Joint Alignment and Modeling of Correlated Behavior Streams

Liliana Lo Presti and Stan Sclaroff
Department of Computer Science,
Boston University, USA
loprest@bu.edu, sclaroff@bu.edu

Agata Rozga
School of Interactive Computing,
Georgia Institute of Technology, USA
agata@gatech.edu

Abstract

The Variable Time-Shift Hidden Markov Model (VTS-HMM) is proposed for learning and modeling pairs of correlated streams. Unlike previous coupled models for time series, the VTS-HMM accounts for varying time shifts between correlated events in pairs of streams having different properties. The VTS-HMM is learned on a set of pairs of unaligned streams and, thus, learning entails simultaneous estimation of the varying time shifts and of the parameters of the model. The formulation is demonstrated in the analysis of videos of dyadic social interactions between children and adults in the Multimodal Dyadic Behavior Dataset (MMDB). In dyadic social interactions, an agent starts an interaction with one or more "initiating behaviors" that elicit one or more "responding behaviors" from the partner within a temporal window. The proposed VTS-HMM explicitly accounts for varying time shifts between initiating and responding behaviors in these behavior streams. The experiments confirm that modeling of these varying time shifts in the VTS-HMM can yield improved estimation of the level of engagement of the child and adult and more accurate discrimination among complex activities.

1. Introduction

Social dyadic interactions encompass the set of reciprocal behaviors of two individuals interacting over time [1, 7]. The study of dyadic interactions is of interest in several domains. In social and cognitive sciences, there is interest in measuring and modeling human behaviors with the goal of automatically analyzing the individuals' social skills for the detection of developmental disorders [23, 19]. In fields such as human-computer interfaces, robotics, entertainment (games, advertising, etc.), understanding and analyzing dyadic interactions is of interest for designing humanlike robots/agents that can communicate in a natural way with the user [6]. Measuring and modeling of social interactions is also important for video indexing and retrieval in meeting analysis [25], as well as in surveillance [10].

During dyadic interactions, each participant tends to initiate or invite reciprocal responses from his peer by means of both verbal (speech/vocalization) and nonverbal (facial expression, eye gaze, body pose, gestures, etc.) behaviors. In this paper, we consider only body gestures, and we refer to the pairs of corresponding initiating and response behaviors within a temporal window as *reciprocal behaviors*. These behaviors may be either communicative gestures (pointing, nodding, waving, etc.) or action gestures (turning pages in a book, rolling a ball, etc.). Fig. 1 shows a simple example of reciprocal behaviors: the two agents are playing with a ball; while the first agent is throwing the ball, the other participant is facing the "thrower", waiting and then catching the ball. In this example, behaviors like the agent asking for the object or directing the eye gaze towards the thrower are considered communicative behaviors, while the act of catching the ball is an action gesture.

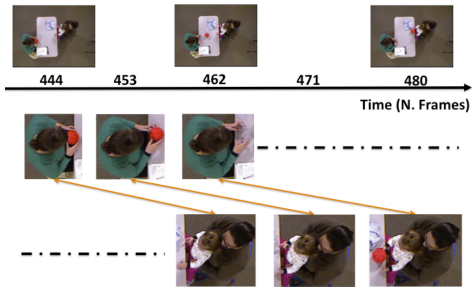

Figure 1. Top row: samples from a video showing a Ball Game. Second and third rows: samples from the two agents' streams. The arrows represent temporal associations discovered by our method.

In reciprocal behaviors, the response time of the agents, the duration of the behaviors, the intensity, rapidity and kind of gestures may vary both within the same pair of individuals and across pairs of individuals. All these issues make the interaction modeling challenging, as the agents' behaviors may be overlapping in time or delayed.

To address the above challenges, we formulate a Variable Time-Shift Hidden Markov Model (VTS-HMM). The VTS-HMM jointly models a pair of correlated streams of observations, where events in the two interacting streams are subjected to both variable time delays and variable durations. To demonstrate our formulation, we apply our model to the analysis of pairs of agents' behavior streams during

dyadic social interactions. Given a set of training videos, our method automatically estimates the temporal alignment of the agents' streams and the VTS-HMM parameters.

In the VTS-HMM formulation, learning the model for temporal alignment of the agents' behaviors includes both adjusting for variable reaction times and accounting for different durations of a behavior that occurs repeatedly.

The main contributions of this paper are:

- a method to model a dyadic social interaction as a mixture of reciprocal behaviors; in our framework, the agents' behaviors are jointly modeled in order to take into account the influence of one agent on the other during the interaction;
- a formulation for temporally aligning the agents' behavior streams based on the VTS-HMM;
- an unsupervised training procedure based on expectation-maximization that puts potentially coupled behaviors in correspondence while learning the parameters of the VTS-HMM.

We adopted the VTS-HMM for the analysis of videos of dyadic social interactions between children and adults in the Multimodal Dyadic Behavior Dataset (MMDB) [23]. These experiments confirm that taking into account the alignment of the correlated behaviors can help in discriminating among interactions and measuring the quality of an interaction, e.g., the level of engagement of the participants.

The plan of the paper is as follows. In Sec. 2, we present some related work. In Sec. 3 we present the probabilistic framework used to model the interaction considering the agents' behavioral reciprocity. In Secs. 4 and 5, we present the method used to infer the temporal alignment and the learning procedure respectively. In Sec. 6, we present our experimental results. Finally, in Sec. 7 we discuss conclusions and future work.

2. Related Work

There is extensive previous work on modeling and analyzing human behaviors during interactions. In works such as [25], the goal is detecting the focus of attention of the participants in a meeting, where eye gaze, speech and facial expressions play an important role. In contrast, we aim to model the reciprocal behaviors of the agents in dyadic interactions with a focus on the body motion.

Other works [10, 15, 24, 20, 16] employ pre-trained event models and some "grammar" to represent complex activities mainly for surveillance applications. Training specific event detectors requires a substantial amount of annotated data. In practice, it is not straightforward to know all possible kinds of events that will be necessary to model an interaction. As also demonstrated in [8, 22], it is possible to use low-level visual events directly to represent an activity.

Previous works focus on discovering the causalities of detected events. In [21], a data-driven approach inspired by Granger's causal analysis for time series is used to analyze causality in video sequences, while in [22] and in [11], the focus is on learning the structure of the causal graph of events detected during an activity. These works focus either on simple single-person activities or on complex activities involving several individuals; in the latter case, the activity is modeled without separating the behaviors of the interacting individuals and ignoring the temporal granularity and delay of the coupled behaviors.

In our work, we model the reciprocal behaviors in pairs of streams representing the agents' behaviors. First the agents are detected and tracked across time. Then, low-level visual features are extracted at each frame. To account for the variable time shifts, we temporally align the agents' streams so as to maximize the co-occurrences of spatio-temporal features in the two time series. In contrast to previous methods, we explicitly model the time delays in the agents' behaviors. During training, the dynamics of the reciprocal behaviors and the temporal warping of the agents' streams is learned jointly.

A closely related model is the Coupled Hidden Markov Model (CHMM). In [4], the CHMM is used to model the movements of two hands: each of the hands is modeled by an HMM, and the states of the HMMs are coupled to model the dependence of the two hands' movements. In [18], the CHMM is used for visual-audio modeling. The method takes unaligned streams as input, but (crucially) does not account for varying time shifts. Indeed, in the CHMM, observations from the two streams are assumed to be conditionally independent; our proposed VTS-HMM formulation does not make this assumption and models them jointly.

We note also that our temporal streams alignment method is related to behavior-based video alignments [9] where two videos taken with different cameras but depicting the same behavior are aligned based on visual similarities. However, in our problem, the behaviors we aim to model are not necessarily similar (as performed by different agents), but probably are correlated. Therefore no visual similarities may be used to account for the sequence alignment. Instead, we rely on the co-occurrences of the visual features in the agents' streams, and we use a maximum likelihood approach to temporally align the streams based on the VTS-HMM. In this sense, our method is related to several recent works that use Dynamic Time Warping (DTW) [26, 27]. In [26], DTW and mutual information are used to align pairs of segmented actions of the same type. In contrast, we do not focus on single actions but on complex activities, and we model the dynamics underlying the interaction. As the two streams are aligned based on the interaction model, then we can state that any pair of agents' streams is aligned to a common reference; thus the streams are made comparable through the interaction model.

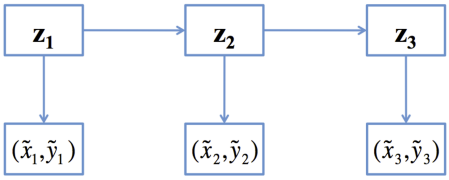

Figure 2. Dynamical model for a social interaction: z is the hidden state, $(\tilde{x}, \tilde{y})$ are the coupled behaviors in the agents' streams.

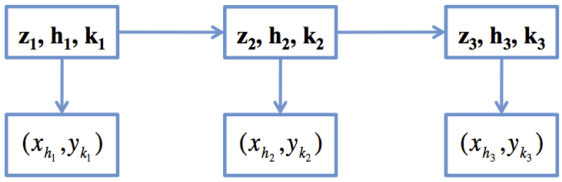

Figure 3. Augmented dynamical model for a social interaction: z is the hidden variable, x_h and y_k are the coupled behaviors detected at time h and k in the agents' streams.

3. Modeling Social Interaction

We assume that two time series of bags of low-level visual events are given to describe the behaviors of each agent. Given a set of similar dyadic interactions, if the reciprocal behaviors of the agents would be perfectly aligned, we would expect that the low-level visual events of the agents' reciprocal behaviors would be maximally associated. However, we do not observe the aligned behaviors directly, but their shifted and warped versions. To account for different velocities and response times, we align the two agents' streams while maximizing the co-occurrences of features in the two time series. The problem of estimating the best alignment entails finding the time-shifted instants in the two agents' streams when correlated bags of spatio-temporal features are observed.

Assuming that the emerging agents' behaviors are reciprocal and may switch between different modes, a natural choice to model an interaction is adopting a mixture model. For aligning the behaviors streams, we propose the Variable Time-Shift Hidden Markov Model (VTS-HMM). Our formulation is more flexible than standard HMM in that it embeds the latent warping of the behavioral streams and accounts for the different time delays that can be observed during the interaction. Our method differs from DTW in that it enables the alignment of streams that are correlated but not necessarily similar. Finally, our formulation is different than the stochastic DTW in [17] in that, by means of the latent variables, it takes advantage of a switching mechanism, which permits us to choose the best model to use for aligning the behavioral streams. In this way, our method accounts for the significant variations that can be observed in the agents' behaviors.

3.1. Interaction Model

In our framework, the behaviors of agent A_1 are described as the stream $X = \{x_1, x_2, ..., x_N\}$, where x_n is a bag of low-level visual events detected at time n. A common representation for a bag of low-level visual events is a frequency histogram of visual words; therefore, $x_n(\alpha)$ will represent the number of occurrences that the α-th visual word is detected at time n for agent A_1. The behavior of agent A_2 is described as the stream Y in a similar way.

In our model, the hidden state z_n represents the evolution of the coupled behaviors $(\tilde{x}_n, \tilde{y}_n)$ during the dyadic interaction across time, while the emission probability models the coupled behaviors of the agents (see Fig. 2).

At each step n, the agents' behaviors are represented as the corresponding bag of visual words in the aligned streams, while the coupled behaviors are represented as the joint co-occurrences of words in the agents' bag of features representation, i.e., a joint histogram of visual words.

Given the aligned streams, the joint probability for our model is:

$$p(\tilde{X}, \tilde{Y}, Z) = \pi(z_1) \cdot \prod_{i=1}^{N} p(\tilde{x}_i, \tilde{y}_i | z_i) \prod_{j=2}^{N} p(z_j | z_{j-1})$$

where $p(\tilde{x}_i, \tilde{y}_i | z_i)$ is the probability of having the joint observation $(\tilde{x}_i, \tilde{y}_i)$ given the state z_i, while $p(z_j | z_{j-1})$ represent the transition model for the social interaction.

As we do not observe the aligned behaviors $\tilde{X}$ and $\tilde{Y}$, but their shifted and warped versions X and Y, we align the agents' reciprocal behaviors by maximizing the likelihood of the streams given the interaction model. Therefore, we modify the interaction model as shown in Fig. 3. In this model, the state has been augmented with the time instants h and k that should be coupled in the streams X and Y. $H = \{h_i\}$ and $K = \{k_i\}$ represent the temporal alignment of the agents' streams, while $Z = \{z_i\}$ represents the sequence of hidden states.

The joint probability for this model is:

$$p(X, Y, Z, H, K) = \pi_z(z_1) \cdot \pi_{h,k}(h_1, k_1) \cdot \prod_{i=1}^{N} p(x_{h_i}, y_{k_i} | z_i, h_i, k_i) \prod_{j=2}^{N} p(z_j | z_{j-1}) \cdot p(h_j, k_j | h_{j-1}, k_{j-1})$$

where $p(x_{h_i}, y_{k_i} | z_i, h_i, k_i)$ is the probability of having the joint observation (x_{h_i}, y_{k_i}) given the state z_i, the h_i-th observation in stream X, and the k_i-th observation in stream Y. Therefore (x_{h_i}, y_{k_i}) is the same as $(\tilde{x}_i, \tilde{y}_i)$.

The conditional probability $p(z_j | z_{j-1})$ represents the transition model for the social interaction, while $p(h_j, k_j | h_{j-1}, k_{j-1})$ represents the temporal dynamics in the stream warping.

3.2. Probabilistic Model

In our implementation, the state transition probability $p(z_n | z_{n-1})$, and the priors π_z and $\pi_{h,k}$ are modeled as multinomial distributions.

We consider that the pair (h, k) is reachable only from a subset of possible states defined as:

$$J(h,k) = \{(h-\alpha, k-\beta)\}_{\alpha=0,\beta=0}^{\alpha=u,\beta=v} - \{(h,k)\} \quad (1)$$

where u and v are the maximal number of observations that may be jumped in each stream. The temporal dynamics are modeled as a uniform distribution:

$$p(h_n, k_n | h_{n-1}, k_{n-1}) = \begin{cases} c, & \text{if}(h_{n-1}, k_{n-1}) \in J(h_n, k_n) \\ 0, & \text{otherwise} \end{cases}$$

where the sum of c over all the possible pairs (h_n, k_n) is 1.

The conditional distribution of the coupled behaviors (x_h, y_k) given the state z is modeled as a joint multinomial distribution. For each value of z, the parameters of this distribution have the form of a matrix ϕ^z whose rows correspond to visual words in the stream X, while the columns correspond to visual words in the stream Y, that is:

$$p(x_h, y_k | z, h, k) = \gamma \prod_{i,j} \phi^z(i,j)^{(x_h(i) \cdot y_k(j))}$$

where γ is the normalization constant in a multinomial distribution, and $\phi^z(i,j)$ is the probability of observing the pair of words (i,j) in the streams X and Y while in state z.

4. Inference of the Temporal Alignment

The streams' temporal alignment is found by inferring the hidden states Z, H, and K. We adopt a maximum likelihood approach and use dynamic programming to maximize the log-likelihood of the alignment.

For each pair of bags (x_h, y_k) in a temporal window, given the state z, the logarithm of the emission probability $M(x_h, y_k | z)$ is defined as:

$$M(x_h, y_k | z) = \log(p(x_h, y_k | z, h, k, \phi^z)). \quad (2)$$

As the probability $p(h_n, k_n | h_{n-1}, k_{n-1})$ is uniform, it may be omitted during inference. For each pair of bags (x_{h_n}, y_{k_n}) in a temporal window, and for each value of the hidden variable z_n, we compute the best hypothesis that could have generated the current pair as:

$$S(x_{h_n}, y_{k_n}, z_n) = M(x_{h_n}, y_{k_n} | z_n) + \max_{\substack{z_{n-1} \\ (h_{n-1}, k_{n-1}) \in \\ J(x_{h_n}, y_{k_n})}} \{S(x_{h_{n-1}}, y_{k_{n-1}}, z_{n-1}) + \log(p(z_n | z_{n-1}))\}. \quad (3)$$

The variable S is initialized considering the prior on the hidden states; therefore:

$$S(x_{h_1}, y_{k_1}, z_1) = M(x_{h_1}, y_{k_1} | z_1) + \log(\pi(z_1)). \quad (4)$$

The best alignment H, K, and the set of hidden variables Z are computed by back-tracking, once the end of the stream has been reached. Algs. 1 and 2 give the pseudo-code for inferring the temporal alignment.

For each video sequence, we consider an initial time delay τ necessary to define the alignment starting time. The meaning of this delay is that one of the two agents' streams (depending on the sign of the delay) must be shifted back in time to align the first reciprocal behavior. Therefore τ represents the value of either h_1 or k_1.

Algorithm 1: Inference of the Temporal Alignment

Input : X and Y, agents' streams;
ϕ, π_z and A_z, parameters of the model
Output: H, K temporal alignment; Z hidden states;
τ initial time delay
for $\tau \leftarrow \tau_{min}$ **to** τ_{max} **do**
  if $\tau > 0$ **then**
    $X_s \leftarrow$ (shift X by τ frames); $Y_s \leftarrow$ Y
  else
    $X_s \leftarrow$ X; $Y_s \leftarrow$ (shift Y by $-\tau$ frames)
  (H_τ, K_τ, Z_τ, S_τ) $\leftarrow$ `Align`$(X_s, Y_s, \phi, \pi_z, A_z)$;
$\tau \leftarrow argmax$(S)
H $\leftarrow H_\tau$; K $\leftarrow K_\tau$; Z $\leftarrow Z_\tau$;

Algorithm 2: Align

Input : X and Y, agents' streams;
ϕ, π_z and A_z, parameters of the model
Output: H, K temporal alignment; Z hidden states;
L likelihood for the alignment
for $h \leftarrow 1$ **to** $\#(X)$ **do**
  for $k \leftarrow 1$ **to** $\#(Y)$ **do**
    for $z \leftarrow 1$ **to** Z_M **do**
      M$(h,k,z) \leftarrow \log(p(x_h, y_k | z, h, k, \phi^z))$;
      S$(h,k,z) \leftarrow -\inf$;
for $h \leftarrow 1$ **to** $\#(X)$ **do**
  for $k \leftarrow 1$ **to** $\#(Y)$ **do**
    for $z_n \leftarrow 1$ **to** Z_M **do**
      if $h = 1$ & $k = 1$ **then**
        S$(1,1,z) \leftarrow$M$(1,1,z) + \log(\pi_z(z))$;
      else
        Compute J as in Eq. 1;
        Compute S(h,k,z_n) as in Eq. 3;
        Store best hypothesis for (h,k,z_n);
Compute best align. (H,K) and Z by back-tracking;
set L to likelihood of the best alignment; L$\leftarrow$ L / #(H);

In our implementation, the time delay assumes values in the range $[\tau_{min}, \tau_{Max}]$. During inference, each of these time delays is tested, and the time delay providing the highest probability is selected.

5. Parameter Estimation

During training, we learn the parameters ϕ^z for the emission probability distribution corresponding to each of the state value z, and the parameters for π_z and for $p(z_n | z_{n-1})$. We define the parameters for $p(z_n | z_{n-1})$ as A_z.

The learning of the parameters for the model in Fig. 3, in which H, K and Z are dependent given the observations, may be achieved considering the Cartesian product HMM where the state space is represented as all the possible com-

binations of H, K and Z. However, this has high time complexity. Instead, we utilize an approximate learning procedure that considers the model in Fig. 2, and we adopt an expectation-maximization (EM) based approach. At each iteration, we infer the temporal alignment of the training sequences as described in Sec. 4 with the given parameter set. Then we treat H and K as given and estimate the parameters of our model on the set of aligned agents' streams via EM. The procedure is repeated until convergence.

In the EM, during the E-step the expected value for the log-likelihood is computed given the current parameter estimate; during the M-step, the parameters are re-estimated by maximizing the expected log-likelihood. Therefore, we re-estimate the parameters π_z, A_z and $\{\phi^z\}_z$. The key difference in the parameter estimation with respect to a traditional HMM is in the multinomial distribution parameters. For each state value z, the corresponding parameters ϕ^z are computed by normalizing the expected value of the co-occurrences of pairs of words (i, j) in the two aligned streams given the time delay h, k and z as follows:

$$\phi^z(i,j) \propto \sum_v \sum_{\tau_v} p(\tau_v|\tilde{X}^v, \tilde{Y}^v) \cdot \sum_n (\#(i,j)^{n,v} \cdot p(z_n = z|\tilde{X}^v, \tilde{Y}^v))$$

where the superscript v is an index over the samples in the training set, n is an index over the observations in the v-th sequence, and $\#(i, j)$ is the count for the pair of words (i, j). This modification comes from the assumption that τ is conditioned only on the aligned bag of visual words and implies that $p(\tau_v, Z|\tilde{X}, \tilde{Y})$ factorizes. The probability $p(\tau_v|\tilde{X}^v, \tilde{Y}^v)$ is assumed to be proportional to the likelihood of the alignment provided by the inference procedure.

The learning procedure is summarized in Algorithm 3. For each pair of agents' streams and for each initial time delay τ, we infer the temporal alignment and compute the pair of aligned agents' streams $\tilde{X}^v$ and $\tilde{Y}^v$. We then weight the joint representation of the streams with the probability $p(\tau_v|\tilde{X}^v, \tilde{Y}^v)$. The effect of this procedure is that of generating a training set (X_{train}, Y_{train}) of possible aligned sequences weighted based on the likelihood of the alignment itself. Finally, we note that in [2], the time-shift is assumed to be constant over time. The inference of this time-shift is performed by testing all the possible time-shifts and selecting the one with the highest probability. In our formulation, the time delay is not constant over time; during inference, the initial time delay is computed in a similar way to [2]. However, during learning, we define a distribution over the initial time delay to make the method robust to the initial choice of parameters.

Algorithm 3: Learning parameters of the model

Input : X^v, Y^v, training set of V videos;
Z_M, number of states;
τ_{min} and τ_{max}, interval for τ
Output: ϕ^z, π_z, A_z, parameters of the model
Initialize ϕ^z, π_z and A_z randomly;
converged $\leftarrow$ false, iter $\leftarrow 0$, $\mathrm{LogL}_p \leftarrow -\inf$;
while *iter<MaxIter & !converged* **do**
  `// Align training set:`
  $\mathrm{X}_{train} \leftarrow \varnothing$; $\mathrm{Y}_{train} \leftarrow \varnothing$; **for** $i \leftarrow 1$ **to** V **do**
    for $\tau \leftarrow \tau_{min}$ **to** τ_{max} **do**
      Shift X^v or Y^v based on τ;
      $(H, K, Z, S_\tau) \leftarrow$
      $\mathtt{Align}(X^v, Y^v, \phi, \pi_z, A_z)$;
      Compute $\tilde{X}^v_\tau$ and $\tilde{Y}^v_\tau$ by (H, K);
    for $\tau \leftarrow \tau_{min}$ **to** τ_{max} **do**
      $S_\tau = \frac{\exp(S_\tau)}{||S||}$;
      Weight observations in $(\tilde{X}^v_\tau, \tilde{Y}^v_\tau)$ by S_τ;
      $\mathrm{X}_{train} \leftarrow \mathrm{X}_{train} \cup \tilde{X}^v_\tau$;
      $\mathrm{Y}_{train} \leftarrow \mathrm{Y}_{train} \cup \tilde{Y}^v_\tau$;
  `// E-step over [`X_{train}`, `Y_{train}`]:`
  Compute exp. log-likel. by fwd-bwd propagation;
  `// M-step:`
  Compute ϕ^z, π_z, A_z by maxim. the exp. log-likel.;
  converged $\leftarrow$ checkConv(LogL, LogL_p);
  iter $\leftarrow$ iter + 1; $\mathrm{LogL}_p \leftarrow$ LogL;

6. Experimental Results

Experiments were conducted using videos from the sessions in the MMDB dataset[1] [23]. Each session follows the same semi-structured protocol where an adult (the examiner) and a child are involved in a series of four games: Ball, Book, Hat, and Tickle. In the Ball game, the examiner initiates a game rolling the ball back-and-forth with the child and then pauses the game to gauge the child's reaction to the break in interaction. In the Book game, the examiner brings out a picture book and encourages the child to flip through the pages. In the Hat game, the examiner puts the book on her head and watches for the child's reaction. Finally, in the Tickle game, the examiner leans in to gently tickle the child several times before pausing the game to see whether the child attempts to re-initiate the interaction.

For each session, the adult interacting with the child assigned a summary rating of the child's level of engagement in each game on a three-point scale ranging from 0 (indicating the interaction with the child required little effort for the adult and/or the child was ready and eager to engage) to 2 (the interaction with the child required extensive effort and/or the child was highly fussy or refused to interact).

In our experiments, we test the hypothesis that modeling the agents' reciprocity may help to predict the engagement level of the child during the interaction. We also evaluate our model's ability to classify different types of games, and we test if taking into account the behavioral reciprocity when modeling the interaction improves activity recogni-

[1]The dataset is publicly available at http://www.cbi.gatech.edu/mmdb/

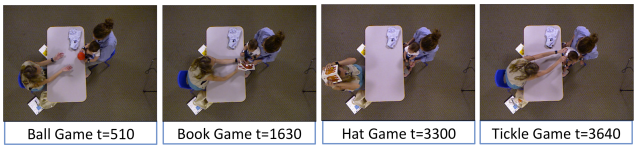

Figure 4. Sample images of the games in a session of the MMDB from the overhead Kinect camera.

tion. Finally, we show qualitative results of our method for the temporal alignment of the agents' behavioral streams.

We use the video captured via an overhead Kinect camera, and use the depth image to detect the agents' bodies. We manually segment the video sequences into clips, so that each clip represents one of the Ball, Book, Hat or Tickle games. Sample images of a video in the data set are shown in Fig. 4. We test our approach on a dataset of 66 sequences. We randomly select 6 sequences to train a codebook. Each session has been manually segmented into 4 clips (one for each task). 30 sequences where the child was scored as engaged in all the tasks are used to train our models. The remaining 30 sequences are used for testing. However, in one test session the child refused to play the Hat and Tickle games, in another the child didn't play the Tickle game. Thus, in total, we have tested on a dataset of 117 clips. Fig. 5 shows an histogram of the engagement scores for the test sequences.

6.1. Implementation details

Agents' Behavioral Representation: As in previous works [21, 8], we adopt spatio-temporal interest points (STIP) [12] for extracting visual features. With respect to other sparse feature representations (SIFT [14], SURF [3]), STIP offers the advantage of considering interest points with a temporal duration. We adopt standard techniques to encode STIP features as visual words. A codebook of visual words on the detected STIP points (represented as HOG [5] and HOF [5]) is learnt by K-means; then, hard coding has been used to associate each STIP feature to a visual word.

Each agent's body is found by background suppression and tracking. The above representation is invariant to the position of the agents as the position of the detected STIP points is not encoded when computing the histograms.

For each agent and for each frame we compute a histogram of the visual words. In our implementation, we used a codebook of 50 visual words. To account for frames where no STIP has been detected, we augment the codebook with a further word representing no motion.

Inference: We set u and v to 1; therefore, $J(h,k)$ in Eq. 1 is composed of only three possible steps. We have limited the interval of time in which (h,k) can vary to 61 frames. Therefore, the matrix M in Eq. 2 is diagonally banded. This makes the implementation faster as it does not require matching all the possible bags in the agents' streams but only the ones in a temporal neighborhood.

In our experiments, we found benefits in normalizing the likelihood in Eq. 2 by the size of the bags of visual words. This corresponds to measuring a geometric average of the emission probability and accounts for bags of features of different sizes across time.

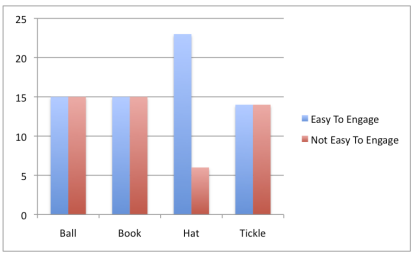

Figure 5. Binary engagement scores in the test set for each game. The vertical axis reports the number of sequences.

The initial time delay (expressed as number of frames) was allowed to vary in the interval $[-25, 0]$ with step 5.

Training: We use the Bayesian Information Criterion (BIC) to select an appropriate number of states for the models ranging from 6 to 10. Based on BIC, we set the number of states to 9, 8, 8 and 9 respectively for the Ball, Book, Hat and Tickle game models.

Baseline Method: To evaluate the effect of accounting for the varying time-shifts, we test against an HMM that jointly models the states of the two streams (implemented as a Cartesian HMM) but without varying time delay. This baseline is equivalent to the CHMM in [18]. To permit fair comparison, this baseline employs the same observation model used in the VTS-HMM. Employing the BIC, the number of HMM states in the baseline is set to 10, 7, 10 and 6 respectively for the Ball, Book, Hat and Tickle game.

6.2. Results

Engagement Estimation Reciprocity is important during interactions with children involved in collaborative tasks [19]. Gestures such as pointing or nodding may indicate the presence of social behaviors, and can permit inference of whether children are paying attention to their partner, and whether they are responding or initiating social interactions. A child who displays social reciprocity is motivated to engage in social interactions with others and participates in long chains of back-and-forth interactions.

As discussed in [13], children with autism display significant impairments in social reciprocity and it is more difficult for them to engage in responsive interactions. Therefore, it is of interest to assign ratings to the level of engagement to help the diagnosis and treatment of developmental and behavioral disorders [23].

It seems natural to hypothesize that modeling the reciprocity may help in predicting the engagement score of the agents during a dyadic interaction. Due to the limited number of samples in our dataset whose engagement score is 2, we considered the class "not easy to engage" that comprises samples with an engagement score of either 2 or 1. Thus, we cast the engagement prediction problem as a binary classification problem, where a predicted score of 0 means "easy

	Ball	Book	Hat	Tickle	Avr.
STIPS, no Align	50	53.33	41.38	53.57	49.57
STIPS + Align	60	56.67	51.72	53.57	55.49
HMM	56.67	70	**58.62**	53.57	59.71
VTS-HMM	**73.33**	**76.67**	55.1	**64.29**	**67.36**

Table 1. Engagement Prediction Accuracy in Leave-1-Out Cross-Validation

to engage", while 1 means "not easy to engage".

We conduct experiments to test if the temporal alignment can improve engagement prediction. We process the test set with our VTS-HMM to infer the temporal alignment of each pair of agents' streams. We then extract features to represent the sequences, and infer the engagement score by a linear SVM. We evaluate the average classifier accuracy using Leave-One-Out cross-validation. We test two cases: feature representations of the temporally aligned agents' streams versus feature representations of the not aligned streams.

In the first experiment, we represent each pair of agents' streams by the joint histogram of STIP words. To reduce the dimensionality of the histograms, we apply PCA and select a number of components covering 95% of the variance (on average 23 components). The engagement prediction accuracies for each game with and without temporal alignment are shown in the first and second rows of Table 1. Overall, the temporal alignment improves the engagement prediction with an increase of the average accuracy of about 10%.

The STIP-based feature representation does not provide any information about the dynamics of the interaction, nor about specific behaviors that may be observed. As the VTS-HMM models an interaction as a mixture of reciprocal behaviors, we test the hypothesis that the distribution of hidden states inferred by our model for a given game is correlated with the engagement score. In this experiment, we represent a pair of agents' streams by the histogram of hidden states inferred by the VTS-HMM. For comparison purposes, we consider also the histogram of hidden states inferred using our baseline method.

The third and forth rows in Table 1 report the results of this experiment. We observe a significant improvement in prediction accuracies for each game. These results suggest that the interaction modes and dynamics learned by our model can reflect the level of engagement of the child. Our method outperforms the baseline for all the activities but the Hat Game. We believe this may be due to the fact that the score for this game is assigned also based on the child's facial expressions. As we are using the overhead camera, we cannot capture this kind of information and this could explain the drop in the accuracy.

Overall, the increase in the average accuracy of our model with respect to the baseline method is of about 13%. With respect to the STIP-based representation without any temporal alignment, we achieve an increase in the average accuracy of about 36%.

True vs Pred.	Ball	Book	Hat	Tickle
Ball	**83.33**	3.33	13.33	0
Book	13.33	**53.33**	33.33	0
Hat	17.24	0	**82.76**	0
Tickle	10.71	3.57	0	**85.71**

Table 2. Recognition with VTS-HMM. **Avr. Accuracy: 76.28%**.

True vs Pred.	Ball	Book	Hat	Tickle
Ball	**90**	0	3.33	6.67
Book	36.67	**40**	10	13.33
Hat	34.48	3.45	**55.17**	6.90
Tickle	10.71	0	0	**89.29**

Table 3. Recognition with HMM. **Avr. Accuracy: 68.61%**.

Activity Recognition In the next experiment, we used our model to classify video sequences representing the four types of games in the MMDB: Ball, Book, Hat and Tickle. Our goal is to demonstrate that, by taking into account the reciprocity of the agents' behaviors, our model is able to describe the interaction better. To classify each test video, we used the log-likelihood of the alignment normalized by the length of the sequence.

Tables 2 and 3 report the confusion matrices for our method and the baseline. As the tables show, there is a clear advantage in considering the temporal alignment when classifying these complex activities. The baseline method provides an average accuracy of 68.61%. Our method provides an average accuracy of 76.28%. Therefore, modeling the variable time-shifts in the reciprocal behaviors leads to an increase of the average classification accuracy of about 10% with respect to the baseline method.

Temporal Alignment: Qualitative Results Figures 1 and 6 show qualitatively some representative temporal alignments obtained by our method. In all the figures, the first line represents samples from the video during the Ball and Tickle games respectively. The second and third lines show manually cropped images of the two agents.

During the Ball game (Fig. 1), the child follows the ball with her eyes (moving her head) and prepares to receive the ball (middle frame in the third line). Our method automatically couples the frames when the adult is going to throw the ball with the frames when the child is preparing to receive the ball. The estimated time delay for this specific example is of -18 frames.

In the Tickle game (Fig. 6), the child is following the hands of the adult (by adjusting her eye gaze and moving her head) and she is moving her body in expectation of the adult's tickle. While the adult's hands approach the child, the time delay decreases from -25 to -8.

7. Conclusions and Future Work

In this paper, we present the VTS-HMM. Our formulation models the underlying process that generates pairs of correlated (but not necessarily similar) streams affected by varying time delays. The model is learned on a set of pairs

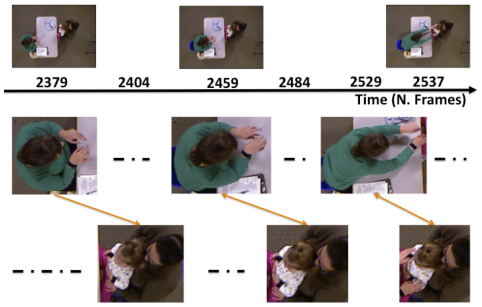

Figure 6. Alignment for the Tickle Game; the estimated delay varies from -25 to -8.

of unaligned streams (rather than a single pair of streams). Learning requires simultaneous estimation of the varying shifts and of the parameters of the model.

We applied our model to dyadic interactions with the goal of finding temporal associations among low-level visual events describing the agents' behaviors. In our experiments, we observed that the temporal alignment of the agents' behaviors improves the prediction of the engagement score of the participants in the interaction, as well the classification accuracy for activity recognition.

In future work, we will model and infer the durations of the coupled behaviors in order to automatically segment them. We will also extend our method to incorporate other sources of data and information, such as speech, eye gaze and facial gestures, to better understand the coordination and reciprocity of the agents' behaviors.

We also expect that the VTS-HMM should have broader applicability to problems where modeling the alignment of correlated streams, which are affected by varying delays, can help to improve classification or regression accuracy. Potential applications include visual-speech modeling, video alignment, modeling of correlated appearances and activities within camera networks, event/gesture modeling and retrieval, etc.

Acknowledgment: This work is supported in part by NSF grants 1029679 and 1029430.

References

[1] P. Allison and J. Liker. Analyzing sequential categorical data on dyadic interaction: A comment on Gottman. *Psychological Bulletin*, 91(2):393–403, 1982. 1

[2] M. Azzouzi and I. T. Nabney. Time delay estimation with Hidden Markov Models. 1999. 5

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. *ECCV*, 2006. 6

[4] M. Brand, N. Oliver, and A. Pentland. Coupled Hidden Markov Models for complex action recognition. *CVPR*, 1997. 2

[5] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. *ECCV*, 2006. 6

[6] Z. Duric, W. Gray, R. Heishman, F. Li, A. Rosenfeld, M. Schoelles, C. Schunn, and H. Wechsler. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proc. of the IEEE*, 2002. 1

[7] E. Ferrer, J. Steele, and F. Hsieh. Analyzing the dynamics of affective dyadic interactions using patterns of intra-and interindividual variability. *Multivariate Behavioral Research*, 47(1):136–171, 2012. 1

[8] U. Gaur, Y. Zhu, B. Song, and A. Roy-Chowdhury. A string of feature graphs model for recognition of complex activities in natural videos. *ICCV*, 2011. 2, 6

[9] I. Junejo, E. Dexter, I. Laptev, and P. Pérez. View-independent action recognition from temporal self-similarities. *PAMI*, 33(1):172–185, 2011. 2

[10] S. Kwak and et al. Scenario-based video event recognition by constraint flow. *CVPR*, 2011. 1, 2

[11] J. Kwon and K. Lee. A unified framework for event summarization and rare event detection. *CVPR 2012*. 2

[12] I. Laptev. On space-time interest points. *IJCV*, 64(2):107–123, 2005. 6

[13] D. Leach and M. LaRocque. Increasing social reciprocity in young children with autism. *Intervention in School and Clinic*, 46(3):150–156, 2011. 6

[14] D. Lowe. Object recognition from local scale-invariant features. *ICCV*, 1999. 6

[15] D. Minnen, I. Essa, and T. Starner. Expectation grammars: Leveraging high-level expectations for activity recognition. *CVPR*, 2003. 2

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

[17] S. Nakagawa and H. Nakanishi. Speaker-independent English consonant and Japanese word recognition by a stochastic dynamic time warping method. *J. of the Inst. of Electronics and Telecommunication Engineers*, 34(1):87–95, 1988. 3

[18] A. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy. A coupled HMM for audio-visual speech recognition. *ICASSP*, 2:II–2013, 2002. 2, 6

[19] L. Ogden. Collaborative tasks, collaborative children: An analysis of reciprocity during peer interaction at key stage 1. *British Educational Research J.*, 26(2):211–226, 2000. 1, 6

[20] M. Pei and et al. Parsing video events with goal inference and intent prediction. *ICCV*, 2011. 2

[21] K. Prabhakar, S. Oh, P. Wang, G. Abowd, and J. Rehg. Temporal causality for the analysis of visual events. *CVPR*, 2010. 2, 6

[22] K. Prabhakar and J. Rehg. Categorizing turn-taking interactions. *ECCV*, 2012. 2

[23] J. Rehg, G. Abowd, A. Rozga, and et al. Decoding children's social behavior. *CVPR*, 2013. 1, 2, 5, 6

[24] M. Ryoo and J. Aggarwal. Recognition of composite human activities through context-free grammar based representation. *CVPR*, 2006. 2

[25] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *Trans. on Neural Networks*, 13(4):928–938, 2002. 1, 2

[26] M. Yamada, L. Sigal, M. Raptis, and M. Sugiyama. Dependence maximizing temporal alignment via squared-loss mutual information. *preprint arXiv:1206.4116*, 2012. 2

[27] F. Zhou and F. De la Torre. Canonical time warping for alignment of human behavior. *NIPS*, 2009. 2