Movement Representation Learning for Pain Level Classification

Temitayo Olugbade*§, Amanda C de C Williams*, Nicolas Gold*, Nadia Bianchi-Berthouze*†
*University College London, UK, §University of Sussex, UK, †n.berthouze@ucl.ac.uk

Abstract—Self-supervised learning has shown value for uncovering informative movement features for human activity recognition. However, there has been minimal exploration of this approach for affect recognition where availability of large labelled datasets is particularly limited. In this paper, we propose a P-STEMR (Parallel Space-Time Encoding Movement Representation) architecture with the aim of addressing this gap and specifically leveraging the higher availability of human activity recognition datasets for pain-level classification. We evaluated and analyzed the architecture using three different datasets across four sets of experiments. We found statistically significant increase in average F1 score to 0.84 for pain level classification with two classes based on the architecture compared with the use of hand-crafted features. This suggests that it is capable of learning movement representations and transferring these from activity recognition based on data captured in lab settings to classification of pain levels with messier real-world data. We further found that the efficacy of transfer between datasets can be undermined by dissimilarities in population groups due to impairments that affect movement behaviour and in motion primitives (e.g. rotation versus flexion). Future work should investigate how the effect of these differences could be minimized so that data from healthy people can be more valuable for transfer learning.

Index Terms—Activity recognition, affect recognition, body movement, chronic pain, representation learning, transfer learning.

1 Introduction

Human movement modelling is an important topic relevant across several disciplines and pivotal for a wide variety of applications [1]. While areas such as action/activity recognition have benefited from extensive interest and a large number of (open) datasets, higher order problems like affect recognition lag behind [1]. For automatic recognition of affective experience from movement, this is despite consistent findings and discussions on the role of the body in experience and expression [2], [3]. In this paper, we explored the possibility of leveraging the larger number of datasets available for movement modelling problems like action/activity recognition in pushing the bounds on affect recognition. We focused on the context of chronic pain, a prevalent condition where pain experience significantly affects engagement in harmless everyday physical activities of value, e.g. activities at home [4].

Findings from human-computer interaction studies highlight the possibility of using technology to support physical rehabilitation and self-management of chronic pain [4], [5]. For example, pain-aware technology could help people with the condition incorporate helpful breaks between activities on lower level pain days when they are more likely to overdo physical activity [4]. This can be valuable as it is important to gradually build capability rather than be caught in a cycle between overdoing and underdoing [4], [6]. The problem of pain level classification during everyday physical activities of people with chronic pain still remains a challenge [7]. Movement in such settings is complex, e.g. with no clear demarcation between actions (e.g. reaching forward and walking) or activities (e.g. washing up dishes and tidying up the kitchen). So, we investigate movement representation learning as an approach to addressing the challenge in pain level classification leveraging well-resourced tasks like action/activity recognition. Representation learning enables harnessing of datasets (including those without pain labels) for modelling unspecified features of movement that are informative for unseen types of labels.

While there have been a number of studies on movement representation learning as will be discussed in the next section, there is limited application to the recognition of affect. Further, the majority of previous work in the area of activity recognition where it has usually been employed focus on datasets captured from healthy population groups to the exclusion of people with movement disorders or conditions that can affect movement behaviour during physical activity. Findings in [8] point to the need to understand the influence of differences in population groups on transfer of representation between datasets. The authors found deterioration by 20% in activity recognition performance when data from healthy people was used to train the model evaluated on data from people with Parkinson's disease. The study in [7] similarly showed considerable deterioration in recognition performance for two activity types when data for healthy people was included in training and test data compared with when only data for people with chronic pain was used. Thus, in the work presented in this paper, we investigated movement representation learning and transfer to movements of people with chronic pain during everyday physical activity, with a threefold contribution:

1) We propose a *Parallel Space-Time Encoding Movement Representation* (P-STEMR) neural network architec-

• *T. Olugbade, A. C. de C. Williams, N. Gold, and N. Bianchi-Berthouze are with University College London, London, United Kingdom. T. Olugbade is also with University of Sussex, Brighton, United Kingdom. E-mail: t.olugbade@sussex.ac.uk*

Manuscript received xxx; revised xxx.

ture that leverages self-supervision at latent layers as well as action recognition for representation learning from body movement data. Beyond the wide availability of open datasets, action recognition represents a fundamental movement perception task [31] and so, it is expected that representations learnt based on action recognition can inform higher order tasks such as pain level recognition. We adapted the model to motion capture data in its evaluation reported in this paper. However, the architecture can easily be extended to other types of movement data, e.g. acceleration, angular velocity.

2) We report our exploration of the performance of the P-STEMR model for action recognition based on end-to-end supervised learning to understand the efficacy of the architecture on this transfer source task. We evaluated performance for two different datasets, one from healthy people and another from people with chronic pain. We additionally analyzed the model to understand the value of our parallel spatio-temporal encoding approach. Further, we examined the effect of two model input types and two model output formats on performance.
3) Finally, we present a rigorous study of the efficacy of the P-STEMR model for movement representation learning. First, we evaluated transfer between different population groups (healthy and chronic pain) for the same task (action recognition) to understand how differences in population groups affects performance. Second, we evaluated transfer between different tasks (action recognition and pain level classification) and settings (lab settings and real world) but for the same population group (people with chronic pain). For further insight into how well representations transferred in the latter case, we also qualitatively analyzed the action recognition performance for the target dataset.

2 Related Studies

2.1 Pain Level Classification

While other modalities like facial expression have been explored for pain level recognition [9], body movement is of particular relevance in physical functioning contexts [10]. Indeed, there have been several studies on pain level classification based on movement behaviour of people with chronic musculoskeletal pain. However, they have all been based on datasets captured in instructed or exercise movements in lab settings [11]. For example, the multilevel hierarchical approach in [12] was evaluated on the EmoPain dataset [13] that was captured during exercise movements (e.g. sit-to-stand, reaching forward) performed in the lab. They obtained 0.79 average F1 score based on hold-out validation with unseen subjects for two pain classes. In [7] where baseline detection based on messy, real world data was investigated, performance was average F1 score of 0.62 for two-level pain classification based on cross-validation with unseen activity instances. The lower performance on the real world data is due to the typical challenges in such settings where dimensions (e.g. number of wearable sensor units) and size (number of data instances) available for modelling are minimal. Further, in addition to the complexity of movements in real life physical functioning, the EmoPain@Home dataset used in [7] has differences between homes that result in differences in the execution of the same activities and the strategies used to deal with increase in pain intensity. In our paper, we explored the use of representation learning for addressing this challenge of such real world data.

2.2 Representation Learning of Human Movement based on Wearable Sensor or Motion Capture Data

Representation learning of human movement beyond computer vision is a relatively recent research area paved by works like Saeed et al.'s [14]. The authors proposed a self-supervised architecture where multiple transformations of movement data are learnt. They explored eight types of transformation across both spatial and temporal dimensions (e.g. scaling, time reversal, subsegment resampling, channel shuffling), with a multilayer perceptron (MLP) decoder for each transformation. The encoder for the learner architecture was based on a convolutional neural network (CNN) and the trained encoder was reused (with the weights of the first two of three layers frozen) with a new MLP for human activity recognition (HAR). Evaluation of their model on six datasets based on angular velocity and acceleration data captured in everyday movement settings, e.g. sitting, walking, showed performance better than representation learning using an autoencoder. Further analysis showed that using multitask learning based on all proposed transformations led to better performance than single task learning with only one transformation at a time even if there were differences in the individual efficacy of each transformation type.

Several studies have since built on this work including the work of Tang et al. [15] who first carefully selected a subset of training data to use for the representation learning. As the target task was to be HAR, they used a model trained for this task in a normal supervised manner to auto-label the unlabelled training data and selected the subset for which there was good HAR prediction confidence. Their representation learning architecture was similar to that of Saeed et al. [14] except that it included an additional task for HAR. Across seven out of eight datasets, the performance obtained with their proposed model was better than performance using the model of Saeed et al. [14] whether or not the data for the representation learning and final HAR modelling came from the same dataset. In Khaertdinov et al. [16], contrastive learning was used to learn the transformations, where different transformations of the same data instance would be a positive pair while transformations of different data instances would be a negative pair. Their model led to better HAR performance than the model of [14].

A number of studies have further extended the model of Khaertdinov et al. [16]. For example, Liu et al. [17] explored transformations in both time and frequency domains. Their transformations in the frequency domain included phase shifts and filtering of high frequencies. Also, in place of the typical CNN encoder, they employed a short-term Fourier CNN (based on convolution of short-term fast Fourier transforms at different timescales) that led to better performance than the vanilla CNN. Unlike in Saeed et al. [14], individual

transformations performed very well across datasets. In Gao et al. [18], investigation was done on full-body 3D joint positions data rather than the angular velocity and acceleration data focused on in prior studies. They applied rotation and translation transformations only, based on rotation matrices. Using a ResNet [19] encoder, they obtained performance better than feature engineering methods, such as skeletal quads [20], histogram of oriented principal components [21], on the NTU RGB+D dataset [22]. The model proposed by Zhang et al. [23] is similarly based on motion capture data although they differently used a graph convolutional network encoder (with convolution done on both spatial and temporal dimensions in serial order) instead of the ResNet. In addition, they applied time cropping together with the rotation transformation technique used by Gao et al. [18]. Further, their network used a twin-branch architecture where two different but identical encoding submodules process two transformations respectively and the branch for one transformation is frozen during training. Their model performed better than both fully supervised and other self-supervised (such as generative adversarial network, autoencoders, and contrastive learners) methods based on two datasets including the NTU RGB+D.

2.3 Leveraging Related Datasets for Affect Recognition

The studies reviewed in the previous section highlight the possibility of learning movement representations that can be transferred between tasks or datasets [1]. Studies such as [24], [25] further demonstrate advantage in applying representation learning to leverage related datasets for affect recognition tasks in particular.

In [24], a speech dataset annotated for speaker recognition tasks (source dataset) was explored to improve affect recognition performance for a different speech dataset (target dataset). In particular, they compared three different applications of a model trained on the source dataset for the target task. Simply using the source model (with the output layer removed) to extract features for the target task did not achieve better affect recognition performance than learning features based on the target task only. Finetuning the source model for the target task instead also did not improve performance. However, combining feature learning based on the target task together with features extracted using the source model led to considerable increase in performance for one target dataset explored and a slight increase in performance for another. In similar work in [25] where the input data was face images, one of the trained object recognition models that is widely used in computer vision tasks was explored for affect recognition. Their findings showed significant improvement in performance on a different target dataset whether the source model was used simply for feature extraction or if finetuned to the target task. However, unlike in [24], finetuning to the target task considerably led to the best performance.

There is limited exploration of such methods in affective movement recognition despite the large number of movement datasets that exist for secondary use [1]. In our work in this paper, we address this gap by exploring movement representation learning for pain level classification. Unlike [24], [25] where the target datasets were acted, captured in controlled lab settings or based on movies and TV shows, our target dataset is messy real-world data (EmoPain@Home [11]) captured from people with chronic pain while they performed everyday physical activities, e.g. changing bedsheets, in their homes. As discussed in the introduction, this is a relevant problem that remains challenging.

3 P-STEMR: Parallel Space-Time Encoding Movement Representation Architecture

The P-STEMR model that we propose is illustrated in Figure 1. This movement representation learning model is made up of: a) an encoder network that processes the spatial and temporal dimensions of the input movement data in parallel (rather than the traditional serial order); b) an action/activity recognition network that is used to train the encoder network; and c) a comparison network that regularizes the model by requiring similarity in the latent output (embeddings) obtained for the input movement data and a transformation. All three networks are trained together end to end in the P-STEMR architecture. Each is described below, and the code for the architecture is available at https://github.com/EnTimeMent/pstemr.

3.1 Parallel Space-Time Encoder Network

The input to the network is a sequence of 3D joint angles $[\theta^d_{j_1}, \theta^d_{j_2}, \cdots, \theta^d_{j_\tau}, \forall j \in J, \forall d \in D]$ where J is the set of joints, D is the set of axes (xy, yz, and xz), and τ is the length of the sequence. Unlike joint positions data, joint angles data enables combination of and/or transfer between datasets captured using different types of motion capture sensors, which would have different reference points underlying the 3D position information that they provide. For uniformity, a custom function could be used to compute the 3D joint angles, instead of using sensor-specific angular data. The 3D joint angles input is first passed through a batch normalization layer before being processed in parallel by the spatial and temporal encoding networks.

We use a graph convolutional network (GCN) [26] as the spatial encoder based on studies such as [23], [27]. The advantage of the GCN is that it leverages the graph structure of the joints in a skeleton. While [23], [27] used joint positions as input for their GCN, similar to [7] we used joint angles instead. Figure 2 shows an example graph input structure based on joint angles data for our GCN. Our time encoder is a long short-term memory neural network (LSTMNN) [28], [29]. The outputs of the spatial and temporal encoders are combined using a dot product.

The encoder network further includes a transformation layer. The rationale for the transformation layer is based on the comparison network discussed in Section 3.3. The layer computes a transformation of the input data after batch normalization. This transformation is also processed through the GCN, LSTMNN, and dot product of the network, similar to the twin-branch approach used in [23] although with a few striking differences. In [23], one of the two network branches that process two different transformations respectively is frozen while training the network. In our model, both the transformation and the original input contribute to the training loss. Further, unlike in [23] where two different (but identical) encoding submodules are used on their two

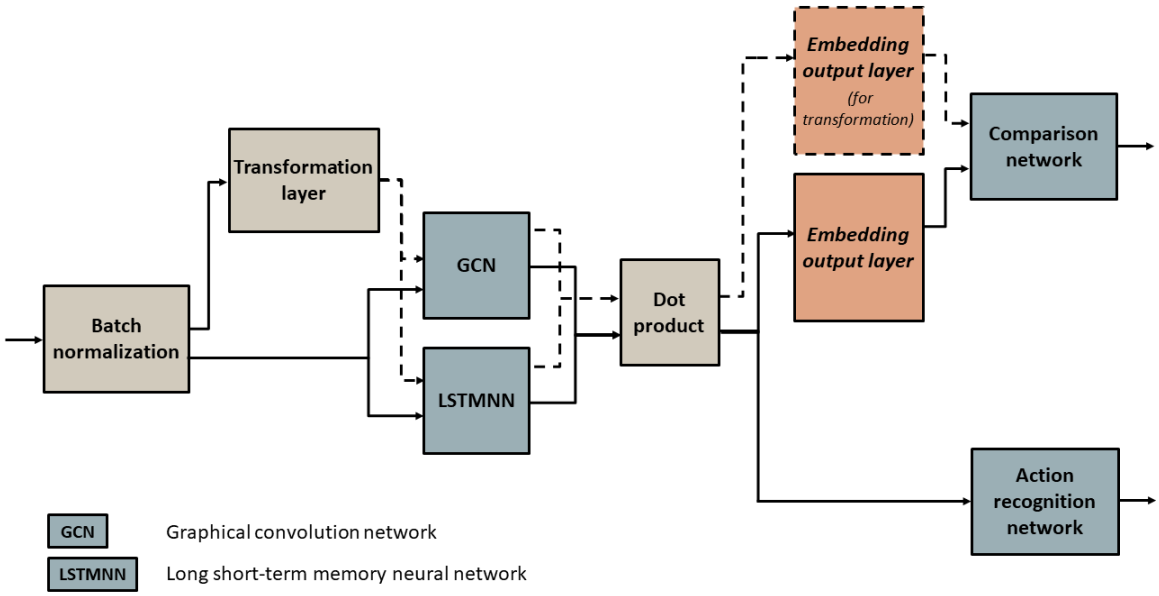

Fig. 1. Our proposed Parallel Space-Time Encoding Movement Representation (P-STEMR) architecture

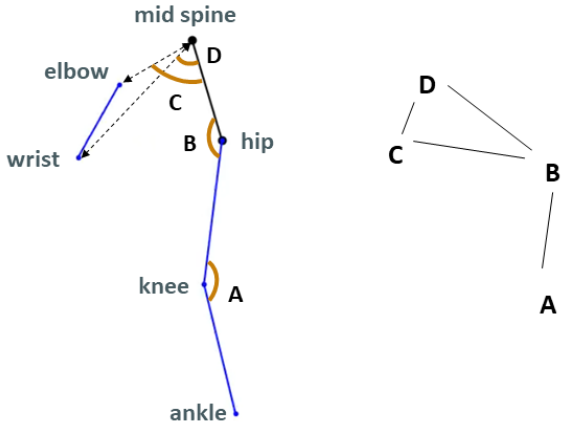

Fig. 2. Illustration of the graph structure for our GCN based on angles input. The graph on the right shows the nodes and edges for our 4-joint angles data, based on position data for 6 joints shown to the left.

transformations, the same encoding submodule is shared by the input and its transformation in our model. The role of the transformation in our P-STEMR model is discussed below with the comparison network. We used a simple scaling (by multiplying with a scalar value) and translation (by adding a scalar value) transformation.

3.2 Activity Recognition Network

The *Activity Recognition Network* is a MLP that takes in as input the output of the encoder network described above and outputs action labels $\hat{y} = \{\hat{a}_c^i, \forall c \in C, \forall i \in n\}$ where C is the set of action classes, n is the set of movement instances in the given batch, and $\hat{a}_c^i$ is the predicted proportion of action class c in data instance i. A loss function l_a is applied on this output. We used a weighted mean square error (MSE) loss that multiplies the MSE by a weight w_i:

$$w_i = \sum_{c=1}^{C}((1 - \frac{\sum_{i=1}^{n} a_c^i}{\sum_{c=1}^{C}\sum_{i=1}^{n} a_c^i}) + \gamma)^{\beta} a_c^i \quad (1)$$

which penalises shorter and less frequent actions in y less than longer or more frequent ones. The $\sum_{i=1}^{n} a_c^i / \sum_{c=1}^{C}\sum_{i=1}^{n} a_c^i$ term in the equation computes the proportional frequency and duration of a_c^i in y. We use two parameters γ and β to control the penalty weighting. $\gamma = 1$ ensures that the penalty term is greater than 1 so that increasing β, $\beta > 1$ will give a stronger weighting.

3.3 Comparison Network

The rationale behind including the *Comparison Network* in the architecture is that $f(\phi)$ should be equal to $f(\phi')$ for movement data ϕ and its transformation $\phi' = g(\phi)$ if $g(.)$ is a function that preserves high level movement characteristics. The comparison network thus takes in as input $f(\phi)$ and $f(\phi')$, which represent the embedding output of the encoder network for the joint angle input and its transformation (through the transformation layer) respectively. The network then performs a subtraction operation to compare the two, with output $\Delta f = f(\phi) - f(\phi')$. A mean absolute error loss l_c between Δf and $\mathbf{0}$ is used such that the full P-STEMR model is trained end to end with loss $l = l_a + l_c$.

4 Experiments

Here, we describe parameter settings for the P-STEMR model in our experiments and datasets used. Details about the experiments are reported in Section 5 with the results.

4.1 P-STEMR Model Parameters

For the scaling transformation layer, we used a random number in the range [-2, 2] for our experiments. We then

applied translation by adding a random number in the range [-0.5, 0.5]. These two ranges were chosen arbitrarily for data normalized to zero mean and unit standard deviation in the batch normalization layer of the P-STEMR model. Both scaling and translation were applied uniformly to all dimensions, i.e. to the xy, yz, and xz axes.

We used a single GCN with 3 layers each of 20 units and a 1-layer LSTMNN in the encoder network. For the action recognition network, we used a 3-layer MLP. We trained the model with an Adam optimizer with no learning rate decay.

4.2 Datasets

We used 3 datasets. The primary target was the EmoPain@Home dataset [11] while the primary source was the EmoPain dataset [13]. We additionally explored the NTU RGB+D dataset [22] as either source or target. Their use for this study was approved by our research ethics committee (ref. 5095/001). All 3 consist of 3D motion capture data but were captured in different settings, from different participants, and with different types of sensors. The datasets and the rationale for selecting them are described below.

4.2.1 EmoPain@Home

The EmoPain@Home was selected as our primary target dataset as it is the only dataset on body movement of people with chronic pain during everyday physical activities. It captures the complexity of real world settings where pain level classification is challenging due to high variations in movement strategies between people as well as the differences in home settings for different participants that can further increase variation in movement strategies.

The dataset was captured from 9 people with chronic pain and 9 healthy people during everyday functioning around the home, e.g. washing up, over multiple days to maximize variability in physical and emotional experience. However, we only used the data for people with chronic pain since our primary interest is in the classification of the pain levels in this group. The dataset was captured using low-cost wearable sensors and is made up of joint positions for the mid spine, hip, right knee, right ankle, right elbow, and right wrist (see Figure 2-Left). A minimal set of sensors was used to minimize the burden on participants. It is representative of real world use of sensors where it is impractical and may be challenging for users to attach multiple individual sensor units [11].

The dataset includes self-reports of pain, worry, and confidence for every activity instance where the researcher was present remotely to record the verbal self-report. We focus on activity instances with self-reports in this paper and as a first step, we investigated classification of pain only. Pain intensity was reported on a 0-to-10 scale every minute during each activity. As pain intensities were provided at roughly every minute of each activity instance, we segmented these instances into segments of 2,400 frames (sampling rate = 40Hz). This led to $N = 226$ instances. In our experiments, the label for each segment was the pain level computed from the pain intensity reported at the end of the given segment. Due to the limited number of instances per intensity on the pain scale, we focused on two levels of pain, *low level pain* for intensity less than 5 and *high level pain* for intensity greater than or equal to 5.

4.2.2 EmoPain

We chose the EmoPain dataset as the primary source dataset as it is the only other dataset on body movement of people with chronic pain, the same population group as the primary target (EmoPain@Home), and it includes action labels which are needed for the use of our P-STEMR model. Thus, this dataset enables investigation of the influence of differences in population group and task on the representation learning transfer between datasets.

The dataset was captured from both people with chronic low back pain and healthy people while they performed exercise movements, e.g. sit-to-stand, in lab settings. It also includes sitting or standing still, which can be also challenging for people with low back pain. In our experiments, we only used the data for participants with chronic pain for a similar population group with the primary target dataset. We used data from 18 participants with chronic pain.

The dataset consists of full-body joint positions and EMG data, but we used joints positions data alone and only for the same 6 joints of the primary target (instead of the 26 available joints). The data was captured using a high-fidelity sensor system based on gyroscopes. Of the annotations in the dataset, we focused on the action labels, provided per frame, which was necessary to train the P-STEMR model. We collapsed the 8 action types in the dataset into 6. First, we combined *forward reach* and *bend* labels given that the two movements are similar and are often combined in some form in real world activities. In addition, there are more variants of the forward reach and bend movements in real life than are represented in the EmoPain dataset and learning the two types of EmoPain movements together might benefit recognition of such variants in the more in-the-wild physical activity. The dataset was then segmented into instances of 240 frames (4 seconds for the EmoPain dataset) with no overlaps. 240 was chosen as a factor of the segment length for the target pain level classification based on the EmoPain@Home dataset. Next, since the primary target dataset was captured from only one side of the body and the *standing on one leg* in the EmoPain dataset is a laterally unsymmetrical movement, we did not use data segments with majority frames having this label, and we also did not include the label in our experiments.

The dataset contains movement sections in which the participant took very short breaks between the instructed action types or transitioned between the end pose for one action type into the start pose for the next. Such transition periods were not labelled with the actual actions that the participants performed but were rather all simply given the label *transition*. As such labelling is not of benefit for action recognition in the real world and can actually be confusing for the machine learning model since it covers several different types of action, we did not use data segments with majority frames having this label. This led to $N = 2,022$ instances in total. The label for each segment was the proportion of frames of each of 6 action types.

4.2.3 NTU RGB+D

We selected the NTU RGB+D as the dataset of healthy people because of its large size and similarity with the EmoPain in actions representative of everyday functioning.

Further, it has been widely used (over 2,000 citations as of early 2023) in action recognition studies, e.g. [18], [23].

The dataset was captured from healthy participants while they performed 60 actions, e.g. drinking water, putting on a pair of glasses, in lab settings. For our experiments, we removed action classes 50 to 60 which include actions that involve a second person (such as 'shaking hands') and action classes 41 to 49 that include acted medical events (e.g. coughing). We focused on the first 27 action classes which further excludes those that are challenging to represent with the 6 joints captured in the primary target dataset alone (e.g. rubbing two hands, shaking the head).

The dataset includes full-body joint positions (25 joints). As with the EmoPain dataset, we only used the same 6 joints available in the EmoPain@Home dataset that is the primary target. The data was captured using a markerless system based on RGB and depth cameras. Each data instance in this dataset is a single action instance. We trimmed each instance to 240 frames (the segmentation window used for the EmoPain dataset) and padded shorter instances by looping the sequence. We obtained $N = 26,398$ instances.

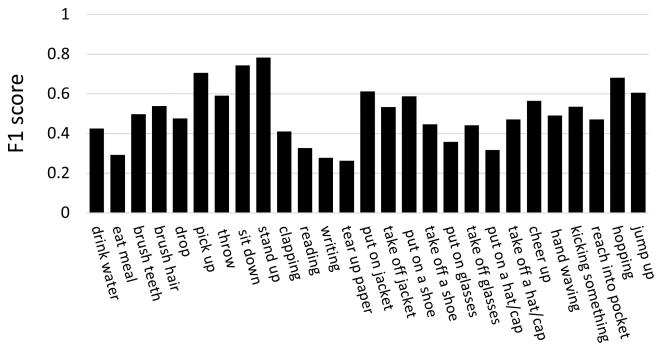

Fig. 3. F1 scores for end-to-end action classification in the NTU RGB+D dataset (based on data from 6 joints only).

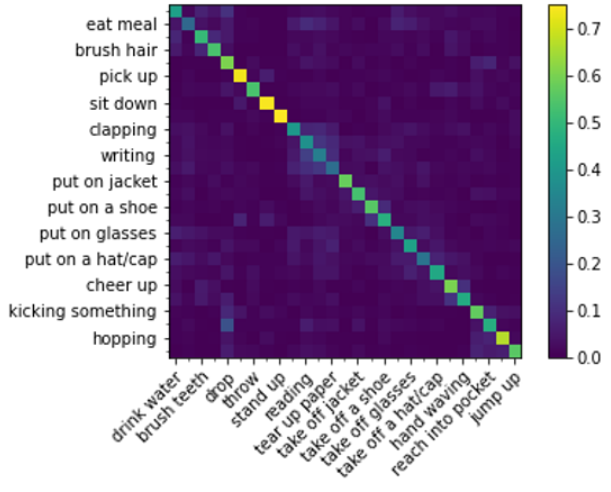

Fig. 4. Confusion matrix for end-to-end action classification in the NTU RGB+D dataset (based on data from 6 joints only). Due to space constraints, alternate action labels are shown on the x and y axes, i.e. action class 1 = 'drink water', action class 2 = 'eat meal', etc.

5 Results

We present here results for 4 sets of experiments. The first set evaluates the performance of the P-STEMR model for the transfer source task (action recognition). The second set further analyzes the model to understand the value of our parallel spatio-temporal encoding approach as well as the effect of alternative input and output formats. The third set of experiments explores the efficacy of the P-STEMR model for representation learning, with both transfer across population groups and tasks investigated. The fourth set of experiments qualitatively evaluates transfer between the primary source and target datasets.

5.1 End-to-End Action Classification with the P-STEMR

This evaluation was done using both the EmoPain and NTU RGB+D datasets. Although the input and output of the model were proportion of each action type, we used majority voting to obtain a single action class per instance for the performance metrics.

5.1.1 *Action Classification for the NTU RGB+D Dataset*

The results reported in this section are based on hold-out validation due to the large size of the NTU RGB+D dataset.

The results are shown in Figure 3. Performance is well above chance level (F1 score = 0.04) with average F1 score of 0.50 across 27 classes. It should be noted that this performance is despite the limited input information, i.e. data from 6 anatomical joints from one side of the body alone. *Pickup*, *Sit down*, and *Stand up* are the best recognized with F1 score above 0.70. This finding is not surprising given that these movement types primarily involve trunk movement which is well captured in the input data. The movement types with the worst performance, *Eat meal*, *Reading*, *Writing*, *Tear up paper*, *Put on glasses*, and *Put on a hat/cap*, either involve minimal motion or mainly require the use of an arm. In the input data, only one arm is captured, and there are occasions where the actor may have used their other arm instead of the one tracked. The confusion matrix (Figure 4) shows that *Reading*, *Writing*, and *Tear up paper* are often confused with one another. Still, performance in recognizing these activities is much higher than chance level.

5.1.2 *Action Classification for the EmoPain Dataset*

The results reported here are based on leave one-subject-out cross-validation to test generalization to unseen subjects.

We obtained average F1 score of 0.91 over the 6 classes. As can be seen in the confusion matrix (Figure 5), the worst performance is for *Sit-to-stand* (F1 score = 0.83) which is sometimes confused with *Stand-to-sit*, *Walking*, and *Standing*. The confusion with these three is not unexpected since the four movement types have standing in common between them. Also unsurprising is the absence of confusion between *Standing* and *Sitting* which both involve little motion and have very distinct hip and knee angle signatures.

We additionally explored the approach used in the MiMT architecture of Olugbade et al. [30] where two different timescales of a label are learnt in parallel using a multitask approach. We modelled both the proportion of the action classes (segment timescale) as well as the frame level action label (frame timescale). We used a MSE loss

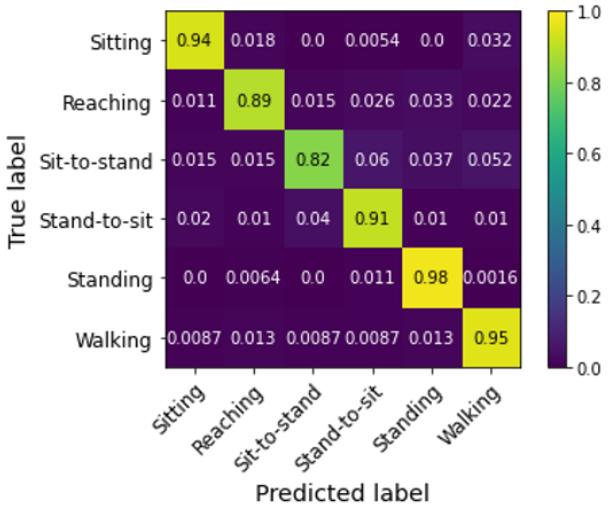

Fig. 5. Confusion matrix (showing proportions across respective rows) for end-to-end action recognition on the EmoPain dataset (with data from 6 joints only) based on the proposed P-STEMR model.

for the frame timescale. We found only marginal increase in (segment action classification) performance, average F1 score of 0.92, although the increase seen supports potential in incorporating multi-timescale learning in the P-STEMR. Higher gain could perhaps be found by also addressing the imbalance in action classes at the frame timescale.

5.2 Ablation Study on the P-STEMR

5.2.1 The Value of Parallel Spatio-Temporal Encoding in the P-STEMR

Serial order spatio-temporal encoding, i.e. encoding structures where spatial encoding is done in succession to temporal encoding or vice-versa, is the traditional approach. Thus, it was important to understand if our approach of instead parallelizing these two structures contributed to the high action recognition performance obtained in Section 5.1 and whether this method outperforms the state of the art. We compared our proposed P-STEMR model with a variant with serial order spatio-temporal encoding (STEMR model). Our evaluation was done on the EmoPain dataset only as this is our primary source dataset for the representation learning experiments (Section 5.3).

The results are shown in Table 1. As can be seen in the table, parallel encoding of the spatial and temporal dimensions has higher F1 score for all 6 action types in the EmoPain dataset, considerably higher for *Walking* especially (0.12 increase in F1 score).

5.2.2 Effect of the Label Form for the Action Recognition Network of the P-STEMR

We aimed to understand if learning a simpler action output format would improve action recognition performance. So, we further compared the STEMR model with a variant (STEMR-vote) where the model was trained on action classes based on the majority action type in a given data instance rather than the proportion of action types used in our P-STEMR architecture. For this STEMR-vote model, we used a categorical cross-entropy loss weighted to address imbalance across the action classes.

TABLE 1
F1 scores for end-to-end action classification on the EmoPain dataset (with data from 6 joints only) comparing serial and parallel spatial-temporal encoding, i.e. the STEMR and P-STEMR models

Encoding	F1 score					
	Sitting	Reaching	Sit-to-stand	Stand-to-sit	Standing	Walking
Serial (STEMR)	0.96	0.84	0.78	0.83	0.90	0.77
Parallel (P-STEMR)	0.97	0.89	0.83	0.88	0.97	0.89

TABLE 2
F1 scores for end-to-end action classification on the EmoPain dataset (with data from 6 joints only), comparing action label forms, i.e. the STEMR-vote and STEMR models

Label form	F1 score					
	Sitting	Reaching	Sit-to-stand	Stand-to-sit	Standing	Walking
Majority action (STEMR-vote)	0.93	0.72	0.74	0.77	0.83	0.52
Proportion of actions (STEMR)	0.96	0.84	0.78	0.83	0.90	0.77

TABLE 3
F1 scores for end-to-end action classification on the EmoPain dataset (with data from 6 joints only), comparing input types based on the STEMR-vote model

Input	F1 score					
	Sitting	Reaching	Sit-to-stand	Stand-to-sit	Standing	Walking
Positions	0.91	0.64	0.71	0.77	0.81	0.64
Angles	0.93	0.72	0.74	0.77	0.83	0.52

The results are shown in Table 2. Use of the proportion labels showed better recognition for all action types with the highest gain occurring for *Walking* with 0.25 increase (16%) in F1 score and also a gain of 0.12 F1 score for *Reaching*.

5.2.3 Effect of the Input Type of the P-STEMR

Unlike previous related studies in [23], [27], [33], we used joint angles input instead joint positions for our GCN encoder. While we had previously explored this GCN input type in [7], we had not evaluated if use of joint angles is superior for action recognition. So, we here compared *joint positions* and *joint angles* using the STEMR-vote model.

The results are shown in Table 3. As can be seen, use of joint angles improves recognition of 4 of the 6 action types (*Sitting*, *Standing*, *Sit-to-stand*, and *Reaching*). However, it considerably lowers recognition of *Walking*. This effect for *Walking* is not surprising given that it is the action that most involves translation: position information is more revealing for translatory movement types. The confusion matrices (not shown here due to space constraints) show that it is confused with *Standing* and this gets worse with angles data.

5.3 Representation Learning with the P-STEMR

To evaluate the performance of the P-STEMR for movement representation learning, it was trained on the EmoPain dataset (source dataset), and the 'Embedding output layer' branch of the trained model was then used as a feature extraction function (referred to as the *P-STEMR latent feature extraction* model for the sake of convenience) for target tasks. We first explored the *P-STEMR latent feature extraction* model in action classification in the NTU RGB+D dataset as target. We then evaluated the same model in pain level classification in the EmoPain@Home dataset, which are our primary target task and dataset respectively.

5.3.1 Representation Learning for Action Classification in the NTU RGB+D Dataset

We used the *P-STEMR latent feature extraction* model to extract embeddings from NTU RGB+D joint angles (computed from the joint positions data available in the dataset). A 3-layer MLP (with 20 units in each hidden layer) was then trained for action classification based on these embeddings as input. We used hold-out validation to evaluate this action classification model similar to Section 5.1.1.

The results are shown in Figure 6. As can be seen in the figure, although recognition is still better than chance level performance, it is much lower than based on end-to-end classification using the P-STEMR, with average F1 score of 0.10. The 5 classes for which performance was below (or about or only marginally better than) chance level classification are: *Reading*, *Writing*, *Tearing up paper*, *Taking off glasses*, and *Waving*. The poor performance for these classes is not unexpected. The first 3 (*Reading*, *Writing*, *Tearing up paper*) were similarly among the classes with the worst performance in the results for end-to-end action classification in Section 5.1.1. *Tearing up paper*, *Taking off glasses*, and *Waving* further commonly involve the arms primarily and they involve arm movements that are not at all represented in the EmoPain dataset classes which only capture flexion/extension movements of the arm.

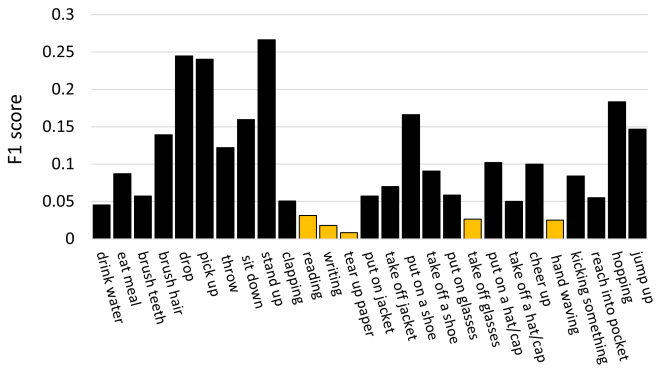

Fig. 6. F1 scores for action classification on NTU RGB+D dataset (with data from 6 joints only) based on the P-STEMR latent feature extraction model trained using the EmoPain dataset. In yellow are performances below (or only marginally better than) chance level recognition.

TABLE 4
F1 scores for pain level classification in the EmoPain@Home dataset comparing hand-crafted features with features based on representations learnt using the P-STEMR model and EmoPain dataset

Features	Pain classifier	F1 score	
		Low level pain	High level pain
Hand-crafted	Bagging	0.63	0.61
Hand-crafted	MLP	0.68	0.74
P-STEMR	MLP	**0.83**	**0.85**
P-STEMR & Hand-crafted	MLP	0.81	0.83
P-STEMR (action output) & Hand-crafted	MLP	0.61	0.69

5.3.2 Representation Learning for Pain Level Classification in the EmoPain@Home Dataset

We used the same *P-STEMR latent feature extraction* model (trained on the EmoPain dataset) to extract embeddings for the EmoPain@Home data. As the input size of the P-STEMR trained on the EmoPain dataset was 240 frames while each EmoPain@Home data instance in our experiments had 2,400 frames, we extracted a sequence of 10 embedding vectors per EmoPain@Home data instance. We concatenated these vectors and used the resulting feature vector as input to a 3-layer MLP (with 10 and 5 units in each hidden layer respectively) for pain level classification. Similar to our work in [7], [11] on the EmoPain@Home dataset, we used leave-one-activity-instance-out cross-validation to evaluate this pain classification model. Results are shown in Table 4.

As can be seen in the table, the use of the *P-STEMR latent feature extraction* (third row of results in the table) resulted in much better performance than the use of hand-crafted features (second row of results), with average F1 score of 0.84 (increase by 18%). A Wilcoxon Signed-Rank Test computed based on accuracy per cross-validation fold showed statistically significant difference, $Z = 2.24, p = 0.025$. We used the same hand-crafted features used in [7] based on joint angles, i.e. speed, jerk, energy, amount of movement, and range of motion over segment and activity timescales. The hand-crafted features with the MLP classifier is a fairer comparison with the *P-STEMR latent feature extraction* & MLP pain level classification. However, for the sake of completeness, we also include results of the hand-crafted features with the Bagging algorithm used in [7] (first row).

Combining both the P-STEMR based features and the hand-crafted features (fourth row of result in the table) did not improve performance beyond this suggesting that the P-STEMR based features cover the important movement characteristics captured by the crafted features. We also experimented with combination of the hand-crafted features with the action output of the trained P-STEMR model (instead of its latent output) as input to the MLP to evaluate the difference in richness of movement information between the action and latent outputs of the P-STEMR model. As the fifth row in the table shows, performance in this case is worse than using the hand-crafted features alone.

Lastly, we evaluated a separate P-STEMR model trained on the NTU RGB+D dataset for feature extraction, to understand the influence of similarities/differences in population

group between target and source datasets on transfer efficacy. The result with the latent output of the NTU RGB+D trained P-STEMR was average F1 score of 0.57, which is only marginally better than chance-level performance and worse than use of the hand-crafted features.

5.4 Qualitative Analysis of Action Labelling of the EmoPain@Home Dataset based on the P-STEMR

We further explored the performance of the P-STEMR model trained on the EmoPain dataset for action labelling of the EmoPain@Home dataset to gain more insight into its efficacy. As the EmoPain@Home dataset does not have ground truth action annotations (it instead has activity labels, e.g. vacuuming, which are of a higher level of abstraction than the action labels in the EmoPain dataset), our evaluation here was purely qualitative. In this analysis, we visually explored a video of plot (see supplementary material) of the half-body joint positions in the EmoPain@Home dataset and the predicted action labels. Although the EmoPain@Home dataset exists as 3D joint positions, the plots were in 2D as this was found to be better for visualization. We also normalized the joint positions to [-1, 1] to aid exploration.

While the movement data and automatic annotations are best explored in video, we include below still images as examples of accurate and inaccurate predictions. Each image shows 6 equally sampled frames from the given data instance together with the predicted action class distribution as well as the larger activity (ground truth) that the instance is a part of.

5.4.1 True Reaching and Sitting Positives with Confusions Due to Missing Context

Reaching movements are valuable to recognize for technology that aims to provide support for people with chronic pain particularly low back pain that is the most prevalent. This is because people with this condition typically find reaching challenging. Our analysis shows that the EmoPain-trained P-STEMR model is able to recognize reach movements in the EmoPain@Home dataset. This is despite differences in reach types between the two datasets. For instance, the model is able recognize EmoPain@Home reach movements in poses not present in the EmoPain dataset, e.g. where the subject is both reaching downward and kneeling in a yoga pose (Figure 7-left). Another example is the upward reaching in Figure 7-right (the EmoPain dataset only has instances of forward and downward reaching).

Sitting postures/actions are also valuable to detect. First, sitting may represent a strategy for coping with challenge in completing an activity. For example, during the capture of the EmoPain@Home dataset, there was a participant who started off washing up dishes at the sink in standing posture but interrupted the activity to sit on a stool to complete washing up. Second, sitting may highlight periods of rest within an activity or between activities. Sit actions were well-detected in the EmoPain@Home dataset and the model was further able to detect transitions between standing and sitting. Figure 8 includes a few examples. In Figure 8-left, the subject is seated on a high stool (or perched at the edge of a seat); in Figure 8-right, the subject is fully seated.

However, the model sometimes confuses reaching with sitting. One example is in Figure 9-left where the subject is squatting. This is an understandable confusion that even a human could make when limited to joint positions information, i.e. without visual background that shows the absence of a seat. Further, such squat actions are limited in the EmoPain dataset used to train the model. A second example (Figure 9-right) is when the trunk is fully flexed such that the hip angle is around right angle. As the action recognition model is based on angles data, it has little way of realizing that the trunk, rather than the upper leg, is at a plane parallel to the floor.

5.4.2 The Problem in Identifying Standing and Walking in Real-Life Physical Activity

We found that the model was able to identify standing and walking actions in the EmoPain@Home dataset, both when they were obvious (such as walking actions during outdoor walking in exercise) and also a number of times when they occurred as brief transitions between two actions or overlapping with another action.

Instances of the latter case led to us to reflect on the consequence of standing and walking not only being standalone actions in themselves but also commonly being part of other actions. Most activities not done seated or lying down are done standing, i.e. upright. Many activities also involve locomotion, e.g. the body moving from one point in a room to another in vacuuming, movement from the sink to the dish drainer that is slightly too awkward to reach without moving the feet during washing up. These make standing and walking especially challenging to label in everyday activities. For example, when should a very minimal forward reaching (such as in washing up) become classified as 'Standing' rather than 'Reaching'? Or when is a point with the feet on the ground in the stance phase of a walk 'Standing' rather than 'Walking'? Does one separate the walk out from the reaching forward done at the same time during vacuuming? These questions are relevant not just at the point of automatic labelling but would also be critical during human annotation, moreso if the latter will be used as ground truth to train an automatic system.

6 Discussion

With the aim of gaining insight into the possibility of movement representation learning that enables improvement in pain level classification, we explored a novel machine learning architecture (P-STEMR) in four sets of experiments. In this section, we bring together the findings from these experiments to highlight the main implications.

6.1 Possibility of Learning Movement Representations that Capture Abstraction Beyond Actions/Activities

The finding of average F1 score of 0.84 for pain level classification compared with 0.71 using hand-crafted features suggests that movement representation learning can indeed improve performance for tasks beyond HAR. This outcome is significant for the field of affective computing where capture of large sets of annotated training data in the wild is not trivial. Our findings are based on transfer of representations learnt from data captured in lab settings to data recorded in real-world activities in homes highlighting

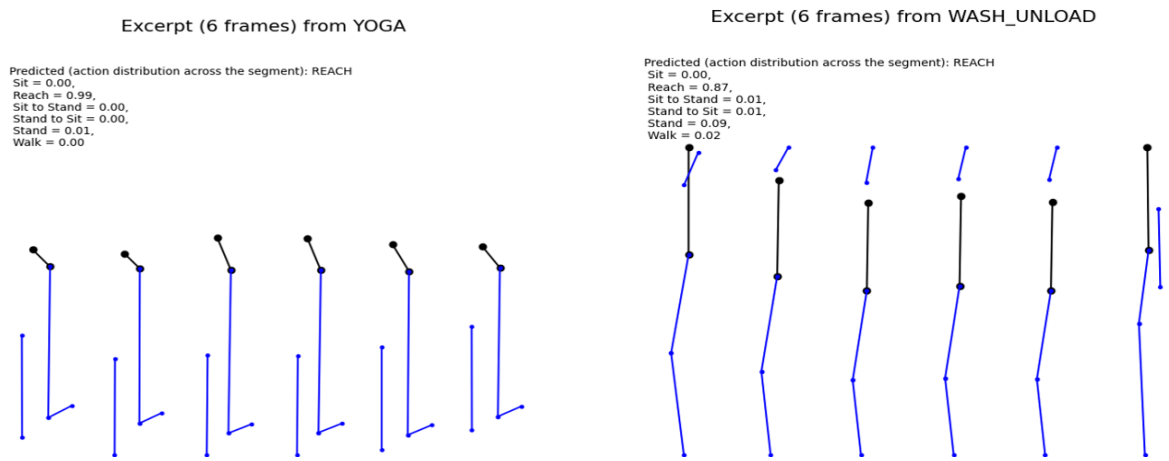

Fig. 7. Example recognition of Reaching actions in the EmoPain@Home dataset. Each still image shows 6 frames that represent the start of each successive second of a 6-second window of movement. As shown in Figure 2-Left, there are two connected groups of joints in the skeletons: the wrist is connected to the elbow; and the mid-spine is connected to the hip (this connection is shown with a black line), which is connected to the knee that is in turn connected to the ankle.

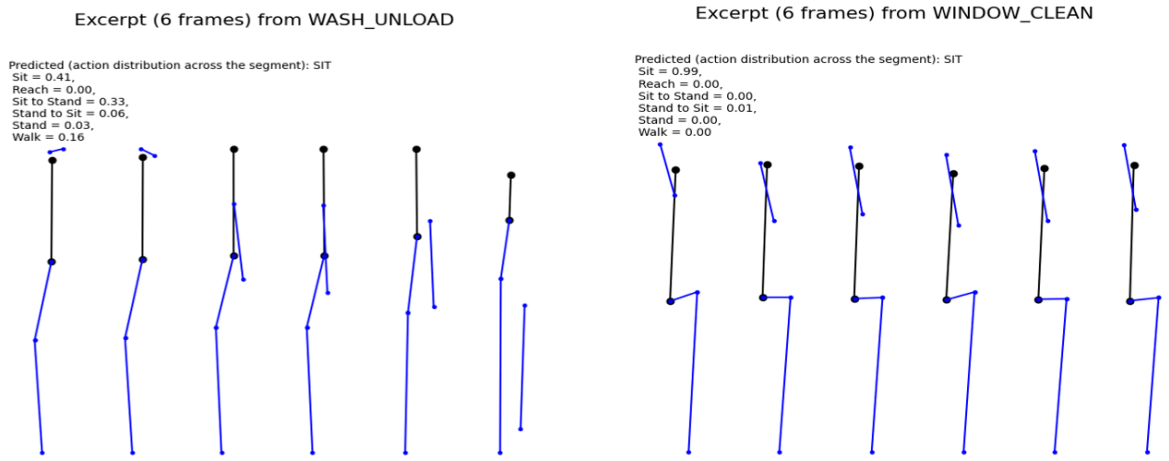

Fig. 8. Example recognition of seated actions in the EmoPain@Home dataset. Each still image shows 6 frames that represent the start of each successive second of a 6-second window of movement. As shown in Figure 2-Left, there are two connected groups of joints in the skeletons: the wrist is connected to the elbow; and the mid-spine is connected to the hip (this connection is shown with a black line), which is connected to the knee that is in turn connected to the ankle.

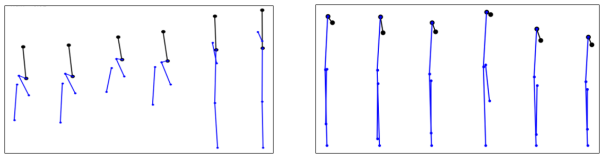

Fig. 9. Example cases of confusion between sitting and reaching actions in the EmoPain@Home dataset. Each still image shows 6 frames that represent the start of each successive second of a 6-second window of movement. As shown in Figure 2-Left, there are two connected groups of joints in the skeletons: the wrist is connected to the elbow; and the mid-spine is connected to the hip (connection shown in black), which is connected to the knee that is in turn connected to the ankle.

possibility of leveraging data captured in more controlled environment for modelling messier real world movement. The considerable improvement in performance using data of relatively small size and limited number of action types further underlines value in employing representation learning for affective movement recognition.

Unlike the state of the art solely based on self-supervised learning [18], [23], our approach combines supervised learning (where learning is supervised for a task other than the target) with self-supervised regularization for the latent representations. The rationale for the use of supervised learning was to explore the possibility of taking advantage of the large number of HAR datasets available for reuse in the research community [1]. The close link between affect assessment and action categorization [31], [32] in movement perception and understanding provides support for this approach. The regularization included in the model was aimed at forcing the model to learn generalized movement representations at the lower layers. Indeed, findings from comparing the latent and action outputs of the P-STEMR for pain level classification (together with hand-crafted features), average F1 scores of 0.82 and 0.65 respectively, show

that the latent outputs contain information beyond what is specific to action recognition. Qualitative findings, in the action labelling of the EmoPain@Home dataset, that the P-STEMR model is able to recognize actions in unseen movement configurations further highlight the generalization power at the level of the action recognition specific output itself. End-to-end action classification performance with the NTURGB+D dataset showing ability to recognize subtle differences between actions (e.g. putting on and taking off shoes, putting on a hat/cap and putting on glasses, etc) with limited input information additionally showcases the level of detail encapsulated in the representations.

Results of our ablation studies highlight value in parallelizing spatial and temporal encoding to capture such rich movement representation. We do not fully understand the reason for the superiority of the parallel encoding but it perhaps characterizes the widely evidenced (e.g. in [33]]) higher status of model-level fusion of multimodal information compared to feature level fusion. In this fusion analogy, temporal and spatial information can be considered separate but interconnected modalities. Comparison of the performance of our P-STEMR in end-to-end action classification in the EmoPain dataset with previous work [27] on the same dataset (although based on a different set of action classes) further demonstrates the merit of the 'model-level fusion' approach to spatial and temporal encoding. In the earlier study [27], the authors obtained average F1 score = 0.76 for 6 action classes (compared to 0.91 for 6 classes in our work) based on a model where a GCN for spatial encoding and an LSTMNN for temporal encoding are connected in serial order. Their model used position data from 7 joints as input.

6.2 Factors to Consider for Preparing Source Datasets

Additional findings in our work highlight factors that may need to be considered to maximize learning of representations that transfer to a given target dataset. Our results particularly highlight differences in population groups which have implication for movement behaviour as a significant factor. Despite the relatively larger size and higher variety in action types of the NTU RGB+D dataset, movement representations learnt based on its data generalized poorly to the EmoPain@Home dataset. This is in stark contrast to the performance obtained with the EmoPain as the source dataset. Similarly, representation learning did not transfer well to the NTU RGB+D from the EmoPain. Evidence points to differences between movements of people with chronic pain and those without the condition, including in range of trunk movement [4], coordination between body parts [34], and gait qualities [35]. These findings suggest that, for maximal performance, the subjects from the source and target datasets need to come from similar population groups especially with respect to characteristics that can have marked effect on movement execution, e.g. a movement disorder.

Major differences in movement types between the NTU RGB+D and EmoPain datasets, particularly in arm motions, may have further undermined transfer of representations between the two datasets. The arm movements captured in the EmoPain (source dataset) are mainly elbow flexion/extension movements unlike the NTU RGB+D actions that include abduction/adduction and rotation. The significance of this difference is evident in the not-better-than-chance recognition, in the target dataset (NTU RGB+D), of all action types where the arm is the primary actor (except for 'brushing of the hair' and 'throwing'). Conversely, the model is able to better recognize other action types which share movement attributes with actions in the source dataset including 'hopping', 'jumping', 'putting on a shoe' which are not action instances themselves present in the source dataset. These findings suggest that it will be of benefit to maximize the types of low-level motion primitives, e.g. flexion/extension, rotation, and lateral movement of given joints, shared between source and target dataset. In so doing, differences in higher-level descriptions such as the types of actions will have limited effect on transfer performance.

Findings of notably good performance in transfer between EmoPain and EmoPain@Home datasets suggest that other differences such as motion capture sensor type or capture settings have minimal effect on transfer efficacy. We used angles data as input to address any effect of sensor type difference. However, comparison of this input type with joint positions input in our experiments showed disadvantages for recognition of walking. Transfer between settings that our findings show to be effective is a critical capability since it enables data captured in constrained and artificial settings to be useful for the real world where movements are messier and emotional expressions are associated with more significant implications for the subjects.

7 Conclusion

We have shown the possibility of leveraging movement representation learning from other tasks for affect recognition. As far as we know, this is the first study to investigate this for affective movement recognition. We propose the use of a novel P-STEMR model that takes advantage of availability of HAR datasets and leads to increase in pain level classification (from average F1 score of 0.71 to 0.84) based on representation learning. The model also shows very good performance in its source task, action recognition: average F1 scores of 0.91 for 6 classes and 0.50 for 27 classes in the EmoPain and NTU RGB+D datasets respectively. Our experiments and analyses highlight differences in population group and motion primitives between source and target datasets as factors that may affect transfer performance. Although several questions still remain, our work is a valuable first step in this area. In future work, we will explore ways of minimizing the effects of differences between source and target datasets on transfer performance. This is an active area of research for machine learning in general and could facilitate a more fine-grained classification of pain level, or even regression on the 0-to-10 pain scale.

Acknowledgments

This work was supported by the EU Future and Emerging Technologies Proactive Programme H2020 (Grant No. 824160: EnTimeMent - https://entimement.dibris.unige.it).

References

[1] T. Olugbade, M. Bieńkiewicz, V. D'Amato et al. 2022. Human Movement Datasets: An Interdisciplinary Scoping Review. ACM Computing Surveys.

[2] B. De Gelder. 2009. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. Philosophical Transactions of the Royal Society B: Biological Sciences 364(1535), pp.3475-3484.
[3] F. Noroozi, C. Adrian Corneanu, D. Kamińska, T. Sapiński, S. Escalera, and G. Anbarjafari. 2018. Survey on emotional body gesture recognition. IEEE Trans. Affect. Comput. 12(2), pp. 505-523.
[4] T. Olugbade, A. Singh, N. Bianchi-Berthouze, N. Marquardt, H. Aung, and A. de C. Williams. 2019. How can affect be detected and represented in technological support for physical rehabilitation?. ACM Trans. Comput.-Hum. Interact. 26(1), pp. 1-29.
[5] A. Singh, N. Bianchi-Berthouze, and A. C. de C. Williams. 2017. Supporting everyday function in chronic pain using wearable technology. Proc CHI Conf. Hum. Factor. Comput. Syst., pp. 3903-3915.
[6] A. Singh, A. Klapper, J. Jia et al. 2014. Motivating people with chronic pain to do physical activity: opportunities for technology design. Proc CHI Conf. Hum. Factor. Comput. Syst., pp. 2803-2812.
[7] T. Olugbade, R. Buono, K. Potapov et al. *2023*. The EmoPain@Home Dataset: Capturing Pain Level and Activity Recognition for People with Chronic Pain in Their Homes. TechRxiv preprint.
[8] M. Albert, S. Toledo, M. Shapiro, and K. Kording. 2012. Using mobile phones for activity recognition in Parkinson's patients. Front. Neurol. 3, p.158.
[9] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. Picard. 2019. Automatic recognition methods supporting pain assessment: A survey. IEEE Trans. Affect. Comput. 13(1), pp. 530-552.
[10] M. Sullivan, P. Thibault, A. Savard, R. Catchlove, J. Kozey, and W. Stanish. 2006. The influence of communication goals and physical demands on different dimensions of pain behavior. Pain 125(3), pp. 270-277.
[11] T. Olugbade, R. Buono, A. de C. Williams et al. 2022. EmoPain (at) Home: Dataset and Automatic Assessment within Functional Activity for Chronic Pain Rehabilitation. Proc. Int. Conf. Affect. Comput. Intell. Interact., pp. 1-8.
[12] M. Uddin and S. Canavan. 2020. Multimodal multilevel fusion for sequential protective behavior detection and pain estimation. Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, pp. 844-848.
[13] M. Aung, S. Kaltwang, B. Romera-Paredes et al. 2016. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal EmoPain dataset. IEEE Trans. Affect. Comput. 7(4), pp. 435-451.
[14] A. Saeed, T. Ozcelebi, J. and Lukkien. 2019. Multi-task self-supervised learning for human activity detection. Proc. ACM Conf. Interact Mob. Wearable Ubiquitous Technol. 3(2), pp.1-30.
[15] I. Tang, I. Perez-Pozuelo, D. Spathis, S. Brage, N. Wareham, and C. Mascolo. 2021. Selfhar: Improving human activity recognition through self-training with unlabeled data. Proc. ACM Conf. Interact Mob. Wearable Ubiquitous Technol. 5(1), pp. 1-30.
[16] B. Khaertdinov, E. Ghaleb, and S. Asteriadis. 2021. Contrastive self-supervised learning for sensor-based human activity recognition. Proc. IEEE Int. Joint Conf. Biometrics, pp. 1-8.
[17] D. Liu, T. Wang, S. Liu, R. Wang, S. Yao, and T. Abdelzaher. 2021. Contrastive self-supervised representation learning for sensing signals from the time-frequency perspective. Proc. Int. Conf. Comp. Commun. Netw., pp. 1-10.
[18] X. Gao, Y. Yang, and S. Du. 2021. Contrastive self-supervised learning for skeleton action recognition. Proceedings of the NeurIPS 2020 Workshop on Pre-registration in Machine Learning, pp. 51-61.
[19] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.
[20] G. Evangelidis, G. Singh, and R. Horaud. 2014. Skeletal quads: Human action recognition using joint quadruples. Proceedings of the International Conference on Pattern Recognition, pp. 4513–4518.
[21] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. 2014. HOPC: Histogram of oriented principal components of 3d pointclouds for action recognition. Proceedings of the European Conference on Computer Vision, pp. 742–757.
[22] A. Shahroudy, J. Liu, T. Ng, and G. Wang. 2016. NTU RGB+D: A large scale dataset for 3d human activity analysis. Proc. IEEE conf. Comput. Vis. Pattern Recognit., pp. 1010–1019.
[23] H. Zhang, Y. Hou, W. Zhang, and W. Li. 2022. Contrastive Positive Mining for Unsupervised 3D Action Representation Learning. Proc. European Conf. Comput. Vis., pp. 36-51.
[24] Y. Xi, P. Li, Y. Song, Y. Jiang, and L. Dai. 2019. Speaker to emotion: Domain adaptation for speech emotion recognition with residual adapters. Proc. Asia-Pacific Signal Information Processing Association Annual Summit Conf., pp. 513-518.
[25] M. Akhand, S. Roy, N. Siddique, M. Kamal, and T. Shimamura. 2021. Facial emotion recognition using transfer learning in the deep CNN. Electronics 10(9).
[26] T. Kipf, and M. Welling. 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint:1609.02907.
[27] C. Wang, Y. Gao, A. Mathur, A. de C. Williams, N. D. Lane, and N. Bianchi-Berthouze. 2021. Leveraging activity recognition to enable protective behavior detection in continuous data. Proc. ACM Conf. Interact Mob. Wearable Ubiquitous Technol. 5(2), pp. 1-27.
[28] S. Hochreiter, J. Schmidhuber. 1997. Long short-term memory. Neural computation 9(8), pp. 1735-1780.
[29] F. Gers, J. Schmidhuber, and F. Cummins. 2000. Learning to forget: Continual prediction with LSTM. Neural Comput. 12(10), pp. 2451–71.
[30] T. Olugbade, N. Gold, A. de C. Williams, and N. Bianchi-Berthouze. 2020. A Movement in Multiple Time Neural Network for Automatic Detection of Pain Behaviour. Companion Publication Int. Conf. Multimodal Interact., pp. 442-445. 2020.
[31] M. Iacoboni, I. Molnar-Szakacs, V. Gallese, G. Buccino, J. Mazziotta, and G. Rizzolatti. 2005. Grasping the intentions of others with one's own mirror neuron system. PLoS biology 3(3): e79.
[32] V. Gallese. 2007. Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. Philos. Trans. R. Soc. B: Biol. Sci. 362(1480), 659-669.
[33] G. Cen, C. Wang, T. Olugbade, A. de C. Williams, and N. Bianchi-Berthouze. 2022. Exploring Multimodal Fusion for Continuous Protective Behavior Detection. Proc. Int. Conf. Affect. Comput. Intell. Interact., pp. 1-8.
[34] M. Shafizadeh. 2016. Movement coordination during sit-to-stand in low back pain people. Human Movement 17(2), pp.107-111.
[35] P. Terrier, J. Le Carré, M. Connaissa, B. Léger, and Fra. Luthi. 2017. Monitoring of gait quality in patients with chronic pain of lower limbs. IEEE Transactions on Neural Systems and Rehabilitation Engineering 25(10): pp. 1843-1852.

Temitayo Olugbade is a Lecturer in Computer Science and AI at University of Sussex and Honorary Research Fellow at University College London (UCL). Her research pursues development and application of machine learning methods to new and challenging affective computing contexts.

Amanda C de C Williams is a professor of clinical health psychology at UCL, consultant clinical psychologist in pain management at UCL Hospital, and Section Editor for Psychology on the journal PAIN. Her research interests include evidence-based medicine applied to psychologically-informed interventions for pain, including systematic review and meta-analysis; behavioural expression of pain and its interpretation; and responsive wearable technology to extend healthcare into patients' own environments.

Nicolas Gold is an Associate Professor at UCL Computer Science. His research interests are in the comprehension and comprehensibility of design representations in a range of disciplines including source code, educational resource design, and research ethics design. He has also published research in computational musicology, creative music systems, research ethics, and music computing applications in healthcare.

Nadia Bianchi-Berthouze is Full Professor in Affective Computing & Interaction at the University College London Interaction Centre. Her research focuses on designing technology that can sense the affective state of its users and use that information to tailor interaction. She has pioneered the field of Affective Computing in investigating how body movement and touch behaviour can be used as means to measure quality of user experience.