D3.5 - Data acquisition and analytic tools for joint actions

Project No	GA824160
Project Acronym	EnTimeMent
Project full title	ENtrainment & synchronization at multiple TIME scales in the MENTal foundations of expressive gesture
Instrument	FET Proactive
Type of action	RIA
Start Date of project	1 January 2019
Duration	48 months

Distribution level	[PU][1]
Due date of deliverable	
Actual submission date	
Deliverable number	3.5
Deliverable title	Data acquisition analytic tools for joint actions
Type	ORDP (Open Research Data Pilot)
Status & version	Draft
Number of pages	15
WP contributing to the deliverable	3
WP / Task responsible	UNIGE / UM_CM
Other contributors	
Author(s)	UM_CN, UNIGE, KTH and Euromov
EC Project Officer	Teresa De Martino
Keywords	Computational models, Data collection, Machine learning, Movement analysis, Software libraries, Sonification

[1] **PU** = Public, **PP** = Restricted to other programme participants (including the Commission Services), **RE** = Restricted to a group specified by the consortium (including the Commission Services), **CO** = Confidential, only for members of the consortium (including the Commission Services).

Contents

Abbreviations

EU	European Union
EC	European Commission
IMU	Inertia Measurement Unit
WP	Work Package

1. Introduction

This deliverable describes the hardware and software tools used for acquiring and analysing data in the context of joint actions. This deliverable is strongly connected to deliverable D3.1 (Phase 1) which is about the hardware and software tools used for acquiring and analysing data in the context of single actions. As such, this document focuses only on the tools of sub projects specific for joint actions.

2. Deep neural network-based movement prediction

KTH has conducted a number of experiments on so-called conversational groups (Yang et al., 2020) in an effort to find solutions based on machine learning for movements to be represented in more compact forms, for movement qualities over time horizons of up to ten seconds to be predicted, and ideally for new movements to be generated, such as the movement of an avatar in virtual reality.

Movement prediction for conversational groups

With the help of 40 human subjects (27F/13M, 22-35 years old), motion capture sequences of a newcomer approaching a group of individuals engaged in a conversation game were collected. Without the subjects being told in advance how to behave, the sequences were later annotated as either Accommodating or Ignoring depending on the behaviour of the group as a whole with respect to the newcomer. Two examples of avatars generated from the captured data can be seen in Fig. 1.

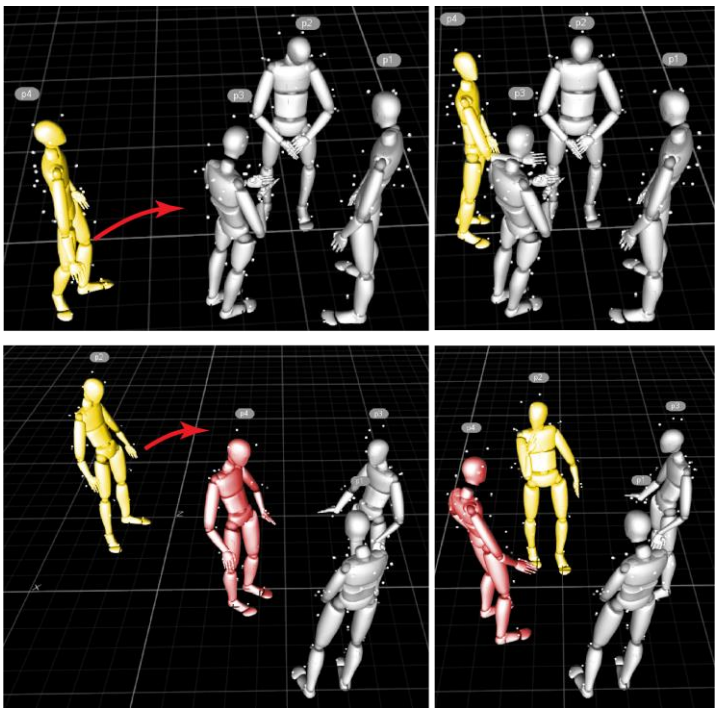

Figure 1. Group approach behaviors with a newcomer being accommodated (above) and ignored (below) as he/she is approaching the group.

A number of machine learning methods for human movement prediction were then tested, based on their ability to model and predict the behaviour of the group. The baseline was an attention-based method (AGNet) that focuses on particular body-markers and events in time that are most indicative, which leads to a representation in which the spatial and temporal dimensions are no longer explicitly represented. Another proposed and tested method was an attention-based method (AGTransformer) that uses so-called Transformer networks that are better at capturing the temporal relations between indicative events.

A potential problem with methods based on attention, at least for some applications, is the reliance on local events that are indicative of the property you like to capture, but in some cases, the overall movement might be what is most indicative. We thus explored Graph Convolutional Networks (GCNs) for the sake of movement representation by extending such networks to the temporal domain. Using a neighbourhood system defined over the set of body markers over time and space, information is gradually spread over a number of layers and stages. By doing so information relevant for the final task is enhanced, while the structure of the sequence in terms of body markers is preserved until the latest possible stage, where a prediction is made.

As can be seen to the left of Fig. 2, information is propagated in a number of stages with the results from each stage concatenated into a long feature vector. Since the spatiotemporal receptive field increases for each stage, the representation is able to capture movements at different temporal resolutions, which is beneficial if the relevant resolution for a particular property of interest is unknown. Each stage consists of a number of layers with information spread spatially (S-GCN) over the body-markers and temporarily (TCN) in an interleaved fashion. For the temporal domain dilated convolutions are used since it allows for properties to be captured over gradually increasing time horizons.

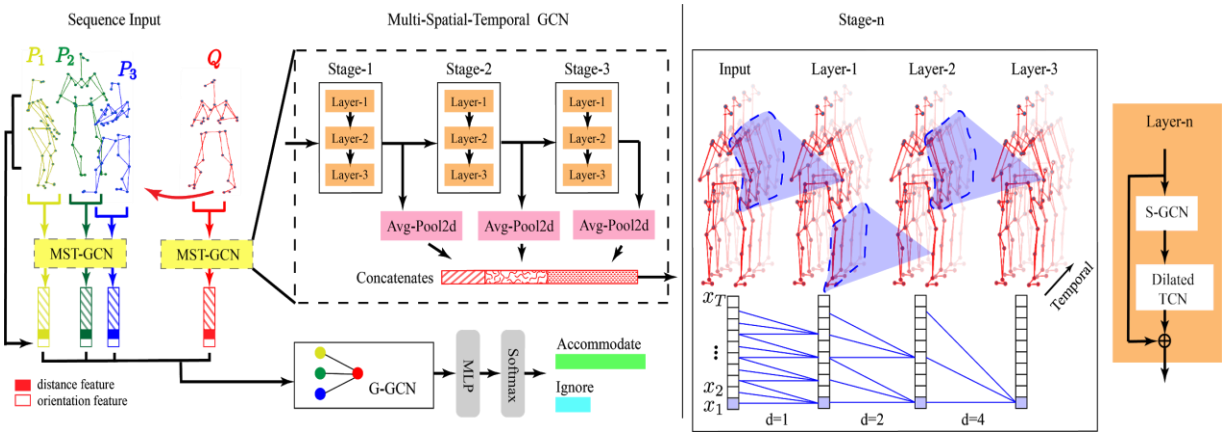

Figure 2. The movement of each agent is represented by a Multi-Spatial-Temporal GCN (MST-GCN), with multiple

such representations combined with a group GCN to model the behaviour of the group as a whole and to predict whether the group is Accommodating or Ignoring the newcomer.

For the task of group behaviour prediction, experiments showed that, with an accuracy of 91.5%, MST-GCN was superior to the tested attention-based methods, among which AGTransformer (79.1%) outperformed AGNet (70.3%). A GCN on group level also showed to be preferable from having an attentional mechanism and a multi-temporal representation improved compared to using a single temporal scale. The best possible results (93.0%) were achieved by also encoding the relative orientation and distance of the agents' heads with respect to the newcomer, but that leads to a considerably less generic representation.

Movement prediction for human-centered collaborative robots

MST-GCN was adopted for a human-centered collaborative robot system (Ghadirzadeh et al., 2020) with which a robot learns to act in accordance with the movement of a human collaborator. The neural network-based framework used by the robot to learn appropriate behaviours can be seen in Fig. 3. The upper half shows an encoder-decoder structure applied for representation learning. Instead of training MST-GCN in a supervised manner to predict behaviours, such as for the conversational groups above, the encoder-decoder is trained to reconstruct the movement data (*bi)* on the input, which for the experiments were 25 frames of motion capture data. Thus, a more compact representation can be learned in a supervised manner, without requiring any annotations.

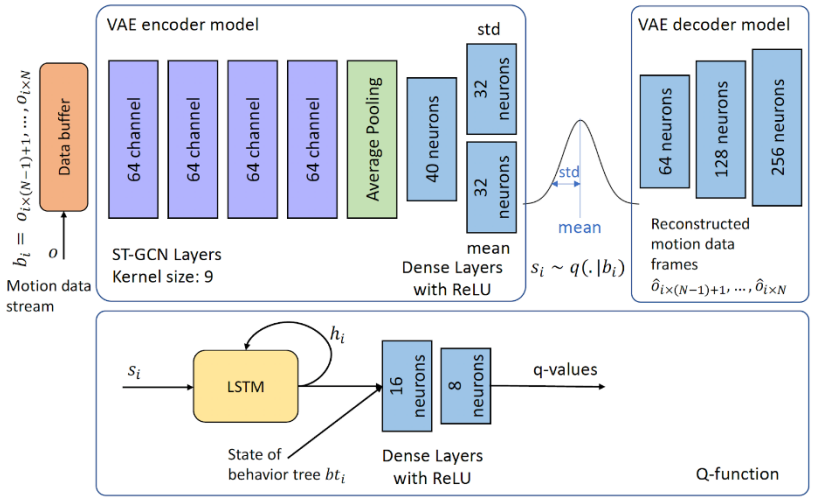

Figure 3. A reinforcement learning based system for a robot to learn to act in accordance with a human collaborator.

The encoder is represented by an MST-GCN followed by a couple of fully connected layers and the decoder by another couple of fully connected layers. Given that the waist of the network consists of only 32 neurons, the network forces the movement data to be represented by a latent space representation *(si)* of a much lower dimensionality, which reduces the redundancy of the

original movement data. In practice, a so-called variational autoencoder is used for the purpose, which regularizes the distribution in the latent space forcing it to be Gaussian. This means that each point in the latent space will appear to be meaningful, if it is reconstructed back into the space of motion capture data, which in turn means that the network used to train a policy for controlling the movement of the robot only needs to consider human movements that appear in practice.

The lower part of the network represents a Q-function that is used for reinforcement learning to learn the policy, given the current state of the world. This state consists of two parts; the latent representation *(si)* of the movement and the state of a behaviour tree *(bti)* that represents how far the partners have come in the collaborative task. Before the latent representation is used, however, it is passed through a recurrent neural network, an LSTM, that further extends the window in time over which movements are considered. This allows the system to be less dependent on the particular choice of window for MST-GCN. The output of the full network, the q-values, represents the predicted accumulated rewards for a discrete set of possible next actions, actions such as “pick up”, “wait”, etc. Once the policy is applied for robot control, the action with the highest q-value is selected in each step of the interaction.

The learning system was tested for a collaborative box packing task, shown in Fig. 4. The study involved 7 human subjects (3F/4M, 24-32 years old) with 430 sessions recorded in total. By observing the movements of a human collaborator, the robot tries to predict what objects to pick up, where to place them and when the box is finished and ready for delivery. Rewards were defined as the time saved by the robot being proactive, instead of letting the robot wait for the human movement to end before it decides what to do. In the graph to the right of Fig. 4, the average reward can be seen as training progresses, with and without supervision. When the robot is supervised, it always knows what the human partner is about to do. Without supervision the partner’s intentions have to be inferred from the movement data. Given the similarity between the two graphs, it can be concluded that the movement data, when compacted into the latent space representation, is rich enough for the robot to infer the intentions of its partner and learn the task, since no supervision was needed.

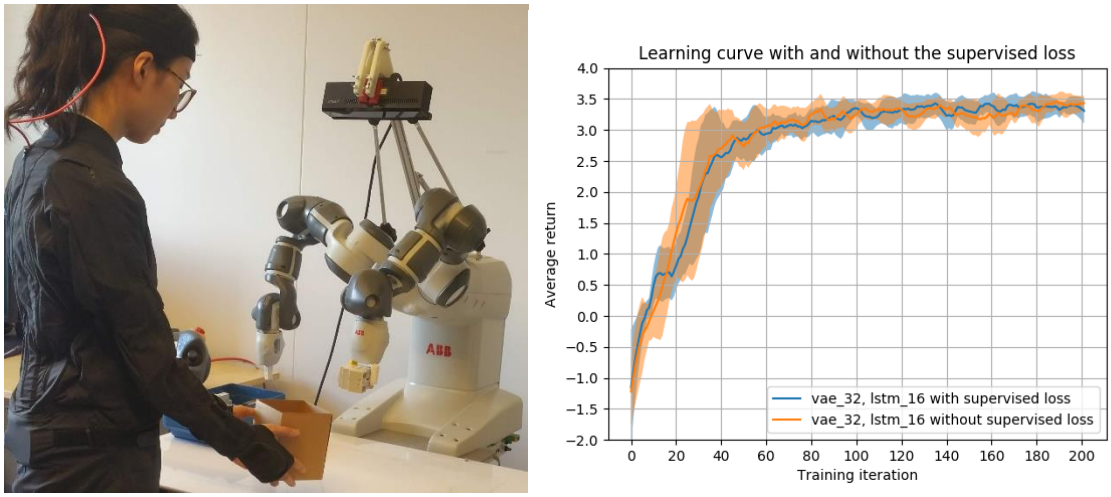

Figure 4: A collaborative box packing task (left) and the learning curve, with and without supervision (right). The indicated average return corresponds to the time saved per interaction step by the robot acting proactively. With supervision the robot knows what the real intentions of the human subject are, but without it, intentions have to be inferred.

References

F. Yang, W. Yin, T. Inamura, M. Björkman, and C. Peters, "Group behavior recognition using attention- and graph-based neural networks", in Proc. European Conference on Artificial Intelligence, 2020.

A. Ghadirzadeh, X. Chen, W.Yin, Z. Yi, M. Björkman, and D. Kragic, "Human-centered collaborative robots with deep reinforcement learning", IEEE Robotics and Automation Letters, vol. 6, no. 2, pp. 566-571, 2020.

3. Group synchronisation

During M7-M20 EuroMov has developed tools and analytical methods for computation of group synchronisation metrics and modelling in different sensorimotor scenarios. Humans interact in groups through various perception and action channels. The continuity of interaction despite a transient loss of perceptual contact often exists and contributes to goal achievement. Here we report a modelling framework used to capture the persistence of synchronization during **group-based pendulum swinging** in novices (n=7) and experts (n=7) when visual coupling is suddenly lost (for rationale please see 2.3.2 Dance to Sync study in D1.2 and D2.5, Bardy et al. 2020).

Data processing. Each pendulum was equipped with a calibrated analog potentiometer to record its angular motion at f_s = 200 Hz. The acquisition was performed using the Matlab software, recording the signals of the seven pendulums simultaneously. The position time series were then smoothed out through a Moving Average filter with a time window of 10 samples (Δt_w = 0.05 s). The Hilbert transform method was applied on the filtered positions to extract the time series of the phases.

Data analysis and relevant metrics. Denoting *T* as the number of samples in each trial and *N* as the number of players, we can define $\theta_i(k)$ as the phase of the *i*-th pendulum at the *k*-th sampling instant, for all *i* = 1,…,N and *k* = 1,…, T. The following set of metrics were used to capture the relevant features of the human group interactions recorded in our experiments:

Individual frequencies and group frequency. At each time step, we computed the angular velocity of each player by applying finite differences (*forward Euler method*) to the extracted phases:

$$\omega_i(k) = \frac{\theta_i(k+1)-\theta_i(k)}{\Delta t}, \qquad i = 1,2,\dots,N, \tag{4}$$

with $\Delta t = 1/f_s$ being the sampling time. This allowed us to characterize the frequency of each participant and its stability. Then, the average frequency of the group, $\omega_{group}(k)$, was extracted as the time-average of $\omega_i(k)$.

Group synchronization metrics. To quantify and characterize the level of synchronization among the players, we used the following metrics:

- *phase-synchronization*: for each trial, we evaluated the extent of synchronization in the group at each sampling time *k* through the order parameter *r(k)*, defined as

$$r(k) = \left|\frac{1}{N}\sum_{i=1}^{N} e^{j\theta_i(k)}\right| \qquad \forall k \in \{1,\dots,T\},$$

where *j* is the imaginary unit. Note that r*(k)* belongs to the interval [0,1], and it is 1 when the phases coincide at time *k*. Then, we computed the average order parameter in the trial $\underline{r}$ and that is,

$$\underline{r} = \frac{1}{T}\sum_{k=1}^{T} r(k)$$

Levels of group phase synchronization: to allow for a proper comparison of the extent of synchronization in the group in the various conditions introduced in the main document, we discretized the order parameters into four phase-synchronization levels (see Figure XX below):

$$\text{Level}_r(k) = \begin{cases} 1, \text{ if } r(k) < 0.70 & \textbf{(not in sync)}, \\ 2, \text{ if } 0.70 \leq r(k) < 0.85 & \textbf{(weak synchronization)}, \\ 3, \text{ if } 0.85 \leq r(k) < 0.95 & \textbf{(medium synchronization)}, \\ 4, \text{ if } 0.95 \leq r(k) \leq 1 & \textbf{(high synchronization)}. \end{cases}$$

Note that *Level$_r$(k)* = 1 (**not in sync**) means that the phase of the pendula at time *k* cannot be grouped in a circular sector of angle π rad.

Time-To-Synchronization and Time-In-Synchronization: Figure below illustrates how data were classified in order to compute the Time-To-Synchronization (*TTS*) and the Time-In-Synchronization (*TIS*).

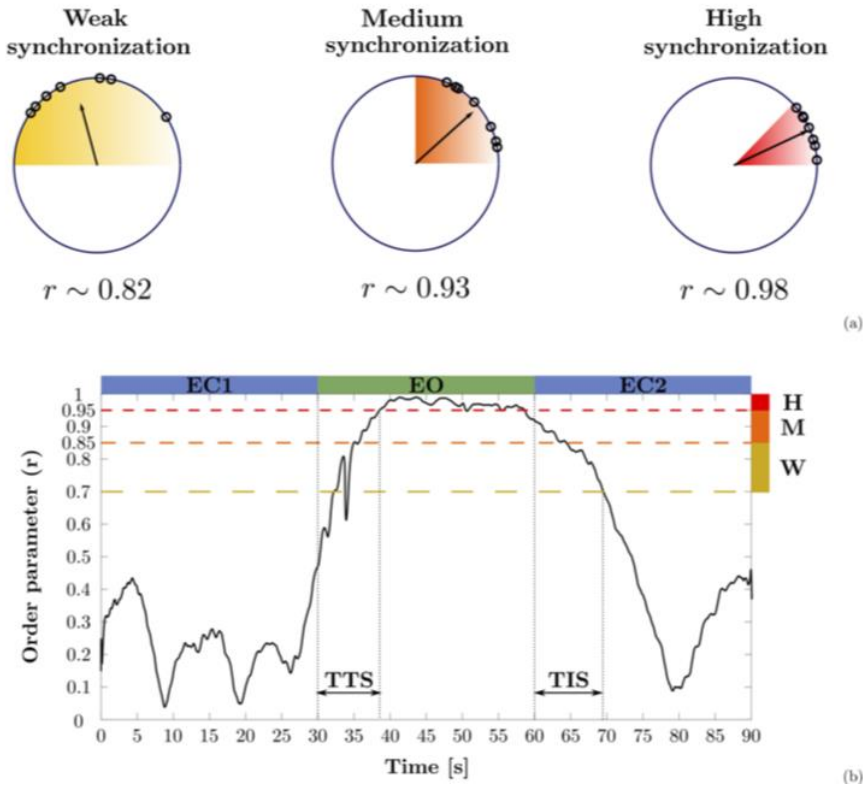

Figure: *Three levels of synchronization — Weak (W), Medium (M), and High (H) — characterized by the value of the order parameter r, used to determine Time-To-Synchronization (TTS) and Time-in-Synchronization (TIS). EO: Eyes Open; EC: Eyes Closed.*

Eyes-open (*EO*): **computing *TTS***. For a given trial, we denoted $T_{sync,i}$ as the number of sampling instants k such that $Level_r(k) = i$, and the corresponding fraction $F_{sync,i} = T_{sync,i}/T$, for $i = 1, \ldots, 4$. We computed *TTS* only for trials in which $F_{sync,1} \leq 0.5$ in order to exclude from the analysis the trials in which synchronization was only occasionally achieved. The remaining trials were classified as follows:

1. If $F_{sync,2} + F_{sync,3} > 0.75(1 - F_{sync,1})$, then the trial was considered as an instance of *Medium* synchronization;
2. If $F_{sync,3} + F_{sync,4} > 0.75(1 - F_{sync,1})$, then the trial was considered as an instance of *High* synchronization
3. If neither condition 1 nor 2 are satisfied, then the trial is considered as an instance of *Weak* synchronization.

Depending on the above classification, *TTS* was defined as the first time instant such that $Level_r$ became 2 (for trials of weak synchronization), 3 (for trials of medium synchronization), or 4 (for trials of high synchronization).

Eyes-closed (EC_2): **computing *TIS***. *TIS* was defined as the first time instant such that $Level_r < 1$ if the players stayed in sync ($Level_r > 1$) after closing their eyes for at least 3 consecutive periods of length $2\pi/\omega_{group}$, where ω_{group} is the mean frequency of the players in the trial. Otherwise, we set $TIS = 0$.

Following (Alderiso et al., 2017), group dynamics were initially modeled as a network of Kuramoto oscillators, coupled through the graph topologies used in the experiment (Complete, Ring, Path, Star, see Figure XX).

Figure: Four topologies during familiar human group cooperation situations, with various coupling modalities. (a) Complete graph: an ordinary organization during everyday working meetings; (b) Path graph: often present in sports, for instance in team rowing where partners are mechanically and visually coupled to two neighbors, except for the first and last rowers; (c) Ring graph: a common structure in many popular dances or among children at play (round dance); (d) Star graph: typical of musical ensembles, for instance when orchestra members are visually

coupled only to the director. Image (b) comes from unsplash.com, all the others come from pixabay.com.

We modeled the transition between 'eyes closed' and 'eyes open' by setting the coupling *gain c* instantaneously to zero so that the motion of each agent in the group is modeled as

$$\dot{\theta}_i(t) = \begin{cases} \omega_i + c\sum_{j=1}^{N} a_{ij}\sin(\theta_j(t) - \theta_i(t)), & \text{if eyes open,} \\ \omega_i, & \text{if eyes closed,} \end{cases} \quad [1]$$

where *N* is the number of players, ϑ_i the phase of the movement of the i-th player, ω_i their natural frequency, and c the strength of the coupling with the other players when visual coupling was established. The coefficients a_{ij} are set equal to 1 if the topology being studied involves a visual connection between players *i* and *j* when eyes are open, otherwise they are set equal to 0. In the following, we will refer to this model as the **Static Coupling** model (**SC**). In Experiment 1, participants' similarity (i.e., homogeneity) was controlled by manipulating the pendula's inertia and hence the natural frequency of the players' oscillatory motion. This enabled us to evaluate the influence of the players' similarity and graph structure on the emergence and quality of group synchronization. Specifically, four conditions were considered, involving (i) individual oscillations (solo), and three collective oscillations (ii) at the same shared frequency (all matched), (ii) at the same frequency for six out of the seven players (all matched but one), and (iv) at seven different frequencies corresponding to each player's preferred pace (natural). In Experiment 2, homogeneity among the players was manipulated at a different scale, by comparing groups of novices with groups of certified dancers (experts). To complete our analyses, we evaluated the effect of homogeneity in individual frequencies on the temporal aspects of the various synchronization regimes. This was performed by focusing on two variables, (i) the time to synchronization (*TTS*), capturing the time necessary for all participants to reach phase synchronization once they had opened their eyes, and (ii) the time remaining in synchronization (*TIS*) after eye closure, quantifying the memory effect. To test the model validity, we parameterized the model from experimental data (Bardy et al. 2020), and then computed the average *TIS* after switching the coupling *c* to zero. We observed that the **SC** model was unable to capture the relatively longer *TIS* measured experimentally, with model predictions being consistently shorter than expected in all conditions except in the natural condition. When used to explain the observations in Experiment 2, the same model did capture the synchronization dynamics of the non-dancers.Therefore, a more sophisticated model is required to adequately capture the experimental observations (Table XX).

Table. Comparison of Static Coupling, Individual Memory and Social Memory models with experimental results. *Average (with standard deviation) experimental Time-In-Sync* $\underline{TIS}_{exp}$ *versus average (with standard deviation) simulated Time-In-Sync* $\underline{TIS}_{sim}$*; **p < 0.01, ***p < 0.001.*

	Conditions	**Experimental results** $\underline{TIS}_{exp}$	**Static Coupling** $\underline{TIS}_{sim}$	**Individual Memory** $\underline{TIS}_{sim}$	**Social Memory** $\underline{TIS}_{sim}$
Exp. 1	Matched	9.95 ± 3.71 s (n=15)	6.52 ± 2.88 s (n=65) **	9.73 ± 3.72 s (n=139)	9.71 ± 3.67 s (n=123)
	Matched-but-one	8.20 ± 1.94 s (n=10)	. 5.94 ± 2.77 s (n=39) **	8.26 ± 2.55 s (n=106)	8.16 ± 3.23 s (n=112)
	Natural	5.32 ± 1.17 s (n=11)	4.74 ± 1.06 s (n=12)	-	-
Exp. 2	Dancers	8.81 ± 3.42 s (n=32)	5.92 ± 2.11 s (n=129) ***	8.90 ± 2.97 s (n=251)	8.97 ± 3.36 s (n=242)
	Non dancers	6.26 ± 2.43 s (n=17)	5.66 ± 2.07 s (n=55)	-	-

More specifically, the longer TIS exhibited in both experimental scenarios suggests that some memory mechanism was present, allowing the groups to stay in sync for longer than predicted by a sudden memory-less transition from eyes-open to eyes-closed. As presented in the Introduction, we contrast below two possible alternatives to model [1].

In the first model extension, the Individual Memory model (IM), we assumed that the motion frequency exhibited by each player at time ta of visual occlusion remains first as similar as possible to the last frequency $\vartheta_i(t_a)$ exhibited with eyes open, and then, after some time lag, relaxes back to the natural frequency of the player, ω_i. The model then becomes:

$$\dot{\theta}_i(t) = \begin{cases} \omega_i + c\sum_{j=1}^{N} a_{ij}\sin(\theta_j(t) - \theta_i(t)), & \text{if eyes open,} \\ \omega_i + \phi(t)(\dot{\theta}(t_a) - \omega_i), & \text{if eyes closed,} \end{cases} \quad [2]$$

with $\phi(t) = exp(-((t - t\_a)) / \tau)$; τ being the estimate of the decay time observed experimentally once visual contact among the participants is lost. We contrasted the model above with the predictions of a different model, the **Social Memory** model (**SM**). In this model, we assumed that participants maintain longer synchronization times at eye closure by internalizing the aggregate group dynamics. These dynamics are captured by the modulus

$r_i(t)$ and phase $\psi^i_{ref}(t)$ of the local order parameter computed by player *i*, using information received from the visually coupled players before closing their eyes. In this case we have

$$\dot{\theta}_i(t) = \begin{cases} \omega_i + c\sum_{j=1}^{N} a_{ij}\sin(\theta_j(t) - \theta_i(t)), & \text{if eyes open,} \\ \omega_i + c\phi(t) r_i(t_a)\sin(\psi^i_{\text{ref}}(t) - \theta_i(t)), & \text{if eyes closed,} \end{cases} \quad [3]$$

where $\psi^i_{ref}(t) = \dot{\psi}^i_{ref}(t_a)(t - t_a) + \psi^i_{ref}(t_a)$ and $\phi(t)$ is the decay function defined above. Both the **IM** model and the **SM** model were found to capture the experimental data . In Experiment 2, the **IM** model was found to better capture the experimental data than the **SM** Model.

Taken altogether, these tools helped to understand why behavioural cohesion is easier to maintain when perceptual exchanges are lost and how perceptuo-motor expertise can reinforce this cohesion. Those tools and findings are currently re-applied and nourishing the design of studies reported as 2.3.3 Time to Sync project; looking into capturing emotional qualities during multi agent scenarios through metrics of Individual and Group Motor Signatures from motion capture data and heart rate variability during joint action, complex tasks.

References

Alderisio, F., Fiore, G., Salesse, R. N., Bardy, B. G. & di Bernardo, M. Interaction patterns and individual dynamics shape the way we move in synchrony. Scientific Reports 7, 6846 (2017).

Bardy, B. G., Calabrese, C., de Lellis, P., Bourgeaud, S., Colomer, C., Pla, S., & di Bernardo, M. (2020). Moving in unison after perceptual interruption. Scientific Reports 10, 18032 (2020).

4. Computed Features of Moving Emotional Bodies in Interactions

Participants

UM_CN has developed tools for the analysis of the mechanisms underlying emotional recognition in social interactions, by looking into single-person and interaction body features. Videos showing an interaction between two actors facing each other were used, with one actor depicting anger (“aggressor”) and the other depicting fear (“victim”). From those videos, body “skeletons” were extracted using state of the art deep learning libraries (OpenPose). Methods were developed to extract meaningful features from the videos including kinematic, postural and interaction features.

Stimuli

The stimuli consisted of 61 videos of 1.5-second duration showing the interaction between two people. As mentioned by Nelson, de Bezerra, Claudio, and Pereira (2013), dynamic stimuli have a higher ecological validity compared to static stimuli. Emotions expressed in dynamic stimuli are better recognised, especially for lower intensities of the emotion. In each video, one of the actors depicted an aggressive behaviour whereas the other one depicted fearful behaviour. The interactions consisted of eight different pairs of two male actors. The actors were dressed in black and acted in front of a neutral light-coloured curtain, standing in front of each other. To prevent interference and better isolate the effect of bodily emotion expression, faces were blurred.

Study Design

Feature Definition

Each video consisted of 39 frames, for which a 2D pose of each actor was estimated using Open Pose (v1.0.1; Cao, Simon, Wei, & Sheikh, 2017). The 2D skeleton consists of 18 key points representing the nose, neck, right and left shoulder, right and left elbow, right and left wrist, right and left hip, right and left knee, right and left ankle, right and left ear. Each key point consisted of an x- and y-coordinate as well as a confidence value, which indicates with which certainly the algorithm determined the key point. The facial blurring resulted in poor estimations of the head key points. Since this study is investigating bodily features, the key points of eyes and ears were dismissed. To be able to reference head positioning, the key point of the nose was kept and corrected if necessary. Other missing or possible incorrectly estimated key points were later added manually to the data. Thus, the resulting data consisted of 61 videos, with 39 frames each for which we have (the x- and y-coordinates of) 14 key points per actor.

From those key points, a total of fourteen kinematic and postural features as well as interaction features were computed, that may contribute to the recognition of emotional movements. The kinematic features are Acceleration, Velocity, Vertical Movement and Forward-Backwards Movement, whereas the postural features consist of Symmetry, Limb Angles, Head Inclination and three further features concerning body contraction (Limb Contraction, Shoulder Ratio and Surface). The interaction features concerned the Distance between the Heads and Hands, respectively, and the closest distance between actors. The features were calculated in MATLAB (vR2017a, The MathWorks Inc., Natick, MA, USA) using the x- and y-coordinates of the 14 key points. The features were first calculated per frame for each video. For the statistical analysis and classification trees, data was averaged across frames and key points.

Velocity was defined by the amount of movement of each key point from the neighbouring frames. Further, Acceleration was computed by the difference in the amount of displacement of each key point. Vertical Movement was computed by the vertical displacement of the key points. The last kinematic feature was Forward Backward Movement, computed by the horizontal displacement of the key points. Furthermore, four more features concerning body contraction were computed. The first one was Limb Contraction, which was computed by the distance between the wrist and the head, and the second Shoulder Ratio, which was the amount of extension of the limbs with respect to the shoulders. The third body contraction feature was Surface which was computed by multiplying the x- and y-axis values. Fourth, Limb Angles which

was the angles between bodily segments. Two more postural features were computed, namely Head Inclination, which was computed by the distance between nose and neck in the vertical axis, Symmetry, which is defined by the symmetrical movement of the left and right limbs. Lastly, three interaction features were calculated to have an indication of the proximity between the two actors: the distance between the hands and between the heads, respectively, and the distance of the key points that get the closest during the whole interaction.

Representational Similarity Analysis

In order to investigate the relationship among the computed features, representational similarity analyses (RSA) (Kriegeskorte, Mur, & Bandettini, 2008) were conducted. The RSA allows the comparison between different representations of stimuli and data modalities, which were computed using MATLAB. The computed relationships between pairs of features/ stimuli are then visualized in representational dissimilarity matrices (RDMs). RDMs contain one cell per feature pair/experimental condition, which reflects the dissimilarity between the two. Thus, an RDM is symmetrical about a diagonal (which represents the dissimilarity between identical conditions/features and therefore equals zero). Each off-diagonal value indicates the dissimilarity between a feature or rating of different pair of videos.

The dissimilarity between feature values for each video pair was computed in Euclidean distance, using the non-averaged data, meaning that it was neither averaged across key points, nor over frames. This was done for each computed feature for each of the 122 single videos. This led to ten 122x122 distance matrices. To investigate the relationship between the computed features and possible correlations among each other, Spearman's rank correlations were conducted between all feature RDMs, resulting in a second order RDM.

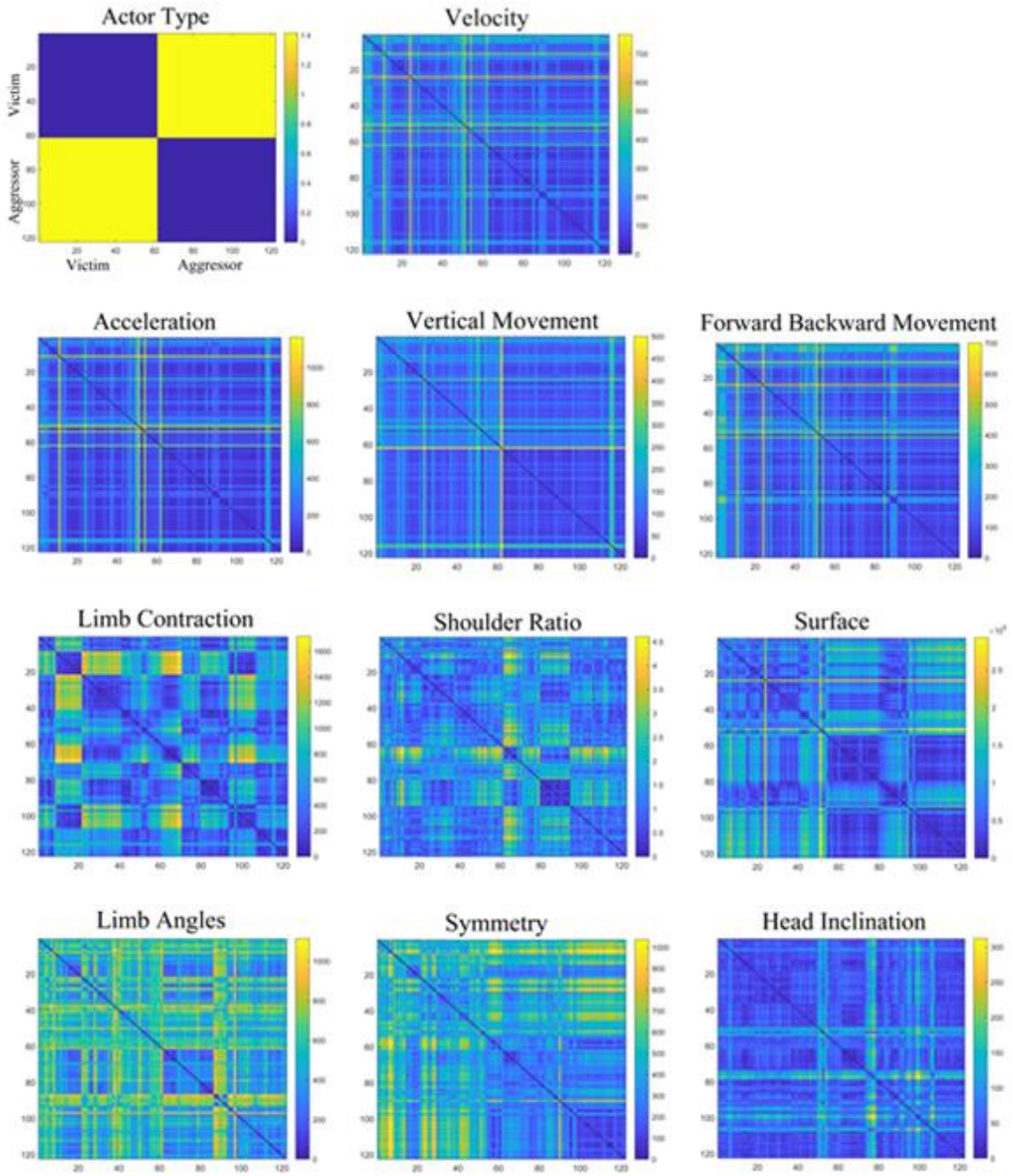

Figure 1. *Representational Dissimilarity Matrices for all computed features concerning single person movements. Data used was not averaged over time or key points. Each RDM reflects a pairwise comparison one feature for all 122 movements. Blue colour indicates more similarity, whereas yellow colour indicate dissimilarity, measured with Euclidean distance. Actor Type (aggressor/victim) represents the organisation of each RDM between aggressor and victim videos. The first two rows represented the kinematic features, with the top left RDM reflecting the structure of all RDMs. The last two rows represented the postural features, with the third row consisting of features related to body contraction.*

Statistical Analysis

The data of the computed features was averaged across key points and frames, per video and feature. Thus, 122 values for each feature were obtained (excluding interaction features), since each of the 61 videos resulted in two values, one for the aggressor, and one for the victim. The statistical analyses were conducted in IBM SPSS Statistics 24.0. An independent-samples t-test was conducted to compare the computed features between the victim and aggressor movements. This was done to determine whether the differences observed in the RDMs are indeed significant.

Classification Trees

To explore which of the computed features were most important in the classification of the aggressors and victim's movement, classification trees were computed in MATLAB. The classification is done by binary splitting the data at each node, choosing the criteria best predicting whether a video depicts an aggressor or a victim movement. A bootstrap-aggregation approach was used to reduce overfitting as well as to improve generalisation. This means that the classification is not based on one, but a weighted amount of decision trees. To determine which computed features can predict the classification between aggressor and victim the best, one classification tree was performed. The classification tree takes all ten computed features into account that concern the single-actor videos, which, for this purpose, were averaged across key points and frames.

References

Nelson, T. A., De Bezerra, I. A. O., Claudino, R. G. E., & Pereira, T. C. L. (2013). Influences of sex, type and intensity of emotion in the ecognition of static and dynamic facial expressions. *Avances En Psicologia Latinoamericana*, *31*(1), 192–199.

Cao, Z., Simon, T., Wei, S.-E., & Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7291-7299.

5. Dyadic Synchronisation using Phase-Amplitude Coupling (UNIGE, WU)

In our recent study titled “Capturing human movement and shape information from small groups to extract expressive and social features using markerless techniques” (Sabharval 2020), we highlight the use of Phase-Locking Values or Phase-Amplitude Coupling as a measure to compute the dyadic synchronization between participants of a musical ensemble. By using a Phase-Amplitude Coupling measure we can evaluate an unconscious body movement synchrony between participants in an experiment. While Phase-Locking Values (PLV) and phase synchrony (Varlet et al. 2020) are also used to understand the synchronization of EEG readings of brain regions, we can apply the same principles to also understand the connectivity or synchronization that exists between two participants in a dyadic setup, by making use of phase synchronization metrics. Hence, PLV values can be utilized to establish a functional level connectivity that is visible between performers in a Musical Ensemble – helping us answer questions raised in the Hypothesis.

Measuring Phase-Amplitude Coupling and compute Phase-Locking Values

To calculate phase-amplitude coupling, the raw signal related information is filtered for a frequency of interest. Then, the real and complex values are extracted from the complex values of the signal being analyzed. After which, the phase angles or amplitude (as the case may be) are extracted from the complex values in the signal.

Phase-Locking Values

Phase-Locking Values as utilized by Mormann et al. (2005) can be computed using the phase values that are extracted from the complex values of the signal, obtained by implementing a Fast Fourier Transform (FFT) - with which we can obtain magnitude values at each frequency bin. Phase locking is an important concept in computing interactions in non-linear and complex systems. Phase-Locking Value (PLV) is the most commonly used interaction measure, and for the purpose of our study, we utilize relative phase values as suggested by previous research (Rosenblum et al. 2000), for assessing the interaction and consequent dyadic synchronization between two co-performers.

Using PLV, we can understand the interaction that exists between two non-linear time- series, and in our case, we use the head movement data (using nose key-point) extracted using HPEs to ascertain the level of interaction between two co-performers (Aydore et al. 2013; Lachaux et al. 1999). For each data point at a specific time interval, phase differences are computed for the signals between which phase-locking values are meant to be computed. The word meant is used very carefully since these signals must be of those two co-performers for whom Phase- Locking Values are meant to computed.

As suggested in the equation below, we compute the absolute value of the mean phase difference between the signals of the two co-performers. This is represented as a complex unit-length vector (Aydore et al. 2013). The absolute value of the mean is then a measure of the magnitude of the vector, which indicates the amount of phase-amplitude coupling, or the coupling strength.

The PLV is calculated using the following formula:

$$\mathbf{PLV} = \left| \frac{\sum_{t=1}^{n} e^{i(\theta_1 - \theta_2)}}{n} \right|$$

Here n is the total number of data points, t is a data point that is available at every equally distributed time stamp. Θ1 and Θ2 are the phase angles of the two signals being analyzed. Thus, by obtaining the instantaneous phase angles of two signals can help us obtain the dyadic synchronization, or the coupling strength, that exists between two performers in a musical ensemble. The degree of synchronization as computed here is in the range of [0,1] where the highest state of synchronization sits at 1.

Future work

While the above method using Phase-Locking Values has given us promising results, we are currently working on developing alternative techniques to compute inter-personal coordination. We will also investigate other areas where these computational models can find their apt use. These include:

1. Cross-spectral coherence
2. Cross-wavelet coherence
3. Multiscale Entropy (Glowinski et al. 2010)
4. Multi-Event Class Synchronization (Volpe 2021)

References

M. Varlet, S. Nozarada.Ln, P. Nijhuis, and P. E. Keller, "Neural tracking and integration of 'self' and 'other' in improvised interpersonal coordination," Neuro Image, vol. 206, p. 116303, 2020

F. Mormann, J. Fell, N. Axmacher, B. Weber, K. Lehnertz, C. E. Elger, and G. Fernandez, "Phase/amplitude reset and theta–gamma interaction in the human medial temporal lobe during a continuous word recognition memory task," Hippocampus, vol. 15, no. 7, pp. 890–900, 2005

M. Rosenblum, P. Tass, J. Kurths, J. Volkmann, A. Schnitzler, and H.-J. Freund, "Detection of phase locking from noisy data: application to magnetoencephalography," in Chaos In Brain?, pp. 34–51, World Scientific, 2000

S. Aydore, D. Pantazis, and R. M. Leahy, "A note on the phase locking value and its properties," Neuroimage, vol. 74, pp. 231–244, 2013

J.-P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela, "Measuring phase syn- chrony in brain signals," Human brain mapping, vol. 8, no. 4, pp. 194–208, 1999

Glowinski, D., Coletta, P., Volpe, G., Camurri, A., Chiorri, C., & Schenone, A. (2010, October). Multi-scale entropy analysis of dominance in social creative activities. In Proceedings of the 18th ACM international conference on Multimedia (pp. 1035-1038)

Volpe, G. (2021). The Multi-Event-Class Synchronization (MECS) Algorithm. Submitted.

Badino, L., Volpe, G., Tokay, S., Fadiga, L. & Camurri, A. (2019). Multi-layer adaptation of group coordination in musical ensembles. Scientific reports, 9(1), 1-10.