Automatic Imitation Assessment in Interaction

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

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

Abstract. Detecting social events such as imitation is identified as key step for the development of socially aware robots. In this paper, we present an unsupervised approach to measure immediate synchronous and asynchronous imitations between two partners. The proposed model is based on two steps: detection of interest points in images and evaluation of similarity between actions. Firstly, spatio-temporal points are detected for an accurate selection of the important information contained in videos. Then bag-of-words models are constructed, describing the visual content of videos. Finally similarity between bag-of-words models is measured with dynamic-time-warping, giving an accurate measure of imitation between partners. Experimental results obtained show that the model is able to discriminate between imitation and non-imitation phases of interactions.

Keywords: Imitation, DTW, unsupervised learning

1 Introduction

Face-to-face interactions are considered as highly dynamic processes [1, 2] based on multimodal exchanges such as turn-taking, backchannels (e.g, head nod, filled pauses...). Sensing, characterizing and modeling interactions are challenging. Various natures of human communication dynamics have to be taken into account: individual (e.g, gesture completion), interpersonal (e.g., mimicking)... In recent years, there has been a growing interest for human communication dynamics in several domains such as Social Signal Processing and Social Robotics. In [3, 4], backchannels are investigated firstly by modeling human-human communication and then the model is employed to generate multimodal feedbacks by an agent (Embodied Communicative Agents/ Robots) during dialogs. Thanks to these dynamical models, agents are able to provide relevant communicative responses and consequently to sustain interactions. Continuously monitoring social exchanges between partners is a fundamental step of social robotics [5].

[*] These authors contributed equally.

Interpersonal dynamics, usually termed interpersonal synchrony [1] is a very complex phenomenon including various concepts such as imitation, mimicking, turn-taking... In this paper, we focus on immediate imitation characterization, which include synchronous and asynchronous reproductions of a demonstrated action within few seconds. Here, imitation is considered as a communicative act allowing to sustain interaction [1]. Being able to automatically assess imitation between social partners (including agents) is required for developing socially intelligent robots [6, 1]. Given this framework, the proposed method is seen to be different to traditional approaches for learning from demonstration in human-robot interaction [7], where imitation metrics usually assess kinematic, dynamic and timing dimensions.

The paper is organized as follows: Section 2 reports recent works on interpersonal synchrony characterization formulated as an action recognition problem. Section 3 briefly describes the approach proposed for imitation assessment. Sections 4 and 5 describe the different steps of our model based on 1) characterization of actions (spatio-temporal interest points, bag-of-words) and 2) similarity metrics (correlation, dynamic time warping). Section 6 presents results on a gesture imitation task. Finally, a conclusion provides a summary of the model discussed throughout the paper and proposes future works.

2 Related Work

Currently, few models have been proposed to capture mimicry in dyadic interactions. Mimicry is usually considered in the larger framework of assessing interactional synchrony, the coordination of movement between individuals in both timing and form during interpersonal communication [8]. Actual state-of-the-art methods to assess synchrony rely on two steps: feature extraction and a measure of similarity.

The first step in computing synchrony is to extract the relevant features of the dyad's motion. We can distinguish between studies focusing on the movement of a single body part and those capturing the overall movement of the dyad. Numerous studies focus on head motion, which can convey emotion, acknowledgement or active participation in an interaction. Head motion is captured using either a motion-tracking device [9] or a video-based tracking algorithm [10, 11]. Many studies capture the global movements of the participants with Motion Energy Images [12, 13] or derivatives [14, 15].

Then, a measure of similarity is applied between the two time series. Correlation is certainly the most commonly used method to evaluate interactional synchrony. After extracting the movement time series of the 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 between the partners [9, 16, 12]. Recurrence analysis is an alternative to correlation [11]. It was inspired by the theory of coupled dynamical systems, providing graphical representations of the dynamics of coupled systems. Recurrence analysis assesses the points in time that two systems show similar patterns of change or movement, called “recurrence points”. These models are often poorly selective for mimicry detection. Indeed, the features (e.g. motion energy) describe rather the amount of movement than the form of the gestures performed. Capturing mimicry entails to have a finer description of the gestures. That can be reached using action recognition techniques.

In the last few years, many researches have emerged in this domain as described in numerous reviews [17], [18], [19]. The first approaches consist to characterize the sequences globally. Davis and Bobick [20] introduced the Motion History Images (MHI) and the Motion Energy Images (MEI) that summarizes in a single image all the motions performed during the sequence. Then, simple moments on these images characterize the sequence. In order to preserve movement kinetic, Mokhber et al. [21] proposed to directly characterize the spatio-temporal volume by geometric moments. Efros et al. [22] characterized individually each image thanks to an optical flow based feature, and then compared the sequences with a measure similar to correlation. Laptev et al. [23] explored the combination of local space-time features histograms and SVM. First, spatio-temporal interest points are detected by extending the Harris detector to the space-time domain. These points are then characterized using several motion representations in terms of spatio-temporal jets, position dependent histograms, position independent histograms, and principal component analysis computed for either spatio-temporal gradients or optic flow. Dollár et al. [24] introduced a new spatio-temporal interest points detector explicitly designed to detect more points and to be more robust. They are then described with spatio-temporal cuboids.

Less works have been made on unsupervised action recognition. Niebles et al. [25] represent a video as a collection of spatial-temporal words by extracting space-time interest points. The algorithm automatically learns the probability distributions of the spatial-temporal words and the intermediate topics corresponding to human action categories. This is achieved by using latent topic models such as the probabilistic Latent Semantic Analysis (pLSA). Rao et al. [26] describe a representation of human action that captures changes in speed and direction of the trajectory using spatio-temporal curvature of 2-D trajectory. Starting without a model, they use this representation for recognition and incremental learning of human actions. Zelnik-Manor and Irani [27] design a simple statistical distance measure between video sequences (possibly of different lengths) based on their behavioral content. It is used to isolate and cluster events within long continuous video sequences, without prior knowledge of the types of events, their models, or their temporal extent.

3 Overview of our approach

In this paper, we propose an innovative approach to measure imitation between two partners, through the use of unsupervised action recognition. Indeed, instead of characterizing a video with global measures by only quantifying movement, we are considering imitation as the similar execution of actions, in which the semantic of actions is not needed. In order to create an imitation rate, as shown in figure 1, the first step is to detect regions of interest, called Spatio-Temporal Interest Points (STIPs) which will be described in section 4.1. They are then described using local histograms, leading to bag-of-words models. Section 5 will give details on the similarity measure which is applied on these models to compare two videos. Finally, as shown in section 6.2 an imitation rate is fitted from the similarity obtained.

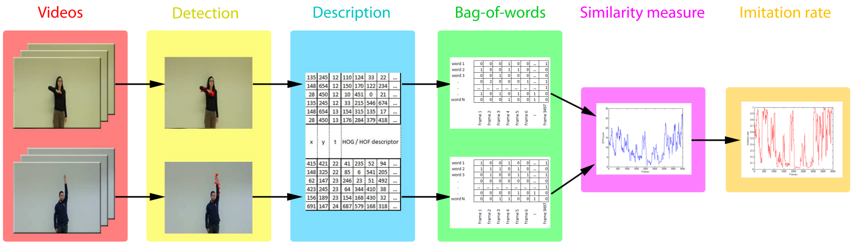

Fig. 1. Video analysis process.

4 Modeling videos with bag-of-words

As it is usually not feasible and impractical to measure correlations between all regions of two videos due to computation limitations, we are using methods that will detect significant areas in the video that are rich in information, also known as Spatio-Temporal Interest Points (STIPs).

4.1 Detection

Detection of the STIPs is based on Dollár's work in [24], where detection of low-level features is performed using 1D Gabor filters. Even though Gabor filters were designed to return high response for periodic motions such as a bird flapping its wings, it has proven to evoke strong response for non-periodic elements such as motion of spatio-temporal corners. It has been preferred to other detectors because of its robustness and for its quantity of correctly detected points (around 12 per frame), allowing a good characterization of the video.

4.2 Description

After obtaining local maxima points from Dollár's response function, each point's spatio-temporal neighborhood (patch) is characterized. However, Dollár's descriptor is based on cuboids, which are expensive both in terms of computation and memory. Indeed, the descriptor uses 19-by-19-by-11 cuboids (9 pixels on each sides and 5 frames before and 5 after), which gives a descriptor with dimension 3971 for each point.

Each point has thus been described by HOG/HOF (Histogram of Oriented Gradients/Histogram of Optical Flow) descriptor as introduced by Laptev in [28]. While HOG has strong similarity to the well known SIFT descriptor, HOF is based on occurrences of orientations of optical flow. Each patch is divided into a grid of cells and for each cell 4-bin HOG and 5-bin HOF histograms are then computed and concatenated into a single feature vector. HOG/HOF descriptors return vectors with 162 elements (72 for HOG and 90 for HOF). HOG/HOF dimensionality is more compact and therefore more convenient than Dollar's cuboid for this application.

4.3 Construction of a vocabulary and bag-of-words model

Similarly to natural language processing, the next step clusters the collection of points in a vocabulary describing the videos. In natural language processing, a cluster would be called a lexical field, thus putting together words like "drive, driving, driver". STIPs are clustered using a k-means technique, forming a k video-words dictionary. The dictionary is formed using training videos.

Then, bag-of-words models are created for each of the two videos. In order to do so, each STIP detected is assigned to the nearest video-word with Euclidean distance. Following representation of each observation as a word in a vocabulary, the entire video is represented as occurrences of words using the bag-of-words model. Every frame is characterized by a histogram of words. This results in a sparse matrix $BOW(w, f)$ of high dimensions : number-of-words by number-of-frames. The bag-of-words models describe the temporal structure of the video, but looses the spatial one. Indeed, the spatial coordinates of a word does not appear in this model, thus the spatial position of a video-word in a frame does not have any impact.

5 Similarity measures

After describing the videos by bag-of-words models, a similarity measure can then be applied to compare them. The first approach coming to mind is to use correlation. However, due to a delay between the imitation initiator and the imitator and to the very sparse nature of bag-of-words, direct correlation is a poor measure. The first idea we present here is to take enlarged analysis windows on which correlation is computed. But since taking the dynamic of the imitation into account is important in interaction, we applied a modified version of Dynamic Time Warping in which similarity is measured.

5.1 Correlation

To allow small variations in time, we represent each instance as a vector and measure sum of words inside a corresponding window as presented in equation

1, where win is the size of the window, w is a word, and BOW_A is the bag-of-words of video A, $a_f(w)$ is the summed vector that defines video A in each instance indexed by the number f of the frame it starts from. Given the fact that a gesture that occurs after more than three seconds can not be considered as imitation, a sliding window of 75 frames has been chosen like shown in figure 5 (since the video frame rate is 25 frames per second).

$$a_f(w) = \sum_{i=f}^{f+win} BOW_A(w,i) \tag{1}$$

After having the corresponding summed vector for each video, the two time series a_f and b_f are compared using normalized correlation coefficient, as recalled in equation 2.

$$normcorr(a_f, b_f) = \frac{a_f^T(w).b_f(w)}{||a_f(w)||_2 * ||b_f(w)||_2} \tag{2}$$

5.2 Dynamic Time Warping

In natural interaction the time-lag of imitation between partners varies all the time and partners continuously change roles. Thus, a straight similarity measurement like correlation is not able to take into account the variations in the time-delay between partners. Thus a dynamic comparison of the imitation between the partners is needed. In this matter, Dynamic Time Warping is a reference to compare two non aligned time series. However, as we are going to present here, we are not using DTW to measure a distance, but to measure similarity.

Whereas the original method from Levenshtein [29] measures a distance between two series, the one developped by Needleman [30] permits to measure similarity and has been widely used with DNA-strand for genome comparison, or sequence alignment. In this last method, the computation of the cumulated similarity matrix follows equation 3, where *normcorr* refers to the normalized correlation defined in equation 2. Detailed explanations on Dynamic Time Warping can be found in Chapter 4 of [31].

$$D(i,j) = \max \begin{cases} D(i-1,j) - (1 - normcorr[x(i-1), x(i)]) \\ D(i,j-1) - (1 - normcorr[y(j-1), y(j)]) \\ D(i-1,j-1) + normcorr[x(i), y(j)] \end{cases} \tag{3}$$

Once the cumulated similarity matrix is computed, the shorter path is searched for. Usually, as one wants to warp two time series together, the matrix is computed such by warping the whole sequence A with the whole sequence B, as shown in figure 2.

But, as illustrated in figure 3, the Overlap Detection variant permits to ignore the beginning of one time series and the end of the other one. This is important for imitation measurement as it permits to have gestures from different lengths, surrounded by gestures which does not fulfill imitation. The only modification to the original algorithm is that the first line and the first column of the cumulated similarity matrix are set to zero, and the maximum score is searched in the whole last column and last line.

The second interesting variation, called Local Alignement, permits to ignore a part of the start and the end for each time series. It is called the Smith-Waterman algorithm (refer [32]). This is illustrated in figure 4. It permits to compare short gestures in two parts of videos. This algorithm is now modified by adding a fourth option to equation 3, so that D(i,j) is the maximum of either those three options or 0. Then if all the numbers are negative, it "resets" the path, permitting to have a new starting point, and leading to results as shown in figure 4. Moreover, the maximum score is now researched in the whole matrix.

During this whole paper, the Local Alignement method has been used.

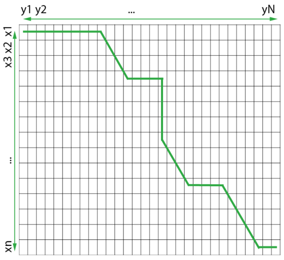

Fig. 2. Classic DTW.

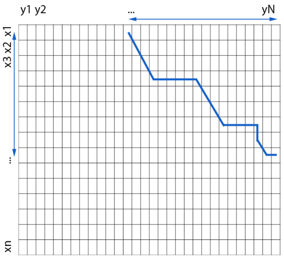

Fig. 3. Overlap DTW.

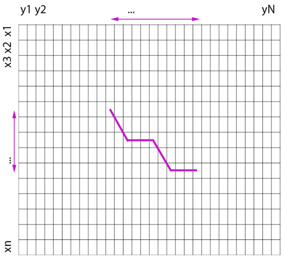

Fig. 4. Local DTW.

In order to allow some time-lag between partners, the two analysis windows can have different lengths. From video A, BOW_A is made with a 75-frames window and is compared with BOW_B made by 125-frames windows (75 plus two times 25 frames), thus allowing a time-lag of one second. The two versions of DTW (75-75 and 75-125) will be compared in the Results section, and are shown in figure 5.

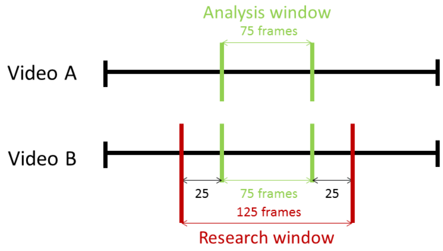

Fig. 5. Windows analysis and research.

6 Experiments

6.1 Data

Current databases often enhance the gesture recognition, or give multiple videos of the same action, done in different contexts. However, few data with synchronized videos are available for interpersonal studies, with annotation. A database of synchronized gestures for two partners has been presented in [33]. Table 1 gives the characteristics of this database, and figures 6 and 7 are illustrations of it.

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 Imitation (S _NBM)	Synchrony and Imitation (S_BM)
20	137[44]	62[19]
25	166[67]	71[28]
30	153[71]	59[27]

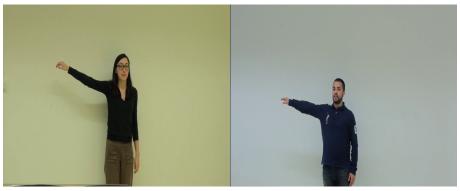

Fig. 6. Imitation dual video.

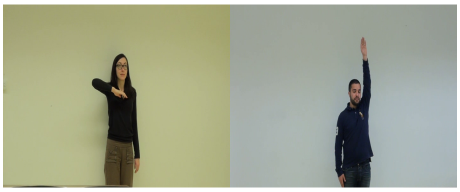

Fig. 7. Non imitation dual video.

6.2 Results

Validation of the protocol : To be sure that the methods actually permits to separate imitation from non-imitation, a first test has been performed with correlation between 75-frames windows. Correlation score is computed at each time for both imitation and non-imitation videos. The results are shown on figure 8 for the two classes, non-imitation videos being longer than imitation videos. Distributions of the two series are shown in figure 9, and a t-test permits to verify that as seen on this figure, the two statistical series are separable (h=1 with p < 0.05). The method is thus suitable for unsupervised imitation measurement.

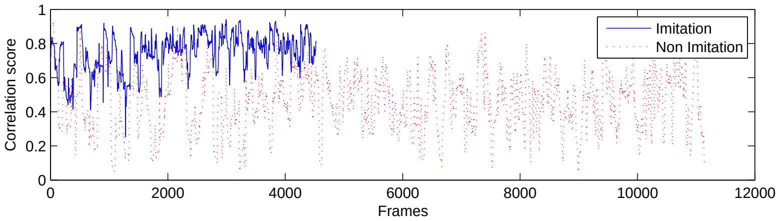

Fig. 8. Correlation score for two imitation videos (solid line) and two non-imitation videos (dotted line).

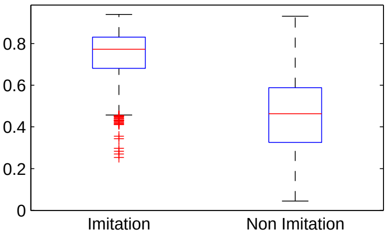

Fig. 9. Distribution of the correlation score.

Comparison of methods : The protocol has been applied to three methods : correlation, DTW 75-75 and DTW 75-125. In order to compare and evaluate the efficiency, Receiver Operating Characteristic (ROC curves) are used. They are estimated using all correlation measures obtained for each time and each video. Results for the three methods are shown in figure 10, where the True Positive Rate (TPR) is plotted as a function of the False Positive Rate (FPR). The Area Under Curve (AUC) is often referred as an efficiency measurement of classifiers.

However, as it can be seen in the figure 10, even if correlation and DTW 75-125 have the same AUC, the curves show better results for DTW 75-125 in the part near the optimal point. This is confirmed by the d' measure, computed by $d' = \max_i [TPR(i) - FPR(i)]$. The d' measure gives the optimal working point, which is significantly higher for DTW 75-125 (figure 12). Moreover, one could note that the results for DTW are not highly superior to correlation. This is explained by the structure of the dataset, which has been created in almost perfect synchrony, and thus where dynamic variations are absent, leading to comparable results for correlation and DTW.

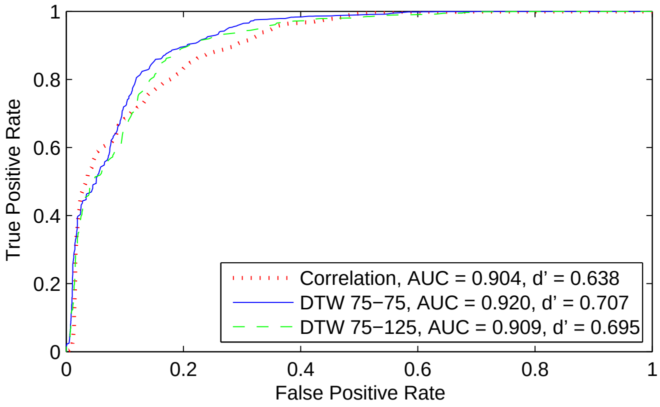

Fig. 10. ROC curves for the three methods.

Robustness : As this protocol is aimed to be used in natural interactions, it has to be robust to shifting. In order to evaluate its robustness, tests have been done by shifting one of the sequences temporally (between -1s and 1s, equivalent to +/- 25 frames). A delay of more than one second has not been envisaged as it cannot be seen as a real imitation. Comparisons were made using ROC, but to summarize results, only AUC and d' measures are presented in figures 11 and 12. Even if DTW 75-75 and correlation seem to have similar results on AUC curves for negative delay, DTW 75-75 outperforms correlation with d' measures, which better represents the real use of the system (near the operating point). The third method DTW 75-125 is robust to delay between sequences which has very little influence on the results. Indeed, as shown in figures 11 and 12, the results are very stable for shiftings between -25 frames and +25 frames.

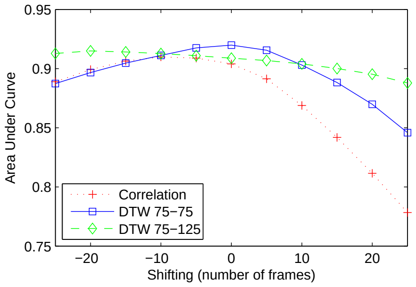

Fig. 11. AUC variations with shifting.

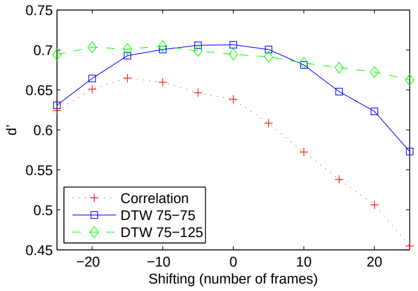

Fig. 12. d' variations with shifting.

7 Conclusion

We have proposed in this paper an efficient process to measure imitation rate during an interaction between partners using unsupervised action recognition. However, the database presented here is strongly synchronized, thus giving more advantage to correlation than to DTW. Moreover, no dynamic appears in the data, such as turn-taking, reducing the interest in terms of interpretation of DTW results. The next step will be to test this method on more natural gestures and interactions in which dynamics will play a large role.

However correlation and DTW each have their interest. Correlation, by its fast computation, permits to have good results as long as data is not too much shifted. On the other hand, even if DTW is a bit more computationally expensive, it permits to take into account the dynamics of interaction, and it will be used in further developments.

Moreover, HOG/HOF descriptor does not take into account the spatial localization of the video-words. Adding information on the relative position of the detected points have been envisaged for the future in order to increase accuracy.

Some further tuning of DTW can be done to improve results, which have not been covered here. The Local Alignement method gave slightly better results than the Overlap Detection, but the influence of the two methods has not been studied here. However, in real interaction videos, some first tests have permitted to see differences between the two algorithms, which will be studied in further developments.

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. Morency, L.P.: Modeling human communication dynamics [social sciences]. Signal Processing Magazine, IEEE **27**(5) (sept. 2010) 112 –116

3. Morency, L.P., Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: Proceedings of the 8th international conference on Intelligent Virtual Agents. IVA '08, Berlin, Heidelberg, Springer-Verlag (2008) 176–190
4. Al Moubayed, S., Baklouti, M., Chetouani, M., Dutoit, T., Mahdhaoui, A., Martin, J.C., Ondas, S., Pelachaud, C., Urbain, J., Yilmaz, M.: Generating robot/agent backchannels during a storytelling experiment. Robotics and Automation, 2009. ICRA '09. IEEE International Conference on (2009) 3749–3754
5. Sidner, C.L., Lee, C., Kidd, C.D., Lesh, N., Rich, C.: Explorations in engagement for humans and robots. Artif. Intell. **166** (August 2005) 140–164
6. Rolf, M., Hanheide, M., Rohlfing, K.: Attention via synchrony : Making use of multimodal cues in social learning. IEEE Trans. Auton. Mental Develop. **1**(1) (2009) 55–67
7. Calinon, S., D'halluin, F., Sauser, E., Caldwell, D., Billard, A.: Learning and reproduction of gestures by imitation: An approach based on Hidden Markov Model and Gaussian Mixture Regression. IEEE Robotics and Automation Magazine **17**(2) (2010) 44–54
8. Bernieri, F., Reznick, J., Rosenthal, R.: Synchrony, pseudo synchrony, and dissynchrony: Measuring the entrainment process in mother-infant interactions. Journal of Personality and Social Psychology **54**(2) (1988) 243–253
9. Ashenfelter, K.T., Boker, S.M., Waddell, J.R., Vitanov, N.: Spatiotemporal symmetry and multifractal structure of head movements during dyadic conversation. J Exp Psychol Hum Percept Perform **35**(4) (2009) 1072–91
10. Campbell, N.: Multimodal processing of discourse information; the effect of synchrony. 2008 Second International Symposium on Universal Communication (2008) 12–15
11. Varni, G., Volpe, G., Camurri, A.: A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. Multimedia, IEEE Transactions on **12**(6) (October 2010) 576 –590
12. Altmann, U.: Studying movement synchrony using time series and regression models. (2011) 23
13. Ramseyer, F., Tschacher, W.: Nonverbal synchrony in psychotherapy: Coordinated body movement reflects relationship quality and outcome. Journal of Consulting and Clinical Psychology **79**(3) (2011) 284 – 295
14. Delaherche, E., Chetouani, M.: Multimodal coordination: exploring relevant features and measures. In: Second International Workshop on Social Signal Processing, ACM Multimedia 2010. (2010)
15. Sun, X., Truong, K.P., Pantic, M., Nijholt, A.: Towards visual and vocal mimicry recognition in human-human interactions. In Tunstel, E., Nahavandi, S., Stoica, A., eds.: IEEE International Conference on Systems, Man, and Cybernetics, SMC 2011: Special Session on Social Signal Processing, USA, IEEE Computer Society (November 2011) 367–373
16. Boker, S.M., Xu, M., Rotondo, J.L., King, K.: Windowed cross-correlation and peak picking for the analysis of variability in the association between behavioral time series,. Psychological Methods **7**(3) (2002) 338 – 355
17. Poppe, R.: A survey on vision-based human action recognition. Image and Vision Computing **28**(6) (2010) 976 – 990
18. Turaga, P., Chellappa, R., Subrahmanian, V., Udrea, O.: Machine recognition of human activities: A survey. Circuits and Systems for Video Technology, IEEE Transactions on **18**(11) (November 2008) 1473 –1488

19. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. Comput. Vis. Image Underst. **104**(2) (November 2006) 90–126
20. Bobick, A., Davis, J.: The recognition of human movement using temporal templates. Pattern Analysis and Machine Intelligence, IEEE Transactions on **23**(3) (March 2001) 257 –267
21. Mokhber, A., Achard, C., Milgram, M.: Recognition of human behavior by space-time silhouette characterization. Pattern Recognition Letters **29**(1) (2008) 81 – 89
22. Efros, A., Berg, A., Mori, G., Malik, J.: Recognizing action at a distance. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. (October 2003) 726 –733 vol.2
23. Laptev, I., Lindeberg, T.: Local descriptors for spatio-temporal recognition. In: In First International Workshop on Spatial Coherence for Visual Motion Analysis. (2004)
24. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, IEEE (October 2005) 65– 72
25. Niebles, J.C., Wang, H., Fei-fei, L.: Unsupervised learning of human action categories using spatial-temporal words. British Machine Vision Conference (BMVC) **3** (2006) 1249 – 1258
26. Rao, C., Yilmaz, A., Shah, M.: View-Invariant Representation and Recognition of Actions. (2002)
27. Zelnik-Manor, L., Irani, M.: Event-based analysis of video. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001. Volume 2., IEEE (2001) II–123– II–130 vol.2
28. Laptev, I., Lindeberg, T.: Space-time interest points. In: IN ICCV. (2003) 432–439
29. Levenshtein, V.: Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady **10** (1966) 707
30. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology **48**(3) (1970) 443 – 453
31. Müller, M.: Information Retrieval for Music and Motion. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2007)
32. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of molecular biology **147**(1) (March 1981) 195 – 197
33. Delaherche, E., Boucenna, S., Karp, K., Michelet, S., Achard, C., Chetouani, M.: Social coordination assessment : Distinguishing between form and timing. In: Multimodal pattern recognition of social signals in human computer interaction. (2012) Submitted.