A Study of Human Interactions By Aligning Behavior Curves

Sagar Kukreja
Advisor: Dr. Ifeoma Nwogu
Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14586
sk3126@cs.rit.edu

***Abstract*—Facial expressions and body movements play an important role in interpersonal communication. These are one of the many key factors that determine the interest and engagement of a person in a communication. Studying facial features and body movements can provide important information about the level of interactional synchrony between individuals. In this work, using the Sayette Group Formation Task (GFT) database which comprises of interacting groups of triads, we computationally detect and measure the extent of interactional synchrony using visual cues only.**

***Index Terms*—Synchrony; Interaction; Pose Tracking**

I. Introduction

People are inherently social. Social interactions play a pivotal role in human behavior.Accessing human social interaction, especially *synchrony*, provides a better understanding of the human behavior. Synchrony refers to the temporal structure of behaviors among interactive partners. The close connection between synchrony and interaction provides researchers promising perspectives to build social interfaces, robots or conversational agents. However, the lack of automatic tools for synchrony discovery limits the exploration in interactive abilities.

In this work, we try to address two questions:
1. Can we computationally detect and measure the extent of interactional synchrony in a conversation involving two or more people, and if so how well do these computational measures compare with human perception of synchrony.
2. What is the best combination of behavioral features that will yield the most human-like measures of interactional synchrony.

A. Motivation

Upto two-thirds of human communication occurs via nonverbal channels such as body movements, facial expression and affective speech prosody. Interactional Synchrony plays a significant role in maintaining healthy social relationships among individuals as it indicates affiliation and feelings of empathy. Understanding the pattern for interactional synchrony can be used to decipher deceptive conversations and can also be leveraged to know the nature of relationships in various social settings.

II. Related Work

With the recent advancements in deep learning methods and computational power, affective computing research has progressed significantly in last decade.

Delaherche et al. [1] presented an extensive survey of synchrony evaluation from a multidisciplinary perspective, focusing on psychologists coding methods, non-computational evaluation and early machine learning techniques. Brand [2] presented a coupled hidden Markov model (cHMM) where the current state is dependent on the states of its own chain and that of a neighboring at the previous time-step and the model was used to classify taichi movements, under the assumption that different parts of the body moving in taichi will be synchronous to each other. The model was used for a 2-class classification of taichi movements. Li et al. [3] presented supervised model used to predict the outcomes of video-conferencing conversations in the context of new recruit negotiations - the model was designed on the assumption that the negotiating parties would exhibit some level of synchrony and modeling this enabled them to predict negotiation outcomes. The drawback of this technique is that it could not provide a useful metric to evaluate the extent of synchrony between the negotiating parties.
Lastly, [4] presented a technique to investigate interactive synchrony in facial expressions and showed using the Pearsons correlation measure, that synchrony features were effective at detecting deception. We expand on this work to develop more robust metrics for interactional synchrony. We provide a comprehensive study to identify the most suitable behavior features.

III. Method

In this section, we present the methods used for feature extraction, feature selection and synchrony measurement via curve matching. We also present the process of obtaining a form of ground truth since there are no natural labels provided for the data.

A. Feature Extraction

The main set of features used in this work are body joint movements.We used an open source to obtain the features for

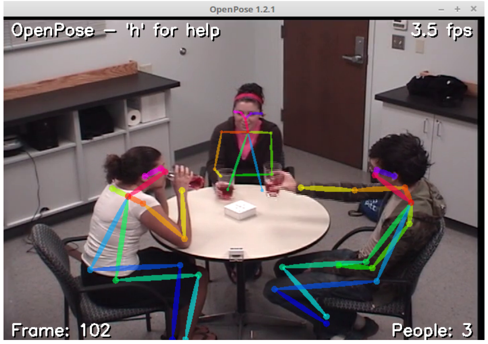


Fig. 1. Openpose output on a sample frame

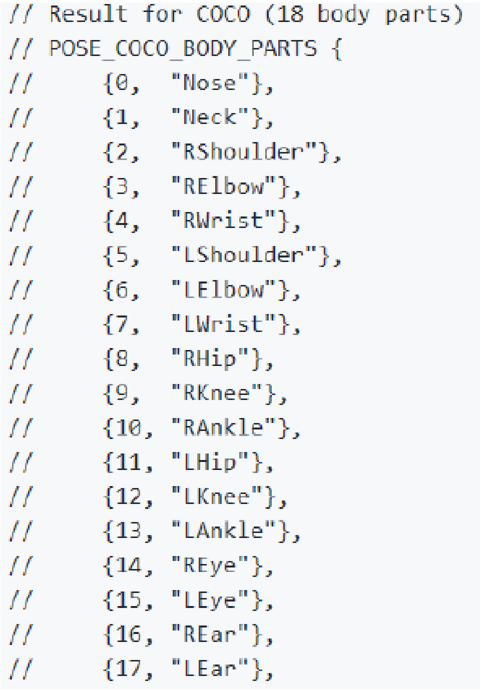
body joints.

1) Openpose: Openpose [5] is a real-time multi-person system used to detect human body, hands and facial keypoints on images and videos. It learns to associate body parts with individuals in image to detect the 2D pose of multiple people in image using Part Affinity Fields(PAFs). It takes image as input and outputs an array of objects. Each object contains an array of pose keypoints for each person in image containing the body part locations and detection confidence.

Openpose uses bottom-up representation of association scores via Part Affinity Fields, a set of 2D vector fields that encode the location and orientation of limbs over the image domain. This system takes color image as input and produces 2D locations of anatomical keypoints (Figure 2) for each person in the image. A feed-forward neural network simultaneously predicts 2D confidence maps of body part locations and a set of vector fields of part affinities, which encode the degree of association between parts.

The architecture [6] of Openpose comprises of a two-branch multi-stage convolutional neural network . A convolution neural network analyzes the input image which generates a set of feature maps. These feature maps are input to the first stage of each branch. Confidence maps are predicted by each stage in the first branch and PAFs in the second branch. In each subsequent stage, original image features and predictions from both stages are concatenated and used to produce refined predictions. Figure 1 shows an illustration of the Openpose output when applied to one of our sample videos.

2) Feature Selection: Openpose produces 2D locations and confidence value for 18 keypoints of a body model. For our dataset, we used only the keypoints from upper body parts as subjects in our dataset are sitting around the table and not moving around; keypoints from lower body, in our case, are not really relevant to study the synchrony among them,especially since the camera could not always have access to the lower body often occuleded by the table or other participants. We then run Correlation and Dynamic Time Warping algorithm separately on x and y coordinates of keypoints. Calculations for these algorithms are done over pairs of individuals.



```
// Result for COCO (18 body parts)
// POSE_COCO_BODY_PARTS {
//     {0,  "Nose"},
//     {1,  "Neck"},
//     {2,  "RShoulder"},
//     {3,  "RElbow"},
//     {4,  "RWrist"},
//     {5,  "LShoulder"},
//     {6,  "LElbow"},
//     {7,  "LWrist"},
//     {8,  "RHip"},
//     {9,  "RKnee"},
//     {10, "RAnkle"},
//     {11, "LHip"},
//     {12, "LKnee"},
//     {13, "LAnkle"},
//     {14, "REye"},
//     {15, "LEye"},
//     {16, "REar"},
//     {17, "LEar"},
```

Fig. 2. The 18 joints captured by OpenPose.We only use the top 7 (except the nose), to represent the joints in the upper body

3) Metrics for synchrony:

Correlation coefficient - Baseline: Correlation quantifies the strength of a linear relationship between two variables. When there is no correlation between two variables, then there is no tendency for the values of the variables to increase or decrease in tandem. Two variables that are uncorrelated are not necessarily independent, however, because they might have a nonlinear relationship.

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.

If two signals are correlated but have a time delay or latency d between them, the latency can be accounted for in form of a sliding window of size d. The correlation coefficient with latency d between a pair of sequences X and Y is therefore given as:

$$r(d) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_{i-d} - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_{i-d} - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of X and Y.

Dynamic Time Warping

Dynamic Time Warping(DTW) is one of the algorithms for measuring similarity between two temporal sequences, which may vary in speed. For instance, similarities in walking could be detected using DTW, even if one person was walking faster than the other, or if there were accelerations and decelerations during the course of an observation. DTW has been applied

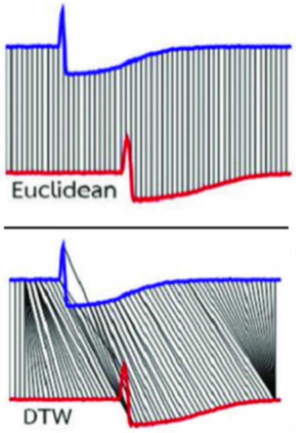


Fig. 3. Top part shows non-correspondence matching while the bottom part shows DTW-based correspondence matching

to temporal sequences of video, audio, and graphics data indeed, any data that can be turned into a linear sequence can be analyzed with DTW. A well known application has been automatic speech recognition, to cope with different speaking speeds. Other applications include speaker recognition and online signature recognition. Also it is seen that it can be used in partial shape matching application.

In general, DTW is a method that calculates an optimal match between two given sequences (e.g. time series) with certain restrictions. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in time series classification. Although DTW measures a distance-like quantity between two given sequences, it doesn't guarantee the triangle inequality to hold.

Figure 3 illustrates the usefulness of DTW when computing the similarity between two sets of sequential points. In Figure 3 top, DTW is not used and the points are simply matched in the order they are recorded, without attempting to search for the most aligned points, which results in a low similarity score. The bottom of Figure 3 shows how DTW matches the most corresponding points, with the result that the two sequences will now have a very high similarity score, since all the points in one sequence have a match in the other. One main advantage with this technique is that we do not explicitly choose a time-window, rather the further away a point is from its corresponding match in the other set of points, the higher the cost incurred when computing the similarity between the two sets of points. Hence when interactional synchrony is present even if the sequences are not exactly aligned, the DTW-based similarity score will be higher, whereas when there is no synchrony, the similarity will be low. We therefore interpret DTW alignment cost as inverse similarity or inverse-synchrony.

Fig. 4. Examples of Video Frames from Dataset

IV. Evaluations and Results

Here, we describe the nature of the data we used for our evaluations and the results obtained when different metrics were applied for measuring synchrony on different visual cues obtained from data. The goal here is it perform a comparative analysis in order to determine a useful combination of behavioral cues and synchrony metrics that will yield the most human-like measures of synchrony.

A. Dataset

The data for this study was obtained from Girard et al. [7] which was drawn from a larger study on the impact of alcohol on group formation processes. All participants in the study met for the first time at the experiment. They were instructed to consume a beverage and then engage with two other participants. The groups on interlocutors were made up of three such subjects who were engaged in about 30-40 minutes of unstructured interactions. For many participants we viewed, this was a rather awkward social setting, hence we consider this as a somewhat abnormal social setting. The data collectors focused on a one minute portion of the entire videos where they believed the participants in a group had become sufficiently acquainted with each other. Separate wall-mounted cameras faced each participant and another camera captured the overall group interaction. For each group, there were four videos, one at the overall group level and three at the individual level. Figure 4 shows the video frames from the overall group interaction as well as the individual participant face videos.

There were a total of 96 participants in this dataset (42% female, 85% white) and all of them consented to having their audiovisual data used in further experiments. They were split in groups of three participants, with 23 mixed-gender groups and 9 same groups, resulting in a total of 32 groups. Groups were randomly assigned to drink an alcoholic beverage (n = 9), a placebo beverage (n = 8), or a nonalcoholic control beverage (n = 15) during the experiment; all participants in a group drank the same type of beverage. Unfortunately, for this work we could not obtain the group labels in order to determine which groups drank what class of beverages. But we compensated for this lack of ground truth as explained in next section.

B. Data Cleaning and formatting

We extract the features from above mentioned dataset using openpose. For each person in video we get 18 feature points. We take seven upper-body feature points and apply DTW and correlation to get the results. But one of the main drawbacks of openpose is that it sometimes interchanges the coordinates of the people in the video. Hence, the data was pre-processed before any algorithm can be applied to it.

C. Human Annotations

In the absence of ground truth data, we requested five individuals to review the videos in the dataset and provide an average group synchrony score based on their perception of how well the group was interacting. The scores were requested to be in the range of 1 to 5, with 1 being the group was not in sync, 5 being completely synchronized and everything else in between. They were instructed to judge overall synchrony so that even if two people in the group appears to be interacting very well with each other, but not with the third person in the group, they could not give the group a high score. The best would be a medium score but again at their own discretion. To account for the subjectivity in the dataset, we computed the total variances for the scores provided by all but one of the labelers, for all five labelers and removed the set of annotations that caused the largest variance in the set. We also spot-checked the variances across each group and when this was larger than a preset threshold, we had additional labelers re-score the video - this was done for only two groups in the entire dataset. The average score obtained from all the labelers was now considered as the human impression of synchrony, the new gold-standard we aim to reach.

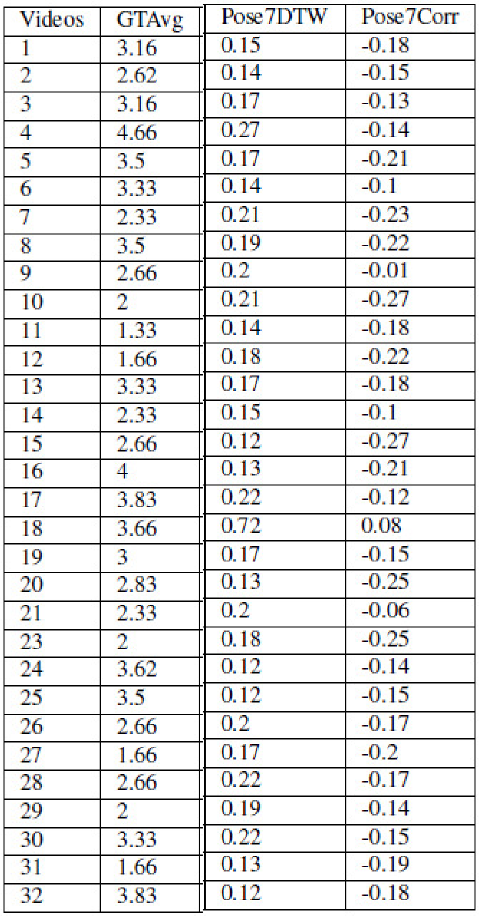
D. Computing synchrony metrics

We compute the correlation coefficients and DTW costs separately on the coordinates of the seven upper body keypoints. The following describe the different measurements computed:

1. Pose7DTW : We considered 7 upper body joints of people present in the video across large number of frames and then calculated the DTW alignment cost between each pair of individuals, i.e Individual A-B , Individual B-c, Individual A-c. We then sum up the costs for all 3 pairs of individuals to get one aggregate value for each group. We do this for each of the 32 videos.

2. Pose7Corr : We perform the same procedure as discussed above but instead compute the correlation coefficient. We sum across the three coefficient values to get one aggregate value for each group. We do this for each of the 32 videos.

The respective coefficient and alignment results are shown in Table 1.



Videos	GTAvg	Pose7DTW	Pose7Corr
1	3.16	0.15	-0.18
2	2.62	0.14	-0.15
3	3.16	0.17	-0.13
4	4.66	0.27	-0.14
5	3.5	0.17	-0.21
6	3.33	0.14	-0.1
7	2.33	0.21	-0.23
8	3.5	0.19	-0.22
9	2.66	0.2	-0.01
10	2	0.21	-0.27
11	1.33	0.14	-0.18
12	1.66	0.18	-0.22
13	3.33	0.17	-0.18
14	2.33	0.15	-0.1
15	2.66	0.12	-0.27
16	4	0.13	-0.21
17	3.83	0.22	-0.12
18	3.66	0.72	0.08
19	3	0.17	-0.15
20	2.83	0.13	-0.25
21	2.33	0.2	-0.06
23	2	0.18	-0.25
24	3.62	0.12	-0.14
25	3.5	0.12	-0.15
26	2.66	0.2	-0.17
27	1.66	0.17	-0.2
28	2.66	0.22	-0.17
29	2	0.19	-0.14
30	3.33	0.22	-0.15
31	1.66	0.13	-0.19
32	3.83	0.12	-0.18

E. Comparisons with ground-truth

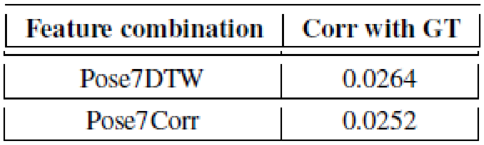
In oder to compare different metrics calculated with the annotators ground truth (GT) values, we again run a correlation between GT and the metrics computed previously. Figure 5 shows the correlation results. Table 2 shows the resulting correlation values indicating how well the metrics match up with ground truth.

Table 1. Comprehensive set of scores computed for the different feature combinations and synchrony metrics

V. Discussion and limitations

From the comparisons provided in the work, we observe that correlation, although a relatively simple comparison metric, gives the closest to human-like synchrony results when compared to DTW. While comparing it with annotated ground truth it gave a correlation value of 0.2634 and 0.2052 for upper body features. Correlation of body features can therefore be used as a quantifiable and repeatable metric for measuring interactional synchrony.

From our observations in the videos, several groups which were considered to be unsynchronized (scoring only a 1 or 2) by most or all of the human labelers involve participants who did not move much at all. To the human eye, it is clear that such a group or pair of subjects is unsynchronized. But to



Feature combination	Corr with GT
Pose7DTW	0.0264
Pose7Corr	0.0252

Table 2. Correlations with human-annotated synchrony scores

the computational methods, when two curves both show little or no movement, this indicates high synchrony. It is therefore possible that the alignment techniques assign low costs (high synchrony) to such data. It might be useful going forward to adjust the algorithms to penalize the costs when the curve shows little or no movements.

Some other limitations of the work include (1) the size of the dataset - deep networks such as variants of recursive neural nets, which have been shown to be successful for prediction tasks, could have been used with a regression cost function to learn the interactional synchrony metrics from pairs of behavioral features, but the overall dataset is rather small to successfully train such a network. (3) Probabilistic graphical models (PGMs) have been used in the past to successfully model synchrony for prediction purposes. But being stochastic in nature, PGMs will not readily provide reliable and repeatable metrics for interactional synchrony across different social constellations.

VI. Conclusion

In this work we discussed several computational methods like DTW and correlation coefficient to detect and measure the extent of interactional synchrony in a social engagement involving two or more interlocutors using visual cues alone. We have also demonstrated that at least one of the metrics we explored correlates well with human-annotated synchrony measures, leading us to the conclusion that the combination of body pose features along with the correlation coefficient as the metric results in a human-like measure of synchrony.

Acknowledgment

Many thanks to Jeffery Girard of University of Pittsburgh for providing us with the Sayette group formation dataset on which we ran these tests and Shravya Bhandari, student at Rochester Institute of Technology for her contributions towards this work.

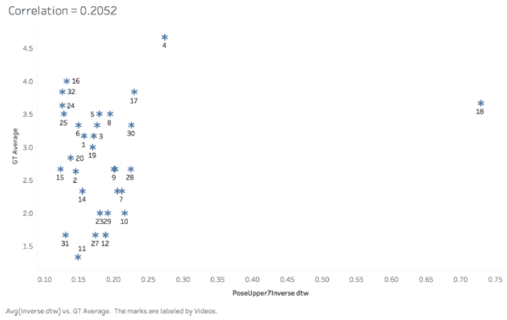


Fig. 5. 7 upper body keypoints using the inverse of DTW.

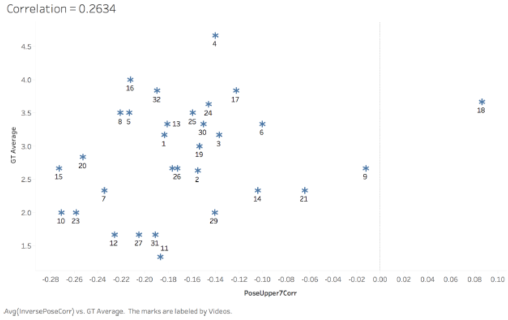


Fig. 6. 7 upper body keypoints using corr. coeff

References

[1] 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. Affect. Comput.*, vol. 3, no. 3, pp. 349–365, Jul. 2012. [Online]. Available: http://dx.doi.org/10.1109/T-AFFC.2012.12

[2] M. Brand, "Coupled hidden markov models for modeling interacting processes," Tech. Rep., 1997.

[3] R. Li, J. Curhan, and M. E. Hoque, "Predicting video-conferencing conversation outcomes based on modeling facial expression synchronization," *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, pp. 1–6, 2015.

[4] 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 *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–6.

[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[6] J. M. Girard, W.-S. Chu, L. A. Jeni, J. F. Cohn, and F. De la Torre, "Sayette group formation task (gft) spontaneous facial expression database," in *2017 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2017.

[7] M. Sayette, K. Creswell, J. Dimoff, C. Fairbairn, J. Cohn, B. Heckman, T. Kirchner, J. Levine, and R. Moreland, "Alcohol and group formation: A multimodal investigation of the effects of alcohol on emotion and social bonding," *Psychological Science*, vol. 23, no. 8, pp. 869–878, 1 2012.