Detecting and Synthesizing Synchronous Joint Action in Human-Robot Teams

Tariq Iqbal and Laurel D. Riek
Computer Science and Engineering, University of Notre Dame
{tiqbal,lriek}@nd.edu

ABSTRACT

To become capable teammates to people, robots need the ability to interpret human activities and appropriately adjust their actions in real time. The goal of our research is to build robots that can work fluently and contingently with human teams. To this end, we have designed novel nonlinear dynamical methods to automatically model and detect synchronous joint action (SJA) in human teams. We also have extended this work to enable robots to move jointly with human teammates in real time. In this paper, we describe our work to date, and discuss our future research plans to further explore this research space. The results of this work are expected to benefit researchers in social signal processing, human-machine interaction, and robotics.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics

1. INTRODUCTION

Humans and robots are increasingly working side-by-side across many disciplines, ranging from manufacturing and agriculture, to education and home healthcare [27]. However, for a robot to be useful to human teammates, it needs to understand the activities occurring around it, as well as their context [19]. If a robot can better interpret its surroundings, its interaction with humans is more likely to reach to a higher level of coordination, and will lead to a *fluent* meshing of their actions over time [5, 8, 3]. Therefore, it is important for robots to possess the ability to interpret the high-level activities happening around them from its sensor data, and to predict future activities in order to inform their own actions [9, 7].

Many fields, such as robotics and computer vision, have been working on improving the automatic understanding of human activities, as well as the dynamics of a group. This work include advances in both exo-centric and ego-centric processing, used to detect activities from buttering toast to social interaction [20, 4]. Recent work has built upon this high-level human activity detection to automatically characterize the dynamics and interaction patterns of a group [26, 2].

This work is useful for many situations; however, in human-robot teamwork, it can be challenging for a robot to understand and contextualize human activity as it happens. Solely relying on high-level human activity detection may not provide robots sufficient contextual information to characterize the situation and environment. Furthermore, it is challenging to sense and respond to people when both are in motion, due to sensor occlusion, camera shake, noise, etc. [9]. However, employing a lower-level, dynamical approach to group behavior and motion understanding may help robots overcome these limitations.

Synchronous joint action (SJA) is a common phenomenon observed during human-human interaction (HHI). This is a form of social interaction where two or more participants coordinate their actions both in space and time while making changes to their environment [21]. Understanding joint action in a human-robot interaction (HRI) scenario is also important, as a robot may make decisions about its actions to improve its overall engagement as a team member. For example, a robot can be a more effective team member by anticipating a human teammate's motion and acting appropriately in collaborative manipulation tasks [23, 22].

Researchers in the fields of cognitive science, psychology and music have explored different methods to measure joint action in HHI [15, 1, 18]. Recently, it also has attracted the attention of the HRI community [13, 6]. Most of this research has focused on dyadic HRI (one human, one robot), and focused mostly on manipulation tasks, such as handovers and assembly. However, it is important for a robot to extend the notion of understanding joint action beyond this dyadic HRI, such as when working in a team with multiple people and robots. In the case of synchronous group activities, such as collaborative climbing, running, or dancing, a robot needs to understand the motion of its co-humans and predict their future actions to move "in-step" with the rest of the group.

This leads us to explore several research questions. First, can we automatically *measure* the degree of synchronous joint action in a group from the high-level activities of the group members? As SJA is an important indicator of the group cohesiveness, and enables an understanding of a group's affective behavior, answering this research question would help machines more accurately characterize group dynamics [12, 25, 18]. Moreover, addressing this question would be helpful for a robot to understand the dynamics of a human group more accurately, even when everyone is continuously in motion.

Second, can this method be used to *synthesize* joint activity in real-time while robots are moving within synchronous human-robot teams? Addressing this question would mean that the robot not only will be able to understand the behavior of the group, but also can inform its own activities appropriately.

Addressing these research questions is important, because this will lead us to build intelligent robots which will both be able to understand human teams, and interact contingently within them.

ICMI '15, November 9–13, 2015, Seattle, WA, USA
© 2015 ACM. ISBN 978-1-4503-3912-4/15/11 ...$15.00.
DOI: http://dx.doi.org/10.1145/2818346.2823315.

In our research, we developed a novel method to automatically detect SJA in a group from multi-modal data streams. This nonlinear, dynamical-systems approach takes multiple types of task-level events into account, and is able to work with non-periodic time series data as it estimates SJA [12, 9]. In the next Section, we describe our SJA measurement method, and briefly present a human-human and human-robot validation study. Then, we describe a method to enable a robot to synchronously move within a human-robot team using the SJA measure. Finally, we discuss our research contribution and plan in the last Section.

2. MEASURING SJA IN A TEAM

Several researchers have worked on measuring synchronization within a group, such as Varni et al. [26]. These methods typically only incorporate a single modality and a single type of event when calculating synchrony of a group (e.g., only gross body motion, or only EEG signals). However, there are cases where multiple types of events are associated with a group activity. Usually to assess these events, analysis of multimodal data streams is necessary. For example, in the case of a repetitive manufacturing process, humans need to perform different types of tasks, and all these types of tasks need to incorporate together if we want to measure the synchronization of the group. In these cases, it is necessary to incorporate all these different types of events together, instead of just one, when measuring group synchrony [12].

To address this gap, we developed a new method that builds on event synchronization work by Quian Quiroga et al. [16] and Varni et al. [26]. Our method considers multiple types of task-level events together to measure SJA (group synchrony), and can also incorporate multiple types of sensor data (i.e., depth, RGB, audio). In addition to this contribution, it is also more accurate than single-event based methods, and able to work with real-world sensor data. The detailed description of the method can be found in [12, 9].

In our method, we employ the following six high-level steps to measure group synchronization [12, 10]:

1. Automatically detect the high-level events during the group activity from the multimodal data stream,
2. Express all of the detected events of each group member with a time series,
3. Measure the pair-wise synchronization index (*PSync*) for each pair of group members, taking all types of events into account together,
4. Build a group topology graph (*CTG*) of the group based on the *PSync* values,
5. Calculate the individual synchronization index (*ISync*) of each group member from the *PSync* and the *CTG*,
6. Determine the group synchronization index (*GSync*) from the *ISync* values and the *CTG*.

To detect the task-level events associated with a group task, we can use different sensor data together. After detecting the multiple types of events, we can present those in a time series. Now, suppose x_n and y_n are the two time series, where $n = 1, \ldots, N$ (N samples). For each event type $e_i \in E$, $m_x(e_i)$ and $m_y(e_i)$ are the number of events occurring in x and y respectively, where E is the set of all events. Now, for each event type $e_i \in E$, we calculate the pairwise synchronization index ($Q(e_i)$) [12, 9].

$Q(e_i)$ represents the synchronization of event event type e_i in two time series. We normalized this value, thus $Q(e_i)$ = 1 means all the events are fully synchronous, whereas, $Q(e_i)$ = 0 means, the events are not synchronous at all. Then, the overall pairwise synchronization index (Q) considering all events is calculated as:

$$\forall e_i \in E : Q = \frac{\sum [Q(e_i) \times [m_x(e_i) + m_y(e_i)]]}{\sum [m_x(e_i) + m_y(e_i)]} \tag{1}$$

After detecting Q, we build a group topology graph (CTG), which is a undirected weighted graph. In this graph, each time series is represented by a vertex, and each pair of vertices are connected by an edge weighted by the Q value of that pair. Based on this graph, and the value of Q, we measure the individual synchronization index (*ISync*) of each member of the group. Individual synchronization index represents how well a member is synchronous with the rest of the group. Then, based the *ISync* values and the *CTG*, we determine the group synchronization index (*GSync*) [12, 9, 10].

3. VALIDATION OF THE METHOD

We validated our method by applying it to two group activities that involve SJA. The first activity involved only humans, while the second activity included both humans and robots. We briefly describe the experimental studies below, though they are explained in detail in [12, 9].

3.1 Human-Human Validation Study

First, we validated the method by applying it to a multiple, event-based, rhythmic group game called "The Cup Game" [12]. During this game, the players stand at a table, perform a sequential and rhythmic activity with their hands and a cup, and pass the cup to their neighbor at the end of an iteration (see Fig. 1-A). We chose this task due to the fact that each player performs a periodic activity, and each player's movements influences the others. This task helped us to address our first research question.

Two Microsoft Kinect sensors were connected to two time synchronized computers running Robot Operating System. We recorded the RGB video and the skeletal data of the players during the games. The events during the game (clap, pass or move the cup) were detected from the hand joint positions of each player and the cup positions. We tracked the hands using the Kinect's skeletal data. The cups were tracked using the standard RGB-based blob tracking techniques from the RGB data. From these events, we measured the *GSync* index for each game using the method described previously [12].

In this study, a total of 22 people participated (50% female) in groups of four players per session, yielding a total of six experimental sessions. Each group had a learning and practice session, followed by two games. After the two games, each group member rated on a Discrete Visual Analog Scale (DVAS) how well-synchronized they felt each game was, and which game was more synchronous [12].

The results suggest that the group synchronization indices produced by our method agreed with the perception of the majority of participants in all of the sessions. The results further suggest that our method is more accurate in estimating SJA than other methods described in the literature. Full experimental details and result analyses are available in [12], though we present an analysis of one of the sessions in Fig 2-A.

3.2 Human-Robot Validation Study

In the second validation experiment, we employed our proposed method on a human-robot teamwork scenario to automatically measure SJA of the group [12, 9]. In this study, two people performed a high-march action down a hallway across four scenarios (see Fig 1-B). The first performer acted as the leader, and wore noise-canceling headphones playing "Stars and Stripes Forever", a march by John Philip Sousa, to help them march at a consistent pace.

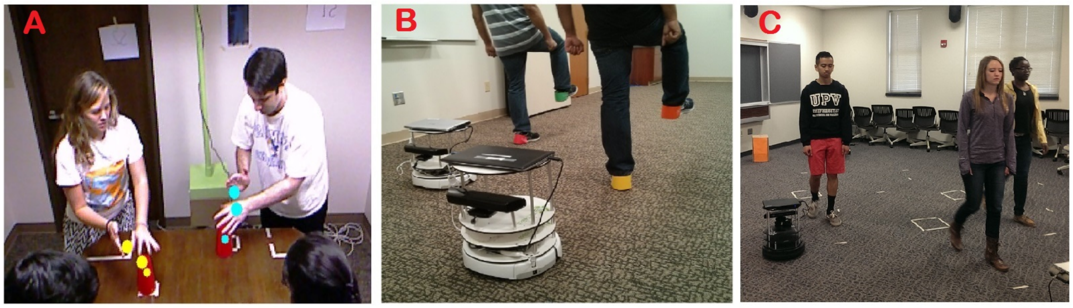

Figure 1: Our research explores psychomotor entrainment between groups of people and robots, where we work toward enabling robots to automatically sense and respond to synchronous group behavior.

The second performer acted as the follower, and was approximately two feet behind the leader to their right. The follower adjusted their steps either synchronously or asynchronously as directed by the experimenter per condition of the experiment. Two autonomous mobile robots (Turtlebots) were programmed to follow the respective performers [9].

We defined two types of task-level events based on the positions of the legs of the performers. These events occurred when a performer began to raise their leg from the ground, and when a leg reached its maximum height. As a result, a total of four types of events occurred during this marching activity (two for each leg) [9].

To detect the events, the feet of each performer were tracked using standard RGB-based blob tracking techniques, as captured from the robots' Kinect sensors. These blobs corresponded to the four unique squares of colored paper attached to the performers' feet. Detecting these events were challenging, as both the humans and the robots were in motion during this activity. From these events, we measured the *GSync* index for each marching session [9].

In the study, we recorded data from a total of four scenarios, which corresponded to the four experimental conditions, and yielded four unique marching patterns. We expected to see high synchronization index values when the performers marched synchronously, and low values when they were asynchronously. As demonstrated by Fig 2-B, our measured experimental results for each scenario reasonably match the expected synchronization indices. The detailed analysis can be found in Iqbal et al. [9, 11].

4. SYNTHESIZING SJA IN A TEAM

It is important for a robot to make appropriate decisions based on its understanding of the activities occurring around it using its sensor data. In the previous Section, we described methods for a robot to understand group activity and behavior from multiple types of task-level events. In this Section, we describe two methods for a robot to anticipate future activities and actions of the other group members. We also describe methods to generate appropriate actions for a robot based on these anticipations. This study was aimed to address our second research question.

We used a synchronous group dance performed by a human-robot team as an experimental test bed for studying SJA tasks. Three human performers along with a mobile robot (i.e., a Turtlebot robot) performed a dance to the song "Smooth Criminal" by Michael Jackson (see Fig 1-C). With the help of an experienced dancer, we choreographed a routine for the participants to perform. The dance is iterative, and performed cyclically in a counter-clockwise manner. Each phase includes four iterations of the following steps in order: move forward, move backward, move forward, move backward, clap, and a 90-degree turn [17].

To acquire the movement data of the humans during the dance sessions, we used four Microsoft Kinect for Windows version 2 sensors. Each Kinect sensor was connected to a computer to capture and process depth, infrared and skeletal data. One of the main challenges of a multi-sensor and multi-client setup is to maintain a consistent time reference across all the client machines. To maintain a time reference for all the machines, we implemented a client-server architecture for communication between the clients with Kinect sensors and the server [17].

The clients periodically synchronized their clocks with the server, giving a global accuracy to the timestamps recorded with detected events. When a client determined that an event had occurred, it sent a message to the server indicating the classification of the event, the time at which it occurred to millisecond precision, and other information relevant to the particular event [17].

Each client processed the data streams captured by the attached Kinect sensor, and sent the processed data to the server. The server received the data, and generated predictions for the movements of the robot client. The robot client was also connected to the server. The robot node ran on the Robot Operating System (ROS) platform to control a Turtlebot's movements based on instructions received from the server [17].

The clients detected five high-level events from the performers' movements during the dance performance: start and stop moving forward, start and stop moving backward, and clap. These events were detected from the change in 3-D skeleton joint positions of the dancers. The start and the stop of the forward and backward motion was detected when there was a sufficient change in the z component of the human performer's skeletal joint coordinates. A clap was detected when the hand joint positions of the human performer reached a sufficiently small local minima within a time window [17].

From these detected events, in the server, we employed two methods to anticipate the future action of the humans, and generate necessary commands for the robot to perform appropriate actions in a timely fashion. The first method that we employed was an *Event Cluster based Anticipation (ECA)* method for predicting the timing of future group events and informing the robot to perform the actions accordingly. Then, relying on our previously proposed event-based method for measuring the degree of synchrony of a group, we used a *Synchronization Index based Anticipation (SIA)* method to inform robot's movements.

The ECA method relied on the assumption that the movement events of one iteration are more or less similar to the events that happened in the previous iteration. This was a valid assumption because this dance is rhythmic and iterative in nature. Thus, the event timing of one iteration was taken into consideration to predict the event timing of the next iteration. First, we clustered all the similar types of events together that happened relatively closer in

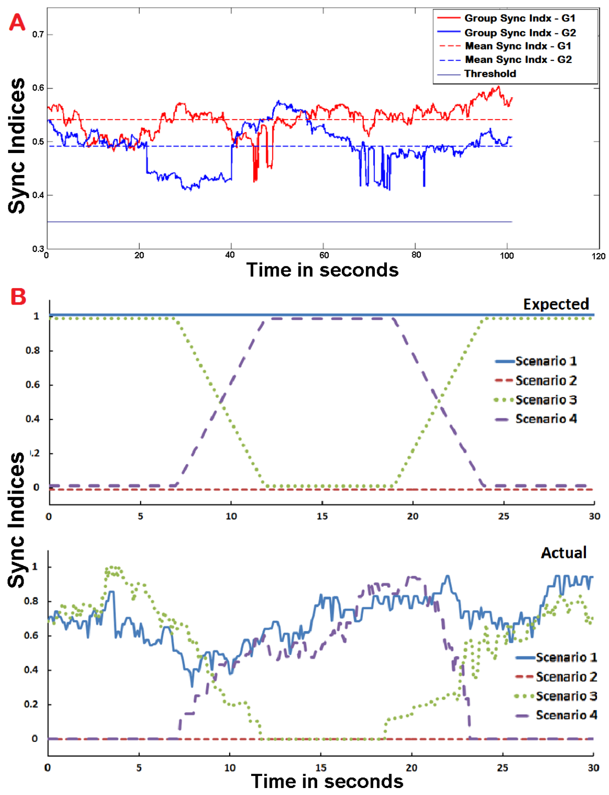

Figure 2: A) Group Synchronization Indices over time during a cup game session [12], B) Expected and Actual synchronization indices for all four experimental scenarios of the human-robot marching study [9].

time. Then, we calculated the average time of all the events and used that time as the timing of the event for the next iteration.

Alternatively, the SIA method takes the group's internal dynamics into account while generating movements for the robot. The method operates under the assumption that a person, who was synchronous with the rest of the group members in the previous iterations, is more likely to be synchronous with the rest of the group members during this iteration of the dance as well. Thus, if the robot follows the movements of that person during this iteration of the dance, it is more likely that the robot would also be synchronous with the rest of the group during this iteration.

We recruited a total of 9 groups (27 participants, 3 persons per group) for our study. Each group had a learning and practice session, and then participated in three dance sessions. During the first dance session, only the humans participated in the dance movements, while the robot joined the humans in the dance during the last two sessions.

During the last two sessions, the robot joined the group. We used the anticipation methods in different sessions. To reduce the bias due to the ordering of the methods, we counterbalanced the anticipation methods. However, we did not disclose which method was in use to the participants during the dance.

Following the experiment, participants completed a short questionnaire asking them to rate which of the two dance sessions they felt was more synchronous. A detailed description of this study can be found in [17].

Our initial data analysis shows some promising results. The results suggest that most of the groups were more synchronous when the SIA method was used as the anticipation method for the robot, than the ECA method. These results might support the robustness of the SIA method over the ECA method, as SIA the method takes the group's internal dynamics into account while anticipating future events.

5. RESEARCH CONTRIBUTIONS AND RESEARCH PLAN

The aim of this research project is to build intelligent robots, able to fluently interact with human teams. To summarize, we have made the following contributions:

1. The creation of a nonlinear, dynamical systems inspired approach to measure the degree of SJA in a group from the multiple types of task-level activities of the group members [12, 10].
2. The employment of this method to a human-robot teamwork scenario, where both the humans and the robots were in motion [9].
3. The application of this method on a robot to anticipate the actions of group members, and to perform appropriate actions depending on their dynamics [17].

Building on this foundation, we plan to develop and incorporate methods to detect high-level activities of the humans using the local multimodal sensors of the robot. Incorporating both local and global sensor data of a robot will be helpful for better high-level event detection, which will lead the robot perceiving group dynamics more accurately. However, as the local sensor data may be more noisy in nature due to camera shake or occlusion, it will be more challenging to incorporate this data to detect high-level human activities. To overcome this problem, we plan to use an activity recognition approach similar to Ryoo et al. [20], which uses a first-person view to detect high level events. By combining the events from both the local and global sensors, we will then use our proposed algorithm to detect the SJA of a group.

Additionally, humans are skilled at synchronizing their movements with event sequences containing continuous tempo changes [24]. Models like ADAM (ADaptation and Anticipation Model) have been proposed in literature to model this behavior by combining adaptation and anticipation during an activity [24]. It will be beneficial for a robot to have this ability to be more acceptable by the human counterparts in a team. We also might explore the adaptation mechanism such as those proposed in ADAM, but in the context of a human-robot team. We might extend the error correction mechanism, and can integrate this with our SIA algorithm. This integration might make the SIA algorithm more robust in anticipating and synthesizing future activities more accurately.

After exploring this, the next step of our research will be to extend our model to work for activities beyond SJA, which are not necessarily synchronous, such as human-robot collaboration tasks in industrial environments, like to Wilcox et al. [27], Nikolaidis et al. [14]. Teams working in these settings have specific sequences of activities to perform over time, some of which might be independent, and might not have to happen synchronously. We are planning to extend our proposed algorithm to work during these kind of activities, by incorporating this ordering of events into account. The extended method might be useful to modify SIA to anticipate future activities of the group.

Our current research will directly support other researchers exploring multimodal interaction in the human-robot interaction domain. This method can enable robots to have an automatic understanding of high-level group behavior by processing multimodal sensor streams, and to inform its own actions in response. This also can play a role in handovers, joint action, and collaborative manipulation in HRI, which are all current topics in the field [9]. Moreover, this research is directly applicable in fields beyond HRI, such as social signal processing [12]. Our hope is that this research will help machines to be more socially aware, as well as be more acceptable to people.

6. REFERENCES

[1] S. M. Boker and J. L. Rotondo. Symmetry building and symmetry breaking in synchronized movement. *Adv Consc Res*, 2002.

[2] O. Brdiczka, J. Maisonnasse, and P. Reignier. Automatic detection of interaction groups. In *Proc. Int. Conf. Multimodal Interfaces*, 2005.

[3] M. Cakmak, S. S. Srinivasa, M. K. Lee, S. Kiesler, and J. Forlizzi. Using spatial and temporal contrast for fluent robot-human hand-overs. In *Proc. of ACM/IEEE HRI*, 2011.

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

[5] G. Hoffman and C. Breazeal. Cost-based anticipatory action selection for human–robot fluency. *IEEE Transactions on Robotics*, 2007.

[6] G. Hoffman and G. Weinberg. Gesture-based human-robot jazz improvisation. In *Proc. of IEEE ICRA*, 2010.

[7] T. Iqbal, M. Gonzales, and L. D. Riek. Mobile robots and marching humans: Measuring synchronous joint action while in motion. In *Proc. of AAAI Fall Symposium Series on AI-HRI*, 2014.

[8] T. Iqbal, M. J. Gonzales, and L. D. Riek. A Model for Time-Synchronized Sensing and Motion to Support Human-Robot Fluency. In *ACM/IEEE HRI, Workshop on Timing in HRI*, 2014.

[9] T. Iqbal, M. J. Gonzales, and L. D. Riek. Joint action perception to enable fluent human-robot teamwork. In *Proc. of IEEE RO-MAN*, 2015.

[10] T. Iqbal and L. Riek. Assessing group synchrony during a rhythmic social activity: A systemic approach. In *In Proc. of ISGS*, 2014.

[11] T. Iqbal and L. D. Riek. Role distribution in synchronous human-robot joint action. In *Proc. of IEEE RO-MAN, Towards a Framework for Joint Action*, 2014.

[12] T. Iqbal and L. D. Riek. A Method for Automatic Detection of Psychomotor Entrainment. *IEEE T on Affective Computing*, 2015.

[13] S. Lallée and et al. Towards a platform-independent cooperative human-robot interaction system: Ii. perception, execution and imitation of goal directed actions. In *IROS*, 2011.

[14] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah. Efficient model learning from joint-action demonstrations for human-robot collaborative tasks. In *Proc. of HRI*, 2015.

[15] L. Noy, E. Dekel, and U. Alon. The mirror game as a paradigm for studying the dynamics of two people improvising motion together. *P Natl Acad Sci USA*, 2011.

[16] R. Q. Quiroga, T. Kreuz, and P. Grassberger. Event synchronization: a simple and fast method to measure synchronicity and time delay patterns. *Phys Rev E*, 2002.

[17] S. Rack, T. Iqbal, and L. Riek. Enabling synchronous joint action in human-robot teams. In *Proc. of ACM/IEEE HRI*, 2015.

[18] M. J. Richardson, R. L. Garcia, T. D. Frank, M. Gergor, and K. L. Marsh. Measuring group synchrony: a cluster-phase method for analyzing multivariate movement time-series. *Frontiers in Physiology*, 2012.

[19] L. D. Riek. The social co-robotics problem space: Six key challenges. In *In Proc. of RSS, Robotics Challenges and Visions Workshop*, 2013.

[20] M. S. Ryoo and L. Matthies. First-Person Activity Recognition: What Are They Doing to Me? *In Proc. of IEEE CVPR*, 2013.

[21] N. Sebanz, H. Bekkering, and G. Knoblich. Joint action: bodies and minds moving together. *T. Cogn. Sci.*, 2006.

[22] K. W. Strabala and et al. Towards seamless human-robot handovers. *JHRI*, 2013.

[23] V. V. Unhelkar, H. C. Siu, and J. A. Shah. Comparative performance of human and mobile robotic assistants in collaborative fetch-and-deliver tasks. In *Proc of. HRI*, 2014.

[24] M. M. van der Steen, N. Jacoby, M. T. Fairhurst, and P. E. Keller. Sensorimotor synchronization with tempo-changing auditory sequences: Modeling temporal adaptation and anticipation. *Brain research*, 2015.

[25] G. Varni, G. Volpe, and A. Camurri. A System for Real-Time Multimodal Analysis of Nonverbal Affective Social Interaction in User-Centric Media. *IEEE T Multimedia*, 2010.

[26] G. Varni, G. Volpe, and B. Mazzarino. Towards a social retrieval of music content. In *IEEE Conference on Social Computing*, 2011.

[27] R. Wilcox, S. Nikolaidis, and J. A. Shah. Optimization of temporal dynamics for adaptive human-robot interaction in assembly manufacturing. In *Proc. of RSS*, 2012.