Decoding Children's Social Behavior

James M. Rehg[†,1], Gregory D. Abowd[1], Agata Rozga[1], Mario Romero[*,4], Mark A. Clements[2], Stan Sclaroff[3], Irfan Essa[1], Opal Y. Ousley[5], Yin Li[1], Chanho Kim[2], Hrishikesh Rao[2], Jonathan C. Kim[2], Liliana Lo Presti[3], Jianming Zhang[3], Denis Lantsman[1], Jonathan Bidwell[1], and Zhefan Ye[1]

[1]Center for Behavior Imaging, School of Interactive Computing, Georgia Institute of Technology, USA
[2]Center for Behavior Imaging, School of ECE, Georgia Institute of Technology, USA
[3]Department of Computer Science, Boston University, USA
[4]Department of High Performance Computing and Visualization, KTH, Sweden
[5]Department of Psychiatry and Behavioral Sciences, Emory University, USA

Abstract

We introduce a new problem domain for activity recognition: the analysis of children's social and communicative behaviors based on video and audio data. We specifically target interactions between children aged 1–2 years and an adult. Such interactions arise naturally in the diagnosis and treatment of developmental disorders such as autism. We introduce a new publicly-available dataset containing over 160 sessions of a 3–5 minute child-adult interaction. In each session, the adult examiner followed a semi-structured play interaction protocol which was designed to elicit a broad range of social behaviors. We identify the key technical challenges in analyzing these behaviors, and describe methods for decoding the interactions. We present experimental results that demonstrate the potential of the dataset to drive interesting research questions, and show preliminary results for multi-modal activity recognition.

1. Introduction

There has been a long history of work in activity recognition, but for the most part it has focused on single individuals engaged in task-oriented activities or short interactions between multiple actors. The goal of this paper is to introduce a novel problem domain for activity recognition, which consists of the decoding of dyadic social interactions between young children and adults. These child-adult interactions are rich and complex, and are not defined by the constraints of a particular task, as in the case of cooking. Nonetheless, these interactions have a detailed structure defined by the patterning of behavior of both participants. Our goal is to go beyond the simple classification of actions and activities, and address the challenges of parsing an extended interaction into its constituent elements, and producing ratings of the level of engagement that characterizes the quality of the interaction. We refer to these problems collectively as *decoding* the dyadic social interaction. We will demonstrate that this decoding problem represents a novel domain for research in activity analysis.

The problem of decoding dyadic social interactions arises naturally in the diagnosis and treatment of developmental and behavioral disorders. Objective approaches to identifying the early signs of a developmental disorder such as autism depend heavily on the ability of a pediatrician to identify a child's risk status in a brief office visit. Research utilizing video-based micro-coding of the behavior of young children engaged in social interactions has revealed a number of clear behavioral "red flags" for autism in the first two years of life [20], specifically in the areas of social, communication, and play skills. Currently, such careful measurement of behavior is not possible (or practical) in the real world setting of a pediatric office or daycare. There is much potential for activity recognition to scale early screening and treatment efforts by bringing reliable, rich measurement of child behavior to real-word settings.

We present an approach to decoding social interactions in the context of a semi-structured play interaction between a child and an adult, called the *Rapid-ABC* [14]. This protocol is a brief (3 to 5 minute) interactive assessment designed to elicit social attention, back-and-forth interaction,

†Corresponding author, rehg@gatech.edu

*This research was conducted while the author was a Research Scientist in the Center for Behavior Imaging at Georgia Tech.

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.438

and nonverbal communication. We have recorded and annotated more than 160 Rapid-ABC sessions. The contribution of this paper is the introduction of the *Multimodal Dyadic Behavior (MMDB)* dataset which contains this interaction data, along with an initial series of single mode and multimodal analyses to segment, classify and measure relevant behaviors across numerous play interactions. We will describe this unique dataset and the challenging analysis tasks that it enables, and present the results of our analysis, which can serve as a baseline for future investigations.[1]

2. Related Work

There is a vast literature on video-based activity and action recognition (some examples include [6, 10, 5, 18, 12]). However, most of these works are focused either on the actions of a single adult subject, or on relatively brief interactions between a pair of subjects, such as the "hug" action in [10] or the fighting activities in [16]. In the case of single person activities such as meal preparation [3], or structured group activities [13], activities can be complex and can take place over a significant temporal duration. However, the structured nature of tasks such as cooking can be exploited to constrain the temporal sequencing of events. In contrast to these prior works, the domain of social interactions between adults and children poses significant new challenges, since they are inherently dyadic, loosely structured, and multi-modal.

Recently, several authors have addressed the problem of recognizing social interactions between groups of people [15, 4, 1, 9]. In particular, our earlier work on categorizing social games in YouTube videos [15] includes many examples of adult-child dyadic interactions. These works have generally focused on coarse characterizations of group activities, such as distinguishing monologues from dialogues. In contrast, our goal is to produce fine-grained descriptions of social interactions, including the assessment of gaze and facial affect and the strength of engagement.

Our approach to analyzing dyadic social interactions is based on the explicit identification of "mid-level" behavioral cues. We extract these cues by employing a variety of video and audio analysis modules, such as the tracking of head pose and arm positions in RGBD video and the detection of keywords in adult speech. Each of these topics has been extensively researched by the vision and speech communities, and it is not practical to cite all of the relevant literature. In this context our contribution is twofold: We show how existing analysis methods can be combined to construct a layered description of an extended, structured social interaction, and we assess the effectiveness of these standard methods in analyzing children's behavior.

[1] Instructions for obtaining the MMDB dataset can be found at www.cbi.gatech.edu/mmdb

3. Challenges

From an activity recognition perspective, the analysis of social interactions introduces a number of challenges which do not commonly arise in existing datasets. First, the dyadic nature of the interaction makes it necessary to explicitly model the interplay between agents. This requires an analysis of the timing between measurement streams, along with their contents. Second, social behavior is inherently multimodal, and requires the integration of video, audio, and other modalities in order to achieve a complete portrait of behavior. Third, social interactions are often defined by the strength of the engagement and the reciprocity between the participants, not by the performance of a particular task. Moreover, these activities are often only loosely structured and can occur over an extended duration of time.

The analysis of adult-child interactions in the context of assessment and therapy provides a unique opportunity for psychologists and computer scientists to work together to address basic questions about the early development of young children. For example, detecting whether a child's gestures, affective expressions, and vocalizations are coordinated with gaze to the adult's face is critical in identifying whether the child's behaviors are socially directed and intentional. Another important challenge is to identify the function of a child's communicative bid. When a child is using vocalizations or gestures, is their intention (a) to request that their partner give them an object or perform an action; (b) to direct the partner's attention to an interesting object; or simply (c) to maintain an ongoing social interaction. Answering these questions in a data-driven manner will require new approaches to assessing and modeling behavior from video and other modalities.

Finally, advances in wearable technology have made it possible to go beyond visible behaviors and measure the activity of the autonomic nervous system, for example via respiration or heart-rate. The autonomic system is closely connected to the production and regulation of behavior, and could be a useful source of insight. In particular, our dataset includes continuous measures of electrodermal activity (EDA) which are obtained using wearable sensors. These physiological signals can be combined with audio and video streams in order to interpret the meaning and function of expressed behaviors [8].

4. The Multimodal Dyadic Behavior Dataset

We introduce the *Multimodal Dyadic Behavior (MMDB)* dataset, a unique collection of multimodal (video, audio, and physiological) recordings of the social and communicative behavior of infants and toddlers, gathered in the context of a semi-structured play interaction with an adult. The sessions were recorded in the *Child Study Lab (CSL)* at Georgia Tech, under a university-approved IRB protocol. The CSL

Figure 1. Child and examiner camera views in the MMDB dataset

is a child-friendly 300-square foot laboratory space which is equipped with the following sensing capabilities:

- Two Basler cameras (1920x1080 at 60 FPS) are positioned to capture frontal views of child and adult
- Eight AXIS 212 PTZ network cameras (640x480 at 30 FPS) are mounted around the perimeter of the room
- A Kinect (RGB-D) camera is mounted on the ceiling and centered on the table.
- An omnidirectional microphone is located above the table and a cardioid mic is in the corner of the room
- Dual lavalier wireless lapel omnidirectional microphones, one worn by the child and one by the adult
- Four Affectiva Q-sensors for sensing electrodermal activity and accelerometry (sampled at 32Hz), one worn on each wrist by the adult and the child.

The interaction follows the Rapid-ABC play protocol, which was developed in collaboration with clinical psychologists who specialize in the diagnosis of developmental delay [14]. This play protocol is a brief (3–5 minute) interactive assessment, in which a trained examiner elicits social attention, back-and-forth interaction, and nonverbal communication from the child. These behaviors reflect key socio-communicative milestones in the first two years of life, and their diminished occurrence and qualitative difference in expression have been found to represent early markers of autism spectrum disorders.

During the play interaction, the child sits in a parent's lap across a small table from an adult examiner. Figures 1 and 3 illustrate the set-up. The examiner engages the child in five activities, which we refer to as the five *stages* of the protocol: *Greeting:* she greets the child while smiling and saying hello; *Ball:* she initiates a game of rolling a ball back and forth; *Book:* she brings out a book and invites the child to look through it with her; *Hat:* she places the book on her head pretending it is a hat; *Tickle:* she engages the child in a gentle tickling game. The behavior of the examiner is structured both in terms of specific gestures (i.e., how the materials are presented to the child) and the language the examiner uses to introduce the various activities (e.g., "Look at my ball!"). Additional *presses* to elicit specific behaviors are built into the assessment. For example, the examiner silently holds up the ball and the book when they are first presented to see whether the child will shift attention from the objects to her face (exhibiting joint attention). She also introduces deliberate pauses into the interaction to gauge whether and how the child re-establishes the interaction. These presses introduce additional structure into the interaction, in the form of *substages*. For example, the ball stage consists of the substages "Ball Present," "Ball Play," and "Ball Pause."

An associated scoring sheet allows the examiner to note whether, for each substage in the activity, the child engaged in specific discrete behaviors, including initiating eye contact and smiling during key moments, looking at the ball/book followed by the examiner, and rolling the ball and turning the book pages. The examiner scores seventeen such behaviors as present or absent at the substage level, immediately following the completion of the assessment. In addition, for each stage of the protocol, she rates the effort required to engage the child using a 3-point Likert scale, with a score of 0 indicating that the child was easily engaged and a score of 2 indicating that significant effort was required. The ratings attempt to capture an overall measure of the child's social engagement, which relates to a core aspect of the behavior of children who may be at risk for an Autism Spectrum Disorder (ASD).

In addition to the scoring sheet, the MMDB dataset also includes frame-level, continuous annotation of relevant child behaviors that occur during the assessment. These annotations were produced by research assistants who were trained to reliability in behavior coding. These additional annotations include precise onsets and offsets of the targets of the child's attention (e.g., gaze to the examiner's or parent's face, ball, book), vocalizations and verbalizations (words and phrases), vocal affect (laughing and crying), and communicative gestures (e.g., pointing, reaching, waving, clapping, etc.).

To date, 121 children between the ages of 15 and 30 months have participated in the Rapid-ABC assessment, and their parents have consented to share their recorded data with the research community. 43 of these children com-

pleted a second session 2–3 months later. The video, audio, and physiological recordings, scoring sheet data, and parent questionnaire results for these sessions are included in the MMDB dataset and available to interested researchers at other academic institutions.

We have explored the automatic analysis of three aspects of the dataset: (1) *Parsing* into stages and substages; (2) *Detection* of discrete behaviors (gaze shifts, smiling, and play gestures); and (3) *Prediction* of engagement ratings at the stage and session level, including some preliminary findings for multimodal prediction. In the following sections, we describe our analysis methods in more detail and present our experimental findings from an initial set of child recordings. Note that our findings are based entirely on the coarse scoring provided the examiner, and do not leverage the additional, more-detailed annotations that we have produced.

5. Parsing Stages

A basic analysis goal is to segment the play interaction into its constituent five stages—greeting, ball, book, hat, and tickle. The ability to parse video and audio records into these major stages and their substages makes it possible to focus subsequent analysis on the appropriate intervals of time. Our approach to parsing leverages the structure which the adult imposes on the interaction. The examiner follows a pre-defined language protocol in which key phrases are used to guide the child into and through each stage. By analyzing the examiner's speech, captured by a lapel microphone, we can identify the beginning of each stage by looking for these key phrases (see Table 1). This is an example of a more general property of many standard protocols for assessment and therapy: By leveraging the statistical regularities of adult speech (and other modalities), we can obtain valuable information about the state of the dyad.

Parsing was done using commercial word- and phrase-spotting technology developed by Nexidia. The Nexidia tool takes as input an audio clip and a phrase of interest. It detects instances of the phrase in the audio stream and outputs the time-stamp locations of the detected phrases and their confidence scores. We first used the tool interactively

No.	Search phrase	Training	Testing	Stage
1	"Hi <name>"	60.00%	71.42%	Greeting
2	"Are you ready to play with some new toys?"	88.00%	100%	
3	"Look at my ball"	80.00%	71.42%	Ball
4	"Let's play ball"	68.00%	92.86%	
5	"Ready, set, go!"	92.00%	92.86%	
6	"Look at my book"	76.00%	78.57%	Book
7	"Where's the yellow duck?"	76.00%	92.86%	
8	"Let's see what's next"	84.00%	100%	
9	"Can you turn the page?"	72.00%	57.14%	
10	"It's on my head! It's a hat"	80.00%	85.71%	Hat
11	"I'm gonna get you!"	92.00%	92.86%	Tickle

Table 1. Search phrases belonging to each stage and the detection accuracy across 39 sessions (25 for training and 14 for testing).

No.	Stage	Error (sec)	Sessions
1	Greeting	0.4653	13
2	Ball	0.8307	11
3	Book	0.8302	11
4	Hat	3.8055	12
5	Tickle	9.8607	13

Table 2. Errors in seconds for predicting the stage starting times.

on a training set of 25 sessions (from our corpus of 39). We used the pronunciation optimization feature to adjust the phoneme sequences associated with our queries, improving detection performance. We used the remaining 14 sessions for testing, giving a total of 70 testing stages. Table 1 gives the detection performance of the tool for each phrase.

If the Nexidia system successfully detects one of the phrases in Table 1, the associated time-stamp can be taken as an estimate of the start of that stage. Table 2 gives the error in seconds associated with this estimate, on our testing set of 14 sessions. The error measure is the average of the absolute difference between the estimated and ground truth starting times across the sessions. The error is largest for Tickle because the key phrase is repeated multiple times during the first 30 seconds of the stage.

Note that in 9 out of 70 stages, the Nexidia tool generated false positives, detecting phrases more than 5 seconds after the true start of the stage. We removed these outliers from the results in Table 2 so that they did not swamp the reported performance. The last column in Table 2 gives the number of non-outlier instances. In a second experiment, we predicted the start times for the substages ball_present and book_present, which occur within the Ball and Book stages, respectively. For 11 sessions, our method achieved an average absolute error of 0.253 seconds for book_present and 0.467 seconds for ball_present.

These results suggest that when the Nexidia tool produces accurate detections, the time-stamps of the detected phrases provide a reliable cue for segmentation. Since the tool is not perfect, additional performance gains could be obtained by incorporating other modalities, such as the overhead Kinect view.

6. Detecting Discrete Behaviors

We have described a procedure for parsing a continuous interaction into its constituent stages and substages. Within each substage, the examiner assesses whether or not the child produced a set of key behaviors (see Section 4), including smiling and making eye contact. We now describe our approach to automatically detecting these two discrete behaviors. The primary challenge stems from the fact that the examiner produces a rating for an entire substage, based on whether or not the behavior occurred at least once. Thus we do not have access to frame level ground truth labels for training purposes. Our approach is to aggregate frame-level

scores to make substage-level predictions.

6.1. Smile Detection

Given a segmented video clip corresponding to a substage in the interaction, our goal is to predict a binary smile label. We employed commercial software from Omron, the OKAO Vision Library, to detect and track the child's face, and obtain measures of face detection confidence, smile degree (the strength of the smile), and smile confidence for each detected face. We used the joint time series of smile degree and smile confidence in each frame as the feature data for smile detection. Features were aggregated from all high-confidence face detections over a single clip into a 2D histogram. Figure 2 gives a visualization of this histogram, averaged over all of the clips with positive (left) and negative (right) smile labels. It is clear that the joint features of degree and confidence have discriminative value.

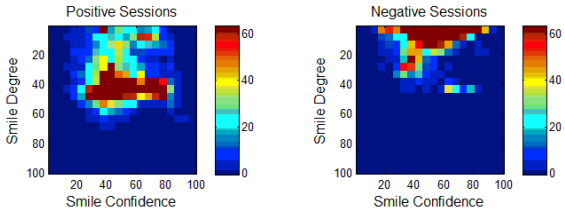

Figure 2. The left and right figures show the 2D histogram of the distribution of positive and negative labels, respectively.

We constructed a temporal pyramid to capture additional structure from the time series. We empirically selected 20×20 bins for the 2D histogram and pyramid level as $[1, 2]$ for all our experiments, yielding a 1200-dimensional feature for each substage. A linear SVM was then trained to perform smile detection.

Experimental Results: First, we present our results given the ground truth segmentation into substages. We used a training set of 39 child participants and a testing set of 17 additional participants. In this ideal case, we correctly predicted 72 out of 90 substage labels, with a balanced accuracy of 79.5%, and a chance performance of 60%. Next, we present the results of combining our smile detection system with the parsing result from Section 5, which yields a fully automated smile detection system. Using 8 sessions with high parser confidence, we correctly detected smiles in 31 out of 40 substages, giving a balanced accuracy of 76.6%. These results suggest that useful predictions can be made in the absence of fine-grained training labels.

We note that children are difficult subjects for automated face analysis, as they are more likely to move rapidly and turn their heads away from the examiner (and therefore the camera). Our experiments suggest that face detection methods which are trained for adult faces incur a 10-15% drop in accuracy when applied to very young children.

6.2. Gaze Detection

Gaze is a fundamentally important element in understanding social interactions, and the automatic non-intrusive measurement of children's gaze remains a challenging unsolved problem. When children's eyes are viewed from significant standoff distances, human coders can make assessments of gaze direction which far exceed the accuracy of any existing automated method. One of the challenges in our dataset is the difficulty of ensuring a continuous view of the child's eyes in the Basler camera, due to occlusion and head rotation. In previous work [21], we used a wearable camera on the examiner to obtain a more consistent viewing angle, but this adds additional complexity to the recording process. Related work by Marin-Jimenez et. al. [11] analyzes cinematic video footage. Our multimodal approach exploits the structured nature of the Rapid-ABC interaction and does not require an active camera system or the need to wear additional hardware.

Given a particular substage within the interaction, our goal is to predict whether the child made eye contact with the examiner at least once. For this analysis, we made use of both head tracking information obtained from the overhead Kinect sensor, as well as information about the child's eye gaze, as measured in the child-facing camera view (see Figure 1). We followed a two-stage approach. First, we used the Kinect to identify moments when the child's head was oriented towards the examiner. We performed head tracking (see Section 7.1 for a discussion) and in addition used template matching to estimate the yaw of the head. Given a within-bounds yaw estimate, the second step was to examine the Basler video to estimate the pitch of the child's gaze. Our goal was to differentiate gaze directed up at the examiner's face from gaze directed down at hands or objects on the table. We used the Omron OKAO software to estimate the vertical gaze pitch. If the estimated pitch was above threshold, we predicted eye contact. If at least 10 frames in the clip received a positive vote, then we predicted a positive label for the clip overall.

Experimental Results: We performed our analysis on 20 hand-picked sessions in which the child remained at the table throughout and the tracker worked successfully. We used five sessions for parameter tuning and 15 for testing. Table 3 gives the percent agreement between the predicted gaze scores and the ground truth for select substages. The algorithm performs better on longer sequences, where there are increased opportunities to observe eye contact. The two main sources of error arise when the child's head pitches down, making it more difficult to estimate yaw, and when the child's eyes are not visible (e.g. when turned away from the examiner). It would be interesting to extend this approach to detect moments when the child is looking at targets other than faces, such as the objects used in the interaction.

Activity	Agreement (%)
Greeting	78.57
Ball Pause	64.28
Book Pause	61.53
Hat Present	85.71
Tickle Play	92.85
Tickle Pause	15.38

Table 3. Accuracy in predicting eye contact

7. Predicting Child Engagement

For each of the five stages of the play protocol, the examiner assessed the difficulty of engaging the child on a scale from 0 (easily engaged) to 2 (very difficult to engage). In this section, we describe methods for predicting the engagement score based on video and audio features. To simplify the task we collapse the categories 1 and 2 together across all five stages, giving a binary prediction problem.

7.1. Engagement Prediction in Ball and Book Play

In this vision-based approach to predicting engagement, we designed engagement features and trained a binary classifier to estimate if the child was easy to engage or not. The features were extracted through analysis of the object and head trajectories, and they leverage the events "ball is shown" and "ball is touched" that may be subsequently detected.

Object Detection and Tracking: We track the objects (ball and book) and the heads of the child and of the examiner using the overhead Kinect camera view. The tracker we adopted does not need human intervention or manual initialization. We detect the heads by searching for local maxima in the depth image. To detect the ball and the book, we use region covariance templates [19] over the RGB-D channels. To deal with detection failures and maintain consistent labeling, we use the tracking-by-detection method proposed in [22], following depth-based background suppression. The tracker keeps an ensemble of appearance templates for each target and uses the tracker hierarchy to automatically handle tracker initialization, termination and tracking failure. Figure 3 shows the output of the tracker for the Ball stage on a sample image.

Event Detection: For the Ball stage, we developed detectors for the events "ball is shown" and "ball is touched". The examiner shows the ball to the child by holding it high and near her head. We detect this event by measuring the relative position of the ball with respect to the head. In order to detect moments during the ball game when the partners touch the ball, we collected a training set of example templates in which the ball is touched and partially occluded. We then computed a rotation-invariant region covariance descriptor for each template, and identified the two that were the most discriminative. During tracking, we detect the ball region, extract its descriptor, and compare it

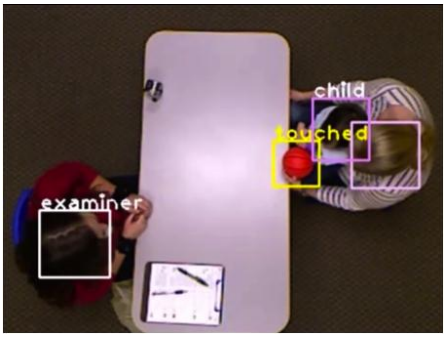

Figure 3. Tracking results for heads and ball

to the top two template descriptors using the Affine Invariance Riemann Metric (AIRM) distance [19]. If the matching score for both templates is below a pre-learned threshold, we predict "ball is touched." We can further classify into "touched by child" and "touched by examiner" by examining the ball location.

Feature Extraction and Engagement Prediction: To estimate the engagement level, we designed and extracted features that intuitively reflect the effort of the examiner to get the attention of the child, and the degree to which the child is participating in the interaction. If the child is easy to engage, we can expect that the examiner will spend less time in prompting the child, and the child will quickly respond to the examiner's initiating behaviors while interacting with the objects.

For the Ball stage, based on the detected ball shown and ball touched events, we extracted the raw features in Table 4, and split them in groups of no more than three. We applied PCA and trained linear SVMs on the coefficients of the projected vectors for each group of features. Finally, the margin from each SVM was treated as a mid-level feature and used to train a decision tree to predict whether the child was easy to engage or not. In our experiments, the groups of features retained by the decision tree for the Ball stage were {Visibility, Near_Examiner}, {Shown, Touched_Examiner} and {Touched_Child}. For the Book stage, we used a linear SVM trained on the raw book features described in Table 4.

Stage	Feature Name	Explanation
Book	Visibility	num. of frames the book is detected
	Touched	changes in presence of skin-color on the book
	Motion_Mag.	of sparse set of corners on the book
	HOF	of a sparse set of corners on the book
Ball	Visibility	num. of frames the ball is detected
	Touched	num. of "ball is touched" events
	Touched_Child	num. of "ball is touched by the child" events
	Touched_Examiner	num. of "ball is touched by the examiner" events
	Shown	num. of "ball is shown to the child" events
	Near_Child	num. of frames the ball is near the child
	Near_Examiner	num. of frames the ball is near the examiner
	Effort_Examiner	Ball Shown + Ball Touched by Examiner

Table 4. Raw features for the Ball and Book stages

Experimental Results: The training set comprises 16 different sequences, while the test set comprises 15 sequences.

During testing, the ball tracker failed in one sequence, and we omitted it from the ball results. The overall accuracy in predicting engagement for the Ball and Book stages was 92.86% and 73.33%, respectively. In all cases, stages with the label "difficult to engage" were predicted perfectly. For the easy to engage cases, we had a false negative rate of 8.3% for the Ball stage and 33.3% for the Book stage. The book interaction involved a deformable object and more complex patterns of occlusion, and was therefore more difficult to analyze.

7.2. Audio-Visual Prediction of Engagement

We have demonstrated that visual features can be used to accurately predict engagement in the Ball and Book stages. In this section, we develop a complementary approach based on acoustic features, and present some initial experimental results for audio-visual prediction.

The first step in our approach is to automatically segment the speech portions of the audio input. We developed a Voice Activity Detector (VAD) which utilized energy-based features, zero-crossing rates, voiced/unvoiced rates, pitch-related statistics, and noise adaptation processing. The VAD was applied to both the child and the examiner's lapel-mounted wireless microphones, thereby identifying the start and end of speech segments and extracting them. Table 5 gives the statistics for the duration and number of extracted speech segments for the examiner (E) and child (C).

	Greet	Ball	Book	Hat	Tickle
Duration (min)	7	51	81	10	30
Dur-Sp-E (min)	4	26	35	9	25
Num-Seg-E	149	880	1338	225	402
Dur-Sp-C (min)	2	8	13	2	5
Num-Seg-C	65	325	553	69	265

Table 5. Speech segmentation results

After obtaining the speech segments, we extracted acoustic features from each one using the openSMILE toolkit [2]. The acoustic features consisted of prosodic, spectral, formant, and energy analyses along with their statistics, regression coefficients, and local minima/maxima [7, 17]. A total of 2265 features were extracted. In addition to acoustic features, we added event-based features such as the duration of cross-talk between child and examiner, the number of turns taken (C-to-E and E-to-C), and the number of speech segments. The features were normalized over duration of stage, durations of segments, and number of segments.

In order to reduce the dimensionality of the feature set, the average accuracy of each individual feature was first calculated using three-fold cross-validation with a Gaussian Mixture Model (GMM) classifier of two mixtures. The features were ranked by the unweighted accuracy, which is equivalent to $\frac{1}{2}(\frac{TP}{TP+FP} + \frac{TN}{TN+FN})$, where TP stands for true positive, FN for false negative, TN for true negative, and FP for false positive. This measure is an appropriate choice for unbalanced problems like ours, where the distribution of scores is highly unequal.

A sequential forward feature selection algorithm was then applied to the 50 highest ranked features, examining them one-by-one in rank order. Starting with the highest ranked feature, any feature x_i that did not result in an improvement in the error rate was discarded. Otherwise, it was added to the working set. At each iteration, the error rate was tested with three-fold cross validation using the GMM classifier. After the sequential forward selection algorithm, 11 features were retained, which included 4 event-based, 3 spectral, 1 energy, 2 formant, and 1 prosodic related features, as shown in Table 6. All analyses utilized 46 sessions for training and another 14 for testing.

Order	Feature	Type
1	Number of Child Speech Segments	Event
2	Number of E-to-C	Event
3	audSpec-Rfilt-sma-de[3]-upleveltime90	Spectral
4	mfcc-sma-de[7]-qregc1	Spectral
5	pcm-RMSenergy-sma-de-percentile1.0	Energy
6	Duration of cross-talk	Event
7	F3-percentile50	Formant
8	Number E-to-C / (number of E segments)	Event
9	mfcc-sma[2]-linregc1	Spectral
10	Bandwidth2-percentile25	Formant
11	F0-sma-qregc2	Prosodic

Table 6. Selected features

Experimental Results, Audio Only: Using the feature set identified in the previous section, the training set was used to train two Gaussian mixtures in 11 feature dimensions, and then testing was performed with a Bayesian Classifier. Table 7 shows the overall results on the test set, which consists of 14 sessions. Our classifier was less effective on the Greeting stage, due to its short duration relative to the other stages.

Experimental Results, Audio-Visual: In addition, we obtained initial experimental results in combining our audio features with the visual features from Section 7.1. We employed a late-fusion approach to combining modalities, and worked directly with the previously-trained audio and video classifiers. Using data from the Book stage, we combined the estimated confidence scores from the two classifiers and made decisions based on the combined score. The audio-based classifier used normalized log-likelihood scores, and the vision-based classifier used a SVM output converted to a probability via a sigmoid mapping. The class with a higher combined confidence score was then selected. Using the 14 overlapping play interaction sessions in the test set, the joint classifier resulted in 10 true positives, 3 true negatives and 1 false positive. This result is an improvement over the previous classification peformance (with the caveat that the sample size is small).

	Greet	Ball	Book	Hat	Tickle
True Positive	60% (6/10)	83% (10/12)	91% (10/11)	86% (12/14)	91% (10/11)
True Negative	25% (1/4)	50% (1/2)	67% (2/3)	- -	33% (1/3)
Unweighted Accuracy	43%	67%	79%	-	62%
Weighted Accuracy	50% (7/14)	79% (11/14)	86% (12/14)	86% (12/14)	79% (11/14)

Table 7. Accuracy of engagement predictions using audio

8. Conclusion

We introduced a new and challenging domain for activity recognition—the analysis of dyadic social interactions between children and adults. We created a new *Multimodal Dyadic Behavior (MMDB)* dataset containing more than 160 examples of structured adult-child social interactions, which were captured using multiple sensor modalities and contain rich annotation. We presented baseline analyses which are a first attempt to decode children's social behavior by determining whether they produce key behaviors, such as looks to their partner, smiles, and gestures, during specific moments of an interaction, and by assessing the degree of engagement.

Our long-term goal is to develop a rich, fine-grained computational understanding of child behavior in these settings. To achieve this goal, we will need to go beyond the detection of discrete behaviors and the prediction of high-level ratings. We must consider many other aspects of these behaviors, such as their coordination (e.g., is the child combining affect, vocalizations, and gestures with looks to the examiner's face), timing (e.g., how does the child time their response to the examiner's social bids), and function (e.g., is the child directing the examiner's attention to an object to share their interest in the object, or only to request it). This endeavor will require new capabilities for face and gesture analysis and new computational models for behavioral coordination. We are making our MMDB database available to the research community to facilitate these advances (see www.cbi.gatech.edu/mmdb for details).

Acknowledgement: Portions of this work were supported in part by NSF Expedition Award number 1029679. Author Ousley acknowledges the support of Emtech Biotechnology Development, Inc.

References

[1] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, 2011. 2

[2] F. Eyben, M. Wollmer, and B. Schuller. openSMILE-The munich versatile and fast open-source audio feature extractor. *Proc. ACM Multimedia*, pages 1459–1462, 2010. 7

[3] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *ICCV*, 2011. 2

[4] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: a first-person perspective. In *CVPR*, 2012. 2

[5] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. *IEEE Trans. PAMI*, 29(12):2247–53, Dec. 2007. 2

[6] Y. Ke, R. Sukthankar, and M. Hebert. Volumetric features for video event detection. *IJCV*, 2010. 2

[7] J. Kim, H. Rao, and M. Clements. Investigating the use of formant based features for detection of affective dimensions in speech. In *Proc. 4th Intl. Conf. on Affective Computing and Intelligent Interaction*, pages 369–377, 2011. 7

[8] A. Kylliäinen and J. K. Hietanen. Skin conductance responses to another person's gaze in children with autism. *Journal of Autism and Developmental Disorders*, 36(4):517–525, May 2006. 2

[9] T. Lan, Y. Wang, W. Yang, and G. Mori. Beyond actions: discriminative models for contextual group activities. In *NIPS*, 2010. 2

[10] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions from Movies. In *CVPR*, 2008. 2

[11] M. J. Marin-Jimenez, A. Zisserman, and V. Ferrari. "Here's looking at you, kid." Detecting people looking at each other in videos. In *BMVC*, 2011. 5

[12] R. Messing, C. Pal, and H. Kautz. Activity Recognition Using the Velocity Histories of Tracked Keypoints. In *ICCV*, 2009. 2

[13] V. I. Morariu and L. S. Davis. Multi-agent event recognition in structured scenarios. In *CVPR*, 2011. 2

[14] O. Y. Ousley, R. Arriaga, G. D. Abowd, and M. Morrier. Rapid assessment of social-communicative abilities in infants at risk for autism. Technical Report CBI-100, Center for Behavior Imaging, Georgia Tech, Jan 2012. Available at www.cbi.gatech.edu/techreports. 1, 3

[15] K. Prabhakar and J. M. Rehg. Categorizing turn-taking interactions. In *ECCV*, Florence, Italy, 2012. 2

[16] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, Kyoto, Japan, 2009. 2

[17] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. The first international audio/visual emotion challenge. In *Proc. 4th Intl. Conf. on Affective Computing and Intelligent Interaction*, 2011. 7

[18] D. Tran and A. Sorokin. Human Activity Recognition with Metric Learning. *ECCV*, pages 548–561, 2008. 2

[19] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. *ECCV*, 2006. 6

[20] A. Wetherby, J. Woods, L. Allen, J. Cleary, H. Dickinson, and C. Lord. Early indicators of autism spectrum disorders in the second year of life. *Journal of Autism and Developmental Disorders*, 34:473–493, 2004. 1

[21] Z. Ye, Y. Li, A. Fathi, Y. Han, A. Rozga, G. D. Abowd, and J. M. Rehg. Detecting eye contact using wearable eye-tracking glasses. In *2nd Workshop on Pervasive Eye Tracking and Mobile Eye-based Interaction (PETMEI)*, 2012. 5

[22] J. Zhang, L. Lo Presti, and S. Sclaroff. Online multi-person tracking by tracker hierarchy. In *Proc. IEEE Conf. on Advanced Video and Signal Based Surveillance*, 2012. 6