Social Coordination Assessment : Distinguishing Between Shape and Timing

Emilie Delaherche, Sofiane Boucenna, Koby Karp, Stéphane Michelet, Catherine Achard, and Mohamed Chetouani

Institute of Intelligent Systems and Robotics,
University Pierre and Marie Curie, 75005 Paris, France.
{delaherche,boucenna,karp,michelet}@isir.upmc.fr
{catherine.achard,mohamed.chetouani}@upmc.fr
http://www.isir.upmc.fr/

Abstract. In this paper, we propose a new framework to assess temporal coordination (synchrony) and content coordination (behavior matching) in dyadic interaction. The synchrony module is dedicated to identify the time lag and possible rhythm between partners. The imitation module aims at assessing the distance between two gestures, based on 1-Class SVM models. These measures discriminate significantly conditions where synchrony or behavior matching occurs from conditions where these phenomenons are absent. Moreover, these measures are unsupervised and could be implemented online.

Keywords: Behavior matching, synchrony, unsupervised model

1 Introduction

Natural conversation is often compared to a dance for the exchange of signals (prosody, gesture, gaze, posture) is reciprocal, coordinated and rhythmic. Rapport building, the smoothness of a social encounter and cooperation efficiency are closely linked to the ability to synchronize with a partner or to mimic part of his behavior. Human interaction coordination strategies, including behavior matching and synchrony are yet delicate to understand and model [1]. However, the close link between coordination and interaction quality bears promising perspectives for researchers building social interfaces, robots, and Embodied Conversational Agents [2].

Many terms related to coordination co-exist in the literature. But we usually distinguish between behavior matching [3] and synchrony. Mirroring; mimicry [4]; congruence and the chameleon effect [5] are related to behavior matching. These concepts concern non-verbal communicative behaviors, such as postures, mannerisms or facial displays, and indicate similar behaviors by both social partners; the analyzed features are isomodal and qualitative.

Synchrony is related to the adaptation of one individual to the rhythm and movements of the interaction partner [3,6,7] and the degree of congruence between the behavioral cycles of engagement and disengagement of two people. In

opposition to behavior matching, synchrony is a dynamic phenomenon and can intervene across modalities.

These definitions are theoretic and in practice both forms of coordination can be observed at the same time. In this paper, we argue that despite the co-existence of both phenomenon in social interactions, a unique system is not adequate to model both forms of coordination. We propose to create two models: one dedicated to characterize synchrony and another system to assess behavior matching. In this paper, we will focus on behavior matching assessment. Synchrony assessment is described succinctly and will be detailed in future work.

2 Previous Works and Proposed Approach

Actual state-of-the-art methods to assess synchrony are based on correlation. After extracting the movement time series of the interactional partners, a time-lagged cross-correlation is applied between the two time series using short windows of interaction. Several studies also use a peak picking algorithm to estimate the time-lag of the predictive association between two time series (i.e., the peak cross-correlation that is closest to a lag of zero) [8]. The main flaw of these methods is the mixing between the temporal and content aspects of coordination. Correlation informs on the temporal relation between events. But the similarity between the shape of events is poorly treated as gestures are often inadequately represented (e.g. motion energy).

In this paper, we propose to differentiate the temporal and the content part of coordination. We propose the following architecture (see Fig 1). A first module detects the onsets of gestures of both partners by identifying a strong increase of motion energy. Two modules receive the timings of the gestures : the synchrony module and the behavior matching module. The synchrony module answers the question : are the two partners in synchrony and is there an interpersonal rhythm between the two partners? Based on the timing of the segmented gestures, several metrics qualifying the respective rhythm of each partner and their interpersonal rhythm are proposed. The behavior matching module answers the question : to which extent two gestures are similar? It first identifies for each gesture of one partner the closest gestures in time from the other partner. Then, it assesses the distance between each pair of gestures. This metric is unsupervised and does not rely on predefined actions. Indeed, we are not interested in categorizing the gestures but only in comparing them.

3 Synchrony module

This module characterizes the dynamics of activation of the dyadic partners. We are interested in the timing of events, regardless of their shape. In this paper, we focus on movement synchrony by analyzing onsets of gestures, but events could also be verbal like back-channel vocalizations for instance. Many studies in psychology underline the importance of synchrony during the interaction between a mother and a baby. For example, babies are extremely sensitive to the

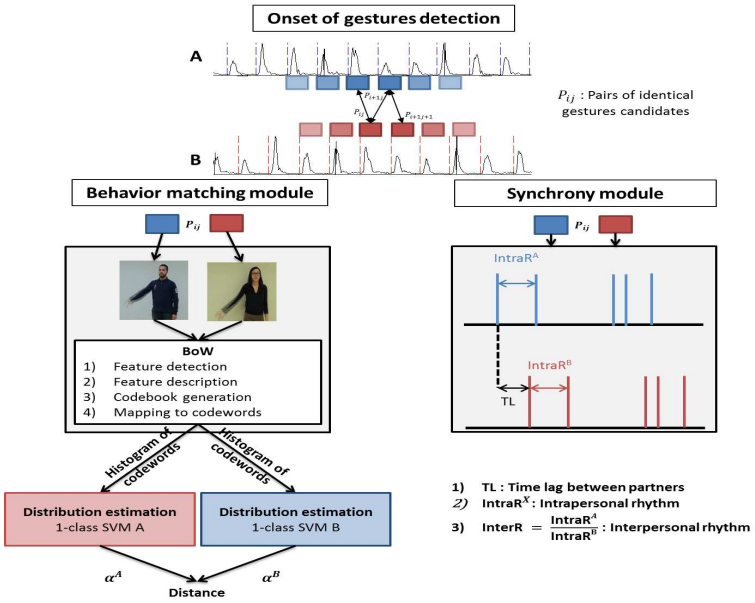


Fig. 1. A first module segments the gestures of both partners by identifying a strong increase of motion energy. Two modules receive the timings of the segmented gestures: the behavior matching module and the synchrony module.

interaction rhythm with their mother [7,6]. A social interaction rupture involves negative feelings (e.g., agitation, tears) while a rhythmic interaction involves positive feelings and smiles.

For each onset of gesture c_n^A at time $t_{c_n^A}$ detected on the partner A, the closest onset of gesture $t_{c_n^B}$ of the partner B is identified. Several features of synchrony can be extracted from these events :

- Time-lag between partners : $TL_n = t_{c_n^A} - t_{c_n^B}$ indicates which partner is leading the interaction at time c_n^A.
- Intrapersonal rhythm : $IntraR_n^A = t_{c_{n+1}^A} - t_{c_n^A}$ assesses the time between two occurrences of events for the same participant.
- Interpersonal rhythm : $InterR_n = \frac{IntraR_n^A}{IntraR_n^B}$ assesses the rapport of intrapersonal rhythm of the two partners. This measure is close to 1 if both partners share the same rhythm (whether there is a time-lag between them or not). The measure is superior to one if partner A rhythm is more important than partner's B.

The time-lag or rhythms at one moment of the interaction are not particularly informative but the variance of these features through the entire interaction, informs on whether the partners were in synchrony most of the time and adopted the same rhythm. More, in the prospect of building social interfaces, rhythm could be used as a reward signal to learn an arbitrary set of sensori-motor rules [9,10].

4 Behavior matching module

This module computes a distance between the dyadic partners gestures, at each time a new onset of gesture is detected. The gestures are represented with histograms of visual words and a metric based on 1-Class SVM is proposed.

4.1 Visual features

Bag of Words models have been successfully applied in computer vision for object recognition, gesture recognition, action recognition and Content Based Image Retrieval (CBIR) [11,12,13]. The method is based on a dictionary modeling where each image contains some of the words of the dictionary. In computer vision, the words are features extracted from the image. Bag of Words models rely on 4 steps : feature detection, feature description, codebook generation, mapping to codebook. In this work, Dollàr detector [14] is used for interest point detection. It was preferred to other detectors for its robustness and for the number of interest points detected was superior, leading to a better characterization of the gesture performed. Histogram Of Oriented Gradient (HOG) and Histogram Of Oriented Flow (HOF) are used for description [12]. These descriptors characterize both shape and motion while keeping a reasonable length (compared to Dollàr descriptors for instance). The size of the feature vectors is 162 (72 bins for HOG and 90 bins for HOF). At last, it is conceivable to construct the codebook on-line with sequential k-means clustering for instance.

4.2 Distance between gestures

We propose to derive an algorithm for novelty detection based on 1-Class SVM, proposed by Canu and Smola to estimate the distance between two gestures [15].

Distribution Estimation (1-Class SVM) 1-Class SVM was proposed to estimate the density of a unknown probability density function [16]. For $i = 1, 2, ..., n$, the training vectors h_i are assumed to be distributed according to a unknown probability density function $P(.)$. The aim of 1-class SVM is to learn from the training set a function f such that most of the data in the training set belong to the set :

$$R_h = \{h \in X \setminus f(h) \geq 0\}$$

and the region R_h is minimal. The function f is estimated such that a vector drawn from $P(.)$ is likely to fall in R_h and a vector that does not fall in R_h is not likely to be drawn from $P(.)$. The decision function is :

$$f(h) = \sum_{i=1}^{n} \alpha_i k(h, h_i) - \rho$$

The kernel $k(.,.)$ is defined over $X \times X$ by : $\forall (x_i, x_j) \in X \times X, k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_H$ where $\langle ., . \rangle_H$ denotes the dot product in H and ϕ is a mapping

from the input space X to a Reproducing Kernel Hilbert Space, called the feature space H. As in our case, h represents an histogram of codewords, we chose the histogram intersection kernel, defined as $k(h_i, h_j) = \sum_{i=1}^{d} min(h_i, h_j)$, where d denotes the size of the histogram.

Distance Let $h_{A_i}, i = 1...n$ and $h_{B_i}, i = 1...n$ be the sequence of codewords histograms for a pair of identical gestures candidates, n denotes the size of the window. Let assume that the sequences are stationary from 1 to n and that h_{A_i} is distributed according to a distribution P_A and h_{B_i} is distributed according to a distribution P_B. To determine if the two gestures are identical, we are interested in testing the following hypothesis:

$$\begin{cases} H_0 : P_A = P_B \text{ (the gestures are identical)} \\ H_1 : P_A \neq P_B \text{ (the gestures are different)} \end{cases}$$

We write the likelihood ratio as follow :

$$L(h_{A_1}, \ldots, h_{A_n}, h_{B_1}, \ldots, h_{B_n}) = \frac{\prod_{i=1}^{n} P_A(h_{A_i}) P_B(h_{B_i})}{\prod_{i=1}^{n} P_A(h_{A_i}) P_A(h_{B_i})} = \prod_{i=1}^{n} \frac{P_B(h_{B_i})}{P_A(h_{B_i})}$$

Since both densities P_A and P_B are unknown the generalized likelihood ratio (GLR) has to be used :

$$L(h_{A_1}, \ldots, h_{A_n}, h_{B_1}, \ldots, h_{B_n}) = \prod_{i=1}^{n} \frac{\hat{P}_B(h_{B_i})}{\hat{P}_A(h_{B_i})}$$

where $\hat{P}_A$ and $\hat{P}_B$ are the maximum likelihood estimates of the densities. The exponential family gives a general representation for many of the most common distributions (normal, exponential, Poisson...). Assuming there exists a reproducing kernel Hilbert space H embedded with the dot product $< ., >_H$ and with a reproducing kernel k, the probability density function of an exponential family can be expressed :

$$P(h, \theta) = \mu(h) exp(< \theta(.), k(h, .) >_H - g(\theta)) \tag{1}$$

where $g(\theta) = log \int_X exp(< \theta(.), k(h, .) >_H) d\mu(h)$, $\mu(h)$ is the carrier density, θ is the natural parameter and $g(\theta)$ is the log-partition function. One-class SVM was proposed to estimate the support of a high dimensional distribution. Assuming that densities P_A and P_B belong to the exponential family and natural parameters θ_A and θ_B are estimated with 1-class SVM model, $\hat{P}_A$ and $\hat{P}_B$ can be written :

$$\begin{array}{l} \hat{P}_A(h) = \mu(h) exp(\sum_{i=1}^{n} \alpha_i^A k(h, h_{A_i}) - g(\theta_A)) \\ \hat{P}_B(h) = \mu(h) exp(\sum_{i=1}^{n} \alpha_i^B k(h, h_{B_i}) - g(\theta_B)) \end{array} \tag{2}$$

where α_i^A (resp. α_i^B) is determined by solving the 1-class SVM on h_{A_i} (resp. h_{B_i}). Thus,

$$L(h_{A_1}, \ldots, h_{A_n}, h_{B_1}, \ldots, h_{B_n}) = \prod_{j=1}^{n} \frac{exp(\sum_{i=1}^{n} \alpha_i^B k(h_{B_j}, h_{B_i}) - g(\theta_B))}{exp(\sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i}) - g(\theta_A))}$$

Two gestures are similar if $L(h_{A_1}, \ldots, h_{A_n}, h_{B_1}, \ldots, h_{B_n})$ is inferior to a given threshold :

$$\sum_{j=1}^{n} \left(\sum_{i=1}^{n} \alpha_i^B k(h_{B_j}, h_{B_i}) - \sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i} \right) < s_A$$

And $\sum_{i=1}^{n} \alpha_i^B k(h_{B_j}, h_{B_i})$ can be neglected in comparison with $\sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i})$. Thus two gestures are similar if :

$$\sum_{j=1}^{n} \left(- \sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i}) \right) < s_A$$

This distance can be interpreted as testing a model learned on h_{A_i} with the data from h_{B_i}. For robustness [17], we adopt the following distance in which the histograms of h_{A_i} and h_{B_i} are alternatively used for learning and for testing.

$$d = \sum_{j=1}^{n} \left(- \sum_{i=1}^{n} \alpha_i^A k(h_{B_j}, h_{A_i}) \right) + \sum_{j=1}^{n} \left(- \sum_{i=1}^{n} \alpha_i^B k(h_{A_j}, h_{B_i}) \right)$$

As the visual words dictionary can be constructed online and the 1-Class SVM models are learned on the fly for each window of interaction, no supervision is required and the system can easily adapt to new gestures.

5 Results and discussion

5.1 Data

An actual issue is the evaluation of synchrony and behavior matching models. Despite the existence of several annotating scheme, the annotation of coordination is often problematic. Indeed, the phenomenon involves the perception of complex and intricate social signals. Consequently, in several studies the measure of coordination is not validated per se, it is the ability of the measure to predict outcome variables that is evaluated.

Fig. 2. Imitation condition : the sequence of gestures.

To circumvent the annotation problem, we constructed interaction data presenting different conditions of rhythm, synchrony and behavior matching. A similar approach of using simple and constructed stimuli was used to evaluate a model of audio-visual synchrony estimation [18]. In all conditions, two partners

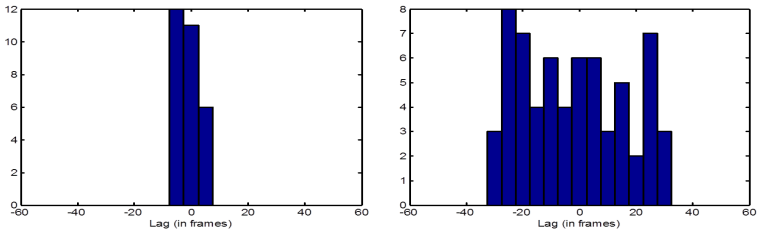
Table 1. Stimuli and conditions. We denote for each sequence its length l in seconds and the number of gestures n in the sequence l[n].

Frequency (in BPM)	Synchrony and No B.Matching (S_NBM)	Synchrony and B.Matching (S_BM)	No Synchrony and No B.Matching (NS_NBM)
20	137[44]	62[19]	
25	166[67]	71[28]	117[NA]
30	153[71]	59[27]	

are standing in front of each other and filmed with a separate Sony camera at 25 fps. The focal length and focus of the cameras were optimized to capture an upper-body view of the participants. In the Synchrony (S) condition, the partners had to perform the gesture of their choice at a given rhythm, they synchronized with each other thanks to a metronome. In the Behavior Matching (BM) condition, the participants had to perform a series of identical gestures represented in Fig. 2. In total, the database contains 256 pairs of gestures including 74 pairs of identical gestures. The non-identical pairs of gestures were more numerous to account for the diversity of non-imitative situations. We voluntarily did not record a video in the NS_BM condition as it is delicate to manipulate the settings to obtain such combination. Moreover, by shifting one of the video of the S_BM condition with a certain time lag, it is possible to recreate such condition. In the NS_NBM condition, one performs at the pace of the metronome while the other is asked to gesture continually. The different conditions are summarized in Table 1.

5.2 Results

Rhythm Detection Module To test this module, we compared the Synchrony (S) and NonSynchrony (NS) conditions. We ran this module on all the videos based on the onset of gestures identified. Figure 3 presents the histogram of time-lags during the 25_NS_NBM and the 25_S_BM conditions. The variance of time-lag is larger in the 25_NS_NBM than in the 25_S_BM condition. We also



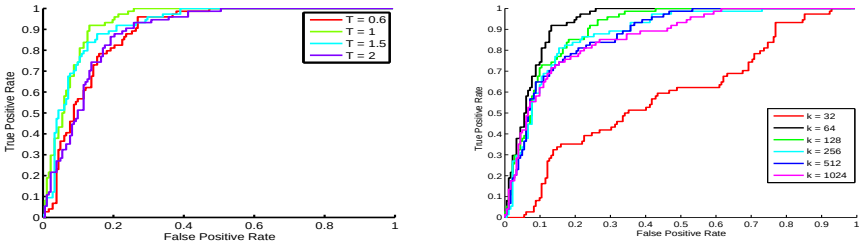
(a) Synchrony (25_S_BM) (b) Non Synchrony (25_NS_NBM)

Fig. 3. Histogram of time-lags

computed the $InterR_n$ for S and NS conditions. We found that the mean and variance of $InterR_n$ were respectively 0.74 and 0.49 for the NS condition and in

average the mean and variance of $InterR_n$ were respectively 0.99 and 0.13 for the S conditions. The S and NS conditions were compared with a Mann-Whitney U-test and the difference between the samples was significant (U=5147,p=7.68e-12). In the S condition, $InterR_n$ is close to 1 and varied less than in the NS condition. Moreover, $InterR_n$ is lesser than 1 in the NS condition showing that the rhythm of partner B is smaller than the rhythm of partner A. This is consistent with our scenario for the NS condition in which partner A was asked to gesture continually while partner B only gestured at the pace of the metronome.

Identical Gestures Detection Module We assessed the measure of distance in the S_BM and S_NBM conditions on the segmented sequences. To test the robustness of the method, the codebook was learned on a different database than the one that serves for testing. This database was constituted with 8 videos of two different subjects performing 5 different actions composed with raising arms and waving sequences. We compared several sizes of codebook $k = 32, 64, 128, 256, 512, 1024$ and several sizes of window to assess the distance $T = 0.6, 1, 1.5$ and 2s. We performed left-tailed t-tests to compare the S_BM and S_NBM conditions. We found that the distance was significantly below in the S_BM condition compared to the S_NBM condition ($p<0.001$) for all k and T. We finally considered a S_BM and S_NBM classification application and drew the ROC curves by varying the threshold on the distance (Fig. 4). The best results were obtained for 64 codewords and windows of 1s. We analyzed the 23 confusions (S_NBM confused for S_BM) corresponding to the best threshold. Among them 9 corresponded to gestures in the same direction but at different levels (e.g.raising arms face /side /up), 4 to partial imitation (one arm performs the same gesture and not the other), 4 were identical gestures, 4 to completely different gestures and 2 were gestures with the same final position but with different initial positions.



(a) Influence of window size. k=64. (b) Influence of codebook size. T=1s.

Fig. 4. Comparaison of the distance measure in the S_BM and S_NBM conditions

5.3 Conclusion

In this paper, we proposed a new framework to assess separately synchrony and behavior matching in dyadic interactions. We proposed several metrics that

discriminate efficiently synchronous from asynchronous situations and behavior matching from non-matching ones. More, assuming the codebook is created with incremental K-means, all the metrics proposed can be computed online given that no prior knowledge or training is required.

References

1. Delaherche, E., Chetouani, M., Mahdhaoui, M., Saint-Georges, C., Viaux, S., Cohen, D.: Interpersonal synchrony : A survey of evaluation methods across disciplines. IEEE Transactions on Affective Computing (2012) to appear.
2. Prepin, K., Pelachaud, C.: Shared understanding and synchrony emergence: Synchrony as an indice of the exchange of meaning between dialog partners. In: ICAART2011 International Conference on Agent and Artificial Intelligence. Volume 2. (January 2011) 25–30
3. Bernieri, F., Rosenthal, R.: Interpersonal coordination: Behavior matching and interactional synchrony. Fundamentals of nonverbal behavior. Cambridge University Press (1991)
4. Sun, X., Nijholt, A.: Multimodal embodied mimicry in interaction. In Esposito, A., Vinciarelli, A., Vicsi, K., Pelachaud, C., Nijholt, A., eds.: Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues. (2011) 147–153
5. Chartrand, T.L., Bargh, J.A.: The chameleon effect: The perception-behavior link and social interaction. Journal of personality and social psychology **76**(6) (June 1999) 893–910
6. Muir, D., Nadel, J.: Infant social perception. In Slater, A., ed.: Perceptual development. Hove: Psychology Press (1998) 247–285
7. Murray, L., Trevarthen, C.: Emotional regulation of interactions between two-month-olds and their mothers. In T.M Field, N.F., ed.: Social perception in infants, Ablex, Norwood, NJ (1985) 177–197
8. Altmann, U.: Studying movement synchrony using time series and regression models. (2011) 23
9. Andry, P., Gaussier, P., Moga, S., Banquet, J., Nadel, J.: Learning and communication in imitation: An autonomous robot perspective. IEEE transactions on Systems, Man and Cybernetics, Part A **31**(5) (2001) 431–444
10. Boucenna, S., Gaussier, P., Andry, P.: What should be taught first: the emotional expression or the face? epirob (2008)
11. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision **73**(2) (June 2007) 213–238
12. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Conference on Computer Vision & Pattern Recognition. (jun 2008)
13. Li, F.F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA (2005) 524–531
14. Dollár, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: In VS-PETS. (2005) 65–72
15. Canu, S., Smola, A.J.: Kernel methods and the exponential family. Neurocomputing **69** (2005) 714–720

16. Schölkopf, B., Platt, J.C., Shawe-Taylor, J.C., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. Neural Comput. **13**(7) (July 2001) 1443–1471
17. Kadri, H., Davy, M., Rabaoui, A., Lachiri, Z., Ellouze, N.: Robust Audio Speaker Segmentation using One Class SVMs. In: Proceedings of the EURASIP EUSIPCO'08, Suisse (2008)
18. Prince, C.G., Hollich, G.J., Helder, N.A., Mislivec, E.J., Reddy, A., Salunke, S., Memon, N.: Taking synchrony seriously: A perceptual level model of infant synchrony detection. In: Proceedings of the Fourth International Workshop on Epigenetic Robotics. (2004) 89–96