Play with Me - Measuring a Child’s Engagement in a Social Interaction

Shyam Sundar Rajagopalan¹ O.V. Ramana Murthy¹ Roland Goecke¹² Agata Rozga³

¹Vision & Sensing Group, HCC Lab, ESTeM, University of Canberra, Australia
²IHCC Group, RSCS, CECS, Australian National University, Australia
³School of Interactive Computing, Georgia Institute of Technology, Atlanta, GA, USA

Abstract—Due to the challenges in automatically observing child behaviour in a social interaction, an automatic extraction of high-level features, such as head poses and hand gestures, is difficult and noisy, leading to an inaccurate model. Hence, the feasibility of using easily obtainable low-level optical flow based features is investigated in this work. A comparative study involving high-level features, baseline annotations of multiple modalities and the low-level features is carried out. Optical flow based hidden structure learning of behaviours is strongly discriminatory in predicting a child’s engagement level in a social interaction. A two-stage approach of discovering the hidden structures using Hidden Conditional Random Fields, followed by learning an SVM-based model on the hidden state marginals is proposed. This is validated by conducting experiments on the Multimodal Dyadic Behaviour Dataset and the results indicate a state of the art classification performance. The insights drawn from this study indicate the robustness of the low-level feature approach towards engagement behaviour modelling and can be a good substitute in the absence of accurate high-level features.

I. INTRODUCTION

A computational understanding of human behaviours has significant potential for applications in multiple domains, including healthcare, security and human-computer interaction. However, modelling and analysis of human behaviours is extremely challenging, as the behaviours are contextual and often social, i.e. in relation to another person. An understanding of the interaction dynamics may be needed to understand the individual behaviour of a person. The complexities are currently tackled by interdisciplinary research involving scientists from psychology, sociology, medicine and computer science in an active emerging research field alternatively referred to as Social Signal Processing [1], [2], Behavioural Signal Processing [3] and Computational Behaviour Science [4].

An important application area of computational behaviour analysis is in characterising the behaviour and developmental change in children diagnosed with autism spectrum disorder (ASD). This condition affects children at their early developmental ages and is growing at a fast rate worldwide, with current prevalence numbers in the USA indicating 1 in 88 children are diagnosed with ASD [5]. Currently, there is no single genetic or biological marker for autism. The diagnosis of this condition relies entirely on the child’s behaviour and a common way of diagnosing is by using behavioural cues of the child [6]. The diagnosis involves clinicians interacting and observing children directly to identify the behaviour cues. Although a number of standardised diagnostic protocols exist for eliciting and scoring these behaviours, fundamentally, the diagnostic process relies on one individual’s subjective assessment of the child’s behaviour. There is significant opportunity for computational analysis tools to augment such diagnostic assessments with objective, quantitative, reliable measures of behaviours.

A social interaction involving two partners is naturally characterised by varying levels of engagement, expressed individually through verbal and non-verbal cues, including head pose, eye gaze, facial expressions, body pose and hand gestures. The use of high-level features, such as head poses and hand gestures, has been studied in the literature to predict the engagement level of a person in a dyadic social interaction. Though there have been significant advancements in computer vision, the automatic extraction of high-level features, particularly in unconstrained conditions, is still a challenging problem. When a person is dominant in a social interaction, the motion information will reflect the behaviour of a dominant person. We explore this hypothesis to study the suitability of using low-level optical flow based features in the absence of robust high-level features. A Hidden Conditional Random Field (HCRF) framework is employed to develop an engagement prediction model. The number of hidden states is crucial in the HCRF framework as an incorrect choice leads to a decline in performance. However, the estimated per-class hidden state marginals can offer additional discriminatory power to improve the robustness of the model. This results in the proposed design of a two-stage approach for improved classification performance.

A. Contributions

The contributions of this paper towards predicting a child’s engagement level in a social interaction are

1) An investigation of the feasibility of using low-level features in the absence of robust, accurate high-level non-verbal features.

2) A two-stage HCRF + SVM model for engagement prediction using the learned hidden structures of the behaviours.

The paper is organised as follows. We discuss the related work in Section II and introduction to HCRF in Section III. In Section IV, we establish the feasibility of using low-level features by comparing engagement level prediction performances. The proposed two-stage approach using low-level features is discussed in Section V, followed by experiments
and results in Section VI. Section VII concludes the paper with potential future directions.

II. RELATED WORK

Engagement is the process by which individuals in an interaction start, maintain and end their perceived connection to one another [7]. Computational models of engagement have been widely studied in Human-Robot Interaction scenarios, where the goal is for the robot to understand the engagement level of humans interacting with it. Both verbal and non-verbal signals are used to develop the models using supervised learning approaches. Recently, Foster et al. [8], have shown that head pose and hand locations are more discriminatory in developing a computational model for an engagement prediction as compared to torso orientations. A survey on social behaviour modelling, analysis and synthesis [9] reviews state of the art methods in the analysis of social relations (role recognition), social emotions (empathy, envy, admiration) and social attitudes (dominance, personality). A survey on the automatic analysis of non-verbal behaviours in small group interactions [10] focussed on issues involved in computational modelling of interaction management, internal states, personality traits and social relationships with relevant pointers to social science literature.

The development of computational models to understand the social-interactive behaviours of children is a relatively new area of study, facilitated by the recent public release of an annotated dataset. Rehg et al. [4] introduced the Multi-modal Dyadic Behaviour Dataset (MMDB), which includes 160 sessions of an interaction between an adult and children aged 15-30 months. The dataset includes the adult’s rating of how easy or difficult it was for the adult to engage the child in the interaction. A computational model using object and head trajectory features together with audio features was proposed to predict these engagement ratings. Presti et al. [11] proposed a variable Time-Shift Hidden Markov Model for learning and modelling pairs of correlation streams and validated their formulation for predicting the engagement level of a child using the MMDB dataset. The communicative or action oriented body gestures obtained from the depth images were used as a feature modality to train a model, which attained an average accuracy of 76.7% based on Leave-One-Out-Cross-Validation (LOOCV) tests.

The electrodermal activity (EDA) of the children, obtained from wearable sensors, has also been used to predict the engagement level of the child [12]. The combined use of child EDA features and EDA features capturing synchrony between the child and the interacting adult were used to train an SVM classifier, with an overall reported accuracy of 81% in predicting child engagement in the MMDB dataset. Finally, acoustic signals have also been evaluated in models aiming to predict child engagement in the MMDB dataset [13]. The spectral, prosodic and duration features obtained from the child and the adult’s vocal data were used to develop a model with a reported binary classification accuracy of 62.9%. Though there have been only few studies reported in the literature so far, it is interesting to note that almost all verbal, non-verbal and physiological modalities have been demonstrated to contribute to predicting a child’s level of engagement in a social interaction. Future work should focus on effective fusion of all these modalities for better understanding of children’s social-interactive behaviour.

III. HIDDEN CONDITIONAL RANDOM FIELDS

In order to model the engagement level – a time varying phenomenon comprising of multiple actions – a framework capable of representing the temporal dynamics of the actions is needed. Hidden Conditional Random Fields [14] are CRFs with latent variables that have been successfully applied to modelling such tasks. The HCRF framework helps in identifying not only the distribution of the hidden structures of each class, but also, by observing the learned weights, helps in understanding the shared structures between the classes.

Following the definition used in [14], the conditional probability distribution of labels $y \in y_1, y_2, \ldots, y_T$ given the observations $x = x_1, x_2, \ldots, x_T$ is modelled using the following latent conditional model

$$P(y \mid x, \theta) = \sum_h P(y, h \mid x, \theta) = \sum_h e^{\Psi(y, h, x; \theta)}$$

where $\theta$ are the parameters of the model and $\Psi(y, h, x; \theta) \in \mathbb{R}$ is a potential function parametrised by $\theta$. The hidden structures of the behaviours are captured using the hidden variables $h \in h_1, h_2, \ldots, h_T$ and are learnt during the training phase. The potential function $\Psi(y, h; x; \theta)$ given by Eq. (2) is used to define the feature functions that provide the compatibility between observations and hidden states $\phi(x_j) \cdot \theta(h_j)$, between hidden states and the category label $\theta(y, h_j)$, and between the pair of adjacent hidden states $\theta(y, h_j, h_k)$.

$$\Psi(y, h, x; \theta) = \sum_j \phi(x_j) \cdot \theta(h_j) + \sum_j \theta(y, h_j) + \sum_{(j,k) \in E} \theta(y, h_j, h_k)$$

The input observations are mapped into feature vectors by using a set of feature functions $\phi(x_j)$, which will capture the domain specific properties and need to be coded manually. The effectiveness of the trained model relies largely on the choice of the feature functions. The model is trained using the feature vectors by optimising the objective function

$$L(\theta) = \sum_i \log P(y_i \mid x_i, \theta) - \frac{1}{2\sigma^2} ||\theta||^2$$

The first term in Eq. (3) is the log-likelihood of the data and the second term is a log of the Gaussian prior with variance $\sigma^2$.

Given new test data $x$ and the learned parameters $\theta^*$ from the training, the predicted label for the test data is given by

$$\arg\max_{y \in \mathcal{Y}} P(y \mid x; \theta^*)$$
IV. HIGH- AND LOW-LEVEL FEATURES

Head pose orientations are a highly discriminating non-verbal cue that has been successfully employed in the automatic analysis of engagement [7], [8]. In this current approach, we evaluate the suitability of using head pose orientations for predicting a child’s engagement in a social interaction. We use the MMDB dataset for our experiments.

A. Multimodal Dyadic Behaviour Dataset

In this dataset, an examiner engages a child in five structured play activities or stages. The stages are: greeting the child by saying hello (Greeting), rolling a ball back and forth (Ball), looking through pictures in a book (Book), placing the book on your head to pretend it is a hat (Hat), and gentle tickling (Tickle). Each activity is designed to elicit various behaviours from the child, including common social-communicative behaviours observed in toddlers. In addition, for each stage, the examiner rates how easy or difficult it was to engage the child in the activity, as follows: 0 = Easy to Engage, 1 = Requires Some Effort to Engage, and 2 = Requires Extensive Effort to Engage. With the initial release of the MMDB dataset, detailed frame-level ground truth annotations of individual child behaviours were available for 59 sessions and were used in the current experiments. The MMDB dataset contains multiple camera views of the child and examiner in every session. Sample screenshots of a play activity across the five stages are shown in Figure 1.

The distribution of the sessions across the three engagement levels for the five stages are given in Table I. The engagement level distribution is biased (> 75%) towards Easy to Engage for all stages except for the Book stage. Hence, the robustness of the computational model can be validated most effectively for this stage and all our experiments were done only on the Book stage. In order to have a balanced dataset, the labels 1 (Requires Some Effort) and 2 (Requires Extensive Effort) are combined to form a single label, resulting in a binary classification problem.

B. Challenges in Automatic Head Pose Extraction

We use a state of the art head pose tracker, the IntraFace library [15], for head pose detection and the MMDB dataset for our experiments. The MMDB includes a high quality (720p resolution) recording from a camera facing the child. The tracker is initialsed manually with the child’s face location in the start frame and also in an undetected frame, if the tracker failed to detect the face continuously for a 1s period. In spite of the high resolution videos and manual guidance, 27% of the videos had a frame level head pose detection success rates of less than 60%. A summary of the frame level head pose detections for all videos is shown in Figure 2. An analysis of the videos with failed detections revealed some of the specific challenges encountered in analysing child interactions. The video recordings were captured in a controlled environment, but many specific child actions that clearly affected the tracker’s performance were observed, such as turning back and climbing up on the caregiver, throwing objects to the side, etc. (see Table II). These behaviours are responsible for additional failures of the automatic head pose estimation, in addition to the usual challenges related to face detection and tracking. Hence, in order to use head pose features to build our model, we needed to devise a reliable scheme to fill in the missing head pose frames. In a dyadic social interaction setting, when one person dominates the interaction, it is highly probable that the main motion in the video reflects the behaviour of the dominant person. This is true in the case of the MMDB interaction sessions because the interaction is structured to elicit specific behaviours from the child. In such a situation, challenges in estimating high-level features such as head poses can be addressed by using low-level optical flow-based features.

![Frame Level Head Pose Estimation](image)

**Fig. 2. Frame Level Head Pose Estimation**

<table>
<thead>
<tr>
<th>No.</th>
<th>Behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Child looking down</td>
</tr>
<tr>
<td>2</td>
<td>Child turning towards other parts of the room</td>
</tr>
<tr>
<td>3</td>
<td>Child climbing up on the mother and looking at her</td>
</tr>
<tr>
<td>4</td>
<td>Child walking away and coming back after some time</td>
</tr>
<tr>
<td>5</td>
<td>Child’s head hair occluding the face</td>
</tr>
<tr>
<td>6</td>
<td>Child’s hand occluding the face</td>
</tr>
<tr>
<td>7</td>
<td>Child looking down at the book</td>
</tr>
<tr>
<td>8</td>
<td>Child’s drinking from a glass occluding the face</td>
</tr>
</tbody>
</table>

**TABLE II**

HEAD POSE DETECTION CHALLENGES IN A CHILD INTERACTION

![Engagement Level Distribution](image)

**Fig. 1. A play activity in the MMDB dataset.**

![Engagement Level Distribution of 59 Annotated Sessions](image)

**TABLE I**

<table>
<thead>
<tr>
<th>Engagement Level</th>
<th>Greeting</th>
<th>Ball</th>
<th>Book</th>
<th>Hat</th>
<th>Tickle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy to Engage</td>
<td>45</td>
<td>51</td>
<td>37</td>
<td>57</td>
<td>47</td>
</tr>
<tr>
<td>Requires Some Effort</td>
<td>11</td>
<td>5</td>
<td>16</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Requires Extensive Effort</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Easy to Engage (%)</td>
<td>76.2%</td>
<td>86.4%</td>
<td>62.7%</td>
<td>96.6%</td>
<td>79.6%</td>
</tr>
</tbody>
</table>

**Engagement Level Distribution of 59 Annotated Sessions**
involving ground truth annotations released as part of the MMDB dataset, manually guided head pose estimations, and optical flow based features to evaluate the feasibility of using the low-level features.

C. Engagement Prediction using Ground Truth Annotations

The MMDB dataset contains frame level annotations of specific child behaviours, including vocalisations, gaze targets and play acts (e.g. rolling a ball to the adult). The interaction consists of five stages and, for each stage, the adult interacting with the child rates how easy or difficult it was to engage the child in the activity. These engagement ratings are available as part of the dataset. In the current analysis, the available frame-level annotations of relevant child behaviours are used to develop a learning model to predict the engagement ratings. Table III lists all the available relevant non-verbal behaviours annotated for this stage.

<table>
<thead>
<tr>
<th>Item</th>
<th>Behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Child looking at the examiner</td>
</tr>
<tr>
<td>2</td>
<td>Child looking at the book</td>
</tr>
<tr>
<td>3</td>
<td>Child looking at the parent</td>
</tr>
<tr>
<td>4</td>
<td>Child looking elsewhere</td>
</tr>
<tr>
<td>5</td>
<td>Child pointing at the book</td>
</tr>
<tr>
<td>6</td>
<td>Child tapping the table</td>
</tr>
<tr>
<td>7</td>
<td>Child turning the book pages</td>
</tr>
<tr>
<td>8</td>
<td>Child closing the book</td>
</tr>
</tbody>
</table>

TABLE III
NON-VERBAL FRAME LEVEL ANNOTATIONS FOR A BOOK STAGE

We parsed the frame level annotations to aggregate the duration of each behaviour w.r.t. the stage duration. To form a baseline for engagement prediction using these annotations, the normalised duration of each behaviour w.r.t. the book stage is fused with all the behaviour durations to form a feature vector. An SVM based binary classifier is trained using these features and LOOCV experiments are conducted using the ground truth engagement scores. The average engagement classification accuracy obtained using all behaviours listed in Table III was 79.7%. The annotated non-verbal behaviours can be broadly grouped into Gaze and Gesture categories. In order to separately study the impact of gaze on overall engagement prediction, we re-ran the experiments using only the gaze related behaviours (i.e. behaviours 1–4 in Table III). The accuracy for a gaze-only model was 71.2%. This indicates a strong discriminatory power of the gaze behaviours in classifying a child’s engagement level. The intent of this experiment is to estimate classification performance based on the available annotations and to study the discriminatory nature of the gaze related features towards developing a model for child engagement.

D. Engagement Prediction using Head Pose Features

The 3 degrees of freedom of the head pose – Pitch, Yaw and Roll – angles were obtained by tracking the child’s face using the IntraFace tracker library [15]. The manually guided successful head pose angles from the frames were used as the features for the detected frames. The estimated features of each frame were clustered along a temporal dimension to form a set of behaviour words. The behaviour words were the cluster centres, which represent the global aggregated head pose estimations for the video. An HCRF framework was employed to learn a model. Every node in the chain was represented by the behaviour word and used as an observation (x). The number of nodes in the HCRF framework was equal to the number of behaviour words. The head pose feature vector dimension used in the experiments was [3x1]. The model was trained using the ground truth engagement score labels (y) using a common set of experiment conditions described in Section VI. The classification accuracy obtained was 69.3%.

E. Engagement Prediction using Optical Flow Features

The motion around the child’s upper body region is analysed by computing Histogram of Optical Flow (HOF) features around the spatio-temporal interest points [16] in every frame. The child’s upper body region is detected using the method proposed in [17]. The HOF features are mapped to a visual vocabulary built using the HOF features of all frames. The visual word representing the maximum number of interest points is taken as the representative feature for a frame. A global clustering technique is employed to capture the motion flow dynamics of the video and the cluster centres obtained are referred to as behaviour words. In order to model the temporal dynamics, a HCRF framework is used, with each node taking the behaviour word as an observation value (x). The feature vector dimension used in the experiments is [90x1] and are used as the node features. The edge features were initialised to 1. The number of nodes in the HCRF framework is equal to the number of behaviour words. The labels of the model are the class label of the video, which is the same for all the nodes as defined in [14]. The model is trained for a fixed number of hidden states that correspond to the hidden structures of the behaviours. A common set of experiment conditions described in Section VI is used to train a model. The classification accuracy obtained was 70.3%.

F. Comparison of Engagement Prediction Performances

The average classification accuracy for child engagement obtained using baseline features consisting of gaze related annotations is 71.2%, while it is 69.3% for the guided head pose features and 70.3% for optical flow based features. These results were obtained using the experiment setup outlined in Section VI for the 9 hidden state configuration scenario. Figure 3 shows the comparison results.

The results show a similar performance in engagement prediction accuracy between head pose and optical flow based features. This confirms our hypothesis that the dominant motions in the video come from the child and reflect the child’s behaviour. In addition, the results of optical flow and head pose features are comparable to the baseline features consisting of gaze related ground truth annotations. This indicates that the features around a child’s head region have powerful discriminatory ability for predicting child
engagement. These results are consistent with findings from previous studies [7], [8].

In summary, in a dyadic social interaction setting, where the child’s behaviour is dominant, the low-level optical flow based features, can act as a good complementary feature for developing an engagement prediction model in the absence of reliable high-level features, such as head pose.

V. TWO-STAGE APPROACH FOR ENGAGEMENT PREDICTION

In this section, we will discuss the proposed two-stage approach for engagement prediction using low-level features. The two stages are

1) Stage 1 – the design of the low-level features for training the HCRF model. The obtained estimated marginal probability distribution of the hidden states for each node of a class are used as an input for the second stage.

2) Stage 2 – joint learning of the hidden structure distribution across the classes using SVM to predict the child’s engagement level.

A. Motivation

The challenges in reliably estimating high-level features from recordings of children prompted the design of low-level features. This resulted in optical flow based features around the upper body regions of the child to model the engagement level.

In the HCRF framework, the class label for a test video is predicted to be the label that corresponds to a minimum negative log-likelihood (NLL). In situations, where the chosen number of hidden states is not optimal, NLL values alone may not be strongly discriminatory. However, the estimated marginal probabilities from the inference procedure can provide additional discriminatory information that can be used to improve the classification performance. This leads to the proposed two-stage approach.

B. Algorithm for Predicting Engagement

In the first stage, a HCRF based model is built using optical flow based low-level features. The behaviour words computed from the optical flow features of each frame provide an aggregation of different motion flow patterns. The details of the learning procedure are the same as described in Section IV-E. The output of this stage is a set of estimated marginal probabilities of a hidden state for a class.

In the second stage, the dynamics of each hidden state across all classes are captured for each video. The sum of estimated marginal probabilities of each hidden state across all classes are computed. The distributions of these values for all hidden states are normalised to form a probability distribution. These are used as input features for the SVM model. We adopt a LOOCV approach and, hence, at every iteration, the training phase consists of 58 videos and a testing phase with one single video that was not present in the training set. At the end of each iteration, a new probability distribution of hidden states is computed for the training videos and used as the training set for the SVM classifier. In a similar way, the new probability distribution of hidden states for the test video is computed and was used to predict the class label from the trained SVM model. An overview of the proposed approach is shown in Figure 4.

VI. EXPERIMENTS & RESULTS

The feasibility of using low-level features and the effectiveness of the proposed two-stage approach are two broad categories of experiments conducted in this work. The experiments related to the comparison of low- and high-level features were discussed in Section IV-F.

A. Experiment Setup

The MMDB [4] is used for the experiments. LOOCV tests are performed on 59 videos and the mean binary classification accuracy is reported over all cross validation runs. The HCRF model was trained using [18], [19]. Experiments are conducted for various hidden state values and using $l2$–regularization with $\lambda = 1$. The number of behaviour words varied from 10 to 50. The clustering technique used is Elkan k-means with an $L1$–distance using the vlFeat library [20]. The clustering is done using two values of random seed initialisations and the average of them is used here to report the accuracy. The upper body configuration is obtained using [17], the STIP/HOF descriptors using [16] and the SVM model was trained using the LibSVM [21] library. An RBF kernel is used and the values for $c$ and $\gamma$ are empirically chosen as 0.005.

B. Results

1) Engagement Prediction using the Two-Stage Approach: The HCRF model trained using the optical flow and head pose features is used to compare the effectiveness of the two-stage approach. The average prediction accuracy with the two-stage approach was 74.4% for the 9-hidden state configuration using optical flow features and the best accuracy obtained was 79.7%. The second stage of augmented training using the hidden state probabilities added an average increase of 5.9% (absolute) to the overall average prediction accuracy. The graphs in Figure 5 show the results of engagement prediction accuracy and the effectiveness of the two-stage approach under two different experiment configurations. It
Stage 1 – Learning the hidden structures of a behaviour

Input Video
Detect child’s upper body region
Compute the STIP-HOF descriptors in the upper body region in each frame
Assign the STIP-HOF descriptors to visual words using a pre-built visual vocabulary

Stage 2 – Training a model using the learnt hidden structures

Compute hidden state marginal distribution across all the classes
Form a feature vector using the marginal for each video
Joint learning of the hidden state marginal using a SVM classifier
Predict the child’s engagement level

Estimated marginal probabilities of the hidden states for all the nodes of all the classes for train and test videos
Train a HCRF model with Behaviour Word as node observation
Form K clusters using feature vectors of all the frames. The cluster centre is the “Behaviour Word”
Select the visual word that contains the maximum STIP-HOF descriptors. This is the feature vector for a frame

Fig. 4. The proposed two-stage method for predicting a child’s engagement level

Fig. 5. Engagement prediction accuracy using only the HCRF model and with the proposed two stage approach (HCRF + SVM).

is worth noting that for both the optical flow and head pose features, the two-stage approach improved the overall prediction accuracy, highlighting its ability to capture more of the underlying relevant information. The current reported results in the literature are 76.7% in [11] and 73.3% in [4]. The exact subset of sessions included in those previous experiments was not reported; hence, it is not possible to directly compare these previous results to our proposed approach. In order to establish a common baseline for future work, a list of sessions used in our experiments is provided in the supplementary material.¹


2) Two-Stage Approach for Action Recognition in Videos: To further test the proposed method more widely and to show its generalisation capability, we conducted experiments for an
action recognition task using one of the benchmark datasets – J-HMDB [22]. The J-HMDB dataset has 21 action classes. As specified in the dataset, for each video, only a few hand-picked frames (15–50), in which the actor was obvious, and the first and last frame roughly correspond to the beginning and end of an action were selected, reducing to as low as 10% of the total frames per video. We hypothesise that these keyframes will tantamount to the number of behaviour words we obtained for the MMDB dataset.

The observations for each node are the features computed around the motion regions in the video. Local descriptors (HOG, HOF, MBH and Trajectory shape) are computed along the improved trajectories [23] that are detected for every successive 15 frames (span). The state of the art results (baseline) [24] were obtained by Fisher Vector encoding of these local descriptors, which is briefly described now. Firstly, PCA is applied on the local descriptors to reduce the dimensionality by a factor of two. Then, $10^6$ descriptors are selected at random from the training set to estimate a GMM with $K (= 256)$ Gaussians. Each video is then represented by a $2DK$-dimensional Fisher Vector, where $D$ is the dimension of the descriptors after PCA. Finally, power and $L2$ normalisation are applied. As the Fisher Vectors are very high-dimensional (100K), these cannot be directly used to represent an observation at each node. Instead, we use the Histograms of Occurrence of $K$ Gaussians to represent each observation.

The histograms are concatenated, yielding a 1280-length ($256 \times 5$) feature vector for every trajectory span. These feature vectors are further reduced to half their dimension by applying PCA. Thus, a video is represented by an $n \times 640$-length feature vector where $n$ is the number of spans (the improved trajectories detected). Each span is represented as a node in the HCRF chain. The HCRF model is trained using $L2$-regularisation with $\lambda = 1$ for different hidden state configurations and the results are shown in Table IV. The average accuracy values over 3-splits of train and test data (as specified in the dataset) were reported for different configurations of the hidden states.

The proposed two-stage (HCRF+SVM) approach improved the average overall accuracy by 3.3% (absolute). They were also found complementary to the state of the art Fisher Vector results reported in [24].

<table>
<thead>
<tr>
<th>Hidden States</th>
<th>HCRF (A)</th>
<th>A + SVM</th>
<th>A + FV</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>23.0%</td>
<td>28.0%</td>
<td>65.3%</td>
</tr>
<tr>
<td>7</td>
<td>30.8%</td>
<td>35.7%</td>
<td>65.5%</td>
</tr>
<tr>
<td>10</td>
<td>36.1%</td>
<td>39.0%</td>
<td>66.4%</td>
</tr>
<tr>
<td>12</td>
<td>37.5%</td>
<td>42.1%</td>
<td>66.1%</td>
</tr>
<tr>
<td>15</td>
<td>35.7%</td>
<td>37.0%</td>
<td>65.0%</td>
</tr>
</tbody>
</table>

| Jhuang et al. [22] (low level features) | 57.6% |
| Fisher Vector (FV) [24]              | 62.8% |

TABLE IV
PERFORMANCE OF THE PROPOSED APPROACH ON THE J-HMDB DATASET.

C. Discussion

If the number of hidden states is less than the optimum, there will be a large overlap of hidden structures between the classes. The per-class likelihood computed in this scenario will not be highly discriminatory. On the contrary, if the number of hidden states is higher than the optimal, it implies that there are additional finer hidden structures. This could lead to two situations: (i) there is a large overlap of finer hidden structures between the classes and/or (ii) hidden structures are well distributed between the classes. In (ii), the per-class likelihood will be discriminatory and the performance from the HCRF model will be optimal. When the number of hidden states is less than the optimum, in (i), the performance of the HCRF model is augmented with the proposed two-stage approach, by using the estimated marginal probabilities of the hidden states. The derived new probability distribution is used as an input feature to an SVM-based classifier. This joint learning of the hidden state class marginals offers additional discriminatory capabilities as compared to per-class likelihood values.

Figure 6 shows the results of the two-stage approach for different configuration of behaviour words. When the number of hidden states is not optimal, the performance of the HCRF-only model remains the same or declines, as shown in Fig. 6(a) and 6(b), and in these scenarios, the proposed two-stage method improved the overall performance. The performance of the two-stage approach was not significantly different when the hidden states were optimal as seen in Fig. 6(c) for the 7 hidden states scenario.

VII. CONCLUSIONS AND FUTURE WORK

Based on this study, we learnt that head pose orientation is strongly discriminatory in recognising a child’s engagement level. When a child dominates in a social interaction, it is highly probable that the motion cues reflect the child’s behaviour. A comparative study involving low-level optical flow features with manually guided head poses and ground truth annotations was carried out and the results support this hypothesis. The optical flow features were used in a HCRF framework to predict the child’s engagement level. We explored using learned hidden structures in an SVM framework to improve the classification performance in situations where the number of hidden states were not optimal. Our experiment results on the MMDB and J-HMDB datasets showed an average 3.3–5.9% improvement in the classification accuracy with the proposed two-stage (HCRF+SVM) approach. Overall, the combined use of low-level features in a HCRF+SVM framework has resulted in a state of the art performance in child engagement prediction in a social conversation but the low-level features are a lot easier to extract robustly than high-level features using current state of the art approaches. The impact of other non-verbal and verbal modalities towards engagement prediction will be studied in the future, while comparing them with the low-level features.

REFERENCES

Fig. 6. The effectiveness of the HCRF+SVM approach in different hidden state configurations


