Unsupervised Synchrony Discovery in Human Interaction

Wen-Sheng Chu[1] **Jiabei Zeng**[2] **Fernando De la Torre**[1] **Jeffrey F. Cohn**[1,3] **Daniel S. Messinger**[4]
[1]Robotics Institute, Carnegie Mellon University [3]University of Pittsburgh, USA
[2]Beihang University, Beijing, China [4]University of Miami, USA

Abstract

People are inherently social. Social interaction plays an important and natural role in human behavior. Most computational methods focus on individuals alone rather than in social context. They also require labelled training data. We present an unsupervised approach to discover interpersonal synchrony*, referred as to two or more persons preforming common actions in overlapping video frames or segments. For computational efficiency, we develop a branch-and-bound (B&B) approach that affords exhaustive search while guaranteeing a globally optimal solution. The proposed method is entirely general. It takes from two or more videos any multi-dimensional signal that can be represented as a histogram. We derive three novel bounding functions and provide efficient extensions, including multi-synchrony detection and accelerated search, using a warm-start strategy and parallelism. We evaluate the effectiveness of our approach in multiple databases, including human actions using the CMU Mocap dataset [1], spontaneous facial behaviors using group-formation task dataset [37] and parent-infant interaction dataset [28].*

1. Introduction

Humans are inherently social. Accessing human social interaction, especially *synchrony*, provides a better understanding of human behavior. Synchrony refers to the temporal structure of behaviors among interactive partners [13]. The close connection between synchrony and interaction provides researchers promising perspectives to build social interfaces [34], robots [6] or conversational agents [18]. However, a lack of automatic tools for synchrony discovery limits the exploration in interactive abilities.

Most prior art emphasizes on learning individual behaviors, and thus requires adequate labeled training data. Successful instances encompass a number of applications, such as action recognition [15, 21, 35], facial expression analysis [12, 14, 24, 26, 40] and sign language interpretation [10]. However, these methods focus on single individuals without considering social behaviors that can be triggered by the perception of actions in others. *E.g.*, during face-to-face interaction between mothers and their infants, they tend to match each other's affective states within lags of seconds. This synchrony improves the infant's experience of social connection during early development. Studying human interaction is crucial, but currently, to the best of our knowledge, no commonly accepted method exists for discovering synchrony among interactive partners.

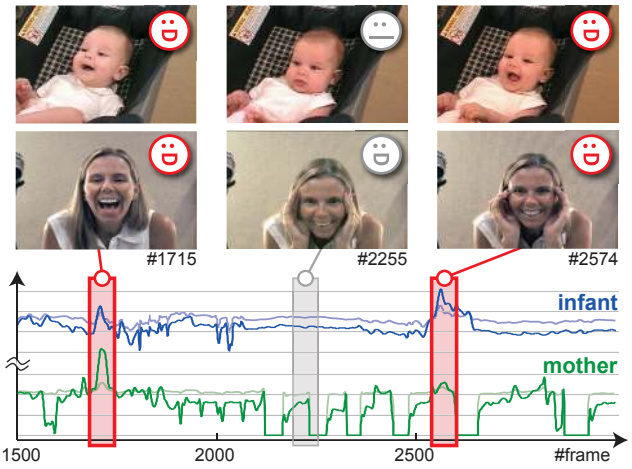

Figure 1. An illustration of *unsupervised synchrony discovery* in mother-infant interaction. Our method automatically discovers dyadic synchronies from multi-dimensional signals. Red bold boxes indicate the engagement in mutual smiles between the infant and the mother. The gray thin box indicate a randomly picked moment, showing an event without synchrony.

This paper presents an unsupervised approach to discover interpersonal synchrony that requires no training data. We term it unsupervised synchrony discovery (USD). Fig. 1 illustrates our main idea in a scenario of mother-infant interaction: Given a two synchronized videos represented as multi-dimensional signals, we aim to find their synchrony within a temporal window. For each behavior produced by one partner, the synchrony is defined as overlapped video frames or segments for the other partner(s) to produce a common behavior. As can be seen, two synchronies were discovered by our approach, where the mother and infant exhibits mutual engagement of smiles.

In specific, USD models the coordination among individuals as a global optimization problem. Unlike a naive approach that exhaustively evaluates temporal regions with

different lengths and locations, USD exploits a branch and bound (B&B) algorithm that allows an efficient search of a large collection of temporal windows. Along with two ways to accelerate the B&B search, USD guarantees to converge to a globally optimal solution with potentially fewer evaluations than exhaustive search. We showed the effectiveness of USD in discovering synchronies of human actions, group-formation tasks, and mother infant interaction.

In summary, our contributions are two-fold: (1) We present a new unsupervised technique for discovering synchrony in human interaction. To the best of our knowledge, our work is the first to match activity among individuals, providing an automatic tool to discover mutual engagement. (2) The proposed algorithm is general in two ways: it takes any signals represented as histograms, which can be bounded with standard metrics or three newly derived ones in this paper; it naturally generalizes to discover synchrony among more than two sequences. The algorithm is optimized to find an exact global solution, and can be further accelerated using a warm-start strategy and parallelism, showing an ability to handle large videos that are computationally prohibitive in exhaustive approaches.

2. Related Work

Synchrony discovery closely relates to human behavior analysis. Below we categorize prior art into supervised and unsupervised approaches, and discuss each in turn.

Supervised behavior analysis: Many techniques in computer vision for individual behavior analysis can be found in the literature, including facial expression recognition [14, 24, 26, 36, 40], surveillance system [16], activity recognition [15, 21, 35], and sign language interpretation [10]. Other works concern about the recognition of behaviors that involve more than one subject interacting in the scene. Brand *et al*. [5] introduced coupled hidden Markov models (CHMMs) to model dynamic interaction between multiple processes. Following up, Oliver and Pentland [33] proposed to recognize interaction between two people using HMMs and CHMMs, and concluded that CHMMs perform better in this task. Hongeng and Nevatia [20] proposed a hierarchical activity representation along with a temporal logic network for modeling and recognizing interaction. More recently, Liu *et al*. [25] proposed to recognize group behavior in AAL environment (nursing homes). A switch control module was performed to alternate between two HMM-based approaches according to the number of individual present in the scene. Messinger *et al*. [27] focused on specific annotated social signals, *i.e*., smiling and gaze, and characterized the transition between behavior states by a maximum likelihood approach. Interested readers are referred to [7] for a review. These techniques, however, require adequate labeled training data, which can be time-consuming to collect and not applicable to our scenario.

Unsupervised behavior analysis: The closest to our study is unsupervised approaches that require no training data. Zheng *et al*. [43] presented a coordinated motion model to detect motion synchrony in a group of individuals such as fish schools and bird flocks. Zhou *et al*. [44] proposed Aligned Cluster Analysis that extends spectral clustering to cluster time series. [44] applied the technique to discover facial events in unsupervised manner. Chu *et al*. [9] proposed a B&B approach to find time boundaries of common events happening in two videos. On the other hand, time series motifs, defined as the closest pair of subsequences in one time series stream, can be discovered with a tractable exact algorithm [29], or an approximated algorithm that is capable of tackling never-ending streams [4]. Some attempts on measuring interactional synchrony include using face tracking and expressions [42], and rater-coding and pixel changes between adjacent frames [38]. Nayak *et al*. [32] presented iterated conditional modes (ICM) to find most recurrent sign in all occurrences of sign language sentences. Recall that a synchrony is defined within a temporal window; it can contain subsequences from different videos that involve a temporal offset and sequence lengths different from each other. Given this structure, it remains unclear how a synchrony can be efficiently discovered using the above approaches.

3. Unsupervised Synchrony Discovery (USD)

3.1. USD for dyadic synchrony

Segment-level feature mapping: To describe the static and dynamic information of a video segment, we extract two types of features as the segment-level feature mapping [8, 19]: *observation features* extracted from a single frame, and *interaction features* extracted from two consecutive frames. Suppose the j^{th} frame is described as a feature vector $\mathbf{x}_j$. We perform k-means to find k centroids $\{\mathbf{c}_k\}_{k=1}^{K}$ as the hidden states. The observation feature $\phi^{\text{obs}}(\mathbf{x}_j)$ describes the pseudo-probability of $\mathbf{x}_j$ belonging to a state, and the interaction feature $\phi^{\text{int}}(\mathbf{x}_j)$ describes transition probability of states between two consecutive frames. As a result, we represent a video segment $\mathbf{X}_i = \{\mathbf{x}_{b_i}, ..., \mathbf{x}_{e_i}\}$ between the b_i^{th} and the e_i^{th} frames by normalizing the sum of the concatenation of the two features, resulting in a feature vector $\phi_{\mathbf{X}_i} = \sum_{j=b_i}^{e_i}[\phi^{\text{obs}}(\mathbf{x}_j); \phi^{\text{int}}(\mathbf{x}_j)]$. See [8, 19] for details about the feature mapping.

Problem formulation: To establish notion, we begin with two synchronized videos S^1 and S^2 with n frames each. The problem of Unsupervised Synchrony Discovery (USD) consists on searching over all possible subsequence pairs and find the one that shows similar patterns of change or movement. These pairwise patterns are known as *dyadic synchrony*. We formulate USD as an integer programming

over two intervals $[b_1, e_1] \subseteq [1, n]$ and $[b_2, e_2] \subseteq [1, n]$:

$$\max_{\{b_1,e_1,b_2,e_2\}} \quad f(\phi_{\mathsf{S}^1[b_1,e_1]}, \phi_{\mathsf{S}^2[b_2,e_2]}), \tag{1}$$
$$\text{subject to} \quad \ell \leq e_i - b_i, \forall i \in \{1, 2\},$$
$$|b_1 - b_2| \leq T,$$

where $f(\cdot, \cdot)$ is a similarity measure between two feature vectors (see details in Sec. 3.2), and ℓ controls the minimal length for each subsequence to avoid a trivial solution. T is a *synchrony offset* that allows USD to discover commonalities within a T-frame temporal distance, *e.g.*, in mother-infant interaction, the infant could start smiling after the mother smiles for a few seconds. Problem (1) is non-convex and non-differentiable, and thus standard convex optimization methods can not be applied. A naive solution is an exhaustive search with complexity $\mathcal{O}(n^4)$, which is computationally prohibitive for regular videos of several minutes.

Algorithm: We adapt a Branch and Bound (B&B) approach that guarantees a globally optimal solution in Problem (1). B&B has shown success in many computer vision problems, *e.g.*, object detection [22, 23], temporal commonality analysis [9], pose estimation [39] and optimal landmark detection [2]. For an event to be considered synchronous, they have to occur within a temporal neighborhood between two videos. For this reason, we only need to search within close regions in the temporal search space. Specifically, we constrain the space before the search begins, instead of exhaustively pruning the search space to a unique discovery (*e.g.*, [9, 22]).

Let $\mathtt{r} = [b_1, e_1, b_2, e_2]$ represent a rectangle in the 2-D search space. A rectangle set $\mathtt{R} = B_1 \times E_1 \times B_2 \times E_2$ in the search space indicates a set of parameter intervals, where $B_i = [b_i^{lo}, b_i^{hi}]$ and $E_i = [e_i^{lo}, e_i^{hi}], i \in \{1, 2\}$ are tuples of parameters ranging from frame lo to frame hi. We denote $|\mathtt{R}|$ as the number of possible rectangles in $\mathtt{R}$. See Fig. 2(f) for an illustration of the notation. Let $L = T + \ell$ be the largest possible period to search, we initialize a priority queue $\mathtt{Q}$ with rectangle sets $\{[t, t+T] \times [t+\ell-1, t+T+L-1] \times [t-T, t+T] \times [t-T+\ell-1, t+T+L-1]\}_{t=1}^{n-T-L+1}$ and their associated bounds (see details in Sec. 3.2). These rectangle sets lie sparsely along the diagonal in the 2-D search space, and thus prune a large portion during the search. Once all rectangle sets are settled, we adapt the Branch-and-Bound (B&B) strategy [9, 22] to find the exact optimum. Algo. 1 summarizes the proposed USD algorithm.

3.2. Measures with bounds

For the sake of using the B&B framework, we need a proper measure for similarity (or distance) between two sequences. This section constructs novel bounding functions for three measures: cosine similarity, symmetrized KL divergence and symmetrized cross entropy. Note that any measure with proper bounds (*e.g.*, ℓ_1, intersection, and χ^2 in [9]) can be directly applied.

Algorithm 1: Unsupervised Synchrony Discovery

input : A synchronized video pair $\mathbf{A}, \mathbf{B}$; minimal discovery length ℓ; commonality period T
output: Optimal intervals $\mathtt{r}^\star = [b_1, e_1, b_2, e_2]$

```
L ← T + ℓ; // The largest possible searching period
Q ← empty priority queue; // Initialize Q
for t ← 1 to (n−T−L+1) do
    R ← [t, t+T] × [t+ℓ−1, t+T+L−1] × [t−T, t+T] × [t−T+ℓ−1, t+T+L−1];
    Q.push(bound(R), R); // Fill in Q
end
R ← Q.pop(); // Initialize R
while |R| ≠ 1 do
    R → R1 ∪ R2;                 // Split into 2 disjoint sets
    Q.push(bound(R1), R1);       // Push R1 and its bound
    Q.push(bound(R2), R2);       // Push R2 and its bound
    R ← Q.pop();                 // Pop top state from Q
end
r* ← rect(R); // Retrieve the optimal rectangle
```

Let S^i denote the i-th sequence and can be represented as an unnormalized histogram $\mathbf{h}^i$ or a normalized histogram $\widehat{\mathbf{h}}^i$. Let h_k^i and $\widehat{h}_k^i$ be the k-th bin of $\mathbf{h}^i$ and $\widehat{\mathbf{h}}^i$, respectively. The normalized histogram is defined as $\widehat{h}_k^i = h_k^i / |\mathsf{S}^i|$, where $|\mathsf{S}^i| = \sum_k h_k^i$. $\|\mathsf{S}^i\| = \sqrt{\sum_k (h_k^i)^2}$ is the Euclidean norm of histogram of S^i. $\mathsf{S}[b, e]$ denotes the subsequence of S that starts from the b-th frame and ends in the e-th frame. Given a rectangle set $\mathtt{R} = B_1 \times E_1 \times B_2 \times E_2$, we denote the longest (shortest) possible subsequence as $\mathsf{S}^{i+}(\mathsf{S}^{i-})$, as illustrated in Fig. 2(f). Let $\mathtt{r} = [b_1, e_1, b_2, e_2] \in \mathtt{R}$ be a rectangle, $\underline{h_k^i} = \frac{h_k^{i-}}{|\mathsf{S}^{i+}|}$ and $\overline{h_k^i} = \frac{h_k^{i+}}{|\mathsf{S}^{i-}|}$, we observe the facts similar to [9]:

$$\text{(a) } 0 \leq h_k^{i-} \leq h_k^i \leq h_k^{i+}$$
$$\text{(b) } \|\mathsf{S}^{i-}\| \leq \|\mathsf{S}^i[b_i, e_i]\| \leq \|\mathsf{S}^{i+}\|,$$
$$\text{(c) } 0 \leq \underline{h_k^i} \leq \widehat{h}_k^i \leq \overline{h_k^i}.$$

Given these facts, below we construct the bounds for similarity (or distance) measures with normalized histograms $(\widehat{\mathbf{h}}^i, \widehat{\mathbf{h}}^j)$, whereas those for unnormalized histograms can be likewise obtained.

Cosine similarity: Treating two normalized histograms $\widehat{\mathbf{h}}^i$ and $\widehat{\mathbf{h}}^j$ as two vectors in the inner product space, we can measure the similarity as their included cosine angle:

$$C(\widehat{\mathbf{h}}^i, \widehat{\mathbf{h}}^j) = \frac{\widehat{\mathbf{h}}^i \cdot \widehat{\mathbf{h}}^j}{\|\widehat{\mathbf{h}}^i\| \|\widehat{\mathbf{h}}^j\|} = \frac{\sum_k \frac{h_k^i h_k^j}{|\mathsf{S}^i||\mathsf{S}^j|}}{\sqrt{\sum_k (\frac{h_k^i}{|\mathsf{S}^i|})^2} \sqrt{\sum_k (\frac{h_k^j}{|\mathsf{S}^j|})^2}}$$
$$= \frac{\sum_k h_k^i h_k^j}{\sqrt{\sum_k (h_k^i)^2} \sqrt{\sum_k (h_k^j)^2}} = \frac{\mathbf{h}^i \cdot \mathbf{h}^j}{\|\mathbf{h}^i\| \|\mathbf{h}^j\|}. \tag{2}$$

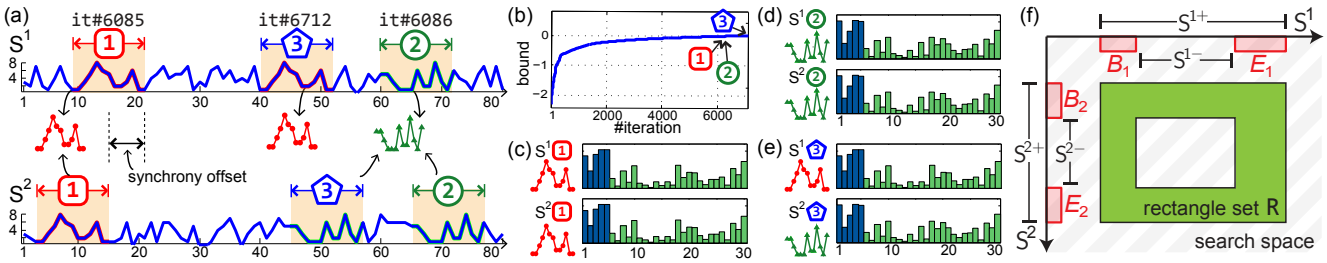

Figure 2. An example of USD on two 1-D time series using $\ell = 13$ and $T = 5$: (a) Top 3 discovered synchronies at different iterations; exhaustive search takes 39151 iterations. (b) The convergence curve w.r.t. bounding value and #iter. (c)∼(e) Discovered synchronies and their histograms, where blue and green bars indicate the segment features ϕ^{obs} and ϕ^{int}, respectively. ϕ^{int} is 10X magnified for display purpose. The ℓ_1 distances between the three histogram pairs are 6.3e-8, 1.5e-7, and 5.8e-2, respectively. (f) An illustration of notation.

Using facts (a) and (b), we obtain the bounds:

$$l_C(\mathtt{R}) = \frac{\sum_k h_k^{i-} h_k^{j-}}{\|\mathtt{S}^{i+}\|\|\mathtt{S}^{j+}\|} \le C(\mathbf{h}^i, \mathbf{h}^j) \le \frac{\sum_k h_k^{i+} h_k^{j+}}{\|\mathtt{S}^{i-}\|\|\mathtt{S}^{j-}\|} = u_C(\mathtt{R}).$$

Symmetrized KL *D*ivergence: As $\widehat{\mathbf{h}}^i$ and $\widehat{\mathbf{h}}^j$ are non-negative and sum to one, they can be interpreted as two discrete probability distributions and measured using the symmetrized KL divergence:

$$\begin{aligned} D(\widehat{\mathbf{h}}^i, \widehat{\mathbf{h}}^j) &= D_{KL}(\widehat{\mathbf{h}}^i || \widehat{\mathbf{h}}^j) + D_{KL}(\widehat{\mathbf{h}}^j || \widehat{\mathbf{h}}^i) \\ &= \sum_k (\widehat{h}_k^i - \widehat{h}_k^j)(\ln \widehat{h}_k^i - \ln \widehat{h}_k^j), \end{aligned} \tag{3}$$

where $D_{KL}(\widehat{\mathbf{h}}^i || \widehat{\mathbf{h}}^j)$ is the KL divergence of $\widehat{\mathbf{h}}^j$ from $\widehat{\mathbf{h}}^i$. From fact (c) and that $\underline{h_k^i} - \overline{h_k^j} \le \widehat{h}_k^i - \widehat{h}_k^j \le \overline{h_k^i} - \underline{h_k^j}$, we have $\ln \underline{h_k^i} - \ln \overline{h_k^j} \le \ln \widehat{h}_k^i - \ln \widehat{h}_k^j \le \ln \overline{h_k^i} - \ln \underline{h_k^j}$. Then, we obtain the bounds for (3):

$$\begin{aligned} l_D(\mathtt{R}) &= \sum_k (\underline{h_k^i} - \overline{h_k^j})_+ (\ln \underline{h_k^i} - \ln \overline{h_k^j})_+ \\ &\le D(\widehat{\mathbf{h}}^i, \widehat{\mathbf{h}}^j) \le \sum_k (\overline{h_k^i} - \underline{h_k^j})(\ln \overline{h_k^i} - \ln \underline{h_k^j}) = u_D(\mathtt{R}), \end{aligned}$$

where $(\cdot)_+ = \max(0, \cdot)$ is a non-negative operator to avoid both terms in (3) being negative.

Symmetrized cross *E*ntropy: The symmetrized cross entropy [30] measures the average number of bins needed to identify an event by treating each other as the true distribution. Similar to KL divergence that treats $\widehat{\mathbf{h}}^i$ and $\widehat{\mathbf{h}}^j$ as two discrete probability distributions, the entropy function is written as:

$$E(\widehat{\mathbf{h}}^i, \widehat{\mathbf{h}}^j) = \sum_k \widehat{h}_k^i \log \frac{1}{\widehat{h}_k^j} + \sum_k \widehat{h}_k^j \log \frac{1}{\widehat{h}_k^i}. \tag{4}$$

Recall the fact (c) and that $0 \le \widehat{h}_b^i \le 1, 0 \le \widehat{h}_b^j \le 1$, we obtain the bounds:

$$\begin{aligned} l_E(\mathtt{R}) &= \sum_b \left(-\underline{h_k^i} \log \overline{h_k^j} - \underline{h_k^j} \log \overline{h_k^i} \right) \\ &\le E(\widehat{\mathbf{h}}^i, \widehat{\mathbf{h}}^j) \le \sum_k \left(-\overline{h_k^i} \log \underline{h_k^j} - \overline{h_k^j} \log \underline{h_k^i} \right) = u_E(\mathtt{R}). \end{aligned}$$

To compute the bounds, we used an implementation of integral image [41] that takes an $\mathcal{O}(1)$ operation per evaluation. We refer interested readers to the supplementary material for detailed derivation of the above bounds. Fig. 2 shows a synthetic example of 1-D sequences with two pairwise synchronies, denoted as red dotted and green triangle segments, where one is a random permutation of another. USD discovered 3 dyads with the convergence curve in (b), and histograms of each dyad in (c)∼(e). Note that the interaction feature distinguishes the temporal consistency for the first and second discovery, maintaining a much smaller distance than the third discovery.

3.3. USD for triadic synchrony and more

We have described above how USD can discover dyadic synchrony with several bounding functions. In this section, we show that the main USD algorithm can be directly generalized and extended to capture mutual attention among a group (*i.e.*, multiple sequences). Specifically, we formulate the discovery among N sequences $\{\mathtt{S}^i\}_{i=1}^N$ by rewriting Eq. (1) as:

$$\begin{aligned} \max_{\{b_i, e_i\}_{i=1}^N} \quad & F\left(\{\phi_{\mathtt{S}^i[b^i, e^i]}\}_{i=1}^N\right) \\ \text{subject to} \quad & \ell \le e_i - b_i, \forall i \in \{1, ..., N\}, \\ & \max(|b_i - b_j|) \le T, \forall i \ne j, \end{aligned} \tag{5}$$

where $F(\cdot)$ is a similarity measure for a set of sequences and defined as the sum of pairwise similarities:

$$F\left(\{\phi_{\mathtt{S}^i[b^i, e^i]}\}_{i=1}^N\right) = \sum_{i \ne j} f(\phi_{\mathtt{S}^i[b_i, e_i]}, \phi_{\mathtt{S}^j[b_j, e_j]}). \tag{6}$$

Given a particular rectangle set $\mathtt{R}$ and sequence pair $(\mathtt{S}^i, \mathtt{S}^j)$, we rewrite their pairwise bounds in Sec. 3.2 as $l_f^{ij}(\mathtt{R})$ and $u_f^{ij}(\mathtt{R})$. The bounds for $F(\cdot, \cdot)$ can be defined as:

$$\begin{aligned} l_F(\mathtt{R}) &= \sum_{i \ne j} l_f^{ij}(\mathtt{R}) \le F\left(\{\phi_{\mathtt{S}^i[b^i, e^i]}\}_{i=1}^N\right) \\ &\le \sum_{i \ne j} u_f^{ij}(\mathtt{R}) = u_F(\mathtt{R}). \end{aligned} \tag{7}$$

Given this bound, Algorithm 1 can be directly applied to discover multiple synchronies.

Comparison with TCD [9]: Although Temporal Commonality Discovery (TCD) also performs unsupervised temporal discovery, this paper bears several technical differences. (1) New bounding functions: we introduce new bounds for cosine similarity, symmetrized KL divergence, and symmetrized cross entropy. These bounds enable applications of the B&B framework to domains where any of the metrics could be applicable. (2) Speed-up strategies: owing to the nature of the proposed problem, this paper introduces a warm-start and a parallelism approach for acceleration. TCD is sequential and thus can be very slow in practice. (3) Discover among >2 sequences: We offer a natural extension of USD for multiple sequences, whose effectiveness is shown in experiments. (4) TCD does not perform synchrony discovery.

4. Extensions of USD

Given the USD algorithm described above, this section describes its extensions to discover multiple synchronies and two accelerate approaches with with warm start (USD$^{\Delta}$) and parallelism (USD$^{\#}$).

Discover multiple synchronies: Multiple synchronies often occur in realistic videos, while the USD algorithm only outputs one synchrony at a time. To discover multiple synchronies, a trivial approach is to repeat USD many times by passing the priority queue `Q` from the previous USD to the next. However, each branching step splits a rectangle set `R` into two, resulting in an exponentially growing `Q` and inefficient search. Here we adapt a pruning strategy to safely discard undesired branches before starting the next USD. Given a previously discovered rectangle `r` and `Q` from the previous USD, we update every `R` using pruning rules that avoids overlapping detection with `r`. Without loss of generality, Fig. 3(a) illustrates the pruning rules for updating E_1 when overlapped with `r`, while the same rule applies for updating B_1. For axes of both S^1 and S^2, all `R` overlapped with `r` is updated according to the illustrated cases, and otherwise discarded. The updated rectangle sets, along with their bounds, are then pushed back to `Q` for the next USD.

This strategy is simple yet very effective. The bounds remain valid because each updated set is a subset of `R`. In practice, it dramatically reduces the number of states for searching the next synchrony. For example, in the example of Fig. 2, the size of `Q` is reduced 19% for the second USD, and 25% for the third USD.

USD with warm start (USD$^{\Delta}$): Due to the B&B nature, USD exhibits poor worst-case behavior, leading to a complexity as high as that of exhaustive search [31]. On the other hand, B&B search can quickly identify the exact solution when a local neighborhood contains a clear optimum [22]. Given this motivation, we explore a "warm start" strategy that estimates an initial solution with high quality, and then initializes USD around the solution. Estimating an initial solution costs only few percentage of total iterations, and thus can effectively prune branches in the main USD algorithm. Fig. 3(b) illustrates the idea. Specifically, we run a sampled sliding window with stepsize=10, sort the visited windows according their distances, and then determine a warm start region around the windows within the lowest one percentile. Then the main USD algorithm is performed only within an expanded neighborhood around the warm start region.

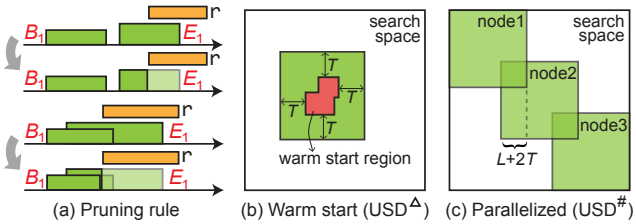

Figure 3. Illustration of USD extensions: (a) pruning rules applied to multi-synchrony discovery, (b) USD with warm start (USD$^{\Delta}$), and (c) parallelized USD (USD$^{\#}$).

Parallelized USD (USD$^{\#}$): The use of parallelism to speed up B&B algorithms has emerged as a way to solve larger problems [17]. Based on the block-diagonal structure in the search space, this section describes an parallelized approach USD$^{\#}$ to scale up USD for larger sequences. Note that the parallelism was not shown possible in previous sequential method [9]. In specific, we divide USD into subproblems, and perform the USD algorithm solve each in parallel. Because each subproblem is smaller than the original one, the number of required iterations can be potentially reduced. As illustrated in Fig. 3(c), the original search space is divided into overlapping regions, where each can be solved using independent jobs on a cluster. The results are obtained as the top k rectangles collected from each subproblem. Due to the diagonal nature of USD in the search space, the final result is guaranteed to be a global solution. The proposed structure enables static overload distribution, leading to an easily programmable and efficient algorithm.

5. Experiments

We evaluated our method on discovering synchronies in a variety of video sources: human actions from CMU motion capture (Mocap) dataset [1], social group interaction from GFT dataset [37] and parent-infant interaction [28].

5.1. Comparison and evaluation metric

To our best knowledge, there is no commonly accepted method that explicitly tackle the USD problem. Instead, we compare USD with a baseline sliding window (SW) approach, *i.e.*, evaluate subsequently rectangles in the search space and take the maximal similarity (or minimal distance) as indicators for the existence of a synchrony. In

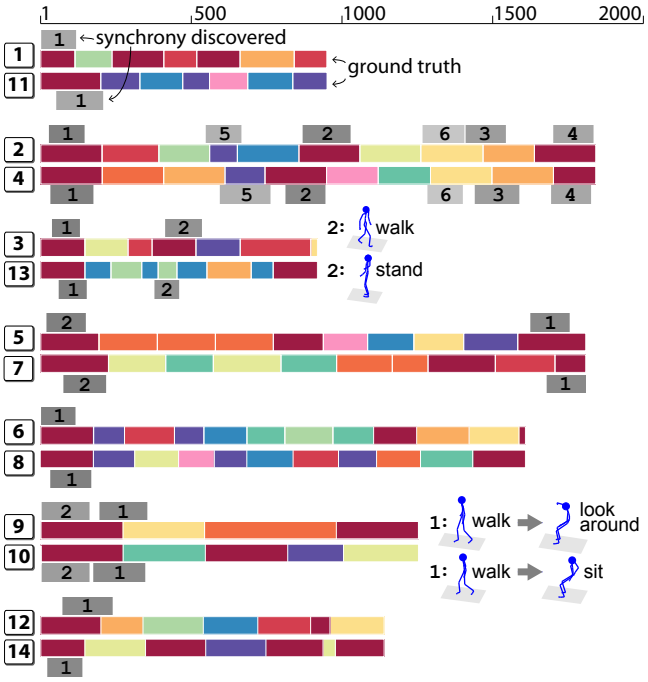

Figure 4. Discovered synchronies on 7 pairs of Subject 86 in CMU-Mocap dataset. Each pair is annotated with ground truth (colorful bars, each represents an action), and synchronies discovered by our method (shaded numbers). Synchronies with disagreed action labels are visualized.

particular, we implemented SW with an initialization of an $\ell \times \ell$ window, and gradually increment the window size along each dimension using a fixed step size s, *i.e.*, multiple window scales were allowed. The window was moved every s frames among the same search region as USD. We compared SW and USD by their discovery speed and quality.

Evaluation of *speed* was computed by the number of function evaluations, to exclude factors in different hardware and implementation. For datasets for which labels are available (*i.e.*, CMU Mocap [1] and GFT [37]), evaluation of *quality* was carried out using the recurrence analysis [13], which was originally designed to analyze a coupled dynamical system based on signal consistency. Let $\{\mathsf{S}^i\}_{i=1}^N$ be a collection of sequences with n frames each, $\mathcal{A} = \{(i,j)|i,j \in \{1,...,N\}, i \neq j\}$ be a pairwise index set of $\{\mathsf{S}^i\}$, $\mathtt{r}$ be a discovered synchrony with n_i frames along sequence S^i, and $\mathbf{Y}_i \in \mathbb{R}^{C \times n}$ be the ground truth labels of S^i, where each column represents labels of C classes. $\mathbf{Y}_i^c[p]$ denote the c-th class labels corresponding to the p-th frame in S^i. For a given $\mathtt{r}$, we define the *recurrent consistency*:

$$\mathcal{Q}(\mathtt{r}) = \frac{1}{C\prod_i n_i} \sum_c \sum_{(i,j)\in\mathcal{A}} \sum_{p,q} I(\mathbf{Y}_i^c[p] = \mathbf{Y}_j^c[q]), \quad (8)$$

where $I(X)$ is an indicator function returning 1 if the statement X is true and 0 otherwise. The quality measures the mutual agreement between each pair of the discovered subsequences, resulting in a score in [0,1]. The score reaches 1 when the discovered synchrony agrees completely on each other's label, and 0 when they completely disagree.

Table 1. Distance and quality analysis on CMU Mocap dataset: (top) χ^2 distance using 1e-3 as unit, (bottom) recurrent consistency. $\mathrm{SW}_s^\star$ indicates the optimal window found by SW_s with step size $s = 5, 10$; SW_s^μ and SW_s^σ indicate average and standard deviation among all windows. The best discovery are marked in bold.

	Pair	(1,11)	(2,4)	(3,13)	(5,7)	(6,8)	(9,10)	(12,14)	Avg.
χ^2-distance	USD	**6.3**	**1.2**	**4.7**	**2.6**	**0.1**	**0.2**	**11.9**	**3.9**
	$\mathrm{SW}_5^\star$	6.5	1.3	6.7	5.4	**0.1**	0.4	12.0	4.6
	$\mathrm{SW}_{10}^\star$	6.7	2.7	6.7	10.1	0.2	0.7	14.3	5.9
	SW_5^μ	97.1	76.9	81.4	64.2	89.3	172.0	334.5	130.8
	SW_5^σ	33.8	74.4	53.8	28.2	79.2	117.7	345.1	104.6
	SW_{10}^μ	94.8	77.3	81.8	63.2	87.1	170.2	327.2	128.8
	SW_{10}^σ	34.3	74.1	54.2	28.3	79.4	117.8	341.5	104.2
Rec. consistency	USD	0.89	**0.85**	0.46	**0.90**	**1.00**	0.64	**0.76**	**0.79**
	$\mathrm{SW}_5^\star$	**0.95**	0.81	**0.50**	0.84	**1.00**	**0.69**	0.73	**0.79**
	$\mathrm{SW}_{10}^\star$	**0.95**	0.75	**0.50**	0.64	**1.00**	0.55	0.00	0.63
	SW_5^μ	0.07	0.32	0.09	0.07	0.08	0.13	0.12	0.12
	SW_5^σ	0.16	0.33	0.25	0.20	0.21	0.29	0.22	0.24
	SW_{10}^μ	0.08	0.31	0.09	0.07	0.09	0.13	0.12	0.13
	SW_{10}^σ	0.19	0.33	0.26	0.21	0.22	0.29	0.23	0.25

5.2. Synchrony in human actions

This section examines the ability of USD to discover synchronies in human actions on the CMU Mocap dataset [1]. Mocap data provides high-degree reliability in measurement and serves as an ideal target for an initial test of our method. We used the *Subject 86* data that contains 14 sequences labeled with action boundaries [3]. To remove the redundancy in action labels, we merged similar actions into 24 categories, *e.g.*, {*arm rotating*, *right arm rotation*, *raise arms*, *both arm rotation*} were categorized as *arm raise*. Each action was represented by root position, orientation and relative joint angles, resulting in a 30-D feature vector. The segment-level feature was used as described in Sec. 3.1. To mimic a scenario for USD, we grouped the sequences into 7 pairs as the ones containing similar number of actions, and trimmed each action to up to 200 frames. USD was performed using $\ell = 120$ and $T = 50$.

Table 1 summarizes the USD results compared with the baseline sliding window (SW). Results are reported using χ^2-distance and the recurrent consistency described in (8). A threshold of 0.012 was manually set to discard discovery with large distance. We ran SW with step sizes 5 and 10, and marked the windows with the minimal distance as $\mathrm{SW}_5^\star$ and $\mathrm{SW}_{10}^\star$, respectively. Among all, USD discovers all results found by SW. To understand how well a prediction by chance can be, all windows were collected to report average μ and standard deviation σ. As can be seen, on average, a randomly selected synchrony can result in large distance over 100 and low quality below 0.3. USD main-

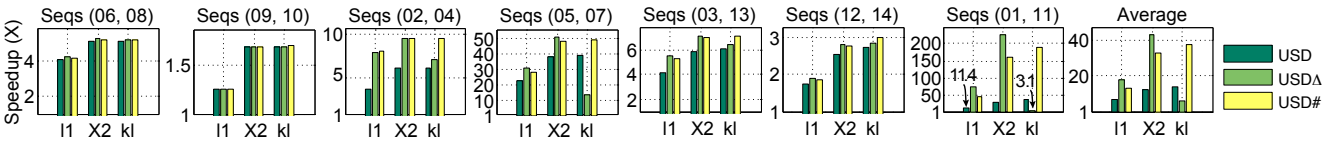

Figure 5. Speedup of our methods against sliding window (SW) in the CMU-Mocap dataset. All 7 pairs of sequences from subject 86 were evaluated. The speedup was computed as the relative number of evaluations $N^{\text{SW}}/N^{\text{USD}}$ using ℓ_1, χ^2 and symmetrized KL divergence.

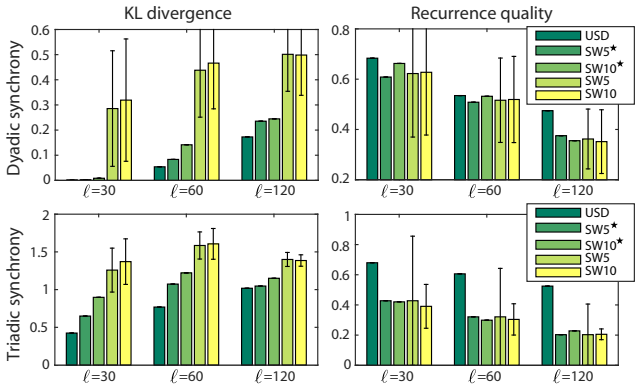

Figure 6. Analysis on top 10 discovered dyadic and triadic synchronies of the GFT dataset. SW denoted with ⋆ indicates the optimal windows discovered, and without ⋆ indicates the average and standard deviation over all visited windows.

tained an exact minimal distance with good qualities as the ones found by exhaustive SW. Note that, because USD is totally unsupervised, the synchrony with minimal distance may not necessarily guarantee the highest quality.

Fig. 4 shows the qualitative results on all 7 pairs, annotated with ground truth and the discovered synchronies. As can be seen, USD allows to discover multiple synchronies with varying lengths. Although some discovered synchronies contain disagreed action labels, one can observe that the discoveries share reasonable visual similarity, *e.g.*, in pair (9,10), the "look around" action in sequence 9 was performed when the subject was seated, sharing the similarity with the "sit" action in sequence 10.

Fig. 5 shows the speed up of USD against exhaustive SW. USD and its extensions demonstrated an improved efficiency over SW. In some cases, USD$^{\Delta}$ improved search speed by a large margin, *e.g.*, in (01,11) with χ^2-distance reached a speed boost over 200 times. Across all metrics, the speed up of USD$^{\Delta}$ was less obvious with symmetrized KL divergence. USD$^{\#}$ was implemented on a 4-core machine; an extension to larger clusters is possible yet beyond the scope of this study. On average, USD$^{\#}$ consistently performed faster across different metrics than the original USD due to parallelism.

5.3. Synchrony in social group interaction

This section describes the discovery of synchronies in social group interaction. We used the GFT dataset [37] that consists of 720 participants recorded during group-formation tasks. Previously unacquainted participants sat together in groups of 3 at a round table for 30 minutes while getting to know each other. We used 2 minutes of videos from 48 participants, containing 6 groups of two subjects and 12 groups of three subjects. USD was performed to discover *dyads* among groups of two, and *triads* among groups of three. Each video was tracked with 49 facial landmarks using IntraFace [11]. We represented each face by concatenating appearance features (SIFT) and shape features (49 landmarks). For evaluating the discovered results, we computed the recurrence quality using the action unit (AU) labels provided in the dataset. In particular, we used AUs (10,12,14,15,17,23,24) that appear most frequently.

As the minimal length ℓ is an empirical parameter to determine, we examined USD with $\ell \in \{30, 60, 120\}$, resulting in synchronies that last at least 1, 2 and 4 seconds; we set the synchrony offset $T = 30$ (1 second). Similar to Sec. 5.2, baseline SW was performed using step sizes 5 and 10. Symmetrized KL divergence was used as the distance function. We evaluated the distance and quality among the optimal window discovered, as well as the average and standard deviation among all windows to tell a discovery by chance. Fig. 6 shows the averaged KL divergence and quality among top 10 discovered dyadic and triadic synchronies. As can be seen, USD always guarantees the lowest divergence because of its nature to find the exact optimum. The recurrence quality decreases while ℓ grows, showing that finding a synchrony with longer period while maintaining good quality is harder than finding one with shorter period. Note that, although the discover quality is not guaranteed in an unsupervised discovery, USD consistently maintained the best discovery quality across various lengths. This result illustrates the power of our unsupervised method that agrees with that of supervised labels.

5.4. Synchrony in parent-infant interaction

Parent-infant interaction is critical for children in early development and social connections. This section attempts to characterize their affective engagement by exploring the moments where the behavior of both the parent and the infant are correlated. We performed this experiment on the mother-infant interaction dataset [28]. Participants were 6 ethnically diverse 6-month-old infants and their parents (5 mothers, 1 father). Infants were positioned in an infant-seat facing their parent who was seated in front of them. We used 3 minutes of normal interaction where the parent plays with the infant as they might do at home. Because

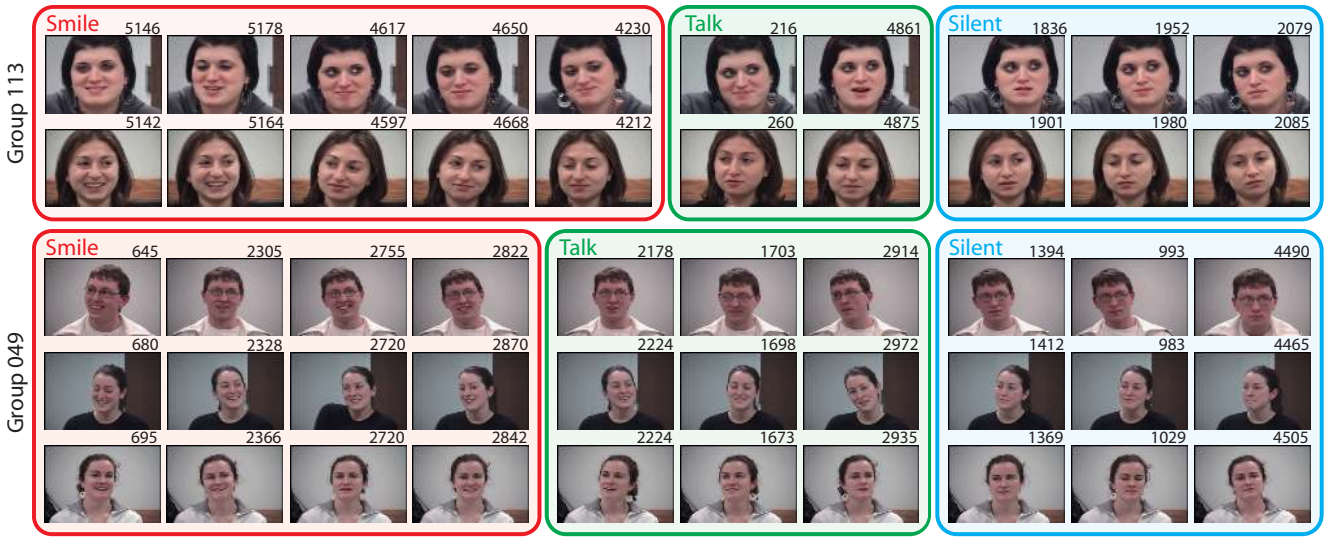

Figure 7. Top 10 discovered synchronies from groups 113 and 128 in the GFT dataset. Each column indicates a discovered synchrony and its frame number. The algorithm correctly discovered *smiling*, *talking* and *silent* moments as different synchrony events.

Figure 8. Discovered sychronies from 6 groups of parent-infant interaction. Each column indicates a discovered synchrony and its #frame.

this dataset does not provide ground truth annotations, we only evaluate the results quantitatively. Similar to Sec. 5.3, we tracked and extracted features on each face. Because the appearance of parents and infants are quite different, we used only the shape feature in this experiment. Throughout this experiment, we set $\ell = 80$ and $T = 40$.

Fig. 8 illustrates three discovered synchronies among all 6 parent-infant pairs. As can be seen, many synchronies were discovered as the moments when both infants and parents exhibit strong smiles, serving as a building block of early interaction [28]. Besides smiles, a few synchronies showed strong engagement in their mutual attention, such as the second synchrony of group ① where the infant cried after the mother showed a sad face, and the second synchrony of the second group where the mother stuck her tongue out after the infant did so. These interactive patterns offered another solid evidence of a positive association between infants and their parents.

6. Conclusion

We presented *unsupervised synchrony discovery* (USD), a relatively unexplored problem that discovers synchrony in human interaction. We formulated USD as a searching problem in time series, and proposed an efficient B&B algorithm, optimized to find the global solution with potentially fewer evaluations than exhaustive search. In addition, we extended our approach to multi-synchrony detection, and two accelerated search—a warm-start strategy and parallelism. Our method can be naturally generalized to discover synchrony among more than two sequences. Our results in discovering synchronies of human actions and interaction illustrate the power of USD that agrees with supervised labels. Moving forward, we plan to extend USD to discover causal-effect synchronies (*e.g.*, question-asking and hand-raising in teacher-student interaction).

Acknowledgment: This paper was supported in part by the National Institutes of Health under Awards Number MH096951 and MH099487.

References

[1] CMU Mocap. http://mocap.cs.cmu.edu/.

[2] B. Amberg and T. Vetter. Optimal landmark detection using shape models and branch and bound. In *ICCV*, 2011.

[3] J. Barbič, A. Safonova, J.-Y. Pan, C. Faloutsos, J. K. Hodgins, and N. S. Pollard. Segmenting motion capture data into distinct behaviors. In *Graphics Interface Conference*, 2004.

[4] N. Begum and E. Keogh. Rare time series motif discovery from unbounded streams. 2015.

[5] M. Brand, N. Oliver, and A. Pentland. Coupled HMMs for complex action recognition. In *CVPR*, 1997.

[6] A. Cangelosi, G. Metta, G. Sagerer, S. Nolfi, C. Nehaniv, K. Fischer, J. Tani, T. Belpaeme, G. Sandini, F. Nori, et al. Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Trans. on Autonomous Mental Development*, 2(3):167–195, 2010.

[7] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. *Expert Systems with Applications*, 39(12):10873–10888, 2012.

[8] W.-S. Chu, Y. Song, and A. Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*, 2015.

[9] W.-S. Chu, F. Zhou, and F. De la Torre. Unsupervised temporal commonality discovery. In *ECCV*, 2012.

[10] H. Cooper and R. Bowden. Learning signs from subtitles: A weakly supervised approach to sign language recognition. In *CVPR*, 2009.

[11] F. De la Torre, W.-S. Chu, X. Xiong, X. Ding, and J. F. Cohn. Intraface. In *Automatic Face and Gesture Recognition*, 2015.

[12] F. De la Torre and J. F. Cohn. *Guide to Visual Analysis of Humans: Looking at People*, chapter Facial Expression Analysis. Springer, 2011.

[13] E. Delaherche, M. Chetouani, A. Mahdhaoui, C. Saint-Georges, S. Viaux, and D. Cohen. Interpersonal synchrony: A survey of evaluation methods across disciplines. *IEEE Trans. on Affective Computing*, 3(3):349–365, 2012.

[14] S. Du, Y. Tao, and A. M. Martinez. Compound facial expressions of emotion. *Proceedings of the National Academy of Sciences*, 111(15):E1454–E1462, 2014.

[15] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009.

[16] R. Feris, R. Bobbitt, L. Brown, and S. Pankanti. Attribute-based people search: Lessons learnt from a practical surveillance system. In *ICMR*, 2014.

[17] B. Gendron and T. G. Crainic. Parallel branch-and-branch algorithms: Survey and synthesis. *Operations research*, 42(6):1042–1066, 1994.

[18] J. Gratch, N. Wang, J. Gerten, E. Fast, and R. Duffy. Creating rapport with virtual agents. In *Intelligent Virtual Agents*, 2007.

[19] M. Hoai, Z.-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *CVPR*, 2011.

[20] S. Hongeng and R. Nevatia. Multi-agent event recognition. In *ICCV*, 2001.

[21] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007.

[22] C. H. Lampert, M. B. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(12):2129–2142, 2009.

[23] A. Lehmann, B. Leibe, and L. Van Gool. Fast prism: Branch and bound hough transform for object class detection. *International Journal on Computer Vision*, 94(2):175–197, 2011.

[24] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006.

[25] C.-D. Liu, Y.-N. Chung, and P.-C. Chung. An interaction-embedded hmm framework for human behavior understanding: with nursing environments as examples. *IEEE Trans. on Information Technology in Biomedicine*, 14(5):1236–1246, 2010.

[26] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPRW*, 2010.

[27] D. M. Messinger, P. Ruvolo, N. V. Ekas, and A. Fogel. Applying machine learning to infant interaction: The development is in the details. *Neural Networks*, 23(8):1004–1016, 2010.

[28] D. S. Messinger, M. H. Mahoor, S.-M. Chow, and J. F. Cohn. Automated measurement of facial expression in infant–mother interaction: A pilot study. *Infancy*, 14(3):285–305, 2009.

[29] A. Mueen, E. J. Keogh, Q. Zhu, S. Cash, and M. B. Westover. Exact discovery of time series motifs. In *SDM*, 2009.

[30] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[31] P. M. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Trans. on Computers*, 100(9):917–922, 1977.

[32] S. Nayak, K. Duncan, S. Sarkar, and B. Loeding. Finding recurrent patterns from continuous sign language sentences for automated extraction of signs. *Journal of Machine Learning Research*, 13(1):2589–2615, 2012.

[33] N. M. Oliver, B. Rosario, and A. P. Pentland. A bayesian computer vision system for modeling human interactions. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):831–843, 2000.

[34] K. Prepin and C. Pelachaud. Shared understanding and synchrony emergence synchrony as an indice of the exchange of meaning between dialog partners. In *Intl. Conf. on Agent and Artificial Intelligence*, 2011.

[35] K. K. Reddy and M. Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.

[36] E. Sangineto, G. Zen, E. Ricci, and N. Sebe. We are not all equal: Personalizing models for facial expression analysis with transductive parameter transfer. 2014.

[37] M. A. Sayette, K. G. Creswell, J. D. Dimoff, C. E. Fairbairn, J. F. Cohn, B. W. Heckman, T. R. Kirchner, J. M. Levine, and R. L. Moreland. Alcohol and group formation a multimodal investigation of the effects of alcohol on emotion and social bonding. *Psychological science*, 2012.

[38] R. C. Schmidt, S. Morr, P. Fitzpatrick, and M. J. Richardson. Measuring the dynamics of interactional synchrony. *Journal of Nonverbal Behavior*, 36(4):263–279, 2012.

[39] M. Sun, M. Telaprolu, H. Lee, and S. Savarese. An efficient branch-and-bound algorithm for optimal human pose estimation. In *CVPR*, 2012.

[40] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *CVPRW*, 2006.

[41] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal on Computer Vision*, 57(2):137–154, 2004.

[42] X. Yu, S. Zhang, Y. Yu, N. Dunbar, M. Jensen, J. K. Burgoon, and D. N. Metaxas. Automated analysis of interactional synchrony using robust facial tracking and expression recognition. In *Automatic Face and Gesture Recognition*, 2013.

[43] Y. Zheng, S. Gu, and C. Tomasi. Detecting motion synchrony by video tubes. In *ACMMM*, 2011.

[44] F. Zhou, F. De la Torre, and J. K. Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(3):582–596, 2013.