2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)

Exploring Multimodal Fusion for Continuous Protective Behavior Detection

Guanting Cen[†]*, Chongyang Wang*, Temitayo A. Olugbade, Amanda C. De C. Williams, Nadia Bianchi-Berthouze[‡]
University College London, UK [†]gtcenn@outlook.com, [‡]n.berthouze@ucl.ac.uk

***Abstract*—Chronic pain is a prevalent condition that affects everyday life of people around the world. Protective behaviors (strategies that are naturally but unhelpfully adopted by people with chronic pain to cope with fear of pain in executing harmless everyday movements) can lead to further disability over time if not recognized and addressed appropriately. In this paper, we build on previous work on unimodal, activity-independent, time-continuous protective behavior detection (PBD) by focusing on the fusion of muscle activity and body movement modalities for characterizing both protective behavior and its physical activity context. We explore different fusion strategies based on consideration of the manner in which protective behavior influences muscle activity and overt body movement, as well as the relationship between the two modalities. We evaluate the various strategies on the multimodal EmoPain dataset containing data from people with and without chronic pain engaged in physical activities that reflect everyday challenges for those with chronic pain. Our results show that a central (model-level) fusion approach leads to better PBD performances than input- and decision-level fusions, or unimodal approaches. We also show that the use of an attention mechanism, typifying shifts in attention characteristic of protective behavior, further improves the sensitivity of the model, i.e. detection of the positive class (which is the minority class). We analyze these results and suggest that fusion in modelling a motor condition should consider how the emotional responses (fear of movement and pain in this case) triggered by a condition affects each of the given modalities and hence their contributions to the modelling task.**

***Index Terms*—chronic pain, deep learning, multimodal fusion**

I. Introduction

Chronic pain (CP), which is defined as pain that persists after the healing period of an injury or appears in the absence of an injury [1], is a common global health problem among adults [2]. Caused by changes in the nervous system, negative emotions and impaired movements due to persistent interference of pain signals are associated with chronic pain. People with musculoskeletal chronic pain tend to exhibit *protective behaviors* (e.g., guarding and hesitation during movement), which are unhelpful strategies used to cope with challenging but harmless everyday physical activities [1]. Unfortunately, such behavior can lead to increased pain, increased difficulty in performing functional activity, negative emotions, and withdrawal from valued physical activities [3]. Physiotherapists watch for protective behaviors in their observation of movement to tailor their feedback to the individual patient and to personalize management strategies that aim to help the patient reduce the use of these behaviors during everyday functioning [4]. However, such support in clinical settings does not easily translate to daily life [5] and is particularly unavailable in situ, i.e. at the moment when the patient encounters the challenge (e.g. while bending down to load the washing machine). Automated and continuous detection of protective behavior (hereafter referred to as PBD for 'protective behavior detection') offers an opportunity to deliver real-time personalized support.

While the number of studies on PBD has increased in recent years, fusion (of the two main modalities that characterize protective behavior, i.e. muscle activity and overt body movement) has not been well investigated. Work has focused on unimodal, simple input-level fusion (i.e. concatenating input data) or late fusion techniques [6] without considering how these two modalities are interactively and separately affected by underlying pain-related affect, especially anxiety about pain. There are overt protective behaviors, such as minimal trunk flexion or the use of support in trunk lifting/lowering [4], that can be captured using sensors that track movement kinematics (e.g. motion capture). The overt behaviors can further be evident in muscle activity patterns, e.g., low activation of lower back muscles in minimal trunk flexion, activation of upper back muscles in the use of the arms as a support. However, there are also muscle activation behaviors that are not well reflected in overt body movement [7] [8]. Various studies [9]–[11] for example show high activation of muscles not (or no longer) involved in the movement being performed, e.g., lower back muscles during walking or at the end of trunk flexion/extension. Hence, although intuitive in movements of healthy people, the relationship between movement kinematics captured in motion capture (MoCap) data and muscle activity captured by electromyography (EMG) is often complex for the movements of people with CP and strongly related to the emotional response. First, people with CP adopt a variety of protective behavior strategies that are dictated by their (unhelpful) perception of danger in movement. In addition, protective behavior embodies a continuous shift in attention between different body parts across specific phases of an activity, with the tendency to avoid the use of those perceived as at risk and instead recruiting others perceived as safe for completing the movement [12], [13]. Such strategies actually make the movement awkward and harder to execute. To add to this complexity, what part of the body people may perceive

This work was supported by the EU Future and Emerging Technologies Proactive Programme H2020 (Grant No. 824160: EnTimeMent). Wang was additionally supported by the UCL Overseas Research Scholarship (ORS) and Graduate Research Scholarship (GRS). *Equal contribution.

978-1-6654-5908-2/22/$31.00 ©2022 IEEE

in danger or safe to use often depends on their own perception of the body rather than their real physical capabilities [12].

Our work builds on the hierarchical HAR-PBD architecture of [14] that shows advantage in leveraging the graphical representation of kinematic body movement data in the form of anatomical joint positions and further integrating activity context in PBD to be able to perform activity-independent time-continuous PBD. However, while the PBD performances obtained with the model of [14] are an improvement on prior state of the art [6], [13], the model uses only overt body movement data (based on motion capture) and does not leverage muscle activity information. Thus, based on an understanding of chronic pain behavior, we make the following novel contributions to the area of PBD in our work: (1) A new architecture named *Central Fusion with Attention* (CFAT) that pushes the state-of-the-art performance in PBD based on fusion of multiple modalities; first, CFAT extends the architecture proposed in [14] by introducing multi-modality using central fusion and additionally incorporating an attention mechanism; second, EMG data is for the first time represented within a graph structure in the modelling of PBD behavior, and our findings suggest that such structure is capable of capturing spatial relations between muscle groups; (2) With in-depth analyses of the CFAT model and comparison with other fusion strategies, we additionally explore how the use of EMG data can facilitate or hinder automatic detection of the activity context for optimization of the PBD performance.

II. Related Works

A. Protective Behavior Detection

Early approaches to PBD and related pain-behavior, e.g. [15] [4] used traditional machine learning models (e.g., Random Forests and Support Vector Machines) and feature engineering. These studies used both kinematic movement data (based on motion capture sensors) and muscle activity data. While these studies were the critical starting point in PBD, they were limited by their use of separately trained models for different activity types. They are additionally limited in their exploration of the fusion of the two modalities as they simply applied late or input-level (i.e. early) fusion based on concatenation of hand-crafted features from the sensor data.

More recently, [6] showed how the use of recurrent neural networks, with Long Short-Term Memory (LSTM) units in particular, can enable activity-independent PBD (i.e. a single PBD model that generalizes to multiple activity types) based on sensor data directly. Their best performance (macro F1 = 0.82) was based on a Stacked-LSTM network that used motion capture and EMG data with early fusion. The main limitation of their model is that it requires pre-segmentation of the body movement data into individual activity types, an approach that is not suitable for real-time detection of PBD where movements or activities performed are unlabeled.

Labels describing the activity context are very relevant for PBD. Indeed, how guarding (one category of protective behavior) is performed differs in sit-to-stand, stand-to-sit, and bending activities. In fact, determination of whether a given behavior is protective requires an idea of what the activity/movement that the person intends to perform is. It is thus important to investigate the application of recognition of the activity context to PBD. Human activity recognition (HAR) is a well-established area of research (e.g., [16]–[18]) but unfortunately minimally investigated in the context of pain-related affective processes which, as discussed in the introduction, make movement highly variable.

Recently, [14] proposed a Hierarchical HAR-PBD architecture comprising HAR and PBD modules built with Graph Convolution Network (GCN) and LSTM networks. Their aim was time-continuous (i.e. no pre-segmentation), activity-independent PBD. In either module, LSTM layers encode temporal dynamics of features passed from GCN layers. The GCN on the other hand strongly captures the relationship between anatomical joints because of their natural graphical representation. The findings in [14] further showed that PBD performance improved when HAR was incorporated: from a F1 score of 0.71 to 0.81. However, their HAR and PBD modules use only kinematic data (MoCap) as input (the PBD module additionally takes in the output of the HAR module for contextualization of the MoCap input). Studies such as [6], [19] highlight the value of PBD based on both overt body movement and muscle activity data. Thus, the question of how EMG data should be integrated in the Hierarchical HAR-PBD architecture, given its biomechanical difference with MoCap data, is pertinent. It is additionally valuable to explore whether EMG data will be informative for both HAR and PBD modules.

B. Attention Mechanism in HAR and PBD

Recent studies in related areas, e.g. HAR based on body movement data, have explored use of the attention mechanism. In [20], attention was used in an LSTM-based architecture to weight input data from multiple sensors at each single timestep as well as weighting the temporal sequence at the LSTM's output layer. This model achieved better performance compared with other previous models. Another notable approach is the BAN model by Cui et al. [21] which captures long-term dependencies using attention embedded in a bidirectional LSTM structure. A different but relevant model called HAMLET was introduced by Islam et al. in [22]. It is a hierarchical architecture that, at the lower level, has multi-head self-attention on spatio-temporal features for each of the modalities. A novel attention mechanism was additionally applied to fuse the multimodal output at the upper layer. Their algorithm obtained high accuracies of 95.12% and 97.45% respectively on the UTD-MHAD and the UT-Kinect datasets, and a F1 score of 81.52% on the UCSD-MIT dataset. In [23], they further proposed a Multi-GAT model that is based on graph networks. A Cross-Modal Graphical Attention (Cross-GAT) was used to capture cross-modal interactions and relations. This approach outperformed the previous multimodal HAR methods.

While the studies above suggest the efficacy of the attention mechanism, PBD works on a different timescale of model reasoning [14] and the size of the available dataset is much

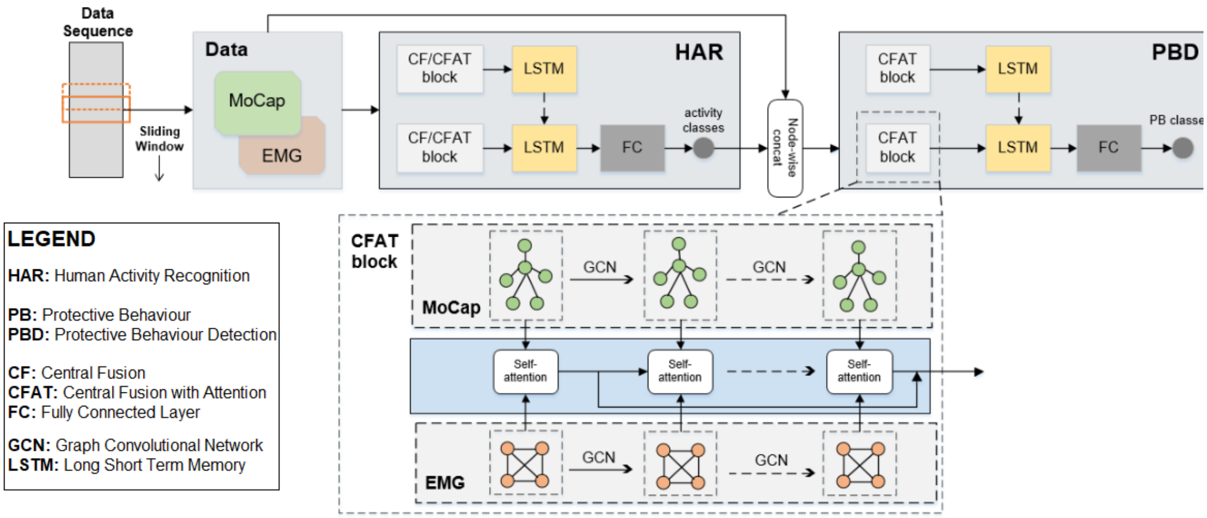


Fig. 1. An overview of the CFAT model. Time-continuous input data (MoCap and EMG data) is segmented with a sliding window. Both input modalities are represented in the form of graphs. (It is possible to use a vanilla central fusion block in the HAR module instead of a CFAT block, as our findings in Table I show that both fusion strategies for the HAR have similar effect on the PBD performance). A CF block differs from a CFAT in that where the self-attention mechanism is used in the CFAT (see Equation 5), a FC layer is instead used in the (vanilla) CF (see Equation 1). Both CF and CFAT integrate hidden states of the GCNs for the two modalities.

smaller than benchmark datasets used in HAR. The latter is because of the difficulties in capturing spontaneous affective experiences [24]. Meanwhile, attention has only been explored in unimodal PBD. [13] explored the use of attention, in a BANet architecture, to fuse input from multiple anatomical joints (MoCap data) and further integrate data from multiple timesteps. The model outperformed previous models with F1 score of 0.84. However, the work considers only MoCap data and works only on pre-segmented activity instances rather than on continuous data as needed for deployment. In this paper, we address these limitations by proposing a novel central fusion strategy with attention (referred to as CFAT) based on both MoCap and EMG data. The proposed method builds on the unimodal Hierarchical HAR-PBD model of [14] by integrating fusion and attention mechanisms informed by understanding of protective behavior.

III. Methodology

In this section, we describe the unimodal Hierarchical HAR-PBD (Baseline Model), our proposed multimodal fusion (CFAT Model), and two other fusion methods for comparison (Early and Late Fusion Models). The backbone comprising GCN and LSTM is used for encoding features from both modalities, which is based on the findings of [14].

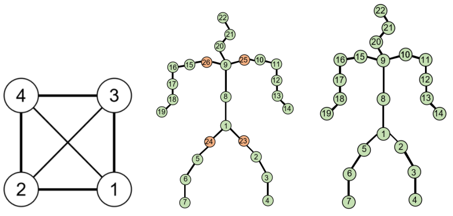
A. Baseline Model (Mocap data only)

The baseline model, i.e. the Hierarchical HAR-PBD architecture of [14], is an unimodal model that comprises an HAR module which takes MoCap data as input and a PBD module which uses a concatenation of the MoCap data and the output of the HAR module as input. Both the HAR and PBD modules are based on a GCN followed by an LSTM network. Thus, the MoCap data at each timestep of the input is represented as a non-directional graph with N_m nodes corresponding to the anatomical joints in the data and with an adjacency matrix corresponding to the natural human body configuration. The LSTM then takes in the graphical convolution outputs across all timesteps, and the computation at each LSTM unit in the sequence is propagated to the next timestep until the last in the sequence. The LSTM is finally followed by fully-connected layers that are used to make HAR or PBD class predictions.

B. Central Fusion (CF) with Attention (CFAT) Model

Our proposed architecture is illustrated in Figure 1. An early (input level) fusion of the MoCap and EMG modalities would not address the biomechanical difference between the two modalities due to pain and fear of movement and a late fusion would not leverage the remaining temporal relationship between them. The novelty of our CFAT approach is the application of weighted fusion in the middle of the GCN for each of the HAR and PBD modules. The weighted fusion is implemented as a central network based on self-attention layers [25] (see Equation 5) which take input from each layer of both the MoCap and EMG GCNs, and additionally has a short-cut connection in its output layer. The (vanilla) CF model has only fully connected layers, i.e. without attention, in its central network (see Equation 1).

MoCap & EMG GCNs. We use the MoCap GCN of the Baseline Model as the MoCap GCN for our CFAT Model. For the EMG GCN, we treat each EMG channel (i.e. each muscle group) as a node and so each GCN processes a graph consisting of N_e nodes, where N_e is the number of EMG channels. An illustration of this EMG graph is shown in Figure 2(a) where the adjacency of the 4 EMG nodes (right and left upper and lower back muscles) is set according to their lateral and vertical spatial relations on the body. Although we assume



(a) (b) (c)

Fig. 2. (a) Our stand-alone EMG graph with 4 nodes representing left and right lower lumbar paraspinal muscles (1 and 2) and upper trapezius muscles (3 and 4); (b) our combined MoCap and EMG graph (the added EMG nodes are in pink); (c) MoCap graph of the Baseline Model. The illustration is based on the EmoPain dataset.

4 EMG nodes in our illustration, our method allows for any number of EMG nodes.

Central Network. We fuse the outputs from the hidden layers of the MoCap and EMG GCNs in a central network (see Figure 2(b)). At layers i, the output H_C^i is computed as

$$\mathbf{H_C^i} = \sigma(\mathbf{w}(\mathbf{H_C^{i-1}} + concat(\sigma(\mathbf{w_1 H_{M_1}^i}), \sigma(\mathbf{w_2 H_{M_2}^i})))), \tag{1}$$

where $\mathbf{H_C^{i-1}}$ is the output of the previous central layer and $\mathbf{H_C^{i-1}} = 0$ for $i = 1$; $\mathbf{w}$, $\mathbf{w_1}$ and $\mathbf{w_2}$ are learnable weight matrices; $\mathbf{H_{M_1}^i}$ and $\mathbf{H_{M_2}^i}$ are the hidden representations at layers i of the modality 1 (MoCap) and modality 2 (EMG) GCNs respectively; σ is the rectified linear activation function.

In [26], central fusion is based on addition and requires identical feature dimension across modalities. In our work, we instead use concatenation, which only requires the dimension of the concatenated axis (i.e. the channel dimension) to be identical. Since the feature dimension of the MoCap and the EMG GCN can be of different lengths, based on [27], we use a 1*1 convolutional kernel to map the feature dimension of each hidden state to a common length. The dimensionality of the hidden states for MoCap and EMG GCNs can be described as (B, τ, N_m, F) and (B, τ, N_e, F) respectively, where $\tau = 1$ since each GCN only processes the data from one timestep at a time, B is the batch size, and F is the common feature dimension. In our central fusion, we concatenated these two modalities along the channel dimension (the third axis), which represents the number of nodes in each modality graph. The dimension of the concatenation is thus $(B, \tau, N_m + N_e, F)$.

Further, inspired by [28], we add a short-cut connection for the final layer of the central network such that

$$\mathbf{\hat{H}_C^{output}} = \mathbf{H_C^{output}} + \mathbf{H_C^1}. \tag{2}$$

Central Network with Self-Attention. As discussed in Section II, the use of the attention mechanism enables weighting of the features of the anatomical joints and muscles according to their relevance in each specific phase of a movement. In addition, we expect the attention mechanism to possibly capture attention shifts evident in limited coordination between anatomical joints in movement execution by people with CP and attempt to protect body parts perceived in danger. A key rationale in applying this method is to characterize the importance and influence of each channel of each modality in PBD. We expect (different) protective behavior strategies adopted by various people on the basis of their perception of what their body can do and what is dangerous will have different effects across the channels.

In this work, we implement a single-head attention mechanism based on the transformer model of [25] in our central network. Consider query $\mathbf{Q_t}$, key $\mathbf{K_t}$, and value $\mathbf{V_t}$ matrices for the attention submodule at timestep t extracted from the intermediate output $\mathbf{H_C^{i^*}}$ of the central layer i, i.e. $\mathbf{Q_t} = \mathbf{H_C^{i^*} W^Q}$, $\mathbf{K_t} = \mathbf{H_C^{i^*} W^K}$, and $\mathbf{V_t} = \mathbf{H_C^{i^*} W^V}$, where $\mathbf{W^Q}$, $\mathbf{W^K}$ and $\mathbf{W^V}$ are learnable matrices derived by implementing 1*1 convolution. The intermediate is computed as

$$\mathbf{H_C^{i^*}} = \mathbf{H_C^{i-1}} + concat(\sigma(\mathbf{w_1 H_{M_1}^i}), \sigma(\mathbf{w_2 H_{M_2}^i})). \tag{3}$$

The output of this layer after applying attention is

$$\begin{aligned} Att(\mathbf{H_C^{i^*}}) &= softmax(\frac{\mathbf{Q_t K_t^T}}{\sqrt{d_k}}) \cdot \mathbf{V_t^i}, \\ &= softmax(\frac{(\mathbf{H_C^{i^*} W^Q}) \cdot (\mathbf{H_C^{i^*} W^K})}{\sqrt{d_k}}) \cdot (\mathbf{H_C^{i^*} W^V}), \end{aligned} \tag{4}$$

where the factor $\sqrt{d_k}$ is the dimension of the key and is used to prevent the small gradient problem for softmax function when $\mathbf{Q_t K_t^T}$ becomes too large in magnitude [25]. We apply the softmax along the channel dimension, i.e. the third axis, whose length is the total number of MoCap and EMG nodes.

Based on Equation 4, the forward equation for the implementation of self-attention at layer i is computed as

$$\mathbf{\hat{H}_C^i} = Att(\mathbf{H_C^{i-1}} + concat((\sigma(\mathbf{w_1 H_{M_1}^i}), \sigma(\mathbf{w_2 H_{M_2}^i})))). \tag{5}$$

C. Early Fusion Model

In order to assess the value of the proposed CFAT architecture, we implement an early fusion version for comparison. Rather than simple feature concatenation of the MoCap and EMG, in our early fusion, we integrate the EMG data into the same graph (and GCN) as the MoCap data. Thus, we extend the MoCap graph of the Baseline Model by treating each EMG channel as an additional node leading to a graph with (N_m + N_e) nodes. An illustration of this multimodal graph based on the dataset used in this paper is shown in Figure 2b. Each of the 4 EMG nodes (right and left upper and lower back muscles) is taken as adjacent to the MoCap joints associated with the corresponding muscle group. As MoCap and EMG nodes may have a different number of features per channel, we use padding (with the same value) to ensure they have the same feature length.

D. Late Fusion Model

For completeness, we also explore late fusion. In our late fusion, for each of the HAR and PBD modules, the MoCap and EMG graphs are processed separately in parallel GC-LSTM networks. Each modality subnetwork outputs a prediction score (dimension = (B, C), where C is the number of classes). The prediction scores from both modalities are then concatenated and input to a fully-connected layer that makes the final prediction with softmax activation.

IV. Experiment Setup

In this section, we present the dataset and the methods used to evaluate the 4 models described in the previous section: the unimodal Baseline Model, our CFAT model, the Central Fusion (CF) model (i.e. CFAT without self-attention), and the Early and Late Fusion models. We report our evaluation based on accuracy, macro F1 score, PR-AUC (Precision Recall Area-Under-Curve), PR curve and confusion matrix.

A. EmoPain Dataset and Data Preprocessing

The EmoPain Dataset [29] is a multimodal dataset used in pain-related research to help design chronic pain (CP) rehabilitation technology. It contains sequences of movement data of 18 people with musculoskeletal chronic low back pain and 12 healthy people performing a variety of typical rehabilitation movements that reflect everyday challenges for people with CP and were selected by physiotherapists. The data was captured using full-body 3D IMU (Inertial Measurement Unit) sensors and EMG sensors placed on the lower and upper back. For the testing, we only used the data for people with CP in this work, since the healthy people were all assumed to exhibit no protective behavior. There were 46 sequences from 30 participants in total. Each sequence includes sit-to-stand, stand-to-sit, reaching forward, bending down, and standing on one foot (as a way to simulate stair climbing) activities. Transition movements between these activities were also included in the dataset. These transition movements include poses that people engaged in to relax their muscles or rest (e.g., standing still, or walking). The protective behavior labels in the dataset are based on annotation by 4 experts who marked the start and end of segments of each of 6 categories of protective behaviors.

To prepare the data, we followed the same approach used in [14]. We used a sliding window of 3 seconds (i.e. 180 samples/timesteps) with 50% overlapping ratio. These are based on findings in [6] that show them to be practical for PBD. Similar to [6], [14], both the HAR and PBD ground truths of a frame were decided by majority voting. That is, a frame was labelled as protective if at least 50% of the samples within it were annotated as protective by at least 2 experts.

To manage the small size (7629 frames of which 5231 from people with CP, in which 4219 are not PB and 1012 are PB) of the EmoPain dataset, jittering and cropping were used as data augmentation [30]. For jittering, we added Gaussian noise (mean=0, standard deviation=0.05 and 0.1 separately) to the input. For cropping, random MoCap joints at random timesteps were set to 0 with selection probabilities of 0.05 and 0.1 separately. After applying the data augmentation methods, the size of the dataset was five times larger than the original one, i.e, it increased from 7,629 frames to approximately 38,145 frames. Leave-One-Subject-Out Cross-Validation (LOSOCV) was used for evaluation.

B. Implementation Details

For each model, we trained the HAR and PBD modules simultaneously. We trained for 100 epochs with batch size 150 using an Adam optimizer [31] with a learning rate of $5e^{-4}$ and a decay of $1e^{-5}$. Based on preliminary experiments, we set the GCN (in both the HAR and PBD modules) to three layers each with 16 units for the MoCap graph and three layers each with 6 units for the EMG graph. For the LSTM, we used three layers with 24 units each for all models except the Late Fusion Model. For the Late Fusion Model, we used three layers each with 24 units for the Mocap subnetwork and three layers each with 8 units for the EMG subnetwork.

The class distribution for the HAR and PBD tasks was skewed, since the majority of the original data sequences are labelled as transition activities (68.29%) and non-protective (78.91%). Thus, we used the CFCC loss introduced in [14] to address the class imbalance problem. We fixed the parameters of the CFCC as γ=2 and β=0.9999, where γ and β control the scale that the focus of a model is balanced between classes.

V. Results

The aim of our experiments was twofold. First, we sought to understand if and how the fusion of EMG and MoCap data leads to improvement in PBD, given the complexity and large variety of movement strategies (which emerge from fear of pain) observed during protective behavior. Second, we investigated if the use of the attention mechanism can lead to further increase in PBD performance. We report the results in the following subsections together with statistical analysis. Given the lack of normality in the data, Friedman tests followed by post-hoc Wilcoxon comparisons with Bonferroni corrections (reported in Table I) were used to evaluate the differences in performances between the various models.

A. Does fusion lead to improvement in PBD performance?

The first five rows of Table I (PBD) show the results of the unimodal Baseline Model, 3 fusion approaches explored (Early, Late, and CF), and use of Central Fusion (CF) in PBD but not in HAR module. This latter model was explored to understand if the fusion is of value to both HAR and PBD modules, or it is useful within the PBD module only. We can see that the CF models (F1=0.89, 0.87 respectively) reached better PBD performance than the unimodal, early and late fusion models (F1=0.83, 0.82, 0.81 respectively). The use of CF in both the HAR and PBD modules reached also better performance than CF in PBD module only. The PR-AUC and PR curve (Figure 3) also reflect these results.

A Friedman test applied to the PBD F1 scores across LOSOCV folds showed statistically significant difference in performances of these 5 models, $\chi^2(4) = 22.080$, $p < 0.001$. Post-hoc analysis with Wilcoxon signed-rank tests with Bonferroni correction further showed that the PBD performance for the HAR(*CF*)PBD(*CF*) model is significantly higher than the PBD performances for the unimodal ($p = 0.016$), early ($p = 0.013$) and late fusion ($p = 0.013$) models. No significant difference was found with the PBD performance

TABLE I

Results for Human Activity Recognition (HAR) and Protective Behavior Detection (PBD). p-values for post-hoc Wilcoxon t-tests between the models highlighted in bold and by +, ++ or +++ and the model in the correspondent row. — = no comparison run, *no diff* = no significance difference. Numbers in bold indicate the best performances. CF=Central Fusion; CFAT=Central Fusion with Attention; *Early* and *Late* = fusion type; *Mocap* = MoCap data only.

	HAR		**PBD** (note: $\ast$ = statistical significance at α=0.05)					
Model	Accuracy	Macro F1	Accuracy	Macro F1	PR-AUC	p-value	p-value	p-value
						vs +	vs ++	vs +++
HAR(*Mocap*) PBD(*Mocap*) - baseline	0.80	0.68	0.89	0.83	0.73	0.061	0.016*	—
HAR(*Early*) PBD(*Early*)	0.82	0.72	0.88	0.82	0.72	0.052	0.013*	—
HAR(*Late*) PBD(*Late*)	0.81	0.71	0.88	0.81	0.69	0.052	0.013*	—
HAR(*MoCap*) PBD(*CF*) +	0.86	0.74	0.92	0.87	0.83	—	*no diff*	0.014*
HAR(*CF*) PBD(*CF*) ++	0.83	0.71	0.93	0.89	0.86	*no diff*	—	0.047*
HAR(*CFAT*) PBD(*CFAT*) +++	0.75	0.57	**0.96**	**0.93**	**0.94**	—	—	—
HAR(*CF*) PBD(*CFAT*)	**0.86**	**0.78**	0.95	0.91	0.92	—	—	*no diff*
HAR(*MoCap*) PBD(*CFAT*)	0.84	0.73	0.96	0.93	0.91	—	—	*no diff*

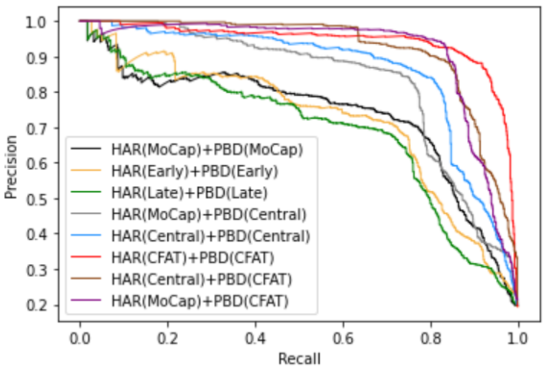


Fig. 3. PBD precision-recall curve of the different models

of the model using CF in the PBD module only, suggesting that the fusion of the modalities may be more important for PBD than the HAR module, as hypothesized. However, PBD performance difference between this latter model (i.e. HAR(*Mocap*)PBD(*CF*)) and unimodal, early fusion, and late fusion models only approached significance ($p = 0.061$, $p = 0.052$, $p = 0.052$ respectively). No significant difference in PBD performance was found between the unimodal, early and late fusion models. We can also see from the confusion matrices of Figure 4 (top) that, with respect to no-fusion or early/late fusion, HAR(*CF*)PBD(*CF*) leads to clear improvement in the recognition of the less represented protective behavior class. These results suggest that the fusion of EMG and MoCap data is valuable for PBD, especially when using model-level fusion.

B. Does attention mechanism further supports PBD?

We additionally compared the PBD performances of the CFAT model (*HAR(CFAT)PBD(CFAT)*), the central fusion model (*HAR(Central)PBD(Central)*), and the model with central fusion in the PBD module only (*HAR(Mocap)+PBD(Central)*). As rows 5-7 in Table I show, the CFAT model (F1=0.93) leads to better performances than the two central fusion models (F1=0.89, F1=0.87). A Friedman test found that the differences in performance are statistically significant ($\chi^2(2) = 10.871$, $p < 0.004$). The Wilcoxon post-hoc test showed that the PBD performance for the CFAT model is significantly higher than the ones of the two models using central fusion models ($p = 0.014$, $p = 0.047$). This result suggests that the attention mechanism further contributes to PBD by weighting the anatomical joints and muscle groups according to their relevance for each different activity phase. The confusion matrix HAR(*CFAT*)PBD(*CFAT*) in Figure 4 (top) shows that CFAT leads to clear improvement in the recognition of the less represented protective behavior class.

C. Do fusion and attention contribute to HAR?

The effects of fusion strategies on HAR can be explored by analyzing the first five rows of Table I. As can be seen, HAR(*Mocap*)PBD(*CF*) shows better HAR performance (F1=0.74) than the unimodal, early, late fusion and HAR(*Mocap*)PBD(*CF*) models (F1=0.68, 0.72. 0.71, 0.71 respectively). However, a Friedman test does not reveal any significant difference between these 5 models HAR performances.To understand the usefulness of the attention mechanism in the HAR module on both PBD and HAR performances, we considered variations of the CFAT modes where the attention mechanism is incorporated in both the PBD and HAR modules or just in the PBD module, i.e. HAR(*CFAT/CF/Mocap*)PBD(*CFAT*). The last 4 rows of Table I show that, although having the CFAT in both PBD and HAR modules leads to the best PBD performance (see also Figure 4(top)), it yields the worst HAR performance (F1=0.57). HAR(*CF*)PBD(*CFAT*) and HAR(*Mocap*)PBD(*CFAT*) reach

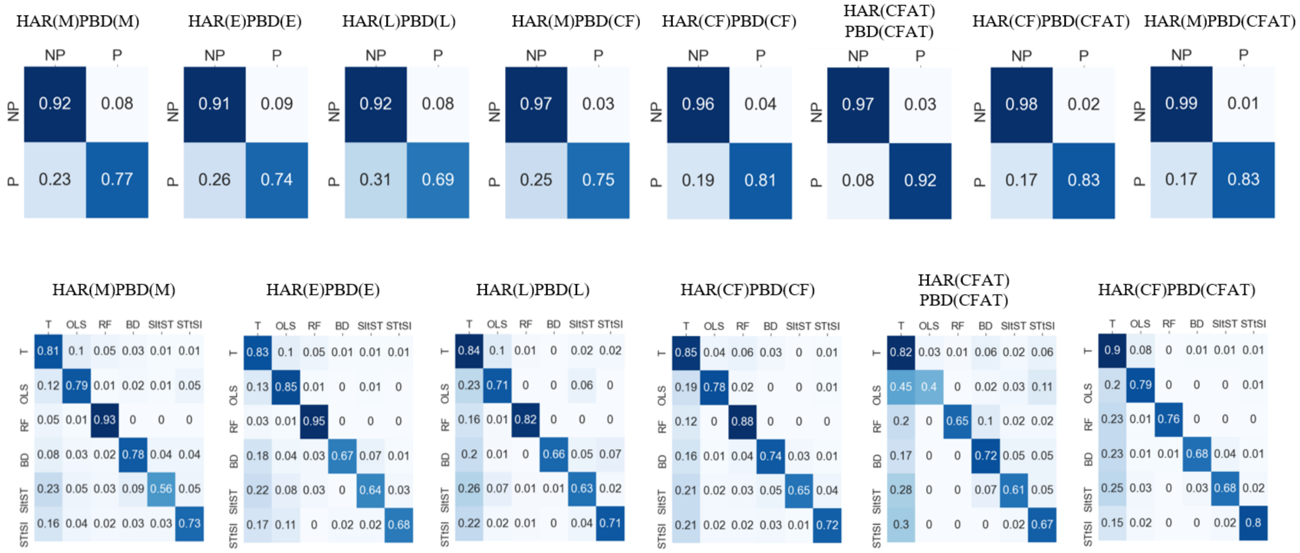


Fig. 4. Confusion matrices for HAR and PBD modules. Matrices titles: M=Mocap data only, E=Early Fusion, L=Late Fusion, CF=Central Fusion, CFAT=Central Fusion with Attention. PBD Matrices' classes: NP=Non-Protective, P=Protective. HAR Matrices' classes: T=Transition, OLS=One-Leg-Stand, RF=Reach-Forward, BD=Bend-Down, SItST=Sit-to-Stand, STtSt=Stand-to-Sit.

much better HAR performances: F1=0.86 and 0.84 respectively. The confusion matrices in Figure 4(bottom) show that this is due to HAR(*CFAT*)PBD(*CFAT*)'s poorer discrimination of standing-on-one-leg and transition activities in particular. The Friedman test however did not reveal any significant differences in HAR performances between the 3 models using CFAT. This suggests that while the inclusion of EMG data and use of attention in the fusion of the two modalities is very informative in detecting behavior triggered by fear of pain, it introduces noise in recognition of the activity context. This could be due to the weakened relationship between muscle activation and the activity performed due to altered muscle recruitment strategies in CP as discussed earlier.

VI. Discussion and Conclusion

We have proposed a novel architecture for PBD. The architecture extends previous work in [14] by exploring the encoding of EMG data using a GCN and through a central fusion with attention approach. Our approach was informed by the understanding about chronic pain behavior emerging from the literature.

Our findings suggest that the use of MoCap and EMG data with an appropriate fusion strategy can improve time-continuous, activity-independent PBD. The finding that only central and not early or late fusion leads to clear increase in performance is in line with the literature in chronic pain that highlights that muscle activity may not be strongly related to activity being performed because of altered muscle activity patterns due to fear of pain or possibly as a result of inability to control those muscles [9]–[11]. The finding that the attention mechanisms also contribute to improvement in PBD perhaps does typify shifts in attention across anatomical segments in people with CP based on (unhelpful) perception of danger.

The findings in [14] clearly highlight the positive effect of integrating HAR in PBD, so why does a worse performing HAR support PBD more than a better performing HAR? In other words, it shows that a HAR module that is better fitted to the manually-labelled activity *ground truth* led to reduction in PBD performance compared to the use of a HAR module that is joint-trained to enable better PBD performance. The most likely explanation is that PBD somehow learns to leverage the knowledge that when a person with CP moves in a way that is intuitive/natural (and so more easily categorizable in terms of the action/activity type), thus they are not behaving protectively. For example, in the dataset we used, a person that aims to bend forward but that is only using a minimal trunk flexion will have *bend* as the activity ground truth even though they have only actually done a *minimal trunk movement* activity. Forcing the HAR module to match the intended (as opposed to the executed) activity may be creating confusion or noise for HAR whereas it provides information that is valuable in PBD. In our architecture (Figure 1), we left the choice of central fusion or central fusion with attention for the HAR module open. The fusion strategy selected for the HAR module would depend on whether the priority is the PBD or HAR performance. However, larger datasets may show benefits of using central fusion with attention for the HAR module, as they may help realize strategy clusters across activity types.

In conclusion, the mid-level fusion of MoCap and EMG data with attention introduced in our CFAT architecture pushes the state-of-the-art in PBD with macro F1 score of 0.93 (PR-AUC=0.94, accuracy=0.96) on the EmoPain dataset from 0.83 (PR-AUC=0.73, accuracy=0.89). Our CFAT paves the way for automatic detection of protective behavior in the wild to provide tailored, real-time support to people with CP during everyday physical activities that they find challenging.

REFERENCES

[1] Turk D C. IASP taxonomy of chronic pain syndromes: preliminary assessment of reliability[J]. Pain, 1987, 30(2): 177-189.

[2] Goldberg D S, McGee S J. Pain as a global public health priority[J]. BMC public health, 2011, 11(1): 1-5.

[3] Della Volpe R, Popa T, Ginanneschi F, et al. Changes in coordination of postural control during dynamic stance in chronic low back pain patients[J]. Gait & posture, 2006, 24(3): 349-355.

[4] Olugbade T A, Singh A, Bianchi-Berthouze N, et al. How can affect be detected and represented in technological support for physical rehabilitation?[J]. ACM Transactions on Computer-Human Interaction (TOCHI), 2019, 26(1): 1-29.

[5] Langhorne P, Bernhardt J, Kwakkel G. Stroke rehabilitation[J]. The Lancet, 2011, 377(9778): 1693-1702.

[6] Wang, C., Olugbade, T.A., Mathur, A., Williams, A.C.D.C., Lane, N.D. and Bianchi-Berthouze, N. Chronic pain protective behavior detection with deep learning [J]. ACM Transactions on Computing for Healthcare. 2021, 2(3): 1-24.

[7] Geisser M E, Haig A J, Wallbom A S, et al. Pain-related fear, lumbar flexion, and dynamic EMG among persons with chronic musculoskeletal low back pain[J]. The Clinical journal of pain, 2004, 20(2): 61-69.

[8] Lelard T, Montalan B, Morel M F, et al. Postural correlates with painful situations[J]. Frontiers in Human Neuroscience, 2013, 7: 4.

[9] Trost Z, France C R, Sullivan M J, et al. Pain-related fear predicts reduced spinal motion following experimental back injury[J]. PAIN®, 2012, 153(5): 1015-1021.

[10] Thomas J S, France C R. Pain-related fear is associated with avoidance of spinal motion during recovery from low back pain[J]. Spine, 2007, 32(16): E460-E466.

[11] Watson P J, Booker C K, Main C J. Evidence for the role of psychological factors in abnormal paraspinal activity in patients with chronic low back pain[J]. Journal of Musculoskeletal Pain, 1997, 5(4): 41-56.

[12] Vlaeyen J W S, Linton S J. Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art[J]. Pain, 2000, 85(3): 317-332.

[13] Wang C, Peng M, Olugbade T A, et al. Learning temporal and bodily attention in protective movement behavior detection[C]//2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW). IEEE, 2019: 324-330.

[14] Wang C, Gao Y, Mathur A, et al. Leveraging activity recognition to enable protective behavior detection in continuous data[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2021, 5(2): 1-27.

[15] Aung M S H, Bianchi-Berthouze N, Watson P, et al. Automatic recognition of fear-avoidance behavior in chronic pain physical rehabilitation[C]//Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare. 2014: 158-161.

[16] Hammerla N Y, Halloran S, Plötz T. Deep, convolutional, and recurrent models for human activity recognition using wearables[J]. arXiv preprint arXiv:1604.08880, 2016.

[17] Zhang P, Lan C, Xing J, et al. View adaptive recurrent neural networks for high performance human action recognition from skeleton data[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2117-2126.

[18] Yan S, Xiong Y, Lin D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]//Thirty-second AAAI conference on artificial intelligence. 2018.

[19] Bao C, Fountas Z, Olugbade T, et al. Multimodal data fusion based on the global workspace theory[C]//Proceedings of the 2020 International Conference on Multimodal Interaction. 2020: 414-422.

[20] Zeng, Ming et al. ”Understanding and improving recurrent networks for human activity recognition by continuous attention.” Proceedings of International Symposium on Wearable Computers (ISWC), 2018, pp.56-63.

[21] Cui Q, Sun H, Li Y, et al. A Deep Bi-directional Attention Network for Human Motion Recovery[C]//IJCAI. 2019: 701-707.

[22] Islam M M, Iqbal T. Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020: 10285-10292.

[23] Islam M M, Iqbal T. Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1729-1736.

[24] Davis M C, Zautra A J, Smith B W. Chronic pain, stress, and the dynamics of affective differentiation[J]. Journal of personality, 2004, 72(6): 1133-1160.

[25] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[26] Vlaeyen J W S, Linton S J. Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art[J]. Pain, 2000, 85(3): 317-332.

[27] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[28] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[29] Aung M S H, Kaltwang S, Romera-Paredes B, et al. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal EmoPain dataset[J]. IEEE transactions on affective computing, 2015, 7(4): 435-451.

[30] Um T T, Pfister F M J, Pichler D, et al. Data augmentation of wearable sensor data for parkinson’s disease monitoring using convolutional neural networks[C]//Proceedings of the 19th ACM international conference on multimodal interaction. 2017: 216-220.

[31] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.