D2.1
Results on prediction in action execution and observation – Phase I

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

Table of Contents

Introduction

This deliverable reports on the progress on the research conducted between M7-M18 of the EnTimeMent project with regards to the individual action execution and observation axis, focused on studies on individual motor behaviour. This part of the project constitute the baseline for research and theoretical work developed in the Phase I of the EnTimeMent project, focused on dyadic studies (n=2) and on group motor behaviour (involving three people or more, n>2) which are reported in D2.3 and D2.5 respectively. The numbering of the studies reported herein, refers to the most recent version of deliverable D1.2 Research Requirements providing an update on the methodological background and know-how of the studies. In this deliverable we report results of studies that have finished the stage of data collection and analysis (*2.1.2, 2.1.6, 2.1.8, 2.1.11, 2.1.15*).

A major theoretical shift in cognitive neuroscience was driven by a new conceptualization of the motor system. In fact, motor processes seem to play a role in perceptual and cognitive functions, challenging the classical sensory versus motor separation and opening the doors to embodied cognition research in both humans and artificial systems. Critically, the recruitment of motor programs, during action/object perception, constrain the active search of specific sensory features that maximize the discrimination between different perceptual hypotheses and support prediction of future information at multiple timescales. The generation of active inferences about future actions of conspecifics is central to our capability to smoothly interact with each other and, therefore, fundamental to the development of human cognition.

In this deliverable we collected all ongoing research, investigating action-perception coupling in single individuals and thus on the neurobehavioral building blocks allowing sensorimotor communication in dyads or groups. Studies presented below are those that have either published or are in an advanced stage close to submission for publication.
The first two studies are based on the same theoretical framework suggesting that Individual Motor Signatures (IMS) characterize action execution. Briefly, IMSs are relatively stable movement strategies that each one of us unknowingly display when moving in our environment. Interestingly, data from the first experiment (*2.1.2*) provide evidence that motor activations during action observation are driven by the mismatch between the observers' and actors' IMSs. The larger the distance the larger is the motor recruitment, thus suggesting that the motor system might indeed act as an inferential engine that compares other's action to our own template. The second study (*2.1.6*) approaches a similar problem from a different perspective. In fact, the goal of this project

is to automatically extract IMSs from arm movement by using both traditional machine learning methods and more recent deep learning strategies.

The third *(2.1.8)* and fourth projects *(2.1.11)* explored the sensorimotor bases of expressivity. Among them the first aimed at extracting expressivity measures from complex individual body motion. In order to do so, a novel computational pipeline has been implemented by integrating graph and game theory towards the analysis of the perceived origin of full-body human movement and its propagation. In fact, the analysis of the origin of movement is an important component in the understanding and modeling expressivity. The data, extracted with the computational method, have then been submitted for evaluation to a panel of dancers with varying degrees of expertise. The following study has instead tried to discover which specific postural and kinematic features could be computed from affective whole-body movement videos and related those to brain responses. By means of state-of-the-art neuroimaging methods it was investigated whether the (dis)similarity of body posture and kinematics between different emotional categories could explain neural responses to body expressions in and beyond body-selective regions.

Finally, the last section reports on the *2.1.14* research activities aimed at the automatic detection of pain and associated behavior from body movements. This research program is a key component of WP4, constituting one of the use case scenarios planned in EnTimeMent. Taken together, Phase I results reported herein pushed forward current state-of-the-art description of human movement from the perspective of individual differences (IMS) and their expressive properties. Studies reported below addressed multiple gaps in the body of research and emphasized the importance of approaching human movement analysis and modeling through the lens of mid-layer features. Research roadmap for Phase II of the EnTimeMent project has been established (D1.2 Research requirements), which will push further the frontiers towards a full understanding of the importance of modeling human movement across multiple timescales.

2.1.2 Action variability in action observation and execution

For a full description please see: Hilt P. M., Cardellicchio P., Dolfini E., Pozzo T., Fadiga L., D'Ausilio A. (2020) Motor recruitment during action observation: effect of interindividual differences in action strategy. Cereb Cortex, 30(7), 3910–3920.

Action Perception

Mirror neurons were originally described as visuomotor neurons that are engaged both during visual presentation of actions performed by conspecifics, and during the actual execution of these actions (Rizzolatti and Craighero 2004). These neurons were first discovered using single-cell recordings in monkey premotor cortex (area F5; di Pellegrino et al. 1992) and later within monkey inferior parietal cortex (PF/PFG; Gallese et al. 2002; Fogassi et al. 2005).

Since then, there has been a growing interest in mirror neurons both in the scientific literature and the popular media. The widespread interest was in particular driven by their potential role in imitation and thus in a fundamental aspect of social cognition (Iacoboni 2005; Rizzolatti and Sinigaglia 2010). In follow-up studies, neurons with mirror properties have been found in different parietal and frontal areas of monkeys and other species, including humans (Rizzolatti and Sinigaglia 2016).

The mirror neuron system has also been associated with action perception. In fact, others' action anticipation and comprehension might be achieved both by the ventral route (Middle Temporal Gyrus – MTG - and the anterior Inferior Frontal Gyrus - aIFG), and the dorsal route (Inferior Parietal Lobule – IPL - and the posterior Inferior Frontal Gyrus - pIFG). The dorsal stream may support this process by reactivating the most likely action needed to achieve the predicted goal. In line with this account, action discrimination could rely on internal forward models (Flanagan and Johansson 2003; Kilner et al. 2004) to anticipate the unfolding of a given action (Schütz-Bosbach and Prinz 2007).

Mirror neuron system in humans

Immediately following the initial reports of mirror neurons in the macaque brain, the existence of an analogous mechanism in humans was discussed. While some authors argued that clear evidence of a human mirror neuron system was still lacking (e.g. Dinstein 2008; Lingnau et al. 2009; Turella et al. 2009), further and numerous results coming from various techniques such as transcranial magnetic stimulation (TMS; Fadiga et al. 2005; Naish et al. 2014), electroencephalography (EEG; Fox et al. 2016), functional magnetic resonance imaging (fMRI; Hardwick et al. 2018) and human single-cell recordings (Mukamel et al. 2010) revealed the existence of a fronto-parietal network with mirror-like properties in humans (Rizzolatti and Sinigaglia 2010).

Based on human brain-imaging data (Rizzolatti et al. 1996; Decety et al. 1997; Iacoboni et al. 1999) and cytoarchitecture (Petrides 2005), the ventral premotor cortex and the pars opercularis of the posterior inferior frontal gyrus (Brodmann area 44) were assumed to be the human homologues of macaque mirror area F5. Later, the rostral inferior parietal lobule was identified as equivalent to the monkey mirror area PF/PFG (Rizzolatti et al. 2001; Rizzolatti and Craighero 2004).

In parallel, EEG research showed that event-related synchronization and desynchronization of the mu rhythm (rolandic alpha band) were linked to action performance, observation and imagery (Pineda 2008; Fox et al. 2016). These results

suggest that Rolandic mu event-related desynchronization (Cochin et al. 1998; Babiloni et al. 2002) during action observation reflects activity of a mirror-like system present in humans (Sebastiani et al. 2014; Fox et al. 2016; Lapenta et al. 2018).

Finally, single-pulse TMS over the primary motor cortex (M1) and motor evoked potentials (MEPs) amplitude were employed as a direct index of corticospinal recruitment (Corticospinal Excitability - CSE). Using this technique, several studies showed a modulation of MEPs amplitude during action observation matching various changes occurring during action execution (Fadiga et al. 1995; for a review please see: Fadiga et al. 2005; Naish et al. 2014; D'Ausilio et al. 2015).

The coordination of our own actions with those of others requires the ability to read and anticipate what and how our partner is about to do. Indeed, when observing someone else moving, we can extract useful information such as future bodily displacements (Flanagan and Johansson 2003; Blakemore and Frith 2005; Falck-Ytter et al. 2006) or infer higher-order cognitive processes hiding behind those actions (Becchio et al. 2008; Soriano et al. 2018). In principle, knowledge about the invariant properties of movement control (Flash and Hogans 1985; Bennequin et al. 2009) could support inferences about the unfolding of other's actions (Dayan et al. 2007; Casile et al. 2010). In this regard, it has been proposed that these inferences may be based on a direct match between actor's sensorimotor activations during Action Execution (AE) and observer's sensorimotor activations triggered by Action Observation (AO; Rizzolatti et al. 2001; Rizzolatti and Craighero 2004; Rizzolatti and Sinigaglia 2016). Indeed, using Corticospinal Excitability (CSE), motor recruitment during AO was shown to replicate the spatio-temporal sequence of motor commands implemented by the actor (for a review please see: Naish et al. 2014).

This idea is however challenged by the redundancy that characterizes the organization of human movement (Kilner 2012; D'Ausilio et al. 2015; Hilt et al. 2017). The abundance of degrees of freedom available during AE suggests that different joint configurations, as well as spatio-temporal patterns of muscle activity, can equally be used to reach the same behavioral goal (Bernstein 1967). In this regard, a strong version of the direct-matching hypothesis (Rizzolatti et al. 2001; Rizzolatti and Craighero 2004; Rizzolatti and Sinigaglia 2016) explains inferences when a direct relationship exists between muscle recruitment, movement kinematics and behavioral goals (e.g. simple finger movements). However, it is less clear how other's complex movements (i.e. multi-joint movements) are transformed onto the observer's motor representations. In this case, any sensorimotor-based inference about other's actions amounts to finding a solution to a many-to-many mapping problem.

Here we suggest that a simpler mapping exists between behavioral goals and the lower dimensionality space of whole-body configurations (i.e. synergies; Hilt et al. 2017). In fact, although a handful of kinematic solutions are biomechanically valid, everyday actions (i.e. reaching for an object on the floor starting from a standing posture) are usually performed via a limited number of possible kinematic configurations of the biomechanical chain (e.g. "ankle" and "hip" strategies for postural control; Horak and Nashner 1986; Berret et al.

2009). On the top of that, each individual carries his own robust and yet unique way of moving (Individual Motor Signature – IMS; Hilt et al. 2016; Słowiński et al. 2016). For instance, in a whole-body reaching task Hilt and collaborators (Hilt et al. 2016) showed low intra-subject motor variability, accompanied by a large inter-subject variability. The inherent lower dimensionality of whole-body postural control and the presence of robust Individual Motor Strategies (IMS) suggest the existence of a simpler AO-AE mapping that may be a function of everyone's individual movement style. Backed by this, we hypothesize that while observing others' multi-joint actions, people build sensorimotor-based predictions by referencing what they see to the motor engrams of their own IMS.
To verify our hypothesis, we asked naive participants to first perform and then observe a whole-body reaching action which could be executed with numerous IMSs generally spread within a continuum between two "extreme" patterns (ankle and knee strategies; Hilt et al. 2016). After characterizing subjects' own IMS during execution, we measured their sensorimotor recruitment (CSE) by administering single-pulse Transcranial Magnetic Stimulation (TMS) on their motor cortex while they observed an actor achieving the same goal by using the two "extreme" patterns of IMSs. CSE was measured from the cortical representation of the Tibialis Anterior muscle (TA) that shows a clearly dissociable pattern while executing the two IMSs. To exclude potential carry-over effects between action execution and observation, the same subjects were also tested several months later in the action observation task only.

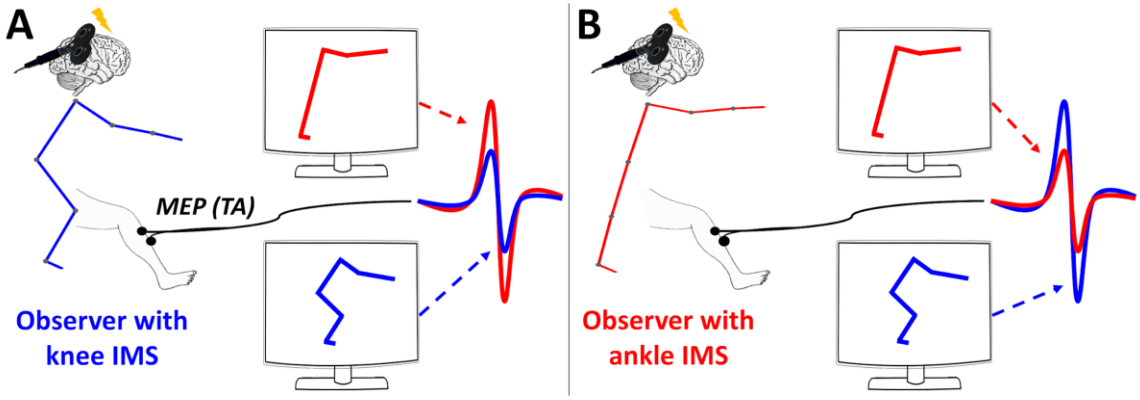


Figure 1: Illustration of the main results. MEPs amplitudes are depicted when observing knee (blue stick figure) or ankle (red stick figure) stimulus, for a subject that performed the knee (A) or the ankle (B) IMS in AE. Our results showed that corticospinal excitability was greater when actor and observer IMSs differ the most. These results agree with the predictive coding hypothesis that hypothesize the existence of a distance computation between observed movement and observer's IMS.

CSE was modulated at the single subject level according to the "distance" between actors' and observer's IMS: larger CSE modulations are associated with the observation of a more different IMS. This result is schematically illustrated in Figure 1 for two hypothetical subjects having extreme IMSs. Importantly, motor priming effects elicited by the action execution task can be excluded considering that the same pattern of results, in the same

subjects, was shown several months later and in the absence of any action execution task.

Our results are at odds with a strictly simulative account of others' actions. Instead, the fact that sensorimotor activities during AO are shaped around a measure of distance between observed and own IMSs, agrees with the predictive coding framework. In this model, prior motor knowledge provides critical top-down signals that are integrated with bottom-up sensory-based processing (Friston 2010; Friston et al. 2011). To do so, a comparison between predicted (own IMS) and observed kinematic information (others' IMS) generates a prediction error signal that is used to update the representation of other's action.

Overall our data suggest that a greater uncertainty about other's action will call for a greater need of trustful predictions and consequently greater sensorimotor recruitment. In this context, the present study adds direct neurophysiological evidence that prediction errors are estimated by accessing IMS-related information. In fact, the many-to-many mapping problem in other's (multi-joint) action discrimination might be solved by accessing knowledge about IMSs. Indeed, the stability of IMSs (Słowiński et al. 2016; Coste et al. 2017) may reflect the implicit control and prioritization of a limited number of internal parameters during action planning and execution, partly solving the motor redundancy problem.

References

Babiloni C, Babiloni F, Carducci F, et al (2002) Human Cortical Electroencephalography (EEG) Rhythms during the Observation of Simple Aimless Movements: A High-Resolution EEG Study. Neuroimage 17:559–572

Becchio C, Sartori L, Bulgheroni M, Castiello U. 2008. The case of Dr. Jekyll and Mr. Hyde: A kinematic study on social intention. Conscious Cogn. 17:557–564.

Bennequin D, Fuchs R, Berthoz A, Flash T. 2009. Movement Timing and Invariance Arise from Several Geometries. PLoS Comput Biol. 5:e1000426.

Bernstein NA (1967) The Coordination and Regulation of Movements, Pergamon P. Oxford

Berret B, Bonnetblanc F, Papaxanthis C, Pozzo T. 2009. Modular Control of Pointing beyond Arm ' s Length. J Neurosci. 29:191–205.

Blakemore SJ, Frith C. 2005. The role of motor contagion in the prediction of action. Neuropsychologia. 43:260–267.

Casile A, Dayan E, Caggiano V, Hendler T, Flash T, Giese MA. 2010. Neuronal encoding of human kinematic invariants during action observation. Cereb Cortex. 20:1647–1655.

Cochin S, Barthelemy C, Lejeune B, et al (1998) Perception of motion and qEEG activity in human adults. Electroencephalogr Clin Neurophysiol 107:287–295

Coste A, Slowinski P, Tsaneva-Atanasova K, Bardy BG, Marin L. 2017. Mapping Individual Postural

Signatures. In: Weast-Knapp JA,, Pepping GJ, editors. Studies in Perception & Action XIV. Taylor & Francis.

D'Ausilio A, Bartoli E, Maffongelli L. 2015. Grasping synergies: A motor-control approach to the mirror neuron mechanism. Phys Life Rev. 12:91–103.

Dayan E, Casile A, Levit-Binnun N, Giese MA, Hendler T, Flash T. 2007. Neural representations of kinematic laws of motion: Evidence for action-perception coupling. Proc Natl Acad Sci. 104:20582–20587.

Decety J, Grèzes J, Costes N, et al (1997) Brain activity during observation of actions. Influence of action content and subject's strategy. Brain 120:1763–1777. doi: 10.1093/brain/120.10.1763

di Pellegrino G, Fadiga L, Fogassi L, et al (1992) Understanding motor events: a neurophysiological study. Exp Brain Res 91:176–180

Dinstein I (2008) Human Cortex : Reflections of Mirror Neurons. Curr Biol 18:956–959. doi: 10.1016/j.cub.2008.09.007

Fadiga L, Buccino G, Craighero L, et al (1998) Corticospinal excitability is specifically modulated by motor imagery: A magnetic stimulation study. Neuropsychologia 37:147–158. doi:10.1016/S0028-3932(98)00089-X

Fadiga L, Craighero L, Olivier E (2005) Human motor cortex excitability during the perception of others' action. Curr Opin Neurobiol 15:213–218. doi: 10.1016/j.conb.2005.03.013

Fadiga L, Fogassi L, Pavesi G, Rizzolatti G (1995) Motor facilitation during action observation: a magnetic stimulation study. J Neurophysiol 73:2608–2611

Falck-Ytter T, Gredebäck G, Von Hofsten C. 2006. Infants predict other people's action goals. Nat Neurosci. 9:878–879.

Flanagan JR, Johansson RS (2003) Action plans used in action observation. Lett to Nat 424:769–771. doi: 10.1038/nature01861

Fogassi L, Ferrari PF, Gesierich B, et al (2005) Parietal Lobe : From Action Organization to Intention Understanding. Science (80-) 308:662–667. doi: 10.1126/science.1106138

Fox NA, Yoo KH, Bowman LC, et al (2016) Assessing human mirror activity With EEG mu rhythm: A meta-analysis. Psychol Bull 142:291–313. doi: 10.1037/bul0000031

Friston KJ. 2010. The free-energy principle: a unified brain theory? Nat Rev Neurosci. 11:127–138.

Friston KJ, Mattout J, Kilner JM. 2011. Action understanding and active inference. Biol Cybern. 104:137–160.

Gallese V, Fadiga L, Fogassi L, Rizzolatti G (2002) Action representation and the inferior parietal lobule. In: Common mechanisms in perception and action. pp 334–355

Hardwick RM, Caspers S, Eickhoff SB, Swinnen SP (2018) Neural correlates of action: Comparing meta-analyses of imagery, observation, and execution. Neurosci Biobehav Rev 94:31–44. doi: 10.1016/j.neubiorev.2018.08.003

Hilt PM, Bartoli E, Ferrari E, et al (2017) Action observation effects reflect the modular organization of the human motor system. Cortex 95:104–118. doi: 10.1016/j.cortex.2017.07.020

Hilt PM, Berret B, Papaxanthis C, et al (2016) Evidence for subjective values guiding posture and movement coordination in a free-endpoint whole-body reaching task. Sci Rep 6:23868. doi: 10.1038/srep23868

Horak FB, Nashner LM. 1986. Central programming of postural movements: adaptation to altered support-surface configurations. J Neurophysiol. 55:1369–1381.

Iacoboni M (2005) Neural mechanisms of imitation. Curr Opin Neurobiol 15:632–637. doi: 10.1016/j.conb.2005.10.010

Iacoboni M, Woods RP, Brass M, et al (1999) Cortical Mechanisms of Human Imitation. Sci New Ser 286:2526–2528. doi: 10.1038/020493a0

Kilner JM, Vargas C, Duval S, et al (2004) Motor activation prior to observation of a predicted movement. Nat Neurosci 7:1299–1301. doi: 10.1038/nn1355

Kilner JM. 2012. More than one pathway to action understanding. Trends Cogn Sci. 15:352–357.

Lapenta OM, Ferrari E, Boggio PS, et al (2018) Motor system recruitment during action observation: No correlation between mu-rhythm desynchronization and corticospinal excitability. PLoS One 13:1–15. doi: 10.1371/journal.pone.0207476

Mukamel R, Ekstrom AD, Kaplan J, et al (2010) Single-Neuron Responses in Humans during Execution and Observation of Actions. Curr Biol 20:750–756. doi: 10.1016/j.cub.2010.02.045

Naish KR, Houston-Price C, Bremner AJ, Holmes NP (2014) Effects of action observation on corticospinal excitability: Muscle specificity, direction, and timing of the mirror response. Neuropsychologia 64:331–348. doi: 10.1016/j.neuropsychologia.2014.09.034

Petrides M (2005) Lateral prefrontal cortex: Architectonic and functional organization. Philos Trans R Soc B Biol Sci 360:781–795. doi: 10.1098/rstb.2005.1631

Pineda JA (2008) Sensorimotor cortex as a critical component of an “extended” mirror neuron system: Does it solve the development, correspondence, and control problems in mirroring? Behav Brain Funct 4:1–16. doi: 10.1186/1744-9081-4-47

Rizzolatti G, Craighero L (2004) the Mirror-Neuron System. Annu Rev Neurosci 27:169–192. doi: 10.1146/annurev.neuro.27.070203.144230

Rizzolatti G, Fadiga L, Matelli M, et al (1996) Localization of grasp representations in humans by PET: 1. Observation versus execution. Exp Brain Res 111:246–252. doi: 10.1007/BF00227301

Rizzolatti G, Fogassi L, Gallese V (2001) Neurophysiological mechanisms underlying the understanding and imitation of action. Nat Rev Neurosci 2:1–10. doi: 10.1038/35090060

Rizzolatti G, Sinigaglia C (2010) The functional role of the parieto-frontal mirror circuit: Interpretations and misinterpretations. Nat Rev Neurosci 11:264–274. doi: 10.1021/am4002502

Rizzolatti G, Sinigaglia C (2016) The mirror mechanism: a basic principle of brain function. Nat Rev Neurosci 17:757–765. doi: 10.1038/nrn.2016.135

Schütz-Bosbach S, Prinz W (2007) Prospective coding in event representation. Cogn Process 8:93–102. doi: 10.1007/s10339-007-0167-x

Sebanz N, Shiffrar M (2009) Detecting deception in a bluffing body: The role of expertise. Psychon Bull Rev 16:170–175. doi: 10.3758/PBR.16.1.170

Sebastiani V, de Pasquale F, Costantini M, et al (2014) Being an agent or an observer: Different spectral dynamics revealed by MEG. Neuroimage 102:717–728. doi: 10.1016/j.neuroimage.2014.08.031

Słowiński P, Zhai C, Alderisio F, Salesse R, Gueugnon M, Marin L, Bardy BG, di Bernardo M, Tsaneva-Atanasova K, 2016. Dynamic similarity promotes interpersonal coordination in joint action. J R Soc Interface. 13:20151093.

Soriano M, Cavallo A, D'Ausilio A, Becchio C, Fadiga L. 2018. Movement kinematics drive chain selection toward intention detection. Proc Natl Acad Sci U S A.

Turrella L, Pierno AC, Tubaldi F, Castiello U (2009) Mirror neurons in humans : Consisting or confounding evidence ? Brain Lang 108:10–21. doi: 10.1016/j.bandl.2007.11.002

2.1.6 Investigate singularity in ellipses drawing

The goal of the experiment is to investigate the singularity which characterizes different people. This high-level feature enables to distinguish different people by analyzing the way they move, write, or perceive an event (for example, auditory or visual). The way these actions are carried out is different from individual to individual and, for this reason, we speak about singularity. The proposed experiment led to measure and investigate this high-level feature. The possibility to measure singularity can contribute to many application fields, from clinical to entertainment and customer applications. The first experiment consists in the analysis of how different people draw an ellipse. The experiment is grounded on the Two-Third Power Law: $\mathbf{V(t) = k * r^{beta}}$ where *v(t)* is velocity, *k* is a constant and *r* is the ellipse radius of curvature. If *beta* were different from each person it could be sufficient for a classification of singularity.

The hypothesis is that *beta* is not enough to provide a measure of singularity, and therefore we need to individuate and measure other features and apply data analysis and machine learning techniques to obtain a correct measure of singularity.

The first scenario analyzed focuses on motor signature: people try several times the drawing of the same ellipse. The final outcomes will be 10 different ellipses for trial. Each participant carries out 6x10 trials. Each trial consist of 10 execution of an ellipse at the same condition. Each participant executes the 10 trials in 6 (2x3) different conditions: 2 hands (right or left) and 3 drawing speeds (slow, normal, quick). For each trial only 7 executions of the ellipse are considered (the first 2 and the last one are discarded). Summarizing the dataset is made up of 10(trials) x 2(hands) x 3(speeds) x 7(ellipses) x 14 (subjects) = 5880 available ellipses. For each ellipse, raw data are:

x	y	Velocity	Curvature	Pressure	Timestamp

As mentioned before, the goal of the experiment is classify different people and identify features able to lead the correct classification. In this experiment, we followed two different approach until now:

1. Traditional machine learning model (study complete);
2. Deep Learning models (currently under study);

Traditional machine learning model

As we know, traditional machine learning models can provide very stable and robust results. In their implementation, they can benefit from steps in a standard Machine Learning (ML) pipeline. A common step is the feature engineering step where statistical features are computed starting from raw data. This process can significantly improve the final prediction of the algorithm used. To perform correctly this step, we need to fix the length of time to analyze. Usually, fixed-sliding windows are selected for this purpose. In this experiment, we cannot select the fixed-sliding window duration because each person draws an ellipse with different timing. Since fixed-sliding windows are not feasible to this experiment, we computed statistical measures on different sections of an ellipse drawn taking into account only raw data. The split considered in the analysis are the following:

- Split = 2: first and second half of the ellipse are considered separately (Figure 1 - top left)
- Split = 2: curves and straight lines are considered separately (Figure 1 – top right);
- Split = 4: each curve and each straight line is considered separately from others (Figure 1 – bottom left);
- Split = 6: each curve is considered separately, and each straight line is divided in two parts (Figure 2 – bottom right).

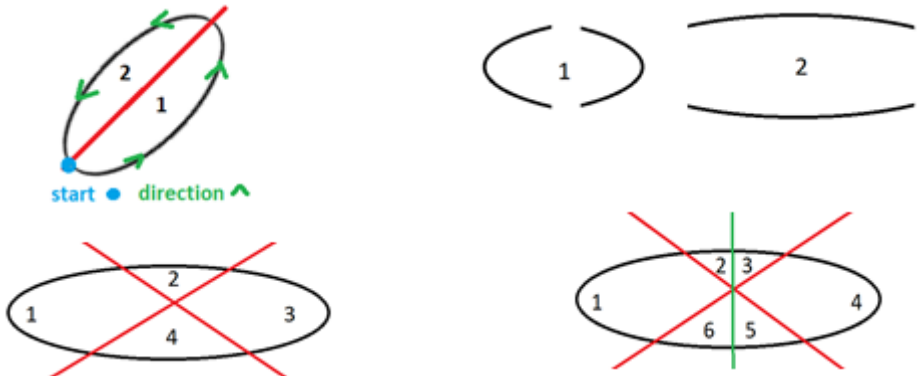


Figure 2: Different sections of the ellipse considered in the analysis. From left to right, split in 2 sections, split in 4 sections and split in 6 sections.

Higher-level features, able to better understand the underlying problem, are obtained considering different splits using statistical measures.

A powerful algorithm, both in terms of theoretical properties and practical effectiveness (Fernández-Delgado et al 2014, Wainberg et al. 2016), for classification is Random Forest (RF) developed in (Breiman et al., 2001) for the first time. RF is composed of the union of multiple Decision Trees (Rokach et al. 2008). Compared to DTs, RF introduces an additional degree of randomness due to the introduction of a bootstrap phase.

From the hierarchy presented in the dataset, we focused our analysis identifying two different scenarios able to determine the behavior of the model trained. These 2 scenarios, providing different sensibilities on data used in training set, can be used to

understand and estimate the algorithm behavior on unseen or new data. The 2 scenarios identified are:

- Leave-One-Hand-Out (LOHO): the learning set of our classifier is made up of all people of the dataset except the information coming from one hand of the tested person. This scenario is interesting because one hand is the dominant one and can be very relevant in the final classification.
- Leave-One-Speed-Out (LOSO): the learning set of our classifier is made up of all people, all hands and all speeds in the dataset, except one speed coming from the tested person. In this case, the learning set contains more information respect to the previous scenario and this will be reflected in higher recognition results.

Finally, a Feature Ranking (FR) step is computed in order to discover the most relevant section of an ellipse. Once a model is built, it is often required to understand how this model exploits, combine, and extract information in order to understand if the learning process has also cognitive meaning, namely it is able to capture the underline phenomena and does not just capture spurious correlation (Calude et al., 2017; Guyon et al., 2003) by comparing the knowledge of the experts with the information learned by the models. FR therefore represents a fundamental phase of model checking and verification, since it should generate results consistent with the available knowledge of the phenomena under exam provided by the experts.
FR methods based on RF are one of the most effective FR techniques as shown in many researches (Genuer et al., 2010; Saeys et al., 2008). Several measures and approaches are available for FR in RF. One method is based on the Permutation Test combined with the Mean Decrease in Accuracy (MDA) metric, where the importance of each feature is estimated by removing the association between the feature and outcome of the model. For this purpose, the values of the features are randomly permuted (Good et al.; 2013) and the resulting increase in error is measured. In this way also the influence of the correlated features is also removed. Note that, in our case, as a feature we do not intend a particular engineered feature but a particular ellipse section (e.g.the first section when split= 6, the second curve sections when split=4, etc.).

Deep Learning models

A parallel approach is related to the use of Deep Learning (DL) models. As we know, these architectures are automatically able to extract the best set of features from raw data. This implies that the feature engineering step and therefore, the sections split, are not needed.
In our analysis we focused attention on a common backbone for all models we will use. This backbone consists of a Convolutional AutoEncoder (CAE) (Masci et al., 2011) able to extract the best set of features. Furthermore, it allows a fair comparison between all models belonging to the state-of-the-art in ML. The usage of the CAE is strictly related to the time-scales of each raw feature. Indeed, in order to better understand multi time-scales we can follow different approaches in DL. Architectures such as Clockwork-RNN (Koutnik et al., 2014) or Multi-LSTM (Liu et al., 2015) are designed to automatically detect multi-time scales information. Other models such as LSTM (Hochreiter et al., 1997), simply RNN, Multi-Layer Perceptron (MLP), are not directly thought to handle this problem

but can provide excellent results if properly used. In particular, the most important thing is regard to the extracted features and not to their final connection which can be performed by several ML models, also traditional ones. The two different approaches discussed before are not necessarily exclusive but can be also considered together. In our analysis, we trained the convolutional autoencoder in order to better reconstruct the input signal coming from raw data. Different kernel size for convolution are used as hyperparameters of the network, and the best set of kernel sizes are chosen. When the CAE achieves the minimum reconstruction error, the weights and nodes of the encoder are frozen. These nodes and their weights are used to extract higher level features from raw data and are used as initial backbone for other models able to compute the final connection between these features and classes to be predicted.
We can now, summarize the models we will used taking into account that convolutional autoencoder is the common backbone for all:

- MLP;
- LSTM;
- Clockwork – RNN.

Recognition performances in LOHO and LOSO

The Table 1 reports the average accuracy. The recognition results are quite high (> 65 %) in LOHO scenario taking into account the 14 people in the dataset. Moreover, we can also observe very high scores in the LOSO scenario (> 85%).
Indeed, it is intuitable that the section with higher scores is the one related to the split of the two halves of the same ellipse. This is a consequence of more consecutive information in those 2 sections. Although this is the most accurate part, it is also the least interesting for a case study. On the contrary, the other splits can be very interesting analyzing if the main information lies in curves or straight lines. Subsequently, we would like to identify which part of each curve or straight line has more information than its counterpart.
Observing the Table 1, we can conclude that, except the split in two halves of the same ellipse, the best sections are provided from split=4.

Table 1. Average accuracy (%) for different splits considered in the analysis (see Figure 3).

SPLIT	SCENARIO	ACCURACY
2 (first half vs second half)	LOHO	70.70
	LOSO	86.49
2 (curves vs straight lines)	LOHO	65.19
	LOSO	82.70
4	LOHO	69.85
	LOSO	85.54
	LOHO	69.71

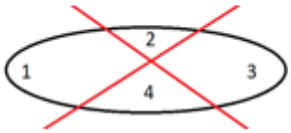
6	LOSO	85.42

Since DL architectures are currently under study we cannot provide detailed results for those models. A first analysis using MLP, achieve excellent result (close to 100% in LOHO scenario). This implies that features learned through the CAE can easily describe how classify different people.
A new analysis is also conducted. The idea of this new analysis was observing how recognition results change removing from dataset information with more noise. Indeed, the not-dominant hand can be very noisy if compared to the other one. Moreover, also "slow" speeds can be noisy. A preliminary study analyzing only "clear" information was conducted. Reasonably, the LOHO scenario loses meaning in this case study. We can observe what happen with a split=6 in the following table:

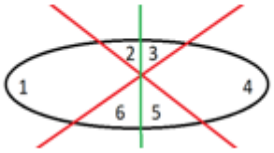
Table 2. Average accuracy (%) for split=6 where not noise information is used in analysis.

SPLIT	SCENARIO	ACCURACY
6	LOSO	71.44

Comparing the two different recognition results, we can observe how "noisy" information is very relevant in the person classification with an accuracy of about 14% higher. On the other hand, the analysis of the dominant hand is very interesting from a neuroscience perspective. As mentioned in the Traditional Machine Learning section, we exploited how different sections of each ellipse influences the recognition performance. Observing rankings in Figure 2, we can easily assert that the main informative section in ellipse analysis is in the first curve (where people start drawing). This is visible in all splits. Moreover, it is true that the split in 2 halves is not focused in curves or straight lines but, also there, the section with higher ranking is the first one (where people start drawing). Therefore, with this analysis we can conclude that the beginning of the draw allows us to capture more insights of the underline problem. Or, in other words, this is the most significant section of all. Next experiment, with DL models, can help us to identify which are the best set of features able to lead this classification problem. As we are observing for first results, MLP can achieve very high recognition scores.



Rank	Section in the plot
1	1
2	3
3	2
4	4



Rank	Section in the plot
1	1
2	4
3	5
4	3
5	6
6	2

Figure 3: Feature Ranking of different splits considered in the analysis.

References

Breiman, L.: Random forests. Machine Learning45(1), 5–32 (2001)

Calude, C.S., Longo, G.: The deluge of spurious correlations in big data. Foundations of science22(3), 595–612 (2017)

Fernandez-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems? The Journal of Machine LearningResearch15(1), 3133–3181 (2014)

Genuer, R., Poggi, J.M., Tuleau-Malot, C.: Variable selection using random forests. Pattern Recognition Letters31(14), 2225–2236 (2010)

Good, P.: Permutation tests: a practical guide to resampling methods for testing hypotheses. Springer Science & Business Media (2013)

Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research3(Mar), 1157–1182 (2003)

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. Koutnik, J., Greff, K., Gomez, F., & Schmidhuber, J. (2014). A clockwork rnn. *arXiv preprint arXiv:1402.3511*.

Liu, P., Qiu, X., Chen, X., Wu, S., & Huang, X. J. (2015, September). Multi-timescale long short-term memory neural network for modelling sentences and documents. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 2326-2335).

Masci, J., Meier, U., Cireşan, D., & Schmidhuber, J. (2011, June). Stacked convolutional auto-encoders for hierarchical feature extraction. In *International conference on artificial neural networks* (pp. 52-59). Springer, Berlin, Heidelberg.

Rokach, L., Maimon, O.Z.: Data Mining with Decision Trees: Theory and Applications, vol. 69. World Scientific (2008)

Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (2008)

Wainberg, M., Alipanahi, B., Frey, B.J.: Are random forests truly the best classifiers? The Journal of Machine Learning Research17(1), 3837–3841 (2016)

2.1.8 Perception of the origin of full body human movement and its propagation

For a full description please see: K. Kolykhalova, G. Gnecco, M. Sanguineti, G. Volpe, and A. Camurri, “Automated analysis of the origin of movement: An approach based on cooperative games on graphs,” IEEE Transactions on Human-Machine Systems, 2020. DOI: 10.1109/THMS.2020.3016085

We developed an approach based on the integration of graph and game theory to contribute towards the analysis of the perceived origin of full-body human movement and its propagation.The analysis of the origin of movement is an important component in the understanding and modeling expressivity. E.g., in rehabilitation: the detection of the origin of movement can help in enabling a patient to learn how to perform a movement (e.g., how to stand up from a chair) correctly to avoid injuries. For example, the leaning forward of an arm can have very different expressive meanings depending on the origin of movement: a “punch” originates from the foot, a “push away” may originate from the shoulder, and a “caress,” from the hand. All these movements are basically a leaning forward of an arm, the very different dynamics of which are explained also in terms of the origin of movement. The approach, which is grounded on the combination of cooperative game theory and graph theory, consists of the following steps.

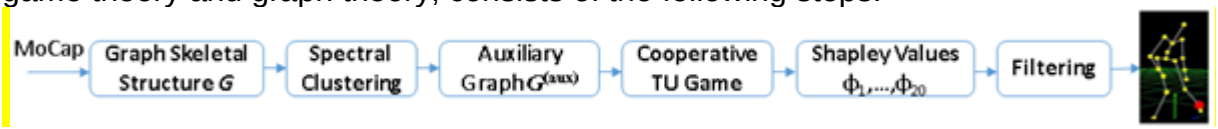


Figure 4: Conceptual architecture of the proposed method

The human body is represented by means of an undirected graph, in which the vertices are the joints and the edges are both physical and non-physical connections between these body joints. Moreover, the edges are associated with weights, the values of which depend on a feature extracted from motion capture data. On the one hand, physical links represent connections between consecutive physical body joints, such as the forearm. On the other hand, non-physical links model the dependencies between joints that are

not physically connected, solely derived from correlations observed in the chosen movement feature between these joints: for example, a hand moving towards the head, followed by a movement of the head in response in the same direction, reveals a non-physical link between the hand and the head. Non-physical links therefore play the role of potential bridges joining body parts that are not directly connected within the skeletal structure but exhibit correlated dynamics during the movement performed. Starting from the graph representing the skeletal structure augmented by non-physical links, we define a suitable mathematical game (Maschler et al., 2013) in which the vertices (i.e., the body joints) are the players and the edges model the communication channels (through which movement can propagate) between these players. Body movement is therefore represented by a game constructed on the graph. A cooperative game model is proposed, since both the vertices and the edges contribute to the overall movement.
Then, the Shapley value (Maschler et al., 2013) – which is a classical solution concept from cooperative game theory able to provide a ranking of the players that represents their relevance in the game - is computed for all the players of the game and adopted as a measure of vertex relevance in the graph to estimate how much each vertex contributes to a shared goal (i.e., to the way in which a specific movement-related feature is transferred among the joints). The possibility to know, moment by moment, which joint(s) is the most representative in the ongoing full-body movement (i.e., those with the highest Shapley value) constitutes precious information for the automated analysis of expressiveness in movement. The joints with the highest Shapley value are candidates to be the perceived origin of movement propagating in the body and they can provide useful cues to detect which parts of the body are most relevant for the analysis of expressive movement and worth a detailed observation by means of further analysis techniques (possibly at a finer scale), as well as to inform automated techniques of movement prediction.
We used a recorded multimodal data set composed of 127 recordings, acquired with the goal of analysing movement, determining the features associated with it, and designing computational techniques for their evaluation. The recordings were acquired using a Qualisys motion capture system with 13 infra-red cameras synchronized with 2 video cameras in the frontal and lateral views. The two professional dancers were equipped with 1 microphone, 5 accelerometers, and 64 infra-red reflective markers. After their acquisition, the data were cleaned and post-processed via the Qualisys Track Manager native software using a cubic polynomial interpolation for trajectories with gaps in the data.
Finally, annotations of the origin, path, and destination of each movement were produced by experts. The expressive movements performed were not related to a specific dance style, being normal full-body movements, e.g., leaning an arm towards a target or turning towards a direction, characterized by a clear origin of movement, enabling the detection of the origin even by a non-expert observer (though its automatic detection is still not a trivial task). The choice of dancers as movement executors was motivated by their full awareness and control of movement details and their higher motor skills with respect to non-trained people, which allowed reducing the amount of noise with respect to alternative performances by non-experts.

Validation of the approach included an on-line survey (based on the data repository) in which participants with different levels of expertise in dance took part. A survey website was developed to collect user ratings. Once the user agrees to participate and perform the task, a series of triplets of videos is presented.

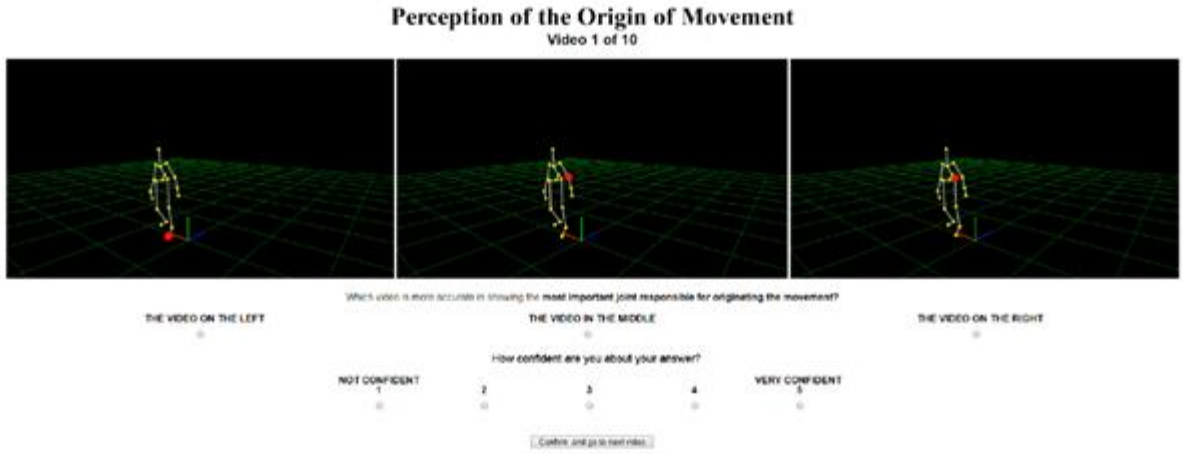


Figure 5: Website for the evaluation of the proposed method: best method selection.

Each of the three videos in a triplet displays a skeletal representation of a dancer performing the same full-body expressive movement. Each video has one highlighted joint (in red). This joint corresponds to the most relevant joint according to one of the following criteria: (i) joint with the maximum Shapley value; (ii) joint with the maximum speed; and (iii) random choice. The identity of the joint highlighted in red is possibly updated by each criterion every second (making it difficult for the user to guess when the criterion applied in a specific video is, e.g., a random choice). The order of the three criteria is randomized among the three videos so that the specific criterion applied to each video is not predictable by the user. For a fair evaluation, the criteria themselves are also completely unknown to the user (i.e., the user has no idea how they are named and how they work). During the survey, the participant is asked to choose the video that better represents the evolution of the most relevant joint responsible for originating the dancer's movement. Once a user has selected one video, he/she is asked to declare how confident he/she is in his/her choice by selecting a value from 1 to 5 on a 5-point Likert scale (levels: not confident, not so confident, neutral, confident, very confident). The participant can see all the videos as many times as desired and has to answer both questions (video choice and confidence level) before proceeding to the next triplet of videos. Each participant has to rate ten triplets of videos proposed from a selection of one hundred triplets using a Latin square selection method. The website was submitted to people with three different levels of expertise in dance: professionals, semi-professionals, and novices/non-dancers. A total of 22 people took part in the evaluation. Each participant self-evaluated his/her own level of expertise. The general information about the participants is as follows: professionals: 8 participants (3 male, 5 female), with a mean age of 42.75 years (std 9.56 years), semi-professionals: 6 participants (3 male, 3 female), with a mean age of 30 years (std 4.47 years), novices/non-dancers: 8 participants (6 male, 2 female), with a mean age of 35.5 years (std 7.4 years).

In the first two cases, the dancers were, respectively, experts and amateurs in contemporary dance.

Some results of the chosen type from among the three types of different stimuli are presented in the Table 1 and in the diagram shown in Figure 4. Both demonstrate that the results of the validation of the proposed method are promising. Indeed, the Shapley value method was selected in the large majority of cases.

Table 3

Participants\Method	Shapley value	Speed	Random
Professionals	90 (±5.35)	8.75 (±3.54)	1.25 (±3.54)
Semi-professionals	83.33 (±8.16)	10 (±8.94)	6.67 (±5.16)
Novices/non-dancers	67.5 (±8.86)	25 (±11.95)	7.5 (±11.65)
All Participants	80 (±12.34)	15 (±11.44)	5 (±8.02)

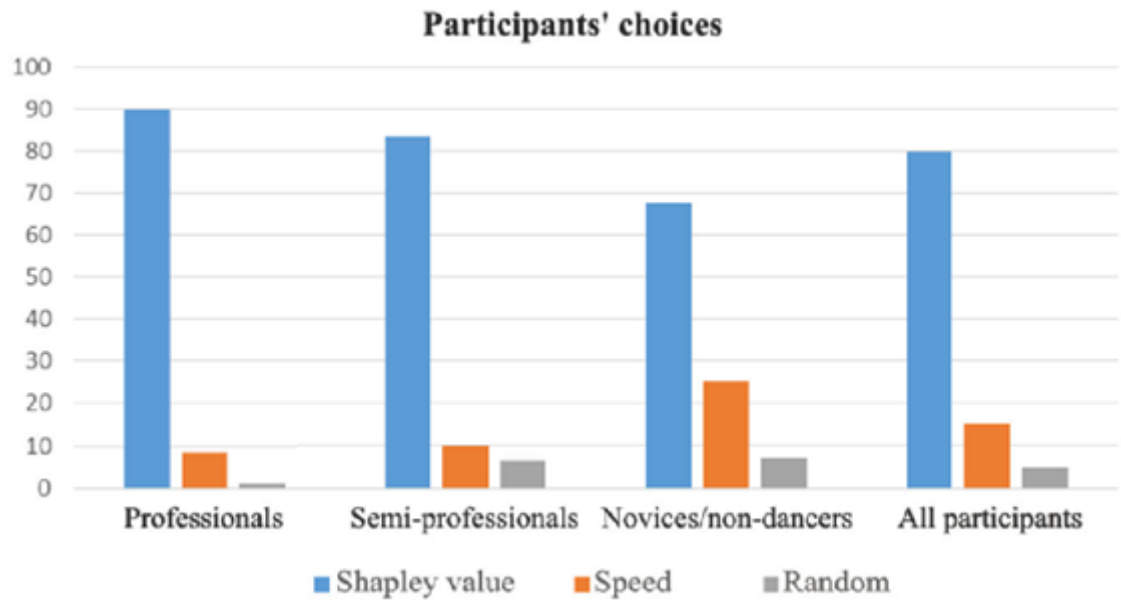


Figure 6: Participants' choices.

Table 4 shows, for two selected frames, the first 5 joints, ordered non-increasingly with respect to their Shapley values, normalized with respect to the maximum Shapley value in each frame. The associated normalized Shapley values are also reported in the table. The movements associated with the two frames are, respectively:
a) a sudden leaning down to the left with the trunk and the head, followed by a rotation to the right, with a final rising of the trunk and head, where the shoulder centre is clearly the origin of movement;

b) a rotation and extension toward the right of the performer, where the right elbow and right shoulder are clearly leading the movement of the whole body.
In both cases, the origin of movement is correctly identified by the proposed approach. The two frames illustrate, respectively, the following situations, which were quite often observed during the analysis of the specific motion capture data set:
a) a case in which the first and second largest Shapley values are very well separated and (at least) the first largest one is uniquely achieved;
b) a case in which there are two joints with the largest Shapley value, but these joints are connected by a physical edge in the body graph, and the second largest Shapley value (which, in the second column, corresponds to the third joint) is very well separated from the first one.

Table 4

1st frame	2nd frame
shoulder centre (1.00)	Right elbow (1.00)
head (0.46)	Right shoulder (1.00)
right ankle (0.40)	Right hip (0.53)
right knee (0.40)	Right knee (0.53)
left elbow (0.38)	left ankle (0.27)

Among possible developments, we would like to exploit a more complex skeletal structure (for which each cluster of joints is associated to a specific joint in the simpler 20-joint skeletal structure), making it possible to analyze movement in parallel at a finer interacting spatio-temporal scale in a multiple-scale approach. Using movement-related features different from speed (or of a higher dimensional feature vector) to compute the Shapley value for a comparison with the results obtained using speed as a feature.
Incorporating multiple temporal scales. For example, one can look at a fast temporal scale at the very first moment of the origin of movement and at a slower temporal scale where one can analyse the origin of movement at a higher level. Applying the developed methodology to analyse the emergence of the origin of movement when two persons or small groups are involved in the movement itself is the next step of this work. We are currently running collaboration with EuroMov looking into multi-person scenarios.

References

Cohen, S.,Dror, G and Ruppin, E. (2007). “Feature selection via coalitional game theory,” Neural Computation, vol. 19, no. 7, pp. 1939-1961.

Kolykhalova, K., Gnecco, G., Sanguineti, M, Volpe, G. and Camurri, A. (2020) “Automated analysis of the origin of movement: An approach based on cooperative games on graphs,” IEEE Transactions on Human-Machine Systems. doi: 10.1109/THMS.2020.3016085

Maschler, M., Solan, E. and Zamir, S. (2013). Game Theory. Cambridge, UK: Cambridge University Press.

Michalak, T., Aadithya, K., Szczepánski, P., Ravindran, B. and Jennings, N. (2013) “Efficient computation of the Shapley value for game-theoretic network centrality,” Journal of Artificial Intelligence Research, vol. 46, pp. 607-650.

2.1.11 Computation-Based Feature Representation of Body Expressions in the Human Brain

For a full description please see: Poyo Solanas, Marta, Maarten Vaessen, and Beatrice de Gelder. "Computation-Based Feature Representation of Body Expressions in the Human Brain." *Cerebral Cortex* (2020).

Humans and other primate species are experts at recognizing body expressions. To understand the underlying perceptual mechanisms, we computed postural and kinematic features from affective whole-body movement videos and related them to brain processes. Using representational similarity and multivoxel pattern analyses, we showed systematic relations between computation-based body features and brain activity. Our results revealed that postural rather than kinematic features reflect the affective category of the body movements. The feature limb contraction showed a central contribution in fearful body expression perception, differentially represented in action observation, motor preparaton, and affect coding regions, including the amygdala. The posterior superior temporal sulcus differentiated fearful from other affective categories using limb contraction rather than kinematics. The extrastriate body area and fusiform body area also showed greater tuning to postural features. The discovery of midlevel body feature encoding in the brain moves affective neuroscience beyond research on high-level emotion representations and provides insights in the perceptual features that possibly drive automatic emotion perception.

It is widely agreed that humans and other primate species are experts at recognizing emotion and intention from face and body expressions (de Gelder 2006; Giese and Rizzolatti 2015). The central importance of nonverbal communication across many social species suggests that the brain is equipped for rapid and accurate face and body movement perception; yet, the mechanisms underlying this ability are still largely unclear. Previous research on face and body expressions has predominantly searched for brain correlates of symbolic emotion categories (Lindquist et al. 2012; Kirby and Robinson 2017), disregarding the visual features that drive movement and emotion perception (e.g., kinematic and postural body features). This is in part due to the fact that methods for fine-grained description of body movements were not yet available. This study used

computational descriptions of body expressions to investigate which features drive emotion and body perception and how they are encoded in the brain.

Previous behavioral and computational studies have provided some indications about relevant features of body posture and movement, and their relation to emotional expressions (De Meijer 1989; Wallbott 1998; Roether et al. 2009; Kleinsmith and Bianchi-Berthouze 2012; Piana et al. 2014; Patwardhan 2017). Some important postural features have been identified, including elbow flexion, associated with the expression of anger, and head inclination, typically observed for sadness (Wallbott 1998; Coulson 2004; Vaessen et al. 2018). Other form-related features that have been investigated are the vertical extension of the body (e.g., upper limbs remain low for sadness but high for happiness), the directionality of the movement (e.g., angry bodies are usually accompanied by a forward movement), symmetry (e.g., the movement of the upper limbs tends to be symmetrical when experiencing joy), and the amount of lateral opening of the body (e.g., hands are close to the body during fear and sadness while extended in happiness) (Kleinsmith and Bianchi-Berthouze 2012).
A central, yet unanswered, question is the relation between candidate features and brain processes. There is sparse evidence in the literature on how particular features may be related to brain processes. One classical proposal is the two-stream model of visual processing with two separate brain pathways for form and movement information (Vaina et al. 1990; Giese and Poggio 2003; Milner and Goodale 2006, 2008). From the primary visual cortex, the dorsal stream leads to the parietal lobe and is specialized in localizing objects in space, processing motion signals and in the visual-spatial guidance of actions. The ventral stream leads to the temporal lobe and is responsible for visual form processing and object recognition. Two areas in this pathway have been identified that sustain a certain level of specialization in the processing of whole bodies and body parts: the extrastriate body area (EBA) in the medial occipital cortex, and the fusiform body area (FBA) in the fusiform gyrus (Downing et al. 2001; Peelen and Downing 2005; Schwarzlose et al. 2005). However, their respective functions are not yet clear and it is also not clear how they, alone or together, contribute to body expression perception. In addition, body shape and movement elicit a widespread neural response beyond the visual analysis of body features in body-category selective areas (de Gelder 2006; Van den Stock et al. 2011), triggering processes related to their affective content, the conveyed action and for the preparation of an appropriate behavioral response (de Gelder et al. 2004; Van den Stock et al. 2011). The present study is the first effort to discover which specific postural and kinematic features could be computed from affective whole-body movement videos and be related to brain responses. By means of representational similarity multivoxel pattern analysis techniques, we investigated whether the (dis)similarity of body posture and kinematics between different emotional categories could explain neural responses to body expressions in and beyond body-selective regions.

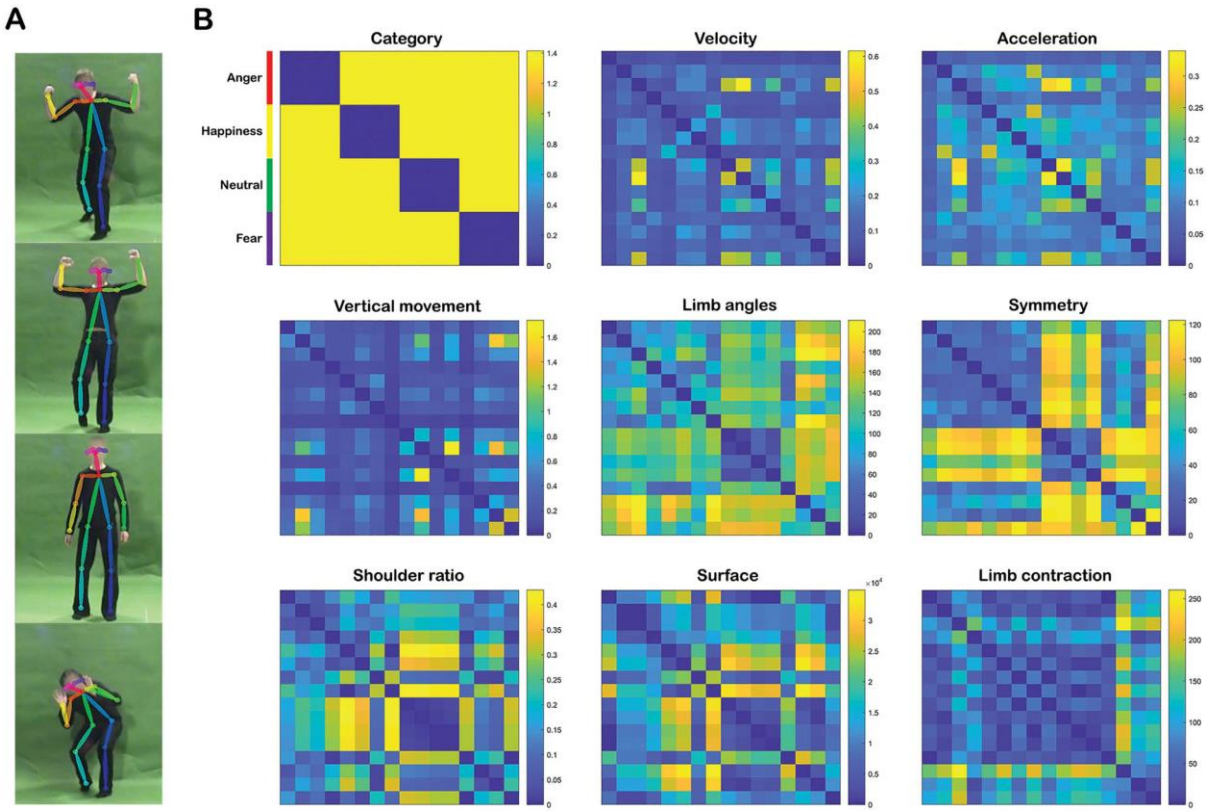


Figure 7: Representational dissimilarity matrices of the kinematic and postural features. (A) Examples of frames from the different affective movement videos with the OpenPose skeleton. Note that participants were shown the videos without the OpenPose skeleton; (B) The RDMs represent pairwise comparisons between the 16 stimuli with regard to the kinematic (i.e., velocity, acceleration, and vertical movement) and postural features (i.e., limb angles, symmetry, shoulder ratio, surface, and limb contraction) averaged over time. The dissimilarity measure reflects Euclidean distance, with blue indicating high similarity and yellow high dissimilarity. Color lines in the upper left corner indicate the organization of the RDMs with respect to the emotional category (anger: red; happiness: yellow; neutral: green; fear: purple) of the video stimuli.

We aimed at investigating whether the (dis)similarity of body posture and kinematics between different emotional categories could explain the neural response of brain regions involved in body processing. For this purpose, several areas were defined as ROI and their neural RDMs were computed and correlated to the emotional and feature RDMs. The ROIs included occipito-temporal areas that have previously shown a certain level of body specificity (three ROIs: FBA, EBA, and pSTS) (Downing et al. 2001; Peelen and Downing 2005; Schwarzlose et al. 2005; Kontaris et al. 2009; Vangeneugden et al. 2014), parietal and temporal areas thought to be implicated in attention and action observation (six ROIs: V7/3a, SPOC, SMG, pIPS, mIPS, and aIPS) (Culham and Valyear 2006; Grafton and Hamilton 2007; Corbetta et al. 2008; Caspers et al. 2010), and frontal areas involved in action observation and other higher cognitive functions (six ROIs: PMv, PMd, SMA, pre-SMA, inferior frontal,and frontal regions) (Grafton and Hamilton 2007; Caspers et al. 2010). See Figure 8 for the full results.

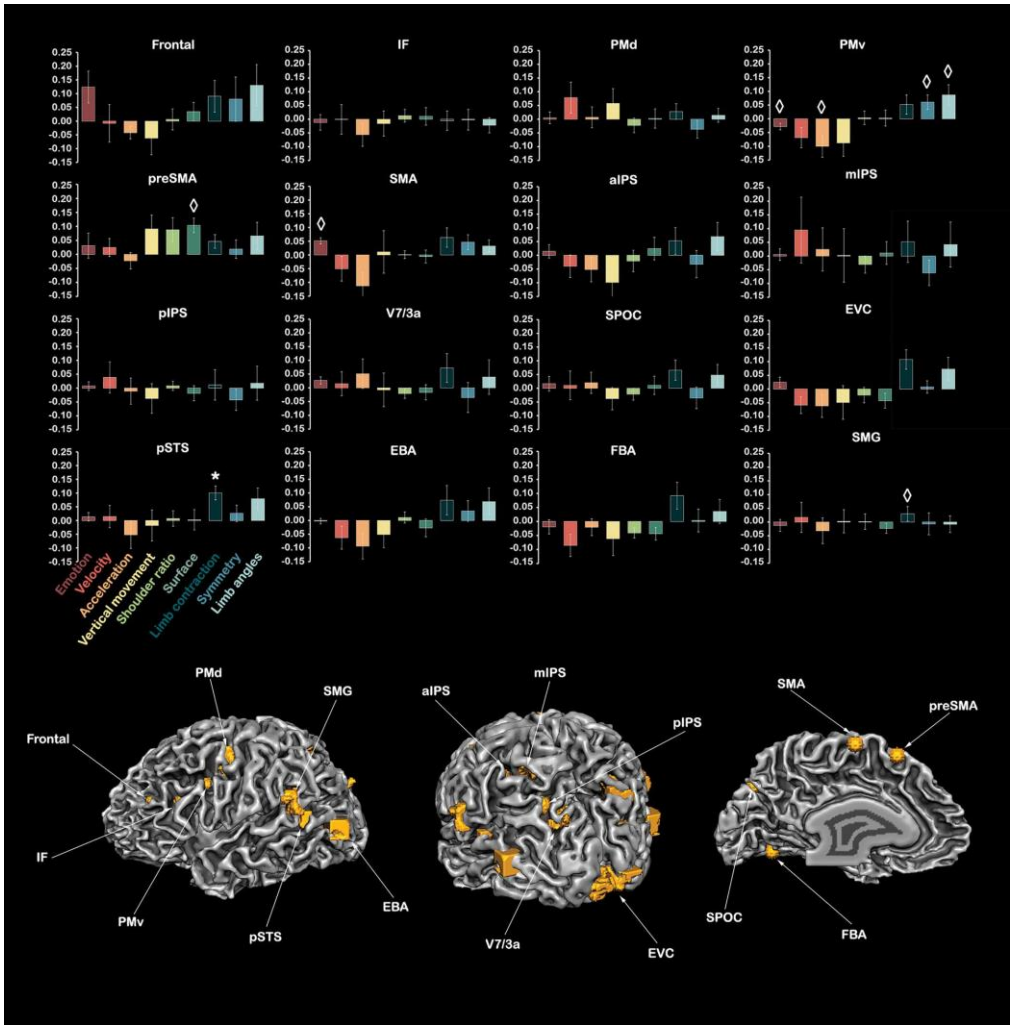


Figure 8: Average Spearman's rank correlation across participants between the kinematic/postural feature RDMs and each ROI matrix. Kinematic features include velocity, acceleration, and vertical movement. Postural features comprise shoulder ratio, surface, limb contraction, symmetry, and limb angles. Positive r values indicate that a high (dis)similarity between a stimulus pair in the feature RDM also has a high (dis)similarity in the neural representation. A negative correlation means that a low (dis)similarity between two stimuli at the feature level would have a higher (dis)similarity in the neural representation. Asterisks and rhombi indicate significant correlations after BHFDR correction and correlations that presented significant uncorrected P-values, respectively (one sample t-test against 0, two-tailed). The error bars denote the standard error of the mean (SEM). Order or relationships across ROIs are not assumed here. Abbreviations: EBA, extrastriate body area; EVC, early visual cortex; FBA, fusiform body area; IF, inferior frontal cortex; IPS, intraparietal sulcus; p, posterior; m, middle; a, anterior; PMd, dorsal premotor cortex; PMv, ventral premotor cortex; pre-SMA, presupplementary motor area; pSTS, posterior superior temporal sulcus; SMA, supplementary motor area; SMG, supramarginal gyrus; SPOC, superior parietal occipital cortex.

We also investigated whether (dis)similarities in body posture and kinematics between different emotional categories could explain the neural response at the whole-brain level. The computed feature RDMs were compared with the multivoxel dissimilarity fMRI

patterns by means of searchlight RSA. The velocity RDM was positively correlated to inferior frontal sulcus and precentral gyrus. Negative main effects for acceleration were found in middle temporal, superior frontal, and postcentral sulci while no positive main effects were observed for this feature. Vertical movement correlated positively with cingulate gyrus, whereas negatively to the frontomarginal and middle temporal gyri. With respect to postural features, limb angles showed a positive main effect in anterior insula and pSTS. Several areas negatively correlated to symmetry in the inferior and middle occipital gyri, precuneus, isthmus, anterior calcarine, intraparietal, and cingulate sulcus. Shoulder ratio negatively correlated to anterior insula, frontal operculum, putamen, ACC, middle frontal gyrus, cingular insular sulcus, claustrum, internal capsule, and parahippocampal gyrus. Surface showed main negative effects in posterior orbital gyrus, thalamus, anterior perforated substance, ACC, inferior and superior frontal sulci, putamen, and internal capsule. Only positive correlations to limb contraction were found in intraparietal sulcus, anterior insula, caudate nucleus, amygdala, superior frontal sulcus and gyrus, precuneus, posterior orbital gyrus, ACC, superior temporal gyrus, inferior precentral sulcus, and SMG (see Figure 9).

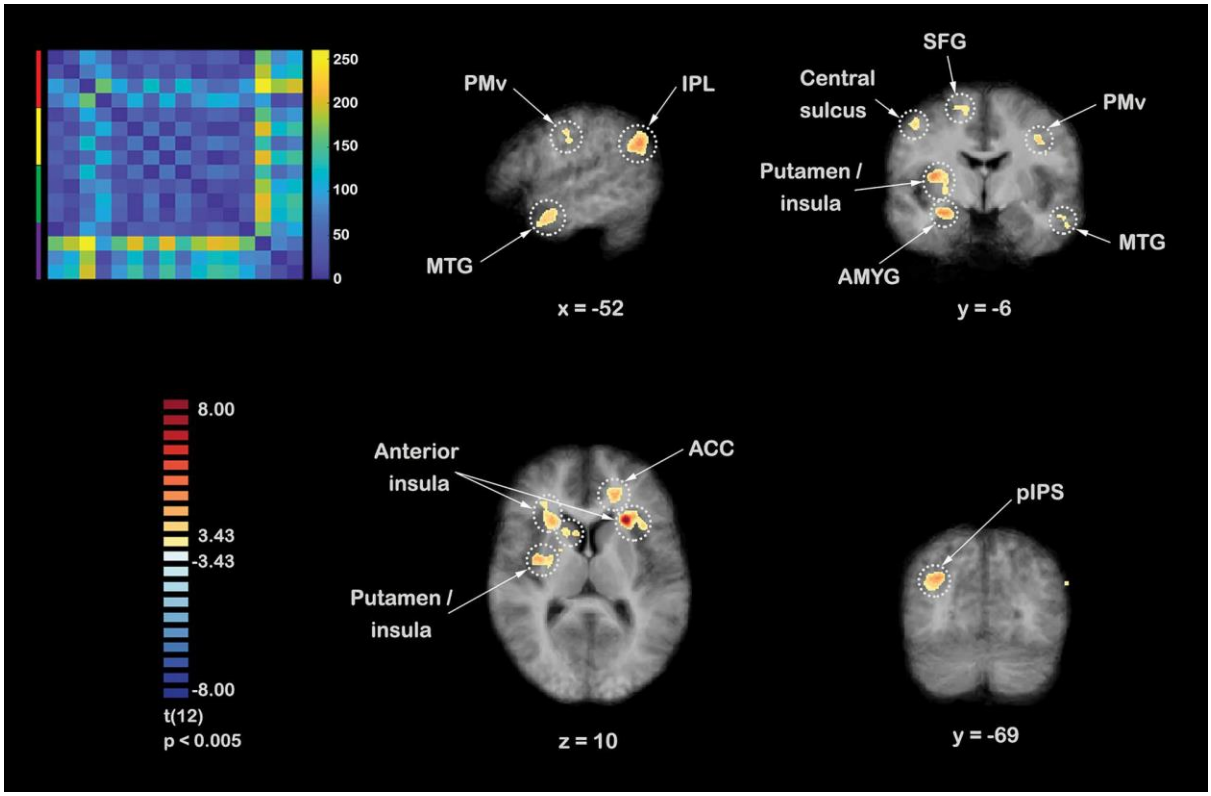


Figure 9: Clusters resulting from the searchlight RSA of the postural feature of limb contraction. The multivoxel fMRI dissimilarity matrices were correlated to the limb contraction RDM (upper left corner). The limb contraction RDM represents pairwise comparisons between the 16 stimuli with regard to limb contraction information averaged over time. The dissimilarity measure reflects Euclidean distance, with blue indicating high similarity and yellow high dissimilarity. Color lines indicate the organization of the RDM with respect to the emotional category (anger: red; happiness: yellow; neutral: green; fear: purple) of the video stimuli. Spearman's rank correlation was used to correlate the limb contraction RDM to the multivoxel fMRI dissimilarity matrices. The resulting maps were z-transformed for each participant. Subsequently, a group-level one-sample t-test against 0was performed (two-tailed, cluster size corrected with Monte-Carlo simulation, alpha level = 0.05, initial P = 0.005, numbers of iterations = 5000). See Supplementary Table R5 in Supplementary Results for more details on location and statistical values of the clusters. Abbreviations: ACC, anterior cingulate cortex; AMYG, amygdala; IPL, inferior parietal lobule; MTG,middle temporal gyrus; pIPS, posterior intraparietal sulcus; PMv, ventral premotor cortex; SFG, superior frontal gyrus.

The present study investigated the mechanisms underlying body expression perception by measuring the brain representation of critical features of body movement and posture. Our results reveal six major findings. First, computationally defined features are systematically related to distributed brain areas. Second, postural rather than kinematic features reflect the affective category structure of the body movements. Limb angles and symmetry were important for differentiating neutral from emotional body movements. Limb angles and especially limb contraction were particularly relevant for distinguishing fear from other body expressions. These two features were represented in several regions including affective, action observation and motor preparation networks. Third, the pSTS differentiated fearful from other affective categories using limb contraction rather than kinematics, despite this area being known for its involvement in biological motion

processing. Fourth, EBA and FBA also showed greater tuning to postural features. Although the pattern of feature representation in these areas was similar, the stimuli representation in EBA was very dissimilar to that of FBA, possibly reflecting their different roles in body processing. Fifth, kinematic and postural feature processing was not segregated into dorsal and ventral streams, with the exception of one feature: velocity. Finally, the brain representation of emotional categories showed a distributed pattern.

By investigating mid level feature processes, this study moves the field of affective neuroscience forward, providing insights into the perceptual features that possibly drive automatic emotion perception. Features at this visual computational level may only partly overlap with feature descriptions used in everyday descriptions of body expressions (Poyo Solanas et al. 2020). Nevertheless, it is important to be aware of the limitations of our findings. For instance, the features defined here were selected due to their relevance in the literature because no feature-based and biologically plausible computational model of naturalistic body expressions is available (Giese and Poggio 2003; Serre 2014). We expect that future studies will also use larger and more diverse stimulus sets with a wider range of affective states and a larger participant sample, also looking into dyadic interactions.

References

Caspers S, Zilles K, Laird AR, Eickhoff SB. 2010.ALE meta-analysis of action observation and imitation in the human brain. Neuroimage. 50:1148–1167.

Coulson M. 2004. Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. J Nonverbal Behav. 28:117–139.

de Gelder B, Snyder J, Greve D, Gerard G, Hadjikhani N. 2004. Fear fosters flight: a mechanism for fear contagion when perceiving emotion expressed by a whole body. PNAS. 101:16701–16706.

de Gelder B. 2006. Towards the neurobiology of emotional body language. Nat Rev Neurosci. 7:242.

De Meijer M. 1989. The contribution of general features of body movement to the attribution of emotions. J Nonverbal Behav. 13:247–268.

Downing PE, Jiang Y, Shuman M, Kanwisher N. 2001. A cortical area selective for visual processing of the human body. Science. 293:2470–2473.

Giese MA, Poggio T. 2003. Neural mechanisms for the recognition of biological movements. Nat Rev Neurosci. 4:179–192.

Giese MA, Rizzolatti G. 2015. Neural and computational mechanisms of action processing: interaction between visual and motor representations. Neuron. 88:167–180.
Grafton ST, Hamilton AF. 2007. Evidence for a distributed hierarchy of action representation in the brain. Hum Mov Sci. 26:590–616.

Kirby LA, Robinson JL. 2017. Affective mapping: an activation likelihood estimation (ALE) meta-analysis. Brain Cogn. 118:137–148.

Kleinsmith A, Bianchi-Berthouze N. 2012. Affective body expression perception and recognition: a survey. IEEE T Affect Comput. 4:15–33.

Lindquist KA, Wager TD, Kober H, Bliss-Moreau E, Barrett LF. 2012. The brain basis of emotion: a meta-analytic review. Behav Brain Sci. 35:121

Milner D, Goodale MA. 2006. The visual brain in action.Oxford (UK): Oxford University Press.

Patwardhan A. 2017. Three-dimensional, kinematic, human Behavioral pattern-based features for multimodal emotion recognition. Multimodal Technol Interact. 1:19.

Peelen MV, Downing PE. 2005. Selectivity for the human body in the fusiform gyrus. J Neurophysiol. 93:603–608.

Piana S, Stagliano A, Odone F,Verri A,Camurri A. 2014. Real-time automatic emotion recognition from body gestures. arXiv preprint arXiv:1402.5047

Poyo Solanas M, Vaessen M, de Gelder B. 2020. The role of computational and subjective features in emotional body expressions. Sci Rep. 10:1–13.

Roether CL, Omlor L, Christensen A, Giese MA. 2009. Critical features for the perception of emotion from gait.J Vis. 9:15–15.

Schwarzlose RF, Baker CI, Kanwisher N. 2005. Separate face and body selectivity on the fusiform gyrus. J Neurosci. 25:11055–11059.

Serre T. 2014. Hierarchical models of the visual system. In: Jung R, Jaeger D, editors. Encyclopedia of computational neuroscience. New York: Springer.

Vaessen M, Abassi E, Mancini M, Camurri A, de Gelder B. 2018. Computational feature analysis of body movements reveals hierarchical brain organization. Cereb Cortex. 1:10.

Vaina LM, Lemay M, Bienfang DC, Choi AY, Nakayama K. 1990. Intact “biological motion” and “structure from motion” perception in a patient with impaired motion mechanisms: a case study. Vis Neurosci. 5:353–369.

Van den Stock J, Tamietto M, Sorger B, Pichon S, Grézes J, de Gelder B. 2011. Cortico-subcortical visual, somatosensory, and motor activations for perceiving dynamic whole-body emotional expressions with and without striate cortex (V1). PNAS. 108:16188–16193.

Wallbott HG. 1998. Bodily expression of emotion. Eur J Soc Psychol. 28:879–896.

2.1.14 Automatic Detection in the Context of Movement with Chronic Pain based on Three Novel Multiple-Timescales Machine Learning Architectures

In this section, we present the results of the exploration of our novel machine learning architectures (*Body Attention Net*, i.e. BANet, *Movement in Multiple Time*, i.e. MiMT, *Global Workspace Network*, i.e. GWN) on the problem of automatic detection of pain and related behaviour from body movement. Please D3.1 for descriptions of the BANet and MiMT which we also refer to as Multi Time neural network (MTNN) or MultiLevNN.

Pain behaviour assessment is an important movement analysis problem in the context of chronic pain physical activities (Cook et al. 2013; Keefe and Block 1982). Automating the assessment of pain behaviour could enable less burdensome, objective measurement as well as open up the opportunity to provide real-time tailoring (e.g. via movement sonification) which fosters engagement of a person with chronic pain with valued physical activities. For example, a sonification framework based on pain behaviour assessment could aim to call the attention of a person with pain to their unhelpful strategies for performing feared or painful movements (Olugbade et al. 2019). Previous bodily-expressed pain behaviour detection studies have focused on classification at single timescales. For example, (Aung et al. 2016) modelled the duration (as a proportion) of pain behaviour in a movement instance based on a simple fusion of motion capture and muscle activity data. Pain level itself could be valuable to assess automatically for the purpose of enabling helpful pacing of everyday physical activities based on the understanding of how pain drives underactivity and overactivity (Olugbade et al. 2019).

For all 3 investigations (on the BANet, MiMT, and GWN respectively), we used the EmoPain dataset (Aung et al. 2016) which contains 3D positions for 26 full-body joints, 13 full-body angles derived from these, and muscle activity data for 4 upper and lower back locations of people with chronic pain and healthy control participants. The data were captured while the participants performed 8 exercise movements (sit-to-stand, stand-to-sit, bend, reach forward, walk, sitting, standing, one leg raised while standing). People with chronic pain provided self-reports of pain after each exercise type on a scale of 0 to 10. Four clinicians further provided annotations for the exercise instances of this group of participants, for 6 pain behaviours (guarding/stiffness, hesitation, bracing/support, abrupt action, limping, rubbing.stimulation) in continuous time and as discrete values 0 for 'not present' and 1 as 'present'.

Study 1: Weighted Fusion of Time and Anatomical Region with The BANet

A contribution of the BANet is its importance weighting of time for each movement dimension (e.g. hip angle) and further weighting of each dimension overall. Further details can be found in the peer-reviewed publication of the study.

Focusing on 5 (sit-to-stand, stand-to-sit, reach forward, bend, one leg raised while standing) of the 8 exercise types in EmoPain dataset, this study was based on the 13 full-body joint angles of the dataset and the angular energies computed from them. Each joint angle is derived from the 3D positions of three consecutive joints while the corresponding joint energy is the square of the change in the angle with respect to the previous timestep. The angle and energy sequences for each exercise type participant were segmented based on fixed window with length = 3 seconds and overlapping ratio = 0.75 based on findings in (Wang et al. 2019), within each exercise instance. Zeroes were used to pad segments at the end of exercise instance and less than the window length. Two data augmentation techniques were then applied to duplicates of these segments to increase the data size, i.e. the number of segments. The first of these adds normalized Gaussian noise (standard deviation = 0.05, 0.1, and 0.15) to the duplicate (based on Wang et al. 2019). In the second, randomly selected (probability = 0.05, 0.1, and 0.15) angles and energies in the duplicate are dropped, i.e. set to 0 (based on Um et al. 2017). The augmentation resulted in 18,653 segments. The ground truth for each segment was set to 'protective behaviour present' if at least 2 of the 4 raters rated any of the pain behaviours as present for half of the segment length and 'absent' otherwise. There were 11,373 *protective behaviour absent* segments and 7,280 *protective behaviour present* segments.

We evaluated the performance of the BANet on automatic discrimination between protective behaviour absent and present classes using leave-one-subject-out cross-validation. To understand the value of approach used in the BANet, we compared its performance on automatic detection of protective behaviour based on these data with the performance of 4 variants of the BANet (BANet-compatibility, BANet-dense, BANet-time-only, BANet-body-only). We also compared the BANet with 3 architectures (bidirectional Long Short-Term Memory neural network i.e. LSTMNN, convolutional LSTMNN, stacked LSTMNN) which are similar to it but do not include weighting, i.e. the machine learning attention mechanism which gives the BANet its name. Table 4 gives an overview of all 8 architectures explored in this study and the hyperparameters used in training the respective models. The Adam optimiser and learning rate of 0.003 was used in all cases.

Table 4
The BANet and 7 peer machine learning architectures that we compared it to.

Architecture	Attention (i.e. weighting) across time	Attention (i.e. weighting) across joint (angle)	Layers Types [number of layers, number of units]	Training batch size
BANet	Yes	Yes but after time attention	1. LSTM [3, 8] 2. 1x1 convolution and softmax (time attention) 3. fully connected [2, 8] and softmax (joints attention) 4. fully connected [1, 2] and softmax	40
BANet-compatible	Yes but after joints attention	Yes	1. LSTM [3, 8] 2. fully connected [2, 8] and softmax (joints attention) 3. 1x1 convolution and softmax (time attention) 4. fully connected [1, 2] and softmax	40
BANet-dense	Yes	Yes but after time attention	1. LSTM [3, 8] 2. fully connected [1, 8] and softmax (time attention) 3. fully connected [2, 8] and softmax (joints attention) 4. fully connected [1, 2] and softmax	40
BANet-time-only	Yes	No	1. LSTM [3, 8] 2. 1x1 convolution and softmax (time attention) 3. fully connected [1, 2] and softmax	40
BANet-body-only	No	Yes	1. LSTM [3, 8] 2. fully connected [2, 8] and softmax (joints attention) 3. fully connected [1, 2] and softmax	40
Bidirectional LSTMNN	No	No	1. bidirectional LSTM [1, 14] and dropout probability of 0.5	40
Stacked LSTMNN	No	No	1. LSTM [3, 28] and dropout probability of 0.5	20
Convolutional LSTMNN	No	No	1. 1x10 convolution, 28 LSTM units, and max pooling	50

The performance of the BANet can be seen in Table 5 in comparison with the other 7 architectures. As can be seen in the table, the BANet outperforms all of the 7 architectures. A paired t test over cross-validation folds, with Bonferroni correction, showed that this is statistically significant ($p<0.05$) for every comparison architecture except the BANet-time-only and BANet-body-only $F(3.072, 89.099)=15.612$, $\mu^2= 0.519$); the significance for the bidirectional LSTM was marginal.

Table 5
Mean F1 score and accuracy of the BANet and comparison architectures (in bold is the best performance and * is used to indicate comparison architectures which performed significantly worse, $p<0.05$, than the BANet).

Architecture	Mean F1 Score	Accuracy	Number of trainable parameters
BANet	0.8440	0.8688	2,131
BANet-compatible	0.5720*	0.6630	6,204
BANet-dense	0.7890*	0.8167	65,430
BANet-time-only	0.7580	0.8060	1,767
BANet-body-only	0.8310	0.8670	2,023
Bidirectional LSTMNN	0.8040	0.8460	14,282
Stacked LSTMNN	0.8120*	0.8534	18,986
Convolutional LSTMNN	0.7370*	0.8059	40,940

Another advantage of the attention weighting of the BANet is that it allows analysis of both temporal and anatomical segment relevance. Figure 9 shows boxplots of the distribution of attention scores (i.e. importance weights) for each joint angle (and its energy) per exercise type. It can be seen that there is a wider distribution of attention scores for the participants with chronic pain, particularly in the exercise segments with protective behaviour absent, compared with the healthy participants. This suggests strong salience of a few antomical segments above the others, perhaps in terms of distinction in timescale, with protective behaviour. The sample plots of temporal attention scores per joint angle and energy in Figure 10 showing larger differences in the timelines for the different joint angles supports this theory.

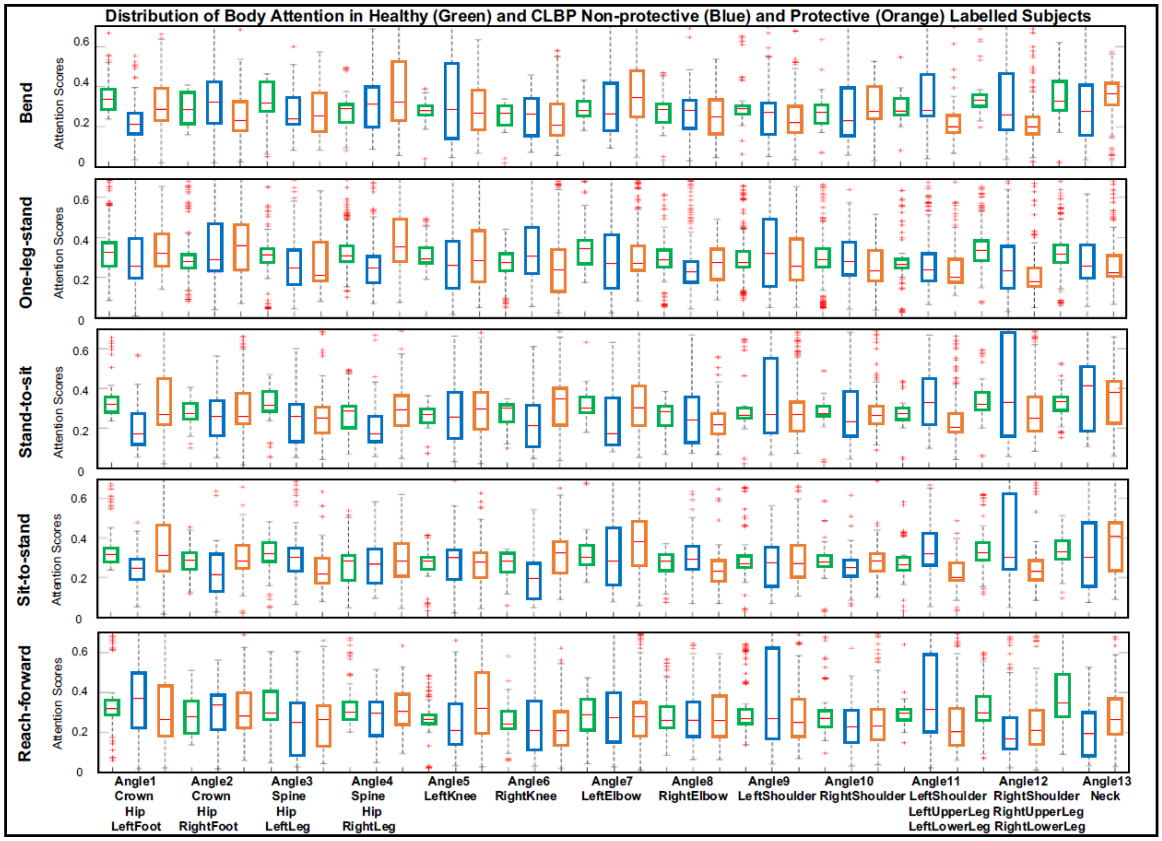


Figure 9: Distribution of attention scores for each joint angle (and its energy) per exercise type. The plots show healthy participants in green, participants with chronic pain and protective behaviour absent in blue, and participants with chronic pain and protective behaviour present in orange.

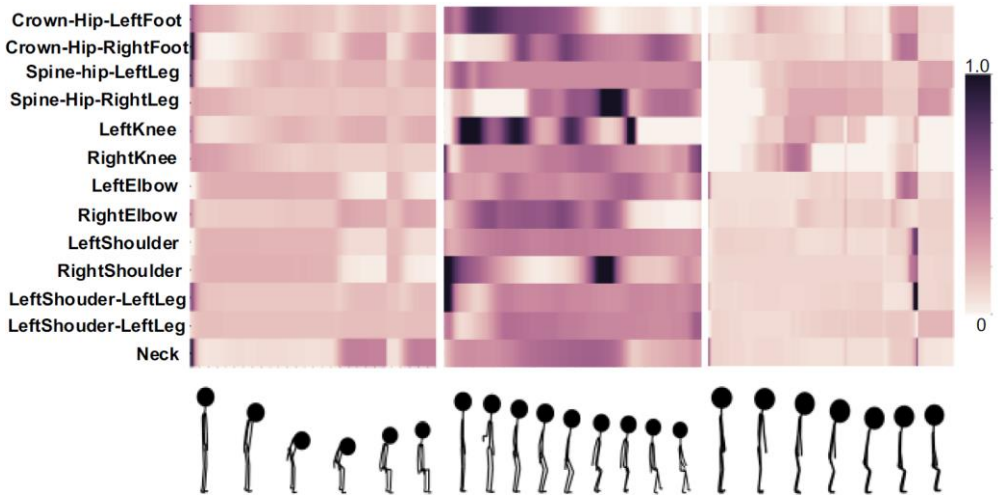


Figure 10: Sample plots of temporal attention per joint angle (and its energy) in a stand-to-sit exercise instance for a healthy participant (left) and two participants with chronic pain (middle and right).

We propose a novel neural network architecture named BANet which performs weighted fusion of movement time and anatomical regions. This approach outperforms similar architectures without explicit weights in fusion, with weights only for time or anatomical region but not both, or with the weighted fusion of anatomical regions before time.

Analysis of these weights, which are learnt by the network based on data, suggests stronger differences in timescales of anatomical segments during anomalous movement behaviour. First, this highlights that multiple timescales occur not just over time itself but also across the different degrees of freedom of movement. We have developed a movement sonification framework that aims to apply multi-dimensionality of time (attention time and the different times of each degrees of freedom of movement) in chronic pain scenarios. On one hand, this could be used to provide self-awareness (attention) in real time to a person with chronic pain about how they are moving. On the other hand, it could serve as to augment an observer's (the person with pain themselves or a clinician) visual assessment of movement. More details about the sonification framework is reported in D3.1. Second, it raises questions about how much the network attention weighting tells us about the timescales involved in the interpretation of movement behaviour by the clinicians who provided protective behaviour labels. We are carrying out further analyses of the attention scores to understand what they imply in this respect. Further, we are additionally conducting an observation study aimed at finding implicit models of pain and movement that physiotherapists use to make clinical observations and interventions.

Study 2: Using The MiMT to Learn Multiple Timescales of Pain Behaviour Labels based on Movement Dimensions with Multiple Timescales

The MiMT models body movement at multiple timescales particularly accounting for independence-cum-coordination between multiple anatomical segments similar to the BANet but also accounting for different timescales of movement interpretation (at level of a single time step and at the level of multiple timesteps). A publication manuscript on this study has been submitted for peer review.

While the joint angles used for Study 1 have the advantage of being location invariant, we chose to use the 3D full-body positions of the EmoPain dataset in Study 2 because they characterise movement execution in a more intuitive way. We excluded eight of the 26 joints (left and right fingertips, ankles, heels, and toes) in our use of the positional data in this study due to the higher level of noise in their position estimates. To minimise the dimensionality of the data, we additionally excluded the crown joint given that the remaining joints include the head and neck. This resulted in 17 full-body joints. We segmented exercise instances in the EmoPain dataset (except the *walking* exercises and for participants with chronic pain alone) using overlapping 3-second windows based on (Wang et al. 2019) (overlap = 0.25 seconds). The label for a frame (timestep) in a segment was set as *of guarding* if at least two raters labelled guarding behaviour as present at that frame, otherwise the label was set as *not of guarding*. The label for a segment (multiple

timesteps) was set as *guarding behaviour* if all the frames in the segment are *of guarding* label, *not guarding behaviour* if all the frames are *not of guarding*, and *mixed* otherwise. We used data augmentation to increase the minority classes at the segment level (*guarding behaviour*, *mixed*) by creating mirror duplicates across permutations of the three axes (based on Olugbade et al. 2020) as well as translated, scaled up/down duplicates. This resulted in 17,185 and 1,394 instances respectively for the training and validation sets.

We evaluated the MiMT on automatic discrimination between of guarding and not of guarding at the frame level and between guarding, not guarding, and mixed classes at the segment level. This evaluation was based on hold-out validation where the subject sets in the training, validation, and test sets are mutually exclusive. To understand the value of the approach of the MiMT (separate but shared time encoding and multiple timescales of the same label), we compared its performance with 3 architectures derived by ablation of the MiMT (MiMT-single-input-time, MiMT-frame-output-time-only, MiMT-segment-output-time-only). Table 6 provides an overview of the differences between the architectures. The time encoder of the MiMT (and the comparison architectures) was based on 3 LSTM layers each with 3 units. Single LSTM and fully connected layers each with 15 units were used for the classifier with additional global average pooling and sigmoid activation for the frame level output and a single layer LSTM and softmax activation after further multiplication with the time encoder output for the segment level output. Each model was trained with the Adam optimizer at learning rate and batch size of 0.005 and 200 respectively.

Table 6 An overview of the architectures compared with the MiMT.

Architecture	Separate but shared time encoding of the input	Frame level output	Segment level output
MiMT	Yes	Yes	Yes
MiMT-single-input- time	No	Yes	Yes
MiMT-frame-output-time-only	Yes	Yes	No
MiMT-segment-output-time-only	Yes	No	Yes

Table 7 shows the performance of the MiMT. As can be seen in the table, the MiMT performs much better than chance level detection (0.5 for the frame label, 0.33 for the segment label). The MiMT further outperforms its three variants suggesting that combination of both the separate but shared time encoding for the input and the multiple output timescales is efficacious.

Table 7 Mean F1 score of the MiMT and comparison architectures (in bold is the best performance).

	Mean F1 score	
Architecture	Frame label	Window label
MiMT	0.63	0.46
MiMT-single-input- time	0.50	0.34
MiMT-frame-output-time-only	0.59	-
MiMT-window-output-time-only	-	0.33

Figure 11 shows two example plots of the activations for each separate (but shared) time encoding. To maximise contrast, we only sampled every 20th frame in these plots. Each band for each group of segments represents the activation for one of the three units of the encoder. Comparing the bands for the lower left and right limb groups of segments clearly show coordination between the two groups of segment yet there are differences in changes in the activations over time further highlighting that different degrees of freedom have different timescales that have moments of synchronization.

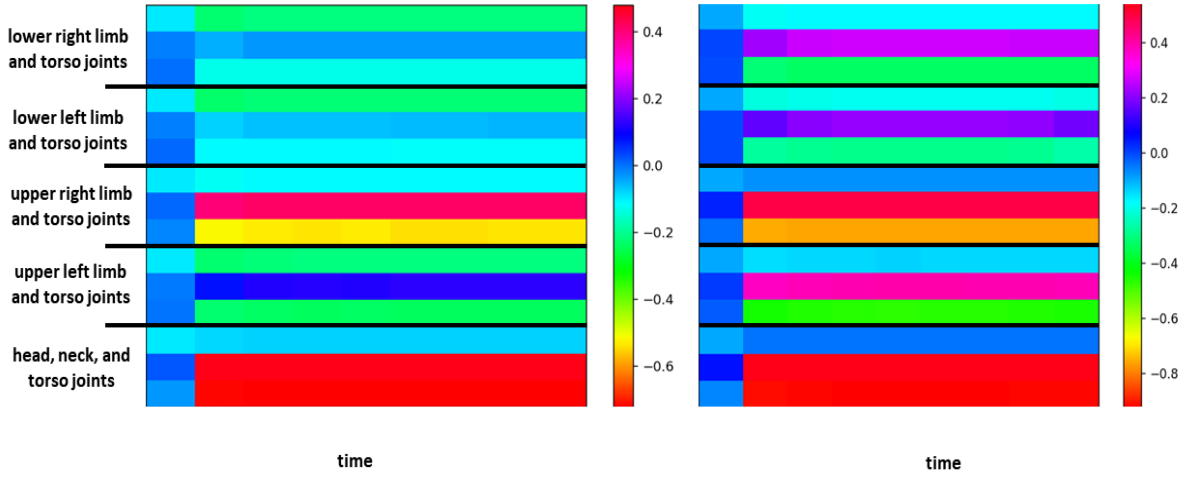


Figure 11: Time encoder activation for two different exercise segments (left, right) and two different exercise types and participants

Building on the BANet which had different but shared time encoding for different groups of anatomical segments, we propose the MiMT architecture that additionally learns multiple label timescales simultaneously. Our findings suggest that the two elements are together valuable for modelling the multiple timescales in movement data. They further highlight the importance of investigating timescales of movement assessment as is one of the aims of our observation study with physiotherapists. We plan to extend the MiMT by integrating it with other multiple timescales machine learning architectures.

Study 3: Multimodal Movement Data Fusion based on the GWN

The GWN addresses the differences in timescales between multiple modalities of movement, using the machine learning attention (i.e. weighting) mechanism for fusion similar to the BANet although the attention module it uses is based on self-attention such that each modality assigns weights to itself and each of the other modalities. In-depth description of the GWN can be found in the peer-reviewed publication of the study (currently under embargo until the publication date).

In this study, we use both the 3D full-body positions and the muscle activity data of the exercise instances in the EmoPain dataset. Since the exercise instances were of varying lengths, zero padding at the start of each instance was used to make them of uniform lengths. To increase the data size, i.e. the number of instances, the same mirror reflection of duplicates used in Study 2 was used here except that the reflection was only done around the y-axis. Three rotation angles were used (90°, 180°, 270°) resulting in 800 data instances in total. The labels for the instances from the healthy control participants was set to *no chronic pain*. The instances from the participants with chronic pain was labelled as *with chronic pain*. The instances from this group of participants was further labelled as *zero pain* if the participant reported pain intensity of 0 for that instance, *low level pain* if the pain intensity was otherwise ≤ 5, and *high level pain* for pain intensity > 5.

We explored the GWN in two separate but related classification tasks: recognition of chronic pain instances and pain level classification. The evaluation of the GWN in these tasks was based on leave-one-subject-out cross-validation. For each task, we compared the performance of the GWN with a baseline architecture where fusion of the multimodal data is based on simple fusion. Table 8 outlines the difference between the GWN and the baseline. In the EmoPain dataset, the two modalities were of the same sampling rate of 60Hz, the muscle activity data having been downsampled from its original 1000Hz. For the *recognition of chronic pain instances* task, both modalities were further resampled to 10Hz to manage the dimensionality of the training data. While the positional data has 3x26=78 dimensions, the muscle activity data has only 4. There were 64 units in LSTM layer which serves as attention time encoder and ordinary time encoder for the GWN and baseline architecture respectively. The Adam optimisation algorithm was used for training the models, with learning rate and batch size of 0.001 and 32 respectively, based on grid search.

Table 8
An overview of the GWN and the baseline used for comparison

Architecture	Maps different sampling rates and/or degrees of freedom in the multiple modalities to a uniform sampling rate and dimensionality	Weighted fusion of multiple modalities (based on self-attention)	Propagation of the weightings over time
GWN	Yes	Yes	Yes
Simple concatenation	No	No	No

The performance of the GWN is shown in Table 9. Both the GWN and the baseline comparison architecture perform much better than chance level classification (0.5 for the recognition of chronic pain instances, 0.33 for pain level classification), the GWN clearly outperforms the baseline architecture. A Wilcoxon signed rank test across the cross-validation folds indeed shows statistically significant difference between their performances for the recognition of chronic pain instances in particular.

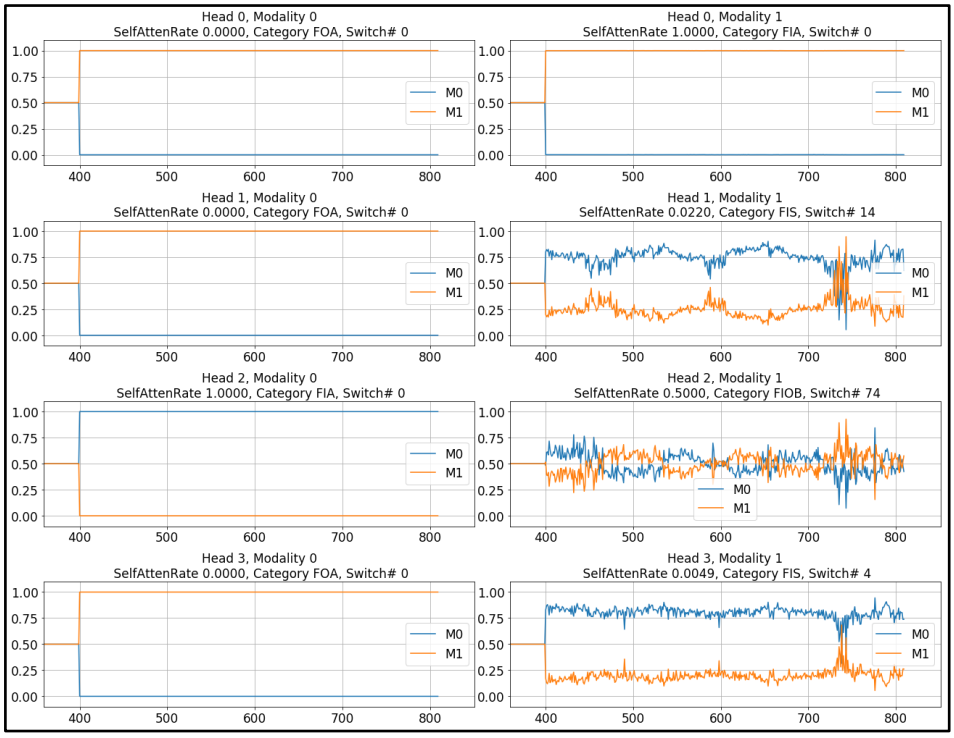
Table 9
Mean F1 scores of the GWN and the comparison architecture (in bold is the best performance and * is used to indicate performance significantly worse, $p<0.05$, than the GWN).

	Mean F1 scores	
Architecture	Recognition of chronic pain instances	Pain level classification
GWN	0.92	0.75
Simple concatenation	0.72*	0.63*

We further analysed the self-attention scores of the GWN and found 5 main temporal patterns of attention. Table 10 gives an overview of these patterns and Figure 12 gives examples of these pattern types.

Table 10
The 5 temporal patterns of self-attention found.

Short name	Long name	Pattern (weighting are between 0 and 1 and add up to 1 by each modality)
FIA	Favours Itself Always	weighting for self > 0.5 100% of the time
FOS	Favours Other Sometimes	weighting for self < 0.5 up to 40% of the time
FIOB	Favours Itself and Other in Balance	weighting for self > 0.5 40-60% of the time
FIS	Favours Itself Sometimes	weighting for self < 0.5 less than 40% of the time
FOA	Favours Other Always	weighting for self > 0.5 0% of the time



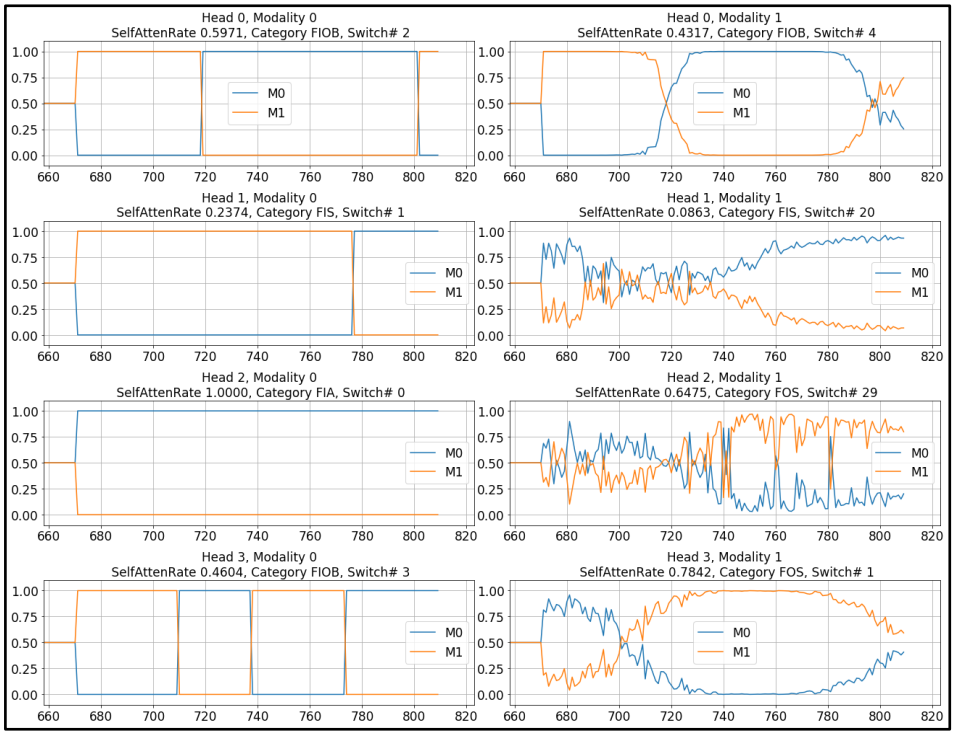


Figure 12: Plots for 2 exercise instances (top and bottom respectively) showing self-assigned attention scores versus time (M0 = positional data, M1 = muscle activity data). Plots on the left and right correspond to attention scores assigned by Modality 0 (M0) and Modality 1 (M1) respectively. The 'Head' identifier refers to the corresponding component of the attention computation ensemble; 'Switch #' refers to the number of attention switches that occur over time; and the category identifier refers to the corresponding attention pattern in Table 7.

One of the merits of the GWN approach is that it can account for noise with an unknown timescale to be accounted for. We demonstrate this by conducting an investigation of the effect of noise on the performance of the GWN and comparing the temporal patterns of self-attention with and without noise. We use Gaussian noise sampled with standard deviation equal to one-tenth of the standard deviation of the respective data modality (i.e. noise standard deviation of 10 for the positional data and 0.001 for the muscle activity data). Table 11 shows the performance of the GWN in pain level classification with and without noise in the modalities. We found no significant difference (p<0.05) between the performance of the GWN in both cases regardless of whether the noise was added to the positional data or to the muscle activity data suggesting that the GWN's approach to multimodal fusion indeed controls the effect of noise on the automatic detection task.

Table 11
Mean F1 scores of the GWN in the pain level classification task with and without noise in the modalities.

	Mean F1 scores		
Architecture	No noise	Noise in 3D position data	Noise in muscle activity data
GWN	0.75	0.72	0.72

Table 9 shows how noise affected the distribution of the 5 temporal self-attention patterns. For the majority of data instances, the positional data modality assigns a higher weight to itself all through the time. For most of the remaining instances, this modality assigns a higher weight to the other modality all through time. The muscle activity modality also assigns a higher weight to itself all through time for a majority of the data instances, but unlike the positional data, for most of the remaining instances it instead shows the FOS pattern where it still assigns a higher weight to itself and not the other modality through most of time. Although this patterns distribution persists when noise is added to the muscle activity data, when noise is added to the positional data the distribution changes such that the positional data assigns higher weights to the muscle activity data all through time for much more data instances than those for which it assigns higher weights to itself all through time. This further highlights that the GWN enables noise in the modalities to be addressed in its fusion of multiple modalities. We speculate that the lack of difference in pattern distribution when noise was added to the muscle activity data is perhaps due to the lower dimensionality (4) of that modality, and so lower impact of noise overall, compared to that (78) of the positional data.

Table 12
The relative frequency of the 5 temporal attention patterns for each modality (M0=positional data, M1=muscle activity data). See Table 7 for the description of the patterns.

	FIA		FOS		FIOB		FIS		FOA	
	M0	M1	M0	M1	M0	M1	M0	M1	M0	M1
No noise	0.51	0.40	0.04	0.29	0.03	0.05	0.05	0.15	0.37	0.11
Noise in M0	0.31	0.43	0.08	0.36	0.02	0.05	0.11	0.10	0.48	0.07
Noise in M1	0.50	0.46	0.02	0.27	0.02	0.05	0.06	0.09	0.41	0.13

We propose the GWN which fuses data from multiple modalities with different sampling rates and/or dimensionalities. We showed that the GWN not only outperforms simple concatenation of these data for pain classification based on positional and muscle activity data but its good performance persists even in the presence of noise of an unknown timescale in either of the two modalities. While the modalities used in our empirical study

had been (re)sampled to the same sampling rate, the timelines and timescales of events in the two modalities could still be different.

References

Aung MSH, Kaltwang S, Romera-Paredes B, Martinez B, Cella M, Valstar M, Meng H, et al (2016) The Automatic Detection of Chronic Pain- Related Expression: Requirements, Challenges and a Multimodal Dataset. IEEE Transactions on Affective Computing 7(4): 1–18.

Cook KF, Keefe F, Jensen MP, Roddey TS, Callahan LF, Revicki D, Bamer AM, et al (2013) Development and Validation of a New Self-Report Measure of Pain Behaviors. Pain 154 (12): 2867–76.

Flash T, Hogans N. 1985. The Coordination of Arm Movements: An Experimentally Confirmed Mathematical Model. J Neurosci. 5:1688–1703.

Keefe FJ, Block A (1982) Development of an Observation Method for Assessing Pain Behavior in Chronic Low Back Pain Patients. Behavior Therapy 13 (4): 363–75.

Olugbade TA, Singh A, Bianchi-Berthouze N, Marquardt N, Aung MSH, Williams A (2019) How Can Affect Be Detected and Represented in Technological Support for Physical Rehabilitation? Transactions on Computer-Human Interaction.

Olugbade T, Newbold J, Johnson R, Volta E, Alborno P, Niewiadomski R, Dillon M, Volpe G, Bianchi-Berthouze N (2020) Automatic Detection of Reflective Thinking in Mathematical Problem Solving based on Unconstrained Bodily Exploration. IEEE Transactions on Affective Computing.

Um TT, Pfister FM, Pichler D, Endo S, Lang M, Hirche S, Fietzek U, Kulić D (2017) Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks. In Proceedings of the 19th ACM International Conference on Multimodal Interaction: 216-220.

Wang C, Olugbade TA, Mathur A, De C Williams AC, Lane ND, Bianchi-Berthouze N (2019) Recurrent network based automatic detection of chronic pain protective behavior using mocap and semg data. In Proceedings of the 23rd International Symposium on Wearable Computers:225–230.