustive search of the feature space [31].

2.1.1 Diverse Density Formulation

The diverse density of a particular concept, c, is defined as its probability given n labeled bags, or, more specifically, l positive and m negative ones, $\mathbf{B} = \{B_1^+, \ldots, B_l^+,\ B_1^-, \ldots, B_m^-\}$. That is,

$$DD(c) \equiv P(c|\mathbf{B}). \tag{1}$$

We then seek to find the concept that maximizes the diverse density. We apply Bayes Rule to perform the maximum likelihood estimation:

$$c^* = \underset{c\in\mathbf{I}}{\operatorname{argmax}}[P(c|\mathbf{B})] = \underset{c\in\mathbf{I}}{\operatorname{argmax}}\left[\frac{P(\mathbf{B}|c)P(c)}{P(\mathbf{B})}\right], \tag{2}$$

where $\mathbf{I}$ is the instance space. We assume equal priors and drop the normalization as it does not affect the max operation. Next we assume independence of the bag instances, then apply Bayes Rule, assume equal priors of the instances, and drop the normalization once again:

$$\begin{aligned} c^* &= \underset{c\in\mathbf{I}}{\operatorname{argmax}}\left[\prod_{i=1}^{l} P(B_i^+|c)\prod_{i'=1}^{m} P(B_{i'}^-|c)\right] \\ &= \underset{c\in\mathbf{I}}{\operatorname{argmax}}\left[\prod_{i=1}^{l} P(c|B_i^+)\prod_{i'=1}^{m} P(c|B_{i'}^-)\right]. \end{aligned} \tag{3}$$

The posterior probability is estimated using the *most-likely-cause* approximation [35], according to:

$$P(c|B_i) \propto 1 - \left|\frac{1+y_i}{2} - \max_{1\le j\le N_i}\left[P(B_{ij}\in c)\right]\right|, \tag{4}$$

where y_i is the label of and N_i is the number of instances in the ith bag. Because the normalization is dropped this is not a proper probability, hence the proportionality. The probability that the jth instance of the ith bag belongs to the concept c is estimated according to their similarity defined by:

$$P(B_{ij}\in c) \propto e^{-||B_{ij}-c||^2}. \tag{5}$$

2.1.2 Point-and-Scaling Concepts

While learning concept points in the instance space we can simultaneously learn a scaling of the features. We define a point-and-scaling concept as:

$$P(B_{ij}\in\{c,s\}) = e^{-\sum_{q=1}^{Q} s_q(B_{ijq}-c_q)^2}, \tag{6}$$

where the instance, B_{ij}, is an Q-dimensional feature vector and s_q is the scaling parameter for the qth feature dimension.

2.2 Learning Multiple Concepts

We define a disjunctive set of multiple concept points as $\mathcal{D} = \{c_1 \vee c_2 \vee \cdots \vee c_d\}$. We can now substitute the disjunctive set of concepts, $\mathcal{D}$, for the single concept, c, in (1)–(4) and maximize the diverse density over multiple concepts. The disjunctive set is determined according to:

$$\mathcal{D}^* = \underset{\mathcal{D}\in\mathbf{I}}{\operatorname{argmax}}[P(\mathcal{D}|\mathbf{B})], \tag{7}$$

and the same operations follow to give the posterior probability of the disjunctive set of concepts:

$$P(\mathcal{D}|B_i) \propto 1 - \left|\frac{1+y_i}{2} - \max_{1\le j\le N_i}\left[P(B_{ij}\in\mathcal{D})\right]\right|. \tag{8}$$

We use the max operator as an approximation of the logical 'or' in order to define the probability that a particular bag instance belongs to the disjunctive set of concepts:

$$P(B_{ij}\in\mathcal{D}) \propto \max_{1\le k\le d}\left(e^{-||B_{ij}-c_k||^2}\right). \tag{9}$$

The assumption is that the probability that the instance belongs to the disjunctive set is equal to the probability of the instance belonging to the single concept with which it is most similar.

2.2.1 Incremental Learning

Unfortunately, learning multiple concepts carries a heavy computational cost: it rises factorially with the number of concepts in the disjunctive set. The authors of [33] proposed a method that greatly reduces the complexity of this search. This method relies on the key approximation:

$$DD(c_1 \vee c_2) \approx DD(c_1) + DD(c_2) - DD(c_1 \wedge c_2). \tag{10}$$

Note that this does not hold in general because of the proportionality of (8). This approximation leads to the measure referred to as the incremental diverse density, which is defined as:

$$\begin{aligned} IDD(\mathcal{D}) \equiv & \sum_{i=1}^{d} DD(c_i) - \sum_{i,j:1\leq i<j\leq d} DD(c_i \wedge c_j) \\ & + \sum_{i,j,k:1\leq i<j<k\leq d} DD(c_i \wedge c_j \wedge c_k) - \cdots \\ & + (-1)^{d-1} DD(c_1 \wedge \cdots \wedge c_d), \end{aligned} \tag{11}$$

for a disjunctive set of d target concepts. This can be more compactly formulated as:

$$IDD(\mathcal{D}) \equiv \sum_{\forall \mathcal{S} \in 2^{\mathcal{C}}} (-1)^{|\mathcal{S}|-1} DD(\mathcal{S}), \tag{12}$$

where $\mathcal{C}$ is the conjunction of hypothesized concepts, $\mathcal{C} = \{c_1 \wedge c_2 \wedge \cdots \wedge c_d\}$ and $2^{\mathcal{C}}$ is the power-set of $\mathcal{C}$. This ensures that concepts will both be of high diverse density as well as dissimilar to one another and thereby offer complementary information about the target class. In order to use this measure we need a definition for the logical conjunction of hypothesized concepts. In keeping with our definition in (9), we define the probability that an instance belong to the conjunctive set $\mathcal{C}$ as:

$$P(B_{ij} \in \mathcal{C}) \propto \min_{1\leq k\leq d} (e^{-||B_{ij}-c_k||^2}). \tag{13}$$

2.2.2 Nearest Concept Features

Once target concepts have been learned, features based on these concepts are used for classification. Nearest concept features are defined for each bag as being the minimum distance of any of the instances in that bag to any of the concepts in the disjunctive set. That is:

$$\phi(\mathbf{B_i}) = \min_{1\leq k\leq d} (\min_{1\leq j\leq N_i} ||B_{ij} - c_k||^2). \tag{14}$$

In this way each bag is represented by a single dimensional feature. The idea is that if the concepts are representative of the positive class then a bag with any instance that is sufficiently close to any one of the concepts will be labeled positive. This threshold is learned using a classifier trained with the nearest concept features. We assume a linear boundary because the nearest concept features are a distance where a shorter distance to the concept indicates the bag belongs to the positive class and further indicates it does not. We use linear support vector machine classifiers due to their robustness and because the nature of the nearest concept features warrants a linear decision boundary. Other classifiers could be used in place of SVMs, however we do not consider them for this work so as to focus on the MIL framework and the insights it is able to lend into behavioral coding domains, rather than a comparison of classifiers.

3 MIL for Behavior Prediction

In order to demonstrate the MIL framework for modeling human behaviors, we apply it to a multi-modal corpus of couples problem-solving discussions as a case study. In this case, the bags are a full session and the instances are segments within the session.

3.1 Case Study: Couples Problem-Solving Discussions

Marital conflict discussions are often used to understand couples' interactions and how their behaviors during these interactions relate to long-term outcomes. For married couples, arguably the most important long term-term outcome is whether the couple divorces. There have been many studies relating communicative behaviors during conflict discussions and divorce [36]. Carrére and Gottman reported the ability to accurately predict divorce using negative and positive affect coded over just the first three minutes of a 15 minute discussion [37]. Furthermore, they found that certain segments from the interaction were better predictors of divorce than others. The serious implications and predictive value of behavioral instances make it an ideal candidate for evaluation with multiple instance learning for behavioral coding.

3.1.1 Data Description

The Couples Therapy Corpus was collected as part of a longitudinal study conducted by a collaboration by psychology researchers at the University of Washington and the University of California, Los Angeles [8]. The study comprised 134 chronically distressed married couples and was collected with the intent of comparing the effectiveness of integrative behavioral couple therapy (IBCT) [8] with traditional behavioral couple therapy (TBCT) [38]. The couples were recorded having two problem-solving discussions, about an issue identified by each spouse. Each discussion lasted 10 minutes. The sessions took place at three different points in time with respect to a one year period during which the couples received marital counseling: before counseling, 26 weeks into counseling, and two years after counseling ended.

The corpus consists of 574 audio-visual recordings of the problem-solving discussions. The video is split screen (704×480 pixels, 29.97 fps) and the audio was recorded with the single channel far-field microphone in the video camera. Because the data were collected for observation by humans, no special care was taken for uniformity in camera angles or audio-visual quality. The data were manually annotated and transcribed by psychology researchers at the word level with the speaker labeled for each utterance. Nonverbal, vocal communication was also transcribed including: laugh, sigh, throat clear, and long pause. Timing however was not included in the transcriptions.

For each video, the spouses were observed and rated according to the Couples Interaction Rating System (CIRS) [39] and the Social Support Interaction Rating System (SSIRS) manuals [40]. Multiple annotators (ranging from 2 to 12) gave ratings for the presence of 33 behavioral codes (13 from the CIRS and 20 from the SSIRS) on a 1-9 scale for each session, where 1 corresponds to not present and 9 to highly present. The ratings were given at the session level, meaning each code describes a subject's behavior throughout the entire interaction with no information as to what particular expressions within the session lead to the resulting judgement. The SSIRS focuses on the emotional content of the session as well as the topic definition. These codes fall into four categories: affectivity, dominance/submission, features of the interaction, and topic definition.

3.1.2 Data Selection and Pre-Processing

Because the data were not originally intended for automatic analysis, the quality of the audio-visual recordings varied widely across sessions. Therefore, some data were rejected from this analysis. We reject sessions based on audio quality according to the methodology of [10], i.e., data with an average estimated signal-to-noise ratio below 5 dB are rejected. After this threshold was applied, 415 sessions remained. The audio data and the manual transcripts were force aligned with the SailAlign tool [41]. This step was applied in order to separate the audio into speaker specific regions and align the text and audio channels. Data were retained if at least 55 percent of the utterances were successfully aligned, leaving 372 sessions which were suitable for audio analysis. We reject sessions based on visual quality according to [42], which removes sessions in which face-tracking fails. This requirement resulted in the rejection of 151 additional sessions, leaving 221 sessions which are suitable for automatic analysis of the audio, visual, and lexical channels.

4 Multimodal Representations of Human Behavioral Data

Researchers have considered many sources of information for modeling human behavior. For this work we focus on three major modalities: lexical, audio, and visual. These features will help capture what is said, how it is said, and the subjects' associated movements. Subsequently, we will discuss methods of fusing information from these sources in order to benefit from the complementary insights they provide. Each modality also presents unique modeling challenges, which must be considered for fusion.

4.1 Audio Features

We represent the acoustic scene with standard low-level speech features based on pitch, intensity, and spectral fluctuations. Each of these low-level descriptors is extracted every 10 ms using a 25 ms Hamming window. They are then mean-normalized in order to reduce speaker specific phenomena from influencing our model. Specifically, we use log-pitch, intensity, and 13 Mel-frequency cepstral coefficients (MFCCs) as our low-level descriptors. The 13 MFCCs are extracted using 15 overlapping triangular filters equally mel-spaced from 20 to 8,000 Hz. The zeroth and 14th coefficients are not included in our feature representation. These low-level descriptors are chosen as they achieved they were shown to be the most informative for this task by Black et al. [10].

Each of these low-level descriptors is normalized using the mean value per speaker across the entire session. Specifically, the low-level descriptors are normalized according to:

$$f_{0_{\log}}^{\text{norm}} = \log_2\left(\frac{f_0}{\mu_{f_0}}\right), \tag{15}$$

$$\text{int}^{\text{norm}} = \frac{\text{int}}{\mu_{\text{int}}}, \tag{16}$$

$$\text{MFCC}_i^{\text{norm}} = \text{MFCC}_i - \mu_{\text{MFCC}_i}, \forall i \in 1, 2, \ldots, 13. \tag{17}$$

Subsequently, we represent each individual segment of speech by a vector of functionals of the low-level descriptors of that segments. We use the functionals referred to by Black et al. as the *six basic functionals*: mean, median, standard deviation, first percentile (robust minimum), 99th percentile (robust maximum), and 99th percentile-1st percentile (robust range) [10]. The six functionals taken of the 15 acoustic features result in a 90 dimensional functional feature representation. The functional feature representation of these speech segments is what we will refer to as instances in our bag-of-instances model. These segments are computed with a two second window which is advanced at a rate of one second (preliminary experiments using windows in the 1-6 second range demonstrated only small variation in results based on the choice of the window length, with two second windows resulting in the best performance). Voice activity detection (VAD) is used to determine the number of speech frames in a window, only windows containing at least 500 milliseconds of speech are retained for modeling. Each bag is comprised of several of these instances taken from overlapping segments of a speaker within a particular session. We show an overview of the audio instance feature extraction process in Fig. 1.

4.2 Visual Features

We use head motion as the base for our visual representation of the sessions [18], [42]. This is because the couples are sitting, which reduces their range of motion, and because face tracking is relatively robust compared to other computer vision techniques. Attempts to model facial expressions were unsuccessful due to the highly heterogeneous and noisy nature of the visual data, e.g., varying sitting positions, camera distance/angle, and lighting conditions. We segment the videos using the same window size (2 seconds) and rate (1 Hz) as the audio features so as to have better comparability between the two feature streams. From each segment we compute the power spectral density (PSD) of the motion vectors in the horizontal and vertical directions, respectively, for each subject. We then use 15

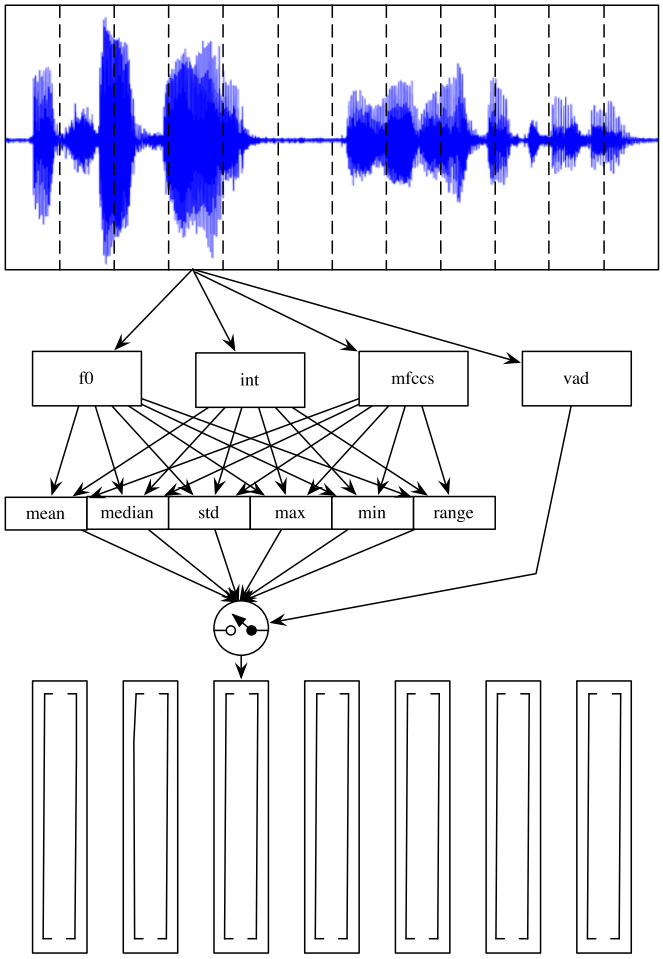


Fig. 1. Diagram of audio instance feature extraction.

frequency bins for each direction, ranging from 0.23 to 3.5 Hz. As reported in [42], these are the frequencies which capture head motions of interest such as nodding and shaking. We show an overview of the visual instance feature extraction process in Fig. 2.

4.3 Lexical Features

We model the language use of the subjects by extracting lexical features from the manual transcriptions. We use term-frequency inverse-document-frequency (TFIDF) to estimate the importance of each word used in the corpus. This representation is chosen due to its previous success in behavior prediction in similar tasks [41] and its compact nature in comparison to other standard linguistic features such as n-grams. We then retain the 50 words with the highest TFIDF scores and use the presence or absence of these words in each utterance as our feature. The resulting utterance feature vector will then be used as an instance in the bag-of-instances model. It is important to note that, due to the clear boundaries of the transcriptions, we use a different method of dividing the transcriptions into instances than we do for the audio and visual features where the divisions between instances are less clear.

4.4 Multimodal Fusion

We utilize two simple multimodal fusion techniques to take advantage of the multiple streams of information extracted from the audio-visual data. The relative accuracy of fusion systems will allow insights into whether the different modalities offer complementary information.

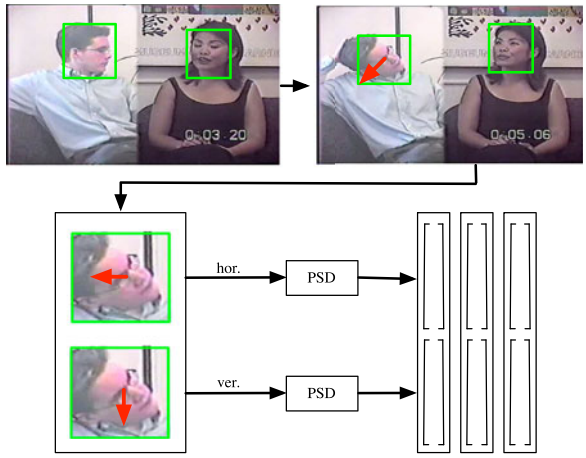


Fig. 2. Diagram of visual instance feature extraction.

4.4.1 Mean Posterior Fusion

The first simple fusion technique we employ is mean posterior fusion. For this system we train a classifier for each modality separately. Each classifier gives a posterior, $p_{y|\phi(\mathbf{B}^r)}$, which corresponds to confidence that the bag of the rth modality belongs to a certain class. Then the classification decision is made according to which class is given the highest mean posterior, according to:

$$p_{y|\phi(\mathbf{B})} = \sum_{r=1}^{M} \frac{p_{y|\phi(\mathbf{B}^r)}}{M}, \tag{18}$$

where M is the number of modalities.

4.4.2 Bag Feature Fusion

Our other adopted fusion technique is bag feature fusion. For this method, we perform fusion before classification. That is, we concatenate the bag features from each stream then train and test the classifier using this augmented feature vector. Thus, $\phi(\mathbf{B}) = [\phi(\mathbf{B}^1), \phi(\mathbf{B}^2), \ldots, \phi(\mathbf{B}^M)]$. We use SVMs with polynomial kernels of degree M to allow for interaction terms between the features from each stream.

5 Experiments and Results

In order to evaluate the efficacy of our proposed methodology, we conduct three experiments. In the first, we compare the predictive accuracy of our proposed methodology with selecting instances for prediction at random. This is what essentially occurs in applications of the thin slices in the literature of behavioral sciences. In the second experiment, we use the incremental diverse density algorithm to estimate multiple concepts for each of the behavioral codes. In the last experiment, we perform multimodal fusion to determine if combining information channels leads to higher predictive accuracy.

5.1 Classification of Behavior Codes

We first evaluate the efficacy of MIL for separating sessions with extreme behaviors. From the 33 behavioral

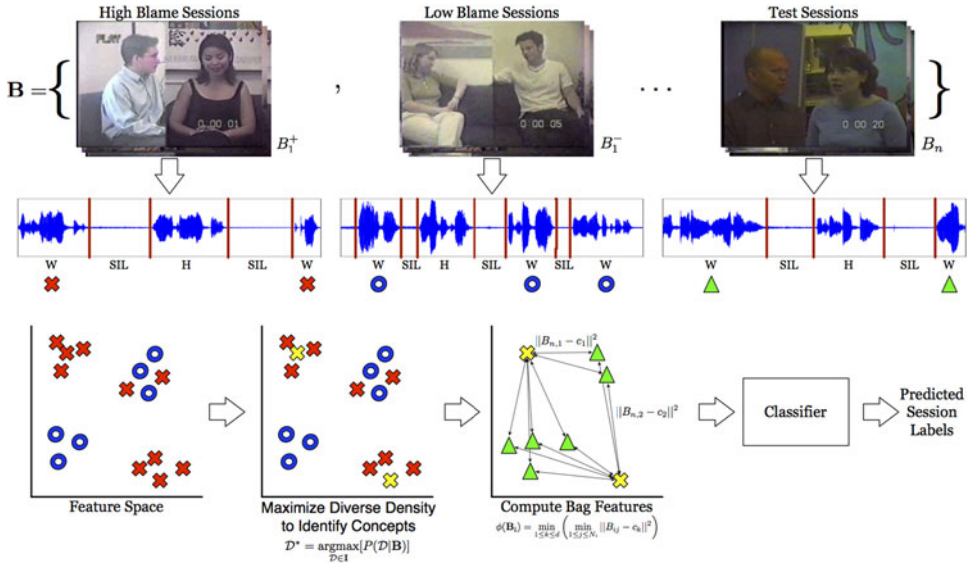


Fig. 3. Classification of couples behavioral codes overview.

codes we choose the four with the highest inter-rater reliability for our classification task (two from the SSIRS and two from the CIRS). They are: *acceptance of other*, *blame*, *global positive affect*, and *global negative affect*. The inter-rater reliabilities (Pearson's correlation) were: 0.751, 0.788, 0.740, and 0.798, respectively.

For each behavioral code, we take the sessions with ratings below the 25 percent percentile and label these sessions *low* and the sessions with ratings above the 75 percent percentile we assign the label *high*. We now have a binary classification problem with balanced classes, where one class contains all the sessions where the behavior of interest is highly present ($y_i = 1$) and the other contains sessions where the behavior is expressed weakly or not at all ($y_i = -1$).

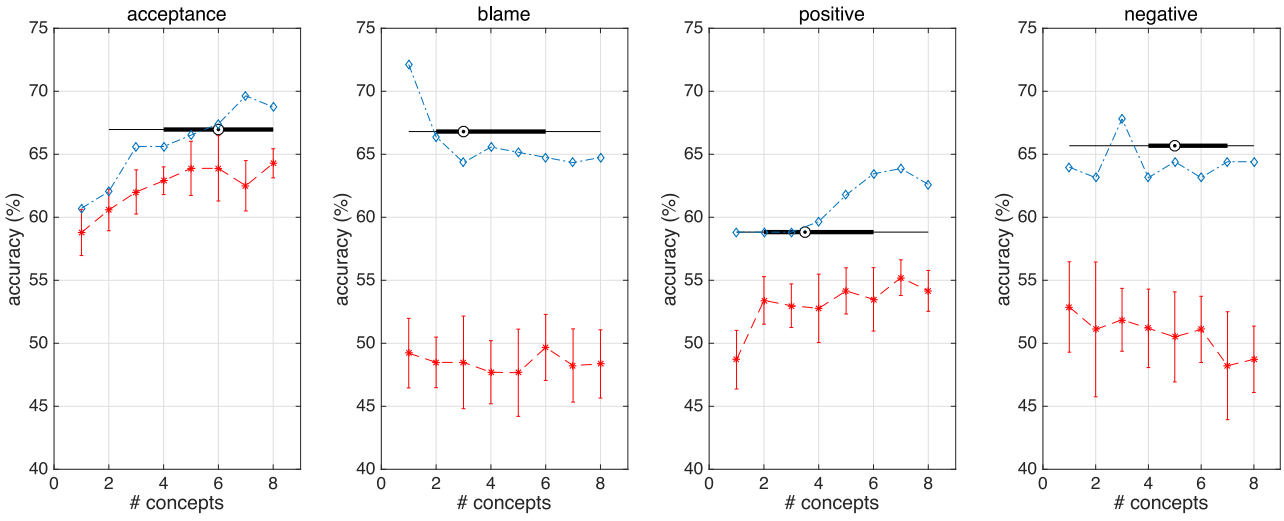
We now compute the bag features for every session as defined in (14). We compute bag features using each of the feature streams extracted from the couples problem-solving sessions. These features will represent the similarity of the subjects' expressions to the behavioral concepts and be used for classification using an SVM classifier. We use linear SVMs because our features are distances to a concept. This makes it reasonable to assume a linear boundary because the more dissimilar a subjects' expressions are from the behavioral concept the lower that subjects' behavioral rating should be. We show an overview of the behavioral code classification approach in Fig. 3.

5.1.1 Experiment 1: Establishing a Baseline

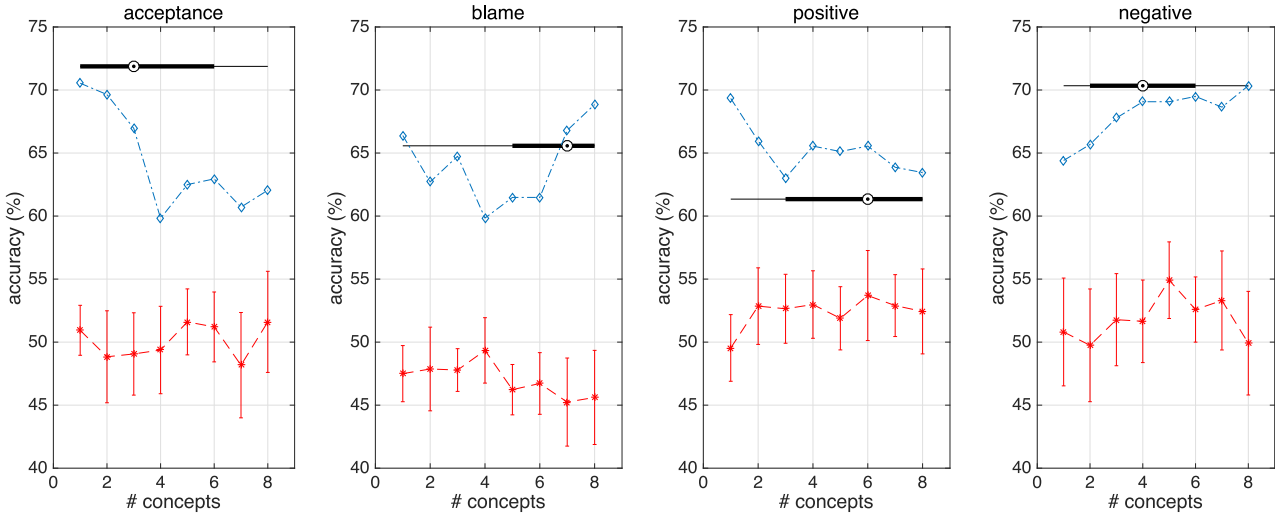
We first see how well we can predict the behavioral ratings if we select concepts at random from the set of positive instances. We then use these concepts for computing the nearest concept features and doing classification as discussed. In these experiments, the number of concepts is the number of random draws we take from our set of positive instances. As previously discussed, this is the methodology used for selecting thin slices to be rated and correlated to judgements made for a full behavioral observation. The difference in accuracy between random concept selection and estimating concepts with the IDD algorithm lends insight into the local versus global nature of the behavioral codes. If any instance chosen at random carries an equal amount of information about a particular behavioral construct, we can infer that the expression occurs more globally throughout the session. However, if the concepts estimated using IDD carry significantly more information, we can infer the opposite, i.e., that behavioral construct is expressed more locally. In Figs. 4a, 4b, and 4c, we show the results of comparing audio, lexical, and visual classification accuracy when concepts are randomly drawn from the bags and that of using concepts learned with the IDD algorithm. We perform 10 trials with the random selection method. The '$*$' represents the mean accuracy across the trials and the bars show the 95 percent confidence interval. The IDD algorithm significantly ($p < 0.05$) outperforms random selection for all behavioral codes and all number of concepts, except for classifying *acceptance* with visual features (there is no significant difference when using four, seven, or eight concepts). The performance margin however differs quite a bit between the behaviors. The average absolute performance difference in the audio channel between the two methods is 3.44 and 7.87 percent for the positive behaviors (*acceptance* and *positive affect*, respectively) versus 17.45 and 13.60 percent for the negative behaviors (*blame* and *negative affect*, respectively). This indicates that, with respect to audio, these positive behaviors generally occur more globally whereas the negative behaviors are more local in nature. It is important to note that while this observation is limited to audio experiments, some results do hold across modalities. For example, the performance margin is always the least for *acceptance* and the most for *blame* for all individual modalities, indicating that *acceptance* is the most globally expressed and *blame* is the most locally expressed of the behavioral constructs being compared.

5.1.2 Experiment 2: Learning Multiple Behavioral Concepts

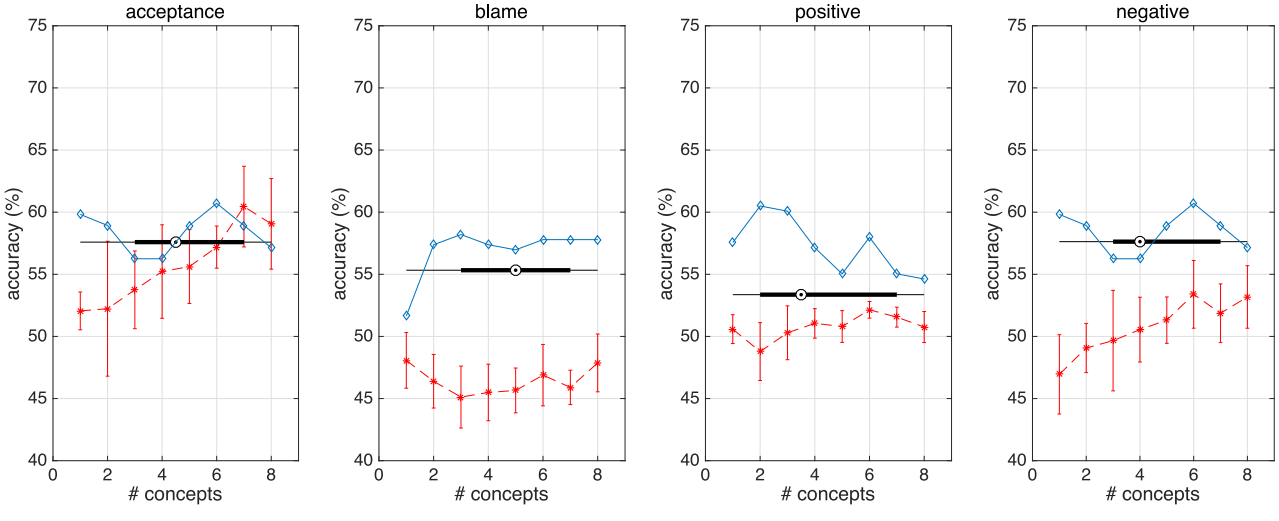
In Figs. 4a, 4b, and 4c we evaluate the accuracy of our behavior classification task with respect to the number



(a)



(b)



(c)

Fig. 4. Audio (a), lexical (b), and visual (c) classification accuracy. $*$ randomly chosen concepts, $\diamond$ IDD, $\odot$ IDD number of ideal concepts estimated ($\odot$ median, ▬ lower/upper quartile, — minimum/maximum).

TABLE 1
Classification Accuracy (Percent) with Audio (aud), Visual (vis), Lexical (lex), and Multimodal Fusion

behavior	aud	vis	lex	aud+vis	aud+lex	vis+lex	aud+vis+lex
			bag feature fusion				
acceptance	66.96	57.59	**71.88**	67.41	67.41	58.04	68.75
blame	66.80	55.33	65.57	65.16	**70.08**	66.39	69.67
positive	58.82	53.36	61.35	60.08	61.77	63.03	**65.13**
negative	65.68	57.63	**70.34**	65.25	65.68	57.63	65.25
			mean posterior fusion				
acceptance	66.96	57.59	**71.88**	67.41	67.86	57.14	67.41
blame	66.80	55.33	65.57	65.16	69.26	66.39	**70.08**
positive	58.82	53.36	61.35	58.82	63.03	58.82	**64.71**
negative	65.68	57.63	**70.34**	65.25	67.37	57.63	66.10

of concepts learned with IDD using audio, lexical, and visual channels, respectively. It is clear that the optimum number of concepts varies between behavioral codes. This gives some indication about the variability in the expression of these behaviors. From these results we can infer that there is less ambiguity in the expression of *blame* (maximum accruacy achieved with one concept) than there is in showing *acceptance* (maximum accuracy achieved with seven concepts), for example. Another important observation is that accuracy does not always increase or decrease monotonically about the optimum number of concepts. This may indicate that our IDD approximation is not selecting individual concepts in order of their usefulness, which means that we may be able to improve this approximation in the algorithm in future work. This may alternatively be due to the fact that an increase in diverse density of a set of disjunct concepts does not necessarily result in increased classification accuracy, which is because the algorithm does not directly optimize for classification accuracy because this is generally very difficult especially in multiple instance problems.

5.1.3 Estimating the Ideal Number of Diverse Density Concepts

Because we do not know the ideal number of concepts in advance of classification, we must treat this as an additional parameter. In order to determine this parameter we perform an internal leave-one-couple-out cross-validation loop and choose the number of concepts that results in the highest classification inner cross-validation accuracy. We show the classification accuracy achieved when using this method with whisker plot in Figs. 4a, 4b, and 4c. The $\odot$ indicates the median number of concepts estimated across all cross validation folds, the thick line indicates the 25th and 75th percentiles, and the thin line indicates the maximum and minimum. In most cases the accuracy achieved when estimating the number of concepts is lower than that achieved by knowing the ideal number of concepts in advance. This indicates that there is room for improvement in the methodology for choosing the ideal number of Diverse Density concepts.

5.1.4 Experiment 3: Multimodal Fusion

For multimodal fusion, we use the behavioral concepts learned in experiment 2. We train and test a classifier for each modality separately. Then we fuse the modalities by weighting the classifier posteriors with the information gain between the training bag features and the training labels. For each modality we use the cross-validation to select the number of concepts that gives the highest accuracy for that modality as discussed in the previous section. Table 1 shows the results of fusing the three modalities. Clearly, when the accuracy of one modality is drastically below that of the modality with which it is being fused the fusion accuracy is typically equal to or below that of the stronger learner.

5.2 Correlating Behavioral Concepts with Annotator Ratings

For the classification experiments it was necessary to binarize the annotator ratings. However, due to the ordinal nature of the ratings typical practice in the behavioral sciences is to report correlation results. To evaluate the efficacy of our method for this framework we present the correlations between the bag features computed using the behavioral concepts learned in the classification experiments with the annotator ratings.

We show the spearman's correlation between the nearest concept bag features and the annotator ratings in Table 2. This is computed with respect to all the sessions, only the sessions with ratings in the top and bottom 25 percentile, and middle 50 percent. We include results for the middle 50 percent independently because they are the most difficult to model with automatic methods [10]. Despite the challenging nature of these data we achieve significant correlations for the *blame* and *positive* behaviors and near significant ($p < 0.1$) correlations for *acceptance* ($p = 0.0846$) and *negative* ($p = 0.0747$). We achieve significant correlations with all the behaviors

TABLE 2
Spearman's Correlation of Bag Features with Annotator Ratings

behavior	all	middle 50%	top & bottom 25%
acceptance	-0.492^{***}	$-0.141^{\dagger}$	-0.616^{***}
blame	-0.478^{***}	-0.160^{*}	-0.596^{***}
positive	-0.476^{***}	-0.229^{**}	-0.580^{***}
negative	-0.407^{***}	$-0.142^{\dagger}$	-0.531^{***}

$^{\dagger}p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

TABLE 3
Features Which Received the Highest Absolute Weights in the Audio Concepts (in Descending Order of Importance) and Whether They Are Positively (+) or Negatively (−) Associated with the Behavior

behvaior	features
acceptance	mean int (+), median int (+), range f0 (−), range int (+), median f0 (−), max f0 (−)
blame	mean f0 (+), std f0 (+), std int (−), range f0 (+)
positive	median f0 (−), mean int (+), min f0 (+), max int (+), mean f0 (−), max f0 (−), min int (−), range int (+)
negative	min f0 (−), median f0 (+), mean int (−), std f0 (+), median int (−), range f0 (+), min int (−)

TABLE 4
Words Which Received the Highest Absolute Weights in the Lexical Concepts and Whether They Are Positively (+) or Negatively (−) Associated with the Behavior

behvaior	words
acceptance	SHE (+), THEY (+), UM (+), AFFECTION (−), TALK (−), YOU'RE (−), NO (−), PROBLEM (−), TOGETHER (−), HOUSE (+)
blame	MM (−), DIDN'T (−), THEY (+), UM (−), YOU'RE (+), NEVER (+), UH (−), MONEY (−), WORK (+), YEAH (+)
positive	SHE (+), UM (+), MONEY (−), RIGHT (−), HOUSE (−), OVER (−), YEAH (−), AFFECTION (−), JOB (−), ASK (−)
negative	YOU'RE (+), WAS (+), MOTHER (+), SAY (+), THEM (−), UM (−), HOME (−), DIDN'T (+), IT'S (−), SAID (−)

using the top and bottom 25 percent, and these are generally stronger than those with all the data because they represent the sessions containing the most extreme displays of these behaviors.

5.3 Analysis of Concepts Learned with IDD

One objective of identifying behavioral concepts is to allow these instances to give insights into the behaviors with which they are associated. We observe that concepts learned for different behavioral codes emphasize different combinations of features. This makes sense as these behaviors are produced in varying ways. As discussed earlier the DD algorithm learns a feature scaling for the concepts. The scaling often produces a quite sparse representation, that is most of the feature dimensions are scaled to nearly zero while a small subset receives substantially higher scaling. Additionally, we find that when learning multiple concepts with the IDD algorithm, the concepts for a single behavior also vary in the features that receive non-negligible scaling. In this way the classification task is gaining varied aspects of the behavioral expressions.

Because audio features such as MFCCs are difficult to interpret directly, we learn concepts using a reduced feature set of functionals of pitch and intensity. This reduced feature set also provided reduced classification accuracy relative to the full audio feature set, however the accuracy of the reduced set is still significantly above chance accuracy (63.84, 63.52, 59.24, 62.29 percent for *acceptance*, *blame*, *positive*, and *negative*, respectively) In Table 3 we show the audio features from the top concept learned using this reduced set that received non-zero scaling weights. Additionally, we indicate whether these features are positively or negatively associated with the behavior using (+) and (−), respectively.

In Table 4 we display the words that received the highest absolute scalings for each behavior. As the scalings can be positive or negative these words can be strongly positively or negatively associated with the behavioral codes. For example, "you're" is strongly associated with three of the four behavioral codes we are evaluating. While it is strongly positively associated with the codes *blame* and *negative*, it is strongly negatively associated with *acceptance*. Many of the words have a clear intuitive association with the behaviors they are modeling. For example, "you're", "work", and "never" all received high scaling for modeling *blame*. Interestingly, all the behaviors were found to be associated with at least one filler, such as "mm", "uh", and "um". Fillers are often used by a speaker to indicate that they have not finished speaking and thus are attempting to hold the floor. In this way these results indicate that a speaker attempting to hold the floor is strongly indicative of the behavioral codes.

Because of the sensitive nature of couples therapy data, we cannot give examples from the actual data. However, as part of the data collection the researchers collecting the data acted couples problem-solving discussion sessions to be used as training examples to give a sense of the nature of the data. To give an example of the types of instances that are selected as prominent we use one of these acted couples conflict discussion session. Below we show the utterance that is minimum distance from the *blame* concept:

> YOU HAVE BEEN LEAVING ME TO DO A **LOT** OF IT AND I DON'T KNOW I **FEEL** SORT OF FRUSTRATED THAT IT ALL KIND OF FALLS ON ME EVEN THOUGH WE BOTH **WORK** A **LOT** OF HOURS AND ESPECIALLY **STUFF** AND AND **SOMETIMES** I **FEEL** LIKE WHEN YOU DO THINGS LIKE I **NEVER** QUITE KNOW WHAT WILL HAPPEN LIKE IF YOU DO THE LAUNDRY I **NEVER** QUITE KNOW HOW THINGS ARE GOING TO TURN OUT YOU KNOW I SO I AM JUST KIND OF

The TFIDF selected words are shown in **bold**. Two of these keywords, "work" and "never" are part of the highly weighted words shown in Table 4. Clearly, "work" is a subject that is often discussed in these sessions and our *blame* concept suggests that simply discussing the subject is indicative of the behavior. Also, absolutes such as the word "never" are strongly associated with the *blame* behavioral construct.

6 Discussion

We now discuss how the three hypotheses presented in Section 1, relate the proposed methodology to saliency measures, and mention points of interest for future work.

6.1 Hypothesis 1

The proposed approach identifies concepts that are significantly better for the classification task than by choosing concepts at random from the set of positive instances. This indicates that not all portions of a given conversation are equally informative for determining the behaviors displayed during the interaction. This is an intuitive result that helps confirm that an instance-based approach is appropriate for analyzing behavioral data. Clearly, as seen in Figs. 4a, 4b, and 4c, the gap between IDD and random guessing varies between behavioral codes. The negatively valenced codes, *blame* and *negative*, benefit much more from the concepts learned with the IDD algorithm. This suggests that the expressions that result in a high rating for these codes are more local in nature, thus not all instances carry an equal amount of information to the behavioral coding process and in fact some carry quite a bit more, hence the large performance gap. While IDD is still better than random selection for the positively valenced codes, *acceptance* and *positive*, the performance gap is much less. This may indicate that the expressions resulting in high ratings for these codes are more global in nature, hence any given instance may carry a similar amount of information towards the behavioral coding process.

6.2 Hypothesis 2

We find that the number of optimum concepts varies between behavioral codes. The optimum number of concepts can be viewed as the number of ways that a behavior can be expressed through that channel. This relates to the level of ambiguity in that particular behavioral expression for a particular mode of expression. For example, with respect to audio, we observe that the negative behaviors benefit much less from learning additional concepts. This indicates there is less ambiguity in negatively valenced expression versus that of positive expressions when being expressed acoustically.

We also find that the optimum number of concepts varies with respect to the modality being modeled. This indicates there are different levels of ambiguity in the expression of a particular behavior in each channel. This is likely because different behaviors will be primarily expressed through only one or two channels hence the remaining modalities will either support or provide little additional information. For example, *acceptance* could be expressed purely through the visual channel through a head nod, while the subject is not speaking, hence the behavior would not be simultaneously expressed through the vocal and verbal channels.

In the lexical channel, we find the opposite trend as in the audio channel: The positively valenced behaviors do not benefit from additional concepts while the negative behaviors do. This makes sense as the two channels are closely linked and that when a behavior is clearly expressed in one channel it is not necessary for it to be clearly expressed in another, allowing for more ambiguity non-primary channels.

6.3 Hypothesis 3

For all behaviors we find that the audio and lexical modalities consistently outperform the video. This is consistent with previous automatic analysis using similar instance based methodology [30]. This is likely in large part due to the quality of the data. The couples therapy corpus was collected and stored on VHS tape and hence it is very difficult to achieve robust modeling using this modality. However, even with these challenges we did achieve above chance accuracy with the visual channel.

We find that particular behaviors are better modeled by different modalities. For example, accuracy for the *positive* code is relatively low compared to that of the other behavioral codes using the audio channel. In this case the lexical channel models this behaviors much better (5.46 percent absolute difference in peak accuracy). This relation is reversed for the *negative* code. We find that in this case the audio channel performs much better than the lexical (2.54 percent absolute difference). These differences may lend some insight into the way these behaviors are expressed, i.e., *negativity* being more associated with tone of voice and *positivity* being more associated with verbal content.

We find that the ability to achieve greater accuracy using fusion of these modalities varies with respect to the behavior and modalities being fused. For example, the highest multimodal accuracy for predicting *blame* is achieved when fusing information from all three channels. This indicates that the three channels are carrying some complementary information about the behavior. However, this was not the case for the other behaviors meaning that these behaviors may be solely or primarily expressed in a single channel. For example, the accuracy of predicting the *positive* behavior using the lexical modality far exceeds that of the audio or visual modalities. In this case, where the disparity in information between the channels is so high, fusion hurts accuracy because the less accurate channels are providing unreliable information.

6.4 Relation to Saliency Measures

The proposed approach attempts to identify and place focus on instances in which behaviors of interest are more strongly displayed. This can be viewed as estimating an instance's *behavioral saliency*. This is different from the signal-based idea of saliency as something that draws attention (e.g., [43], [44], [45]) in that it attempts to find instances that stand out from the background with respect to the behavior of interest rather than an instance that stands out in terms of its relative entropy with respect to the rest of the data. In this way we can focus only upon instances which carry information about subjects' behaviors. For instance, a subject coughing loudly would be *salient* with respect to drawing the attention of the observer but would not be *behaviorally salient* with respect to a code such as positive affect as does not carry information about that behavior and therefore should not be given prominence for behavioral modeling. In the future, we would like to compare the proposed approach with information theory based approaches and determine if we can advance our methodology by attempting to directly optimize information metrics.

6.5 Future Work

We would like to further investigate instance based analysis of behavioral observation data. We would like to compare

our methodology with fully unsupervised methods such as signal-based saliency [43]. As we previously discussed, this may give insight into the overt versus covert nature of behavioral expressions as well as how much certain behaviors rely upon context versus being directly observable in isolated segments. In contrast to a purely signal-based approach, we are interested into developing a methodology based on treating the instances as samples in a traditional fully supervised setting then using fusion techniques to predict the session level behavioral rating. Such an approach could offer understanding of how judgements of behavioral expressions are integrated over time. Also, this methodology would support hierarchical learning methods enabling analysis at many temporal granularities simultaneously.

7 Conclusion

In this work, we investigated evaluating behavioral observation data using a multiple instance learning framework. This framework allows for experimentation which reveals the local/global nature of behaviors as expressed through human communicative channels (e.g., speech). With respect to speech, we found that negative affect behaviors (*blame* and *negative affect*) are more locally displayed than behaviors conveying positive affect (*acceptance of other* and *positive affect*). Furthermore, we used the Incremental Diverse Density algorithm to learn multiple behavioral concepts. This methodology allows for estimating the level of ambiguity presented via a particular channel. For example, we found that with respect to speech, learning additional concepts for the negative behaviors did not provide additional information about the behavior, meaning there is little ambiguity in the expression of these behaviors. However, the opposite is true for expression of positive behaviors in vocal expressions, meaning there is much more ambiguity of expression in this case. Additionally, we find that this relation is reversed in the lexical channel: i.e., there is more ambiguity in expression of negative behaviors and less in positive behaviors, which is likely do to the close relation of these two modes of expression. We also used this methodology with multimodal fusion and found that in certain cases combining information from multiple channels provided increased classification accuracy over using that of only a single modality.

Acknowledgments

The authors would like to thank Andrew Christensen and the Couple Therapy research group for sharing the data. This work was supported by the National Science Foundation.

References

[1] G. Margolin, P. Oliver, E. Gordis, H. O'Hearn, A. Medina, C. Ghosh, and L. Morland, "The nuts and bolts of behavioral observation of marital and family interaction," *Clinical Child Family Psychol. Rev.*, vol. 1, no. 4, pp. 195–213, 1998.

[2] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychological Bull.*, vol. 111, no. 2, p. 256, 1992.

[3] N. Ambady and R. Rosenthal, "Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness," *J. Personality Social Psychol.*, vol. 64, no. 3, p. 431, 1993.

[4] P. Borkenau, N. Mauer, R. Riemann, F. M. Spinath, and A. Angleitner, "Thin slices of behavior as cues of personality and intelligence," *J. Personality Social Psychol.*, vol. 86, no. 4, p. 599, 2004.

[5] K. A. Fowler, S. O. Lilienfeld, and C. J. Patrick, "Detecting psychopathy from thin slices of behavior," *Psychological Assessment*, vol. 21, no. 1, p. 68, 2009.

[6] C. Lord, M. Rutter, P. DiLavore, S. Risi, K. Gotham, and S. Bishop, *Autism Diagnostic Observation Schedule: ADOS-2*. Western Psychological Services Torrance, 2012.

[7] W. R. Miller and G. S. Rose, "Toward a theory of motivational interviewing," *Amer. Psychologist*, vol. 64, no. 6, p. 527, 2009.

[8] 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," *J. Consulting Clinical Psychol.*, vol. 72, no. 2, pp. 176–191, 2004.

[9] S. Narayanan and P. G. Georgiou, "Behavioral signal processing: Deriving human behavioral informatics from speech and language," *Proc. IEEE*, vol. 101, no. 5, pp. 1203–1233, May 2013.

[10] 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 Commun.*, vol. 55, pp. 1–21, 2013.

[11] R. W. Picard, "Affective computing," M.I.T. Media Laboratory, Cambridge, MA, USA, no. 321, 1995.

[12] C. M. Lee and S. S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 2, pp. 293–303, Mar. 2005.

[13] B. Schuller, S. Steidl, and A. Batliner, "The Interspeech 2009 emotion challenge," in *Proc. Interspeech*, 2009, pp. 312–315.

[14] E. Mower, A. Metallinou, C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *Proc. 3rd Int. Conf. Affective Comput. Intell. Interaction*, 2009, pp. 1–8.

[15] E. M. Provost, "Identifying salient sub-utterance emotion dynamics using flexible units and estimates of affective flow," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 3682–3686.

[16] 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, pp. 89–92.

[17] P. G. Georgiou, M. Black, A. Lammert, B. Baucom, and S. Narayanan, "That's aggravating, very aggravating: Is it possible to classify behaviors in couple interactions using automatically derived lexical features?" in *Proc. 4th Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 87–96.

[18] B. Xiao, P. G. Georgiou, B. Baucom, and S. S. Narayanan, "Data driven modeling of head motion towards behavioral analysis in couples' interactions," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, 2013, pp. 3766–3770.

[19] C.-C. Lee, A. Katsamanis, M. P. Black, B. R. Baucom, A. Christensen, P. G. Georgiou, and S. S. Narayanan, "Computing vocal entrainment: A signal-derived PCA-based quantification scheme with application to affect analysis in married couple interactions," *Comput. Speech Language*, vol. 28, no. 2, pp. 518–539, 2014.

[20] B. Xiao, P. G. Georgiou, C.-C. Lee, B. Baucom, and S. S. Narayanan, "Head motion synchrony and its correlation to affectivity in dyadic interactions," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2013, pp. 1–6.

[21] T. Dietterich, R. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, nos. 1/2, pp. 31–71, 1997.

[22] O. Maron and A. Ratan, "Multiple-instance learning for natural scene classification," in *Proc. Int. Conf. Mach. Learn.*, 1998, vol. 15, pp. 341–349.

[23] Y. Chen and J. Wang, "Image categorization by learning and reasoning with regions," *J. Mach. Learn. Res.*, vol. 5, pp. 913–939, 2004.

[24] Y. Chen, J. Bi, and J. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[25] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," *Adv. Neural Inf. Process. Syst.*, vol. 15, pp. 561–568, 2002.

[26] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, Feb. 2010.

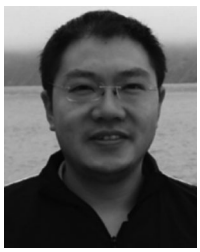
[27] B. Schuller and G. Rigoll, "Recognising interest in conversational speech-comparing bag of frames and supra-segmental features," in *Proc. Interspeech*, 2009, pp. 1999–2002.

[28] 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. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1561–1564.
[29] A. Katsamanis, J. Gibson, M. P. Black, and S. S. Narayanan, "Multiple instance learning for classification of human behavior observations," in *Proc. 4th Int. Conf. Affective Comput. Intell. Interaction*, 2011, pp. 145–154.
[30] J. Gibson, B. Xiao, P. G. Georgiou, and S. Narayanan, "An audio-visual approach to learning salient behaviors in couples' problem solving discussions," in *Proc. IEEE Int. Conf. Audio, Speech Signal Process.*, 2013, pp. 1–4.
[31] Q. Zhang and S. Goldman, "EM-DD: An improved multiple-instance learning technique," *Adv. Neural Inf. Process. Syst.*, vol. 2, pp. 1073–1080, 2002.
[32] J. Foulds and E. Frank, "Speeding up and boosting diverse density learning," in *Proc. 13th Int. Conf. Discovery Science*, 2010, pp. 102–116.
[33] J. Gibson and S. Narayanan, "Learning multiple concepts with incremental diverse density," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 4558–4562.
[34] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 1998, pp. 570–576.
[35] O. Maron, "Learning from ambiguity," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 1998.
[36] J. M. Gottman and R. W. Levenson, "Marital processes predictive of later dissolution: Behavior, physiology, and health," *J. Personality Social Psychol.*, vol. 63, no. 2, p. 221, 1992.
[37] S. Carrere and J. M. Gottman, "Predicting divorce among newlyweds from the first three minutes of a marital conflict discussion," *Family Process*, vol. 38, no. 3, pp. 293–301, 1999.
[38] D. H. Baucom, V. Shoham, K. T. Mueser, A. D. Daiuto, and T. R. Stickle, "Empirically supported couple and family interventions for marital distress and adult mental health problems," *J. Consulting Clinical Psychol.*, vol. 66, no. 1, p. 53, 1998.
[39] C. Heavey, D. Gill, and A. Christensen, *Couples Interaction Rating System 2 (CIRS2)*, Univ. California, Los Angeles, CA, USA, 2002.
[40] J. Jones and A. Christensen, *Couples Interaction Study: Social Support Interaction Rating System*, Univ. California, Los Angeles, CA, USA, 1988.
[41] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Proc. Workshop New Tools Methods Very-Large Scale Phonetics Res.*, 2011.
[42] B. Xiao, P. G. Georgiou, B. Baucom, and S. S. Narayanan, "Power-spectral analysis of head motion signal for behavioral modeling in human interaction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2014, pp. 4593–4597.
[43] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[44] C. Kayser, C. I. Petkov, M. Lippert, and N. K. Logothetis, "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biol.*, vol. 15, no. 21, pp. 1943–1947, 2005.
[45] O. Kalinli and S. Narayanan, "Prominence detection using auditory attention cues and task-dependent high level information," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 5, pp. 1009–1024, Jul. 2009.

James Gibson received the BS degree (magna cum laude) in electrical engineering from the University of Miami, in 2010, and the MS degree in electrical engineering from the University of Southern California (USC), in 2012. He is currently working toward the PhD degree in the Signal Analysis and Interpretation Laboratory (SAIL) at USC. His research interests are focused in human-centered signal processing. He is especially interested in how signal processing and machine learning techniques can enhance computational modeling of human communication in the context of mental health and well-being domains. He is a member of the IEEE signal processing society and Eta Kappa Nu. He received the USC Annenberg Fellowship, 2010-2012.

Athanasios Katsamanis received the diploma in electrical and computer engineering (with highest honors) and the PhD degree from the National Technical University of Athens, Athens, Greece, in 2003 and 2009, respectively. He is currently a research associate in the Computer Vision, Speech Communication and Signal Processing Group, National Technical University of Athens. His research interests lie in the area of speech and multimodal signal analysis and processing for modeling of human behavior, as well as image, acoustic, and articulatory data processing for speech production modeling. He is a member of the IEEE.

Francisco Romero received the BS degree in electrical engineering with a minor in physics from the University of Southern California, in 2015. He is currently working toward the MS degree in electrical engineering at Stanford University. His general research interests include multicore architectures, reconfigurable computing, machine learning, and pattern recognition.



Bo Xiao received the bachelor's and master's degrees from the Electronic Engineering Department, Tsinghua University, Beijing, in 2007 and 2009, respectively. He is currently working toward the PhD degree in electrical engineering at the University of Southern California, Los Angeles. His current research focuses on multimodal and multimedia signal processing towards analysis and modeling of human interactive behaviors. He is broadly interested in multimedia signal processing, speech and language processing, machine learning, and human-centered computing. He is a student member of the IEEE.

Panayiotis Georgiou received the BA and MEng degrees (with Honors) from Cambridge University, Cambridge, United Kingdom, in 1996, and the MSc and PhD degrees from the University of Southern California (USC), Los Angeles, CA, in 1998 and 2002, respectively. Since 2003, he has been with the Signal Analysis and Interpretation Lab, USC, where he is currently an assistant professor. He has authored or coauthored over 100 papers in the fields of behavioral signal processing, statistical signal processing, alpha stable distributions, speech and multimodal signal processing and interfaces, speech translation, language modeling, immersive sound processing, sound source localization, and speaker identification. His current research interests include the fields of multimodal and behavioral signal processing, computational mental health, multimodal environments, and speech-to-speech translation. Dr. Georgiou has been a PI and Co-PI on federally funded projects. He is currently an editor of the *EURASIP Journal on Audio, Speech, and Music Processing* and *Advances in Artificial Intelligence*. He served as a guest editor of *Computer Speech and Language* and as a member of the Speech and Language Technical Committee. He is currently the technical chair of InterSpeech 2016, area chair for InterSpeech 2015, and has served on the organizing committees for numerous conferences. Papers co-authored with his students were the recipients of Best Paper Awards at the International Workshop on Multimedia Signal Processing 2006, Interspeech 2010, and the International Conference on Cross-Cultural Design 2013.

Shrikanth (Shri) Narayanan is Andrew J. Viterbi professor of engineering at the University of Southern California (USC), and holds appointments as a professor of electrical engineering, computer science, linguistics, psychology neuroscience and pediatrics and as the founding director in the Ming Hsieh Institute. Prior to USC, he was with AT&T Bell Labs and AT&T Research from 1995-2000. At USC, he directs the Signal Analysis and Interpretation Laboratory (SAIL). His research focuses on human-centered signal and information processing and systems modeling with an interdisciplinary emphasis on speech, audio, language, multimodal and biomedical problems and applications with direct societal relevance. He is a fellow of the Acoustical Society of America and the American Association for the Advancement of Science (AAAS) and a member of Tau Beta Pi, Phi Kappa Phi, and Eta Kappa Nu. He is also an editor for the *Computer Speech and Language Journal* and an associate editor for the *IEEE Transactions on Affective Computing*, *IEEE Transactions on signal and information processing over networks*, *APSIPA Transactions on Signal and Information Processing* and the *Journal of the Acoustical Society of America*. He was also previously an associate editor of the *IEEE Transactions on Speech and Audio Processing (2000-2004)*, *IEEE Signal Processing Magazine (2005-2008)* and the *IEEE Transactions on Multimedia (2008-2011)*. He received a number of honors including Best Transactions Paper awards from the IEEE Signal Processing Society in 2005 (with A. Potamianos) and in 2009 (with C.M. Lee) and selection as an IEEE Signal Processing Society Distinguished Lecturer for 2010-2011 and ISCA distinguished lecturer for 2015-2016. Papers co-authored with his students have won awards including the 2014 Ten-year Technical Impact Award from ACM ICMI and at Interspeech 2015 Nativeness Detection Challenge, 2014 Cognitive Load Challenge, 2013 Social Signal Challenge, Interspeech 2012 Speaker Trait Challenge, Interspeech 2011 Speaker State Challenge, InterSpeech 2013 and 2010, InterSpeech 2009-Emotion Challenge, IEEE DCOSS 2009, IEEE MMSP 2007, IEEE MMSP 2006, ICASSP 2005 and ICSLP 2002. He has published over 650 papers and has been granted 17 US patents.