The Static Multimodal Dyadic Behavior Dataset for Engagement Prediction

P. Daphne Tsatsoulis[1](✉), Paige Kordas[1], Michael Marshall[1], David Forsyth[1], and Agata Rozga[2]

[1] University of Illinois at Urbana-Champaign, Champaign, USA
{tsatsou2,pkordas2,mtmarsh2,daf}@illinois.edu
[2] Georgia Institute of Technology, Atlanta, Georgia
agata@gatech.edu

Abstract. The Rapid-Attention, Back and Forth, and Communication (Rapid ABC) assessment is a semi-structured play interaction during which an examiner engages a child in five activities intended to elicit social-communication behaviors and turn taking. The examiner scores the frequency and quality of the child's social behavior in each activity, generating a total score that reflects the child's social engagement with her during the assessment. The standard Rapid ABC dataset contains a daunting amount of detail. We have produced a static version that captures the action-reaction dynamic of the assessment as frames. We have conducted a user study on our dataset to see if subjects can predict the engagement of a child in the video. We presented subjects both frames from our staticMMDB dataset and the full video of the original MMDB dataset and found little difference in their performance. In this paper we show that computer vision methods can predict children's engagement. We automatically identify the ease-of-engagement of a child and provide evaluation baselines for the task.

1 Introduction

A child's early developmental period is crucial and being able to capture characteristic differences between children‘s behaviors is important. The Rapid ABC assessment is a standardized play assessment for eliciting early social-interactive behaviors in toddlers [1]. Such assessments are common in developmental psychology research, where they are used both to understand typical behavior and to identify children who may be experiencing delays, such as autism. The Rapid ABC scores children on how they participated in the assessment by checking for specific behaviors and the overall ease of engagement. During this assessment an examiner prompts a child and measures the child's response. The examiner then scores the child.

We use the Multimodal Dyadic Behavior Dataset (MMDB), a dataset with videos of children undergoing the Rapid ABC assessment [2]. This dataset contains hours of videos in which examiners and children are interacting. The dataset has annotations of behavior, speech, and the child's engagement in tasks. It is

G. Hua and H. Jégou (Eds.): ECCV 2016 Workshops, Part III, LNCS 9915, pp. 386–399, 2016.
DOI: 10.1007/978-3-319-49409-8_31

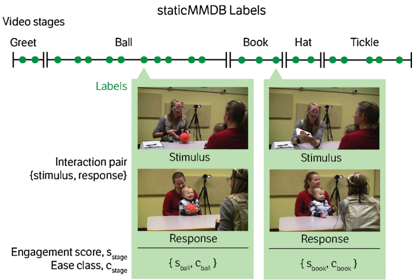


Fig. 1. The original MMDB dataset contains videos of scripted interactions with children that are broken into five stages. During each stage, the examiner scores the engagement of the child. We have created the staticMMDB dataset by capturing the most important interactions during the assessment. We extracted frames to represent the stimulus-response pairings of each interaction.

a rich resource for structured examiner prompts and spontaneous child reactions. An examiner follows a script and will prompt a child identically in every Rapid ABC assessment making the examiner actions across assessments equivalent. The reactions from children however, are unscripted and varied.

Our goal is to summarize the MMDB Dataset (visualized in Fig. 1) so that visually important interactions are highlighted. We label the videos in the MMDB dataset so that every important visual interaction (when an examiner prompts a child) is represented by a stimulus and a response. Our resulting dataset, the staticMMDB dataset, supplements the original dataset and is comprised of time-stamps that identify important interaction frames. Our dataset gives indices into the data that correspond to the most interesting visual events. The staticMMDB dataset gives the opportunity to explore very hard questions, like how engaged a child is during interactions and what is the appropriate response to a stimulus.

Our labelling supplements the original dataset. It is comprised of time-stamps that identify important interaction frames and gives the vision community indices into the dataset that correspond to the most interesting visual events. Our dataset still encapsulates the hard aspects of the data and removes distractions. Even though we select frames to represent the video we still retain the ability to compute video features. It is easy to extract video data surrounding those frames to use for experimentation.

We automatically identify the engagement of children from the frames in the staticMMDB dataset. Our predictions agree with those of the examiner. We can predict the engagement of children not present in the training set, and most importantly, we can identify disengaged children. Our dataset and the evaluation task serve as baselines for future work to build on.

In order to verify that the staticMMDB dataset is as informative as the original MMDB dataset we conducted a user study. The user study tested whether subjects could accurately predict a child's engagement in an activity from clips of video (the original MMDB) and from frames of video (our staticMMDB). The subjects in our study did very well at identifying engaged children but had difficulty identifying disengaged children. This highlights how difficult this task is. We also compared subject performance when using just videos or just frames to make their decision. The study showed that frames and video are equally informative for predicting engagement. This indicates that our dataset contains the same information for this task as the original dataset.

Contributions:

1. We have summarized an extensive video dataset, the MMDB, to provide a series of standard interactions that emphasize visual events. Our summary, called the staticMMDB dataset, provides time stamps of frames and is a structured annotation for the MMDB dataset.
2. We have conducted a user-study that demonstrates the difficulty of the task and validates our claim that the staticMMDB dataset is as informative for engagement prediction as the original MMDB dataset.
3. We demonstrate a method that is able to identify the engagement of children and beat baseline evaluations for the task.

2 Background

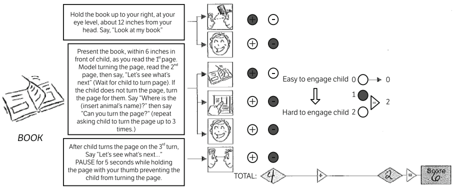
2.1 The Multimodal Dyadic Behavior Dataset

The Multimodal Dyadic Behavior (MMDB) dataset presents a unique opportunity to investigate real interactions between two people. It is different from existing behavior datasets because the actions that we see are not drawn from a curated collection like the internet. Children's responses are open-ended and do not fall naturally into categories of behavior.

Engagement. For every child's video there exist a fixed number of important interactions when the examiner prompts the child and the child reacts. The examiner scores the child's engagement based upon that reaction. The examiner tries to engage the child during five substages: greeting, ball play, book, putting on a hat, and tickling. Each substage is identically structured and contains many interactions that occur in the same order in each video. The examiner scores the level of engagement of the child during each substage. An example of the evaluation script an examiner uses is shown in Fig. 2.

2.2 Engagement Prediction

There is little work on the Multimodal Dyadic Behavior dataset and in engagement prediction. We are aware of three works, Rehg et al. [2], Presti et al. [3], and Gupta et al. [4] that predict engagement on this data. The three works address the question of ease-of-engagement classification and approach the problem as a binary ([2,3]) or three-class [4] classification problem.



Hold the book up to your right, at your eye level, about 12 inches from your head. Say, "Look at my book"

Present the book, within 6 inches in front of child, as you read the 1st page. Model turning the page, read the 2nd page, then say, "Let's see what's next" (Wait for child to turn page). If the child does not turn the page, turn the page for them. Say "Where is the (insert animal's name)?" then say "Can you turn the page?" (repeat asking child to turn the page up to 3 times.)

After child turns the page on the 3rd turn, Say "Let's see what's next..." PAUSE for 5 seconds while holding the page with your thumb preventing the child from turning the page.

Fig. 2. A portion of the scoring form an examiner completed during a Rapid ABC assessment. The assessment is broken into five stages, one of which is the Book stage. An examiner scores explicit actions of the child during the assessment and also the overall ease with which she engaged the child. These two values are summed to produce the engagement score of the stage. In this example the ease class is 1 and the engagement score is 6. In this work we predict the ease-of-engagement of the child as a binary task. Engaged children have an ease-of-engagement score of 0 and disengaged children have a score of 1 or 2.

Rehg et al. [2] introduce the MMDB dataset. They detect gaze direction, smiles, and predict the ease-of-engagement in the videos. When using vision-features, they predict whether a child is engaged (binary task) in two substages ('ball' and 'book') on 14 test videos. After incorporating speech features, they are able to predict whether a child is engaged during all five substages for those 14 videos. Gupta et al. [4] perform a 3-way classification task on the ease of engagement (0 - easy to engage, 1 - moderately easy, 2 - hard to engage) using prosodic speech cues. They report their performance on 74 test videos using a stage-wise per-class recall. Presti et al. [3] predict a binary ease-of-engagement class using vision features. They propose a Hidden Markov Model for predicting the engagement class on 33 test videos. They also classify the stage (greeting, ball, book, hat, tickle) of each video clip.

None of the existing work is comparable. Each work uses different subsets of the original dataset that each contain a different distribution of disengaged children. The dataset we have created solves this problem. Our dataset summarizes the original MMDB dataset in a way that allows for comparisons between different methods and between various time points in the exam.

3 The staticMMDB Dataset

The original MMDB dataset contains a lot of extraneous video, and our summarization removes unnecessary noise while preserving the structure of the assessment. The reduced dataset we create is still a very hard dataset. We labeled the original dataset to create a more concise representation of the two-person interactions occurring in the videos. For each video we selected frames to represent key interactions between the examiner and child. The staticMMDB dataset summarizes every video and creates a more tightly defined task to which future work can compare.

We summarize the dataset so that every important interaction (when an examiner prompts a child) is represented by a stimulus and a response. The decisions we made when selecting interactions were based on the original dataset's annotations. We chose examiner speech and motion cues that would elicit visual responses from the child. We have identified 22 stimuli in the examination that represent the most significant examiner prompts (seen in Fig. 4).

For each video we use two camera angles: the camera positioned behind the child and the camera positioned behind the examiner. We will call frames from the camera positioned behind the examiner 'child-frames' because they are focused on the child. In a child-frame, the child and the adult holding the child are in the middle of the frame. The back of the examiner is also visible in the child-frame so that the frame captures the whole interaction. An 'examiner-frame' is focused on the examiner with part of the child's back visible.

The examiner-frame was extracted at the start of a stimulus and the child-frame was selected to capture the child's response to the examiner's prompt. To find an accurate portrayal of the child's interaction around a time-slot, we sampled a number of frames for 10 s after the examiner-frame. We then looked at the subset of frames and chose the one we thought best described the child's interaction.

The dataset frames were selected by a labeler who was shown a subset of frames for every interaction and asked to select the best child-response frame. They selected 22 child and 22 examiner frames for each subject. We used 98 subjects and their corresponding engagement scores for our experiments.

3.1 Stimulus Selection

When making the staticMMDB dataset, we wanted the examiner's speech and actions to be identical across the videos we used. Every video in the original MMDB dataset has been annotated with 11 elements (seen in Fig. 3) [2]. The

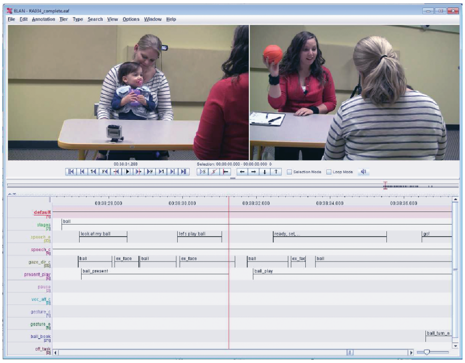


Fig. 3. A screenshot to show both the child and examiner video along with the annotation provided with each video in the MMDB dataset. The annotations are rich and many overlap making it unnecessary for us to use all of them in our dataset. For example, 'ball present', 'look at my ball' and 'let's play ball' all occur at about the same time. We chose to represent all three by using the single stimulus 'look at my ball'.

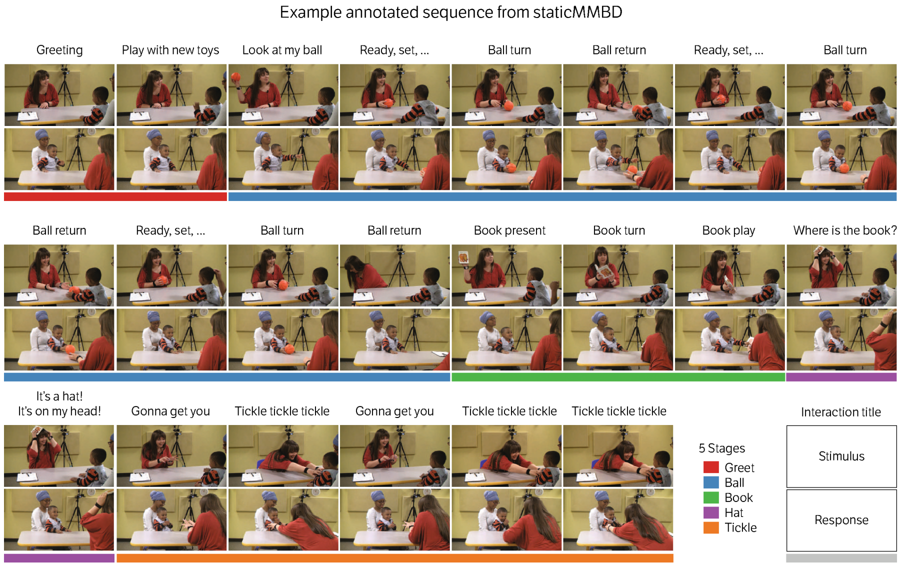


Fig. 4. This is an example of one annotated video in the staticMMDB dataset. Each stimulus-response pair represents one interaction. The color of the line under the interaction represents the stage in which the interaction occurred. The title of the interaction is the action of the examiner or the speech of the examiner that serves as the stimulus. The examiner-frame chosen to represent a stimulus is the first frame in the video labeled as a specific stimulus. The child-frame was selected within 10 s of the examiner's stimulus. (Color figure online)

figure shows each element as a row in which a professional annotator has indicated if a specific type of action is occurring. We are interested in visually interesting actions like the child's gestures (reaching, pointing) and the child's gaze direction (at the examiner, the book, the ball). Of the videos given to us we chose the 98 videos with identical speech cues. For example, we excluded videos that used 'one, two, three' instead of 'ready, set, go' in the ball phase. We also chose to use videos for which the cannon cameras had been annotated. This was to ensure the same view direction and resolution between subjects.

The MMDB dataset contains a plethora of annotations that do not all correspond to visually interesting events. In order to create the staticMMDB dataset we selected key interactions by dropping some annotations, consolidating others, and creating new annotations to capture important instances.

Dropping Annotations: We used the five stages listed in the engagement evaluation form excluding 'name_s' and 'name_f'. We used the annotations in speech_e_cv, ball_book_cv, and present_play_cv as stimuli since these are scripted and identical across videos. Speech and actions that were not the same across videos were not used. For example, 'can you turn the page?', occurs during the book stage and the examiner asks the child to turn the page of the book. This speech cue can be repeated up to three times depending on the child's initial reaction making it different in each video.

Annotation Consolidation: We do not include all speech and actions of the examiner because they may fall immediately after each other; for example, saying "go!" and passing the ball to the child have been consolidated into the stimulus 'ball turn e'. Generally, if two examiner prompts occurred within a few seconds of each other (or simultaneously) in most videos we picked one to represent the overlapping group. We tried to pick the first stimulus in the group though this sometimes varied slightly between videos and so we selected to the best of our ability. By grouping overlapping stimuli we avoided having too few response frames to choose from and we avoided having very similar stimulus-response pairs.

New Annotations: We also created three new stimuli by breaking the provided stimuli into two. We split 'ball turn e' into two stimuli ('ball turn e' and 'ball return e') to capture the pass to the child and the return from the child separately. Since this action occurs three times in the video we had to create three new instances. We did this by splitting the time block for 'ball turn e' into two equally sized time blocks.

Some stimuli (such as 'ready set ...' and 'tickle tickle tickle') occur multiple times throughout the video. It is important to distinguish between each instance of the same stimulus because children will react differently every time. For example, 'tickle tickle tickle' is repeated three times and children's reactions escalate throughout the video because they know what to expect after the first instance. Similarly, 'ready set ...' is followed by 'ball turn e' (which represents the phrase 'go!' and 'ball turn e') three times. On the last 'ready set ...' the examiner waits a few seconds before releasing the ball to the child. In this case both the stimulus and reaction are different highlighting the need for stimuli with the same speech to be unique.

3.2 Response Selection

After selecting the 22 stimuli we extracted the examiner frame at the start of the stimulus to represent the response. The original MMDB dataset provides timestamps corresponding to the start of each of the stimuli. We then showed a labeler child-frames that occurred in the 10 s after the examiner's stimulus. If a new stimulus started before 10 s had elapsed we truncated the 10 s so that the child's response would not overlap with a new stimulus.

Labeler: During labeling we showed 15 frames from the first 5 s and 5 frames from the second 5 s for each response. The labeler then picked the child frame that best captured the child's response to the stimulus. Many of the labelling decisions made were motivated by the data itself. For example, if the examiner asked "look at my ball" and 10 of the 11 frames showed the child looking down, and only 1 frame where the child looked up, the labeler chose a frame with the child looking down.

The labeler also made the decision to be consistent among videos for which frames he chose to represent certain scenes. For example, if the examiner said, “catch the ball!”, resulting frames showed a child looking in anticipation, reaching their hands out, and finally holding the ball. In these scenarios, the labeler chose a specific frame for each child that behaved in the same way. For example, in the case of the ball being passed, he chose the frame with the child holding the ball. This made sure that children that performed the same interactions were indeed represented in the same way in our resulting frames.

3.3 Contributions and Challenges

The staticMMDB supports a wide variety of studies. In Sect. 4 we explore the task of engagement prediction in both the video and frames of children interacting. This task is very hard to perform on this dataset. A number of sometimes subtle actions are used to score a child. There is no single clear indicator for predicting engagement. Children of varying engagement will sit close to an adult, will look at the examiner, and will react to the presence of an object. Furthermore, for each stage there are very few examples of low-engagement children making the dataset very biased as seen in Fig. 5.

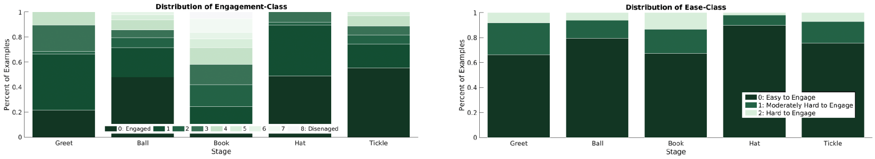


Fig. 5. The staticMMDB dataset has engagement scores and ease classes for every interaction. We have plotted the scores for each of the five stages in the 98 videos. The datasets are strongly biased towards high-engagement children (low-scoring) making the tasks of predicting engagement and ease class very challenging.

The staticMMDB dataset gives the opportunity to explore pairing stimuli and responses. Very hard questions, like what the best response to a stimulus is, can be explored using this data.

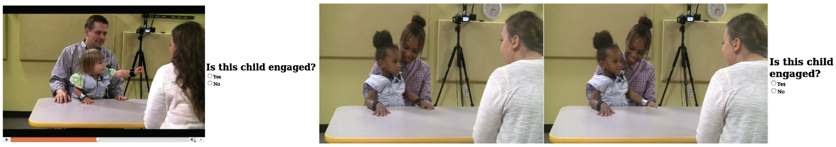
4 User Study

The purpose of this study was to test whether subjects can predict a child’s engagement in an activity from clips of video and from frames of video. Since engagement is not a concrete action (as is smiling or gaze direction) it is difficult to define and predict. Our user study indicates how well a user can match the gold-standard engagement scores provided by expert examiners.

Our goal was also to compare the performance of the two sources (clips and frames) in order to confirm which is more informative or if they are equally informative. We have introduced the staticMMDB dataset, a summary of the MMDB dataset. We claim that our dataset is a cleaner version of the original dataset and captures the information of the original MMDB dataset with less noise. By showing that frames and video are equally informative for predicting engagement we have shown that our dataset contains the same information for this task as did the original dataset.

4.1 Method

We collected examples of disengaged and engaged children for each of the five stages. We then had users look at example videos (from the original MMDB dataset) and example frames (from our staticMMDB dataset) during each stage. Please see Fig. 6 for an example of the interface subjects used.



Is this child engaged?
○ Yes
○ No

Is this child engaged?
○ Yes
○ No

Fig. 6. Examples of the screens that would be shown to a subject during our user study. On the left is a screen-shot of a video from the 'Greeting' stage. Subjects were asked to watch the video (without sound) and then decide whether the child in the video was engaged. On the right is a screen-shot of the frames from the 'Greeting' stage. Subjects were again asked to decide if the child was engaged. Each subject was shown frames and videos from different children and from different engagement classes.

We recruited 23 users for our study. The subjects were randomly split into three groups; each group was shown the same data to annotate. For every stage a subject was shown 9 video examples and 9 frame examples. Six of the nine examples were children who scored a '0' (engaged), two of them scored a '1' (disengaged) and one child who scored a '2' (disengaged). Subjects were given examples of dis/engaged frames and videos before making their own predictions in order to train them on the task. They were told that engagement is determined by a number of factors including eye gaze direction, gestures, and reciprocity.

The study was divided across five days: one day for every stage (greet, ball, book, hat, tickle). The dataset contains 98 children with very few disengaged children. We spread the study across multiple days in order to prevent subjects remembering the behavior of a child across stages since they could potentially see the same child twice.

4.2 Results

Engagement Prediction: Subjects were very good at correctly identifying engaged children as shown in the class confusion matrices in Fig. 7. They did

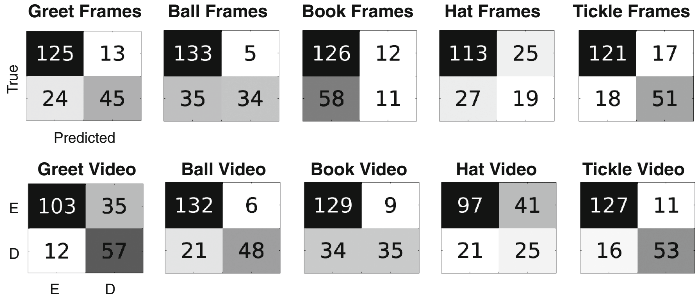


Fig. 7. Class-confusion matrices for the frame (top row) and video (bottom row) predictions for all examples. Subjects were very good at identifying engaged children (E) using both frames and videos for all stages. They were not as precise when identifying disengaged children (D) especially in the Book and Hat stages.

not do as well when predicting disengaged children. Engagement can represent itself subtly and can vary greatly between children. For certain activities, like greeting a person, it is much more obvious what cues to look for when making the decision (waving, eye contact). When engaging in a book however, it can be much more difficult to see if the child is interested and participating. Another particularly difficult stage for subjects was the Hat stage since very few children are disengaged (see Fig. 5 for statistics). All children in our dataset made eye-contact with the examiner during the Hat stage but it was not enough to decisively determine that the child was engaged.

Video vs. Frames: Correctly identifying disengaged children is critical when using this dataset. Since disengagement is a strong indicator of developmental delays such as autism it is most important to identify low-engagement children. We use the F1 score to report our results. The F1 score $\in [0, 1]$ captures the method's performance using both precision and recall $F1 = 2 \times \frac{p\ r}{p+r}$. The scores for every stage, divided by performance on the engaged children (class 0) and disengaged children (class 1), can be seen in Fig. 1.

There is little performance difference when seeing videos (the original MMDB dataset) and our selected frames (the staticMMDB dataset). This supports our claim that the staticMMDB dataset is a useful summary of the MMDB dataset since it captures the important engagement information with far less data. We report the Kolmogorov-Smirnov test's p-value above every frame-video comparison. This test assumes that the data (frame predictions and video predictions) are sampled from the same distribution. The hypothesis is true for all our comparisons (Fig. 8).

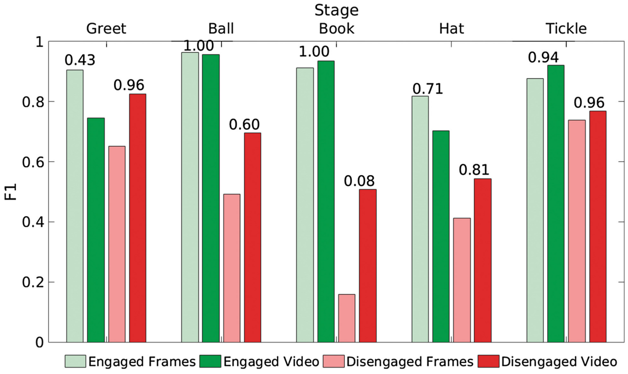


Fig. 8. For every stage we compare the F1 score of subjects when predicting the engagement for engaged children (green) and disengaged children (red). There is very little performance difference when seeing videos and frames (with the exception of disengaged children in the book stage). This supports the fact that the staticMMDB dataset captures the important information for predicting engagement in a much smaller dataset. We report the Kolmogorov-Smirnov test's p-value above every frame-video comparison. (Color figure online)

5 Engagement Prediction

5.1 Method

In our work we focus on representing the child's face because facial expressions, vocalizations, and gaze direction are of particular interest to examiners as outlined in their evaluation form.

Face Detection: Before extracting faces from the frames we cropped out a region around the child's face. We did this because our face detector either failed to detect faces in the full frame or only succeeded at detecting the parent's face.

For each child, we ran the frontal and profile Viola-Jones face detectors in MATLAB on each of the 22 annotated child-frames in the staticMMDB. If the algorithm detected more than one face we filtered out detections that were too low in the frame (since the children sit at a table), that had significant overlap with other detections, or that had very small bounding boxes. If two unique faces remained in the image, we selected the lowest face to represent the child (since the higher face is that of the adult). Face detections were manually checked and corrected to ensure that a child face had been identified.

We then expanded the face detection in the 22 frames to a larger box around the child's face. We applied the same box to surrounding frames. For frames located between two annotated face-box frames we applied a weighted (based upon temporal distance) average of the crop locations. This way we had a crop containing a child-face for every frame in the video. A visualization of our pipeline can be seen in Fig. 9. We passed the crop from every frame through the dlib face detector provided by [5].

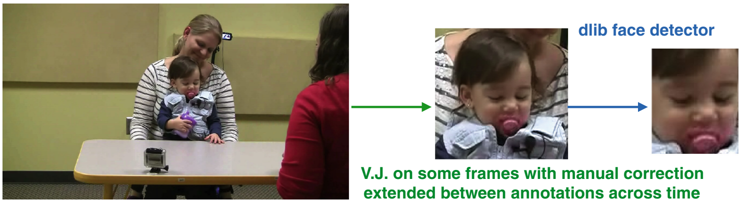

dlib face detector
V.J. on some frames with manual correction extended between annotations across time

Fig. 9. Our method's pipeline. Frames are reduced to crops around the child's face using 22 corrected Viola-Jones face detections in each video. A face detector is run over the crops of all frames in the videos. We extract VGG Face features for every face and train a SVM to predict engagement.

Train - Test Split: We split our data into train and test sets based upon children. Since the scores of children across frames and stages are correlated we could not include the same child in training and testing. We split the children into 10 groups - eight of which contained 10 children and two of which contained 9 children. The groups were split so that every stage would contain at least one disengaged child. This allowed us to perform a 10-fold cross validation on the data. For each fold we randomly chose 8 groups of children to use in training and 1 to use in validation. The last group was reserved for testing.

VGG Face and SVM: Face identification networks extract features that describe the face and could be useful in predicting engagement. We chose the VGG Face network because of its state-of-the-art performance on face identification. The VGG-Face descriptors are computed using an implementation based on the VGG-Very-Deep-16 CNN architecture as described in [6]. We passed our detected child faces through the VGG Face network and extracted the features from the second-to-last layer. We normalize the feature representation and train a SVM to predict the engagement of the child in each frame. The SVM has a uniform prior as to not over-predict engagement.

Prediction: Children were given five ease-of-engagement scores, one score per stage. We report our method's performance based upon votes over the predictions for a child in every stage. For each stage the staticMMDB contains a set number of annotated frames (the Book stage has 3 frames for example). The annotated frames vote to predict the ease-of-engagement for the child in that stage and the final decision is what we use in evaluation.

5.2 Results

Baselines: We present two baselines for comparison. All results are computed and reported per-stage.

Prior: The prior performance on the dataset is analogous to what would happen if we drew a label (engaged or disengaged) from the training data's distribution (think of flipping a weighted coin). This strongly favors the engaged class.

Table 1. The F1 results of our method and two baselines: a draw from the prior distribution and human performance on frames. All methods perform much better at predicting engaged children than disengaged children. Humans have the most difficulty in the Book and Hat stages. Our method outperforms the Prior baseline in all stages and outperforms humans in the Book stage (disengaged) and the Hat stage (engaged).

F1 Disengaged Class

	Greet	Ball	Book	Hat	Tickle
Prior	31	17	30	8	22
Human	71	63	24	42	74
VGG Face	45	17	31	15	32

F1 Engaged Class

	Greet	Ball	Book	Hat	Tickle
Prior	65	79	66	89	75
Human	87	87	78	82	87
VGG Face	80	82	76	94	84

Human: We compare our method's performance to that of human prediction. It is not our goal to outperform humans at this task but rather to understand how much potential we have for improvement.

VGG Face Results: Our method performs strongly when predicting engaged children. It clearly outperforms the prior in all stages and outperforms humans in the Hat stage. Our method and humans perform worse when considering disengaged children. Our method beats human performance on the book stage and beats the prior baseline in all stages. Table 1 lists the F1 results on engaged and disengaged children for the baselines and our method.

Greeting Results: Our method performs strongly on the Greeting stage. The Greeting stage is a short and clear stage in which the child is expected to make eye-contact with the examiner, smile, or wave. This type of engagement is easily captured by our face features.

Ball Results: Our method struggles to rank disengaged children in the Ball stage. The ball stage is strongly determined by the child's interaction with the ball itself, an interaction that is difficult to capture using face features.

Book Results: Our method performs well at this stage and outperforms humans on the disengaged class. Humans struggle with the Book stage because of its subtlety and because of the stark contrast in a child's behavior between the ball and book stages. The Book stage's evaluation also depends on physical interactions with the book not visible in the face.

Hat Results: The hat stage is a difficult stage as reflected by the human performance on engaged children. We are able to predict engaged children more accurately than humans but do not perform as well as humans on disengaged children. The hat stage is evaluated using two prompts: "Where is the book?" and "It's a hat! It's on my head!". The examiner then looks to see if the child looked at her and if the child laughed. In the data, all children (including disengaged children) looked at the examiner. This stage also has the smallest number of disengaged children making it a difficult stage to predict engagement.

Tickle Results: We perform well on this stage. Children are most likely to hide their faces and move a lot in this stage but we are able to predict their engagement regardless. Humans still strongly outperform our method on disengaged children leaving room for improvement.

Conclusions: We have created a dataset that summarizes the very complex MMDB dataset. Our static version captures the action-reaction dynamic of the original assessment by highlighting the most important interactions. Our user study shows that our dataset preserves the information needed for predicting engagement while being much simpler than the original dataset. Our dataset provides the opportunity to work on the important task of engagement prediction and we have provided a baseline for future work to compare to.

Acknowledgements. This material is based in part upon work supported by the National Science Foundation under Grant Numbers IRI-1029035. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

1. Ousley, O.Y., Arriage, R.I., Morrier, M.J., Mathys, J.B., Allen, M.D., Abowd, G.D.: Beyond parental report: findings from the rapid-abc, a new 4-minute interactive autism. Technical report, Georgia Institute of Technology, September 2013
2. Rehg, J.M., Abowd, G.D., Rozga, A., Romero, M., Clements, M.A., Sclaroff, S., Essa, I.A., Ousley, O.Y., Li, Y., Kim, C., Rao, H., Kim, J.C., Presti, L.L., Zhang, J., Lantsman, D., Bidwell, J., Ye, Z.: Decoding children's social behavior. In: CVPR, pp. 3414–3421. IEEE (2013)
3. Lo Presti, L., Sclaroff, S., Rozga, A.: Joint alignment and modeling of correlated behavior streams. In: The IEEE International Conference on Computer Vision (ICCV) Workshops, December 2013
4. Gupta, R., Bone, D., Lee, S., Narayanan, S.: Analysis of engagement behavior in children during dyadic interactions using prosodic cues. Comput. Speech Lang. **37**, 47–66 (2015)
5. Amos, B., Ludwiczuk, B., Satyanarayanan, M.: Openface: a general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science (2016)
6. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Proceedings of the British Machine Vision Conference (BMVC) (2015)