EMOPAIN Challenge 2020: Multimodal Pain Evaluation from Facial and Bodily Expressions

Joy O. Egede[1†], Siyang Song[1†], Temitayo A. Olugbade[2†], Chongyang Wang[2†], Amanda C. De C. Williams[2], Hongying Meng[4], Min Aung[3], Nicholas D. Lane[5], Michel Valstar[1], Nadia Bianchi-Berthouze[2*]

[1] School of Computer Science, The University of Nottingham
[2] University College London
[3] Department of Computer Science, University of East Anglia
[4] Department of Electronic and Computer Engineering, Brunel University London
[5] Department of Computer Science, University of Oxford

***Abstract*— The EmoPain 2020 Challenge is the first international competition aimed at creating a uniform platform for the comparison of multi-modal machine learning and multimedia processing methods of chronic pain assessment from human expressive behaviour, and also the identification of pain-related behaviours. The objective of the challenge is to promote research in the development of assistive technologies that help improve the quality of life for people with chronic pain via real-time monitoring and feedback to help manage their condition and remain physically active. The challenge also aims to encourage the use of the relatively underutilised, albeit vital bodily expression signals for automatic pain and pain-related emotion recognition. This paper presents a description of the challenge, competition guidelines, bench-marking dataset, and the baseline systems' architecture and performance on the Challenge's three sub-tasks: pain estimation from facial expressions, pain recognition from multimodal movement, and protective movement behaviour detection.**

I. INTRODUCTION

The EmoPain 2020 Challenge [1] is the first international competition in automatic pain recognition aimed at benchmarking the performance of machine learning methods designed to recognise or quantify chronic pain from behavioural—face and body—cues, and also recognise pain-related movement behaviours. Chronic pain (CP) is a widespread distressing problem that not only restricts body activities but significantly impacts on the mental, psychological, social and economic status of people with chronic pain. A 2016 study [1] showed that over 40% of the UK population are affected by chronic pain with this number going up to 62% for people over 75 years. A similar study for the United States puts the former figure at 25% [2]. Beyond the individual, CP has dire consequences on socio-economic growth and development. Amongst other medical conditions, chronic pain was responsible for most medical consultations and costs the US approximately $560 billion dollars each year [2]. The escalating socio-economic costs of CP, as well as its detrimental effect on the quality of life of individuals and their families, buttress the urgent need for efficient chronic pain interventions.

* Corresponding author
† These authors made equal contributions
[1] https://mvrjustid.github.io/EmoPainChallenge2020/

Technological interventions present a plausible solution, but the first step towards a workable system requires accurate identification and interpretation of pain-associated expressions and behaviours. Consequently, technology-driven methods (see survey in [3]) utilising clinically certified behavioural and physiological pain indicators for pain assessment have been proposed within the machine learning and computer vision research community. Although machine-assisted pain assessment methods have advanced considerably, their practical application has been constrained by data-related and design issues.

One major problem is that there are few publicly accessible pain datasets that meet requirements for effectively training such predictive systems. Secondly, pain expression is multi-faceted, yet there is an over-reliance on unimodal clues, particularly the face, whereas body movements are critical to effective chronic pain assessment [4]. Although facial expressions give a good indication of affect intensity, without the body context, its discriminative property of affective states diminishes [5]. In contrast, pain-related movement behaviour provides more information about the distress level of a pain stimulus (physical activity) and what form of support is required [6], [4]. Thus, pain literature [7], [4] strongly advocates the use of multiple, rather than isolated behavioural cues for pain assessment. Lastly, existing bench-marking pain corpus [8], [9] predominantly feature pain expressions induced in constrained environments and by non-threatening stimuli which are not fully representative of real-world distressing physical activities encountered by people with chronic pain; whereas, for technological interventions to be beneficial, it should be developed on data which represent the everyday body functions of the target population. Also, some of these datasets [9], [10] provide only uni-dimensional—facial cues—behavioural chronic pain characterisations.

The EmoPain 2020 challenge aims to address the above gaps by creating a platform to foster multi-modal automatic pain recognition research within the machine learning community. The challenge is based on the multi-modal EmoPain dataset, which for the first time, is opened up to the community in a competition framework to benchmark

978-1-7281-3079-8/20/$31.00 ©2020 IEEE
DOI 10.1109/FG47880.2020.00078

automated pain assessment methods. The EmoPain dataset [7] consists of audiovisual, motion data and muscle activity captured from chronic lower back pain (CLBP) and healthy participants engaged in both instructor-led and self-directed physical exercises which replicate everyday body functions. Utilising the visual and movement data dimensions, the EmoPain 2020 challenge presents three pain recognition tasks: (i) Pain Estimation from Facial Expressions Task, (ii) Pain Recognition from Multimodal Movement Task and (iii) Multimodal Movement Behaviour Classification Task.

Participants could choose to compete in all or some of the tasks. Data for each task is split into training, validation and a held-out test partition. To ensure a fair comparison, participants were given the same training and validation data to develop their algorithms/models, which was then sent to the organisers for evaluation on the held-out test set. Participants did not have access to the test data partition. Papers accompanying the challenge submissions were presented at the FG2020 International Workshop on Automated Assessment of Pain.

The rest of the paper is organized as follows: Section II discusses relevant work in automatic pain recognition; Section III gives a full description of the EmoPain dataset and the three sub-tasks as well as the metrics used for ranking participants' submissions; Section IV describes the baseline features and models developed for each task, and the results obtained. Lastly, section V summarises the contributions and concludes the work.

II. RELATED WORK

This section describes current approaches to automatic pain recognition with a focus on pain-associated face and body expression synthesis, processing, analysis and interpretation. Relevant pain literature will be discussed in three groups building on the challenge's task categorisation. An extended survey is provided in [3].

A. Automatic Pain Detection based on Facial Expressions

The face is a key medium for communicating pain in human interactions, particularly when pain expression is not actively suppressed by the individual. Facial expressions of pain have been shown to have distinctive properties from other basic emotions [11], [12], lending credence to its pertinence to pain recognition. Due to its relative ease of accessibility and utilisation in daily social interaction, faces have been explored extensively for automatic pain recognition. Early work based on facial actions was limited to binary classification of face images into *pain* or *no pain* [13], [14] or distinguishing real pain from posed pain [15]. However, this outcome was not adequate for clinical applications as evidenced by the self-report pain assessment scales [16] which aim to quantify pain rather than identify its occurrence. Consequently, recent studies moved on to estimating pain levels from facial expressions using either a multi-class classification set-up [17] or regression framework [18], [19], [20]. This shift was also propelled by the introduction of pain datasets [8], [9] which provide discrete pain annotations of face images. Most of these studies [20], [19] predict pain on the 16-point Prckachin and Solomon Pain Intensity (PSPI) [21] scale or a condensed version [17], while others [22], [23] focus on recognising observer reported or patients' self-reported pain ranging from two to five pain levels.

To discriminate pain expressions, face shape and appearance descriptors have been widely employed due to their proven effectiveness in facial expressions analysis. Appearance features encode facial deformations due to expressions (e.g., wrinkles) while shape features describe the spatial localisation of facial components (i.e., eyes, mouth and nose). In terms of facial features used, previous work on pain recognition can be classified into three: (i) handcrafted feature methods [22], [17], [13], [20], (ii) data-learned feature methods[24], [25] and (iii) hybrid-feature methods [19], [26]. Handcrafted facial descriptors are statistical measures computed from a face image using human-designed algorithms. Commonly used features in this category include gradients features [22], Gabor features [15] , Active Appearance Models (AAM) [13], [17], Local Binary Patterns (LBP) [20], facial landmarks and associated distance metrics [22] amongst others. Data-learned features are offshoots of neural network applications to pain recognition and are automatically generated within the network. Hybrid features, on the other hand, are an integration of traditional and data-learned features and have been shown to significantly improve the predictive ability of recognition models on small datasets [27].

Although pain recognition from faces has witnessed tremendous progress, there is still ample scope for improvement. Current work has concentrated on facial data collected in constrained, ideal settings where several video cameras are positioned at strategic positions to capture face images from all possible angles. Thus, captured images are usually high resolution, near frontal and unobstructed faces, whereas this is not always the case in typical everyday settings, e.g., performing rehabilitation exercise at home. Another open challenge is insufficient data representation for higher pain levels in existing pain corpus, which limits the performance of recognition models on these pain classes [27]. Hence, novel methods that make the most of existing data, and more focus on the creation of representative chronic pain facial data are required.

B. Automatic Pain Detection based on Bodily Expressions

Despite findings in [4] that the body may be more expressive of pain experience than the face or vocal modality, which are more dependent on social context, it has not been as widely explored for automatic detection of pain levels as the face. Most of the early studies [28], [29] and a number of more recent work [30], [31] focused on discrimination between people with chronic pain and those without. Other studies have similarly investigated differentiation between two levels of pain [32], [33]. One exception is [34] where 11 levels of pain were detected. While studies such as [35], [36] have also gone beyond binary classification, unlike the afore-mentioned, they are based on experimentally-induced

pain which is transient and not usually perceived a threat [37].

The bodily expressions used in the investigations carried out in these studies have typically depended on the pain location and the activity being performed. For example, in the work of [31], automatic detection of knee pain was based on gait characteristics and ground force reaction during walking tasks. Similarly, the automatic detection of neck pain in [30] used neck movements measured while participants performed neck exercises. For low back pain, where participants are usually being assessed during physical activities involving the trunk, features of trunk [28], [29], [32], spine [34], knee [29], and hip [29] movement, corresponding back muscle activity, and force and centre of gravity [29] have been used for pain (level) detection.

Another work in the area related to body movement is the one of Rivas et al. [38]. In their work, the authors explore the use of hand pressure and joystick manipulation to detect stroke patients' pain level by personalising the model to each patient by using data from 10 different sessions. In [39], the authors extend the work by combining multiple modalities (hand pressure, gesture and facial expressions) to investigate the relationship between affective states and pain during rehabilitation. Again, individual models are built by taking advantage of the multiple sessions.

In a recent study [40] on automatic discrimination between healthy participants, low-level pain, and high-level pain based on complete movement instances in the EmoPain dataset, we explored features of the trunk, knee, head/neck, and arm movements computed from full-body positional data as well as features from shoulder and lower back muscle activity. We used two separate sets of features for trunk flexion and sit-to-stand movements respectively, given the considerable differences in the temporality of the two movements and the anatomical regions recruited in performing them. We additionally built a separate model for each movement type for this reason and especially to manage the limited data size available. For full and forward trunk flexion, we extracted the range of trunk and neck movement, the amount of unsteadiness in arm movement, and the time and amplitude of high-to-low muscle activity change; for sit-to-stand, we extracted range of trunk and neck movement, knee and pelvic angles at the point of buttocks lift, speed and duration of the lift phase, and the time of high-to-low muscle activity change and muscle activity range. We obtained 0.90 F1 score (0.90 accuracy) on average, over the three classes and three movement types, based on leave-one-subject-out cross-validation.

C. Automatic Detection of Protective Movement Behaviour

Aside from the pain estimation on bodily expressions, the movement behaviour presented therein is informative not only of pain level but also of the emotional state and engagement level of people with chronic lower back pain (CLBP). Specifically, the protective behaviour, e.g., hesitation, guarding, stiffness, the use of support and bracing [41], expression of fear or low-efficacy of movements, is currently adopted by physiotherapists in tailoring their feedback and interventions [42], [43]. As the rehabilitation for CLBP people is moving towards self-management outside the hospital, researchers started to work on the establishment of a virtual physiotherapist, where the first step is about the automatic detection of protective behaviour. Early studies in this direction mainly focused on feature-engineering methods to extract discriminative features from motion capture (MoCap) and surface electromyographic (sEMG) data with shallow classifiers like Random Forests and Support Vector Machine applied on top of them [7], [44], [45]. To name a few, features used include the range of joint angle, the mean of the angular velocity and the mean of the upper-envelope of the sEMG data. One limitation of these works is the lack of generalisability across different types of movement. Recently, efforts are also seen in using deep learning for the detection of protective behaviour. A comparison of different vanilla neural networks is provided in [46], while some data augmentation techniques were also explored. The result achieved is much higher than previous feature-based methods, on the data pooled from different movement types. Later on, a collaboration of LSTM network with attention mechanism is presented in [47], where better and explainable results are reported. However, challenges still exist, such as the dependence on the pre-segmented activity sequences which is not able to provide real-time encouragements and feedback, and the lack of exploitation of the bio-mechanical nature of MoCap and sEMG data especially, resulting from the traversal data processing strategy.

III. Challenge Description

This section describes the data collection protocol for the benchmark data (EmoPain database), the Challenge's tasks, task data partitioning, and proposes real-world applications of each task to clinical pain management.

A. EmoPain Dataset

The EmoPain dataset [7] provided for the challenge originally comprised of audiovisual, motion-capture and muscle activity data, collected from 18 CLBP and 22 healthy participants. Here need to note that, the real number of participants provided for each challenge task differs. Each participant went through at least one trial of the data collection, either the normal or the difficult trial. Within a trial, the participant performs a sequence of activities, namely one-leg-stand, reach-forward, stand-to-sit, sit-to-stand and bend-down. These activities are connected by transition activities, like standing still, sitting still and self-preparation. In the difficult trial, participant has to follow instructions set by the experimenter and carry a 2Kg weight in each hand during the performance of reach-forward and bend-down. There are no such limitations in the normal trial.

For the facial expression video, several sets of features are extracted for the challenge participant, which will be described in detail in the next section. For the body movement data, the joint angles and respective angular velocities are

computed. The dataset for the challenge is split into training, validation and a held-out test partition. The participant partition are shown in Table I. The class distribution is not considered for the partition of the dataset, but we ensure each partition has sufficient representation of healthy participants and CLBP patients' data.

B. Challenge Tasks

The EmoPain Challenge consists of three main tasks namely: (i) pain estimation from facial expressions, (ii) pain recognition from multi-modal movement, and (iii) protective movement behaviour detection. Participants were expected to compete in at least one or more tasks.

The ***Pain Estimation from Facial Expressions Task*** aims to develop technology to automatically quantify pain from face images of CLBP and healthy participants performing physical activity. These technologies could potentially support real-time pain assessment for patients who are unable to self-report pain, e.g., unconscious patients, and in constrained settings, e.g., ICUs, where continuous recording of a person's face is possible. Anchoring on facial properties deemed suitable for facial expression analysis [20], [27], data for this sub-task consists of anonymized face shape and appearance features extracted from the EmoPain video images (see details in IV-A), as well as observer pain annotations for each face image on an 11-point scale ranging from 0 (no pain) to 10 (maximum possible pain intensity).Due to data protection and ethical constraints, we did not provide the original video images.

Note that the values of the original pain annotations for the face range from 0 to 1000. These labels are heavily unbalanced, as the value of most labels are zero and for some other values, only less than 10 frames have such pain level. To alleviate this problem, we re-sampled all labels into 11 bins, from 0 to 10. Specifically, the values of all original labels were divided by 100, and then allocated to the bin whose value corresponds to their integral part, e.g., a label value of 232 will be assigned to *bin* 2. The distribution of the final provided labels are detailed in Table II. Participants' submissions to this task were ranked using the *Concordance Correlation Coefficient (CCC)* [48] which measures the temporal association between the model predictions and ground truth pain labels. CCC is preferred over similar measures— *Pearson's CC* and *Spearman's CC*— because it encodes precision and accuracy metrics in a single measurement and is robust to location and scale variations [48].

The ***Pain Recognition from Multimodal Movement Task*** aims to detect and classify levels of pain experienced by a person with chronic pain during movement activities. Technology with this capability could help a person with chronic pain more helpfully pace physical activity performance [40]. Data for this sub-task comprises of muscle activity data, 13 joint angles and angular energies (see full description in [47]) captured from CLBP and healthy participants while performing physical activities. Each activity instance is accompanied by a three-class pain annotation: no pain, low pain and high pain, which will serve as ground-truth labels for the task.

TABLE I
PARTICIPANT DISTRIBUTION IN EACH DATA PARTITION. CLBP - CHRONIC LOWER BACK PAIN; HP - HEALTHY PARTICIPANTS

Partitions	Face Tasks	Body Tasks
Train	8 CLBP and 11 HP	10 CLBP and 6 HP
Validation	3 CLBP and 6 HP	4 CLBP and 3 HP
Test	3 CLBP and 5 HP	4 CLBP and 3 HP

The submissions for this task were evaluated using F1 scores and accuracy, but final ranking was done based on Matthew Correlation Coefficient (MCC) [49] which better accounts for the negative classes.

The ***Multimodal Movement Behaviour Classification Task*** aims to develop technology that can detect and classify protective behaviours (e.g., rigid movement) in people with chronic pain. Such technologies could provide immediate and appropriate feedback or support to users, e.g., notifying the user to adopt a correct posture if the use of maladaptive strategy is detected [40], [46]. Data for this task consists of 13 bodily joint angular features and muscle activity for each movement frame with corresponding activity-type labels and binary protective behaviour annotations by 2 physiotherapists and 2 psychologists. For this task, macro average F1 score and the F1 score for each class (i.e. protective and non-protective) were used for ranking participants' submissions.

IV. BASELINE FEATURES AND SYSTEMS

In this section, we describe the features extracted from each pain expression modality, the baseline models implemented for each task, and present the results obtained from the performance evaluation of the models.

A. Pain Estimation from Facial Expressions

For the pain estimation from face sub-challenge, we extracted four facial descriptors using the OpenFace 2.0 toolkit [50], and two deep-learned emotion-oriented feature representations [51]. The detailed descriptions of these features are as follows:

- *Facial landmarks*: 68 2-D and 3-D fiducial facial points.
- *Head pose*: Pitch, yaw and roll angles.
- *Gaze*: 3-D gaze directions.
- *HOG*: a 4464-D Histogram of Oriented Gradients (HOG) features.
- *Action Unit (AU) occurrence*: 18 AUs whose values are 1 (present) or 0 (absent).
- *AU intensities*: 17 AUs whose values range from 0 to 5 (max intensity).
- *VGG-16 feature*: 4096-D deep features extracted from the second fully-connected layers of the VGG-16 network [52].
- *ResNet-50 feature*: 2048-D deep features extracted from the fully-connected layers of the ResNet-50 network. [53]. The VGG-16 and ResNet-50 network are pre-trained on the Affwild dataset [54] with valence and arousal labels.

TABLE II

LABEL DISTRIBUTION OF THE PAIN ESTIMATION FROM FACE SUB-CHALLENGE

Label value	0	1	2	3	4	5	6	7	8	9	10
Training	646634	39694	31032	61148	41286	17122	16958	9140	3734	626	2078
Development	475717	20731	31697	25613	20765	15416	7425	9972	198	176	218

Although the data labels are significantly imbalanced as seen in Table II, we do not perform any data augmentation, to enhance the reproducibility of the reported results. While the task can be solved as an 11-class classification problem, in this challenge, we treated it as a regression problem.

The face baseline system employed four different feature sets: 2 hand-crafted features including geometric features (a combination of 2-D facial landmarks and gaze directions) and 4464-D HOG feature; and 2 emotion-oriented deep-learned feature sets including 4096-D VGG-16 features and 2048-D ResNet features. Note that the 2-D facial landmarks are transformed into a 136-D dimension feature vector for each frame. The training process starts with feature normalisation. For each dimension of the input feature, the training set was normalised using z-score as shown in Equation 1.

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ and σ are the mean and standard deviation of the feature values over the entire training data. The obtained mean value and standard deviation were then applied to normalize the validation and test set. In this sub-challenge, we trained an Artificial Neural Network (ANN) for each feature subset. The employed ANNs follow the set-up presented in [55], which consists of 4 fully connected hidden layers. A dropout [56] with probability 0.5 and a ReLU layer is placed after each fully-connected layer. RMSprop is used as the training method, while Mean Square Error (MSE) is employed as the loss function. The hyper-parameters and topology chosen for the baseline systems are shown in Table III. These hyper-parameters were determined by grid search on validation set.

The baseline results of the Pain Estimation from Faces sub-challenge are given in Table IV. They show that amongst the single-feature models, the best correlation (CCC) on the development set results was achieved by VGG-16 feature, which also obtained good RMSE and MSE results. However, while VGG-16 feature also achieved solid performance on the test set in terms of the RMSE and MSE, its predictions are not highly correlated with the ground-truth of the test set. Instead, the combination of facial landmarks and eye gaze features produced excellent RMSE and MSE results on both development and test set, and also generated predictions with the highest correlation (PCC) to the labels in the test set. These results indicate that the pain level can be partially reflected by the geometric information of the face and eyes.

The decision-level fusion of all modalities gave the best results on both the development set (RMSE = 1.69, PCC = 0.25, CCC = 0.18) and test set (MAE = 0.91, RMSE = 1.41, PCC = 0.10, CCC = 0.06), except the MAE returned on the development set (MAE = 1.26) is slightly higher than the best one (MAE = 1.24). Based on the fusion results, we can argue that though the individual features were not very informative for pain intensity estimation when simple ANNs are used as the back-end, their fusion still seems to provide more valuable and **positive** information for pain estimation. Based on all results, the recognition of pain intensities from the face is still challenging when only combining existing standard hand-crafted or deep-learned features with a simple back-end. This observation opens interesting research questions about how to extract pain-related cues from complex facial expressions and emotions.

B. Pain Classification based on Body Movement and Muscle Activity

Due to the limited data size available in this task, we chose to build a single model for all movement types in the dataset so as to maximise the training data. The features that we extracted (see Table V) were based on findings in [40]. We extracted range of joint angles, to characterise the range of movement across anatomical regions relevant to the movement types. We additionally computed speed of movement over all joint angles and over each movement. While it might ordinarily be valuable to compute speed separately for each joint, it was necessary for us to constrain feature dimensionality in order to further address the data

TABLE III

THE CHOSEN HYPER-PARAMETERS OF ANNS FOR FACIAL CHALLENGE BASELINE SYSTEMS

Feature	Hidden Layers Size	Learning Rate	Batch Size
FL+Gaze	(128, 64, 32, 32)	0.001	128
HOG	(2000, 512, 256, 64)	0.001	256
VGG-16	(1024, 256, 64, 64)	0.005	128
ResNet-50	(1024, 256, 64, 64)	0.001	256

TABLE IV

BASELINE RESULTS FOR THE PAIN ESTIMATION FROM FACIAL FEATURES. BEST RESULTS ARE HIGHLIGHTED IN BOLD

Modality	Partition	MAE	RMSE	PCC	CCC
FL+GAZE	Valid.	1.51	1.74	0.04	0.003
FL+GAZE	Test.	1.37	1.56	**0.10**	0.003
HOG	Valid.	**1.24**	1.91	0.05	0.04
HOG	Test.	0.93	1.61	0.03	0.02
VGG-16	Valid.	1.34	1.82	0.24	**0.18**
VGG-16	Test.	0.92	1.43	0.02	0.004
ResNet-50	Valid.	1.42	2.08	-0.08	-0.04
ResNet-50	Test.	1.14	1.74	-0.09	-0.06
Fusion	Valid.	1.26	**1.69**	**0.25**	**0.18**
Fusion	Test.	**0.91**	**1.41**	**0.10**	**0.06**

TABLE V

BODILY FEATURES USED FOR PAIN CLASSIFICATION

Features		Formulae	Dimension
Range of joint angle		$\Delta J_i = \max_t J_i - \min_t J_i$	11
Speed	max	$\max_i \max_t \frac{\delta J_i}{\delta t}$	1
	min	$\min_i \min_t \frac{\delta J_i}{\delta t}$	1
	mean	$\frac{\sum_i \frac{\sum_t \frac{\delta J_i}{\delta t}}{T}}{I}$	1
Range of muscle activity		$\Delta E_k = \max_t E_k - \min_t E_k$	4
where i = 2, 3, ..., I; I = 13; t = 1, 2, ..., T; k = 1, 2, 3, 4			

size limitation. Finally, we computed the range of activity for each of the four muscle groups in the sEMG data.

Each data instance is made up of one or more iterations (up to 6) of a complete movement type, and so it was important to additionally incorporate the dynamics within each instance in the feature set. We addressed this by extracting the 18 above-mentioned features in 4 identically-sized non-overlapping window segments that together cover the data instance. 4 was a compromise between limiting the number of features and characterising movements which had the maximum number of repetitions. This led to 72 dimensions for the feature vector for each data instance.

We explored three main algorithms for the three-level classification of pain based on body movement and muscle activity data: Random Forest (RF) [57], Support Vector Machines (SVMs) [58], and k-Nearest Neighbours (kNN). The algorithms were evaluated using leave-one-subject-out cross-validation, based on the challenge training set alone. The hyperparameters for the algorithms were set based on grid search using an inner validation set within each validation fold, and among: 1, 5, 10, and 50 trees for the RF, and one, square root of the total amount, and the total amount for the number of features used to split each node in the RF; 1 to 5 degrees for the polynomial SVM, Gaussian or sigmoid kernels for the SVM, and 0.001, 0.01, 0.1, 1, 10, and 100 as the box constraint size for either of the three SVMs; k between 1 and 5, and *minkowski*, *euclidean*, *manhattan*, or *chebyshev* distances for the kNN. Note that in the SVMs and kNN setup, the feature set was normalised to zero and unit variance.

The kNN, and sigmoid and Gaussian SVM, which emerged as not worse off than chance-level detection based on the cross-validation, were further evaluated in hold-out validation, with the challenge training, validation, and test sets for training, validation, and testing respectively. Table VI shows the data sizes across the three pain classes (healthy, low-level pain, and high-level pain) for both the leave-one-subject-out cross-validation (LOSO-CV) and the hold-out validation. Table VII shows the F1 scores, Matthews Correlation Coefficients (MCCs) [49], and accuracies of the SVM, RF, and kNN, for three-level pain classification based on leave-one-subject-out cross-validation with the training set. Both the RF and polynomial SVM perform worse than chance-level detection (F1 score = 0.33; MCC = 0; accuracy = 0.33). As can be seen in Table VIII, although the non-polynomial SVM has the best performance in the cross-validation, it performs much poorly in further evaluation on the test set, whereas the kNN has more or less the same performance in both the cross-validation and the hold-out validation, albeit only about as good as chance-level detection. In the cross-validation, the kNN performs worst in detection of the high-level pain class (F1 score = 0.16, MCC = -0.02) compared with the healthy class (F1 score = 0.44, MCC = 0.1) and the low-level pain class (F1 score = 0.41, MCC = 0.08). However, in hold-out validation, its performance is worst for the low-level pain class (see Table VIII).

TABLE VI

DATA SIZES FOR MOCAP AND SEMG DATA FOR PAIN CLASSIFICATION

Pain Class	Training Set	Validation Set	Test Set
Healthy	34	25	25
Low-Pain	44	30	4
High-Pain	35	5	26

TABLE VII

LOSO-CV BASELINE RESULTS FOR PAIN CLASSIFICATION FROM MOCAP AND SEMG DATA

Algorithm	F1 Score*	MCC*	Accuracy
Sigmoid/Gaussian SVM	0.41	0.19	0.44
kNN	0.34	0.05	0.37
RF	0.26	-0.10	0.27
Polynomial SVM	0.15	-0.16	0.26

C. Protective Movement Behaviour Detection

To leave enough space for explorations, a stacked-LSTM network adapted from [46] is used as the baseline for the movement behaviour detection task. The architecture stays the same, where three LSTM layers with 32 hidden units are used together with a softmax fully-connected layer for classification. The input to the network is a frame with size of NxTxD, where N is the number of samples, T is the length of timesteps and D is the dimension of features. The data used is the 13 angles and their respective square of angular velocities as well as the upper envelope of the sEMG data. As a result, the data matrix has the dimension D=30. A sliding window of 180 timesteps long and a 0.75 overlapping ratio is used to extract consecutive frames from each activity type. To enable the training of stacked-LSTM, we further applied two augmentations: i) jittering, where Gaussian noise with standard deviation of 0.05, 0.1 and 0.15 are globally applied to the raw data; ii) cropping, where samples at random timesteps and body parts are set to 0 with probability of 0.05, 0.1 and 0.15. Augmentation is only applied to training data. The number of frames after segmentation is 6623 (with protective frames totalling 1,330), which is augmented to 33,115 (with protective frames totalling 6,650). The hold-out validation stays the same with the other two tasks. The

TABLE VIII
HOLD-OUT VALIDATION BASELINE RESULTS FOR PAIN CLASSIFICATION USING MOCAP AND SEMG DATA

	kNN (k=1, manhattan distance)		Sigmoid/Gaussian SVM (Gaussian kernel, box constraint=0.1)	
Metric	F1 Score	MCC	F1 Score	MCC
Healthy (0)	0.39	-0.04	0.00	-
Low-Pain (1)	0.09	-0.06	0.14	-
High-Pain (2)	0.44	0.16	0.00	-
Average	0.31	0.02	0.34	-
Accuracy	0.35		0.07	

TABLE IX
BASELINE HOLD-OUT VALIDATION RESULTS FOR PROTECTIVE MOVEMENT BEHAVIOUR DETECTION WITH MOCAP AND SEMG DATA

Method	Partition	Class	Acc	F1 score
stacked-LSTM	Valid	Non-protective (0)	-	0.9622
		Protective (1)	-	-
		Average	0.4636	0.4811
	Test	Non-protective (0)	-	0.9029
		Protective (1)	-	0.2465
		Average	0.828	0.5747

groundtruth of each frame is determined by majority-voting: a frame is labelled as protective if at least half of the samples within it were coded as protective, and vice versa.

The results achieved by the stacked-LSTM network are reported in Table IX. We can see from the result that all the frames in the validation set are detected as non-protective. This can be due to the fact that the protective and non-protective samples included in the training set are very imbalanced, while the baseline method does not apply any technique to solve it. On the other hand, the size of the training data is still limited. The result on the test set is slightly better with some frames correctly detected as protective (F1 score of protective class=0.2465). This proved the feasibility of using deep learning for the detection of protective behavior. Except for processing the MoCap and sEMG in a traversal way that ignored the biomechanical connectivity, challenges remain on i) how to deal with the imbalance problem in the data set; ii) how to design better data augmentation approaches.

V. CONCLUSION

In this paper, we introduced the first EmoPain 2020 Challenge on automatic pain recognition from multimodal face and body expressions based on the EMOPAIN dataset and guidelines for participation in the competition. It featured three tasks: (i) pain estimation from face shape and appearance features, (ii) pain recognition from muscle activity and joint angle statistical features, and (iii) classification of protective body movement behaviour. For each task, we described the expressive behavioural features extracted, the baseline system implementations and performance on the benchmark dataset. In this challenge, participants only received the extracted expression features rather than the video data, thus the baseline implementations do not employ feature optimisation or augmentation methods to allow for reproducibility of the results. Lastly, the baseline program code, results and participant rankings can be found on the EmoPain2020 Challenge's webpage: *https://mvrjustid.github.io/EmoPainChallenge2020/*.

VI. ACKNOWLEDGMENTS

This work was funded by the EPSRC grant Emotion & Pain Project EP/H017178/1 and the NIHR Nottingham Biomedical Research Centre.

REFERENCES

[1] A. Fayaz, P. Croft, R. Langford, L. Donaldson, and G. Jones, "Prevalence of chronic pain in the uk: a systematic review and meta-analysis of population studies," *BMJ open*, vol. 6, no. 6, p. e010364, 2016.

[2] J. Dahlhamer, J. Lucas, C. Zelaya, R. Nahin, S. Mackey, L. DeBar, R. Kerns, M. Von Korff, L. Porter, and C. Helmick, "Prevalence of chronic pain and high-impact chronic pain among adults—united states, 2016," *Morbidity and Mortality Weekly Report*, vol. 67, no. 36, p. 1001, 2018.

[3] P. Werner, D. Lopez-Martinez, S. Walter, A. Al-Hamadi, S. Gruss, and R. Picard, "Automatic recognition methods supporting pain assessment: A survey," *IEEE Transactions on Affective Computing*, 2019.

[4] M. J. L. Sullivan, P. Thibault, A. Savard, R. Catchlove, J. Kozey, and W. D. Stanish, "The influence of communication goals and physical demands on different dimensions of pain behavior," *Pain*, vol. 125, no. 3, pp. 270–277, 2006.

[5] H. Aviezer, Y. Trope, and A. Todorov, "Body cues, not facial expressions, discriminate between intense positive and negative emotions," *Science*, vol. 338, no. 6111, pp. 1225–1229, 2012.

[6] P. Watson, C. Booker, and C. Main, "Evidence for the role of psychological factors in abnormal paraspinal activity in patients with chronic low back pain," *J Musculoskelet Pain*, vol. 5, no. 4, pp. 41–56, 1997.

[7] M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh *et al.*, "The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset," *IEEE transactions on affective computing*, vol. 7, no. 4, pp. 435–451, 2015.

[8] S. Walter, S. Gruss, H. Ehleiter, J. Tan, H. Traue, P. Werner, A. Al-Hamadi, S. Crawcour, A. Andrade, and G. da Silva, "The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system," in *International Conference on Cybernetics*, 2013, pp. 128–131.

[9] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, S. Chew, and I. Matthews, "Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database," *Image and Vision Computing*, vol. 30, no. 3, pp. 197–205, 2012.

[10] S. Brahnam, C.-F. Chuang, F. Shih, and M. Slack, "Machine recognition and representation of neonatal facial displays of acute pain," *Artif Intell Med*, vol. 36, no. 3, pp. 211–222, 2006.

[11] J. Kappesser and A. C. de C Williams, "Pain and negative emotions in the face: judgements by health care professionals," *Pain*, vol. 99, no. 1-2, pp. 197–206, 2002.

[12] D. Simon, K. D. Craig, F. Gosselin, P. Belin, and P. Rainville, "Recognition and discrimination of prototypical dynamic expressions of pain and emotions," *PAIN®*, vol. 135, no. 1-2, pp. 55–64, 2008.

[13] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon, "The painful face–pain expression recognition using active appearance models," *Image and vision computing*, vol. 27, no. 12, pp. 1788–1796, 2009.

[14] S. Brahnam, C.-F. Chuang, R. S. Sexton, and F. Y. Shih, "Machine assessment of neonatal facial expressions of acute pain," *Decision Support Systems*, vol. 43, no. 4, pp. 1242–1254, 2007.

[15] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1797–1803, 2009.

[16] H. Breivik, P. Borchgrevink, S. Allen, L. Rosseland, L. Romundstad, E. Breivik Hals, G. Kvarstein, and A. Stubhaug, "Assessment of pain," *BJA: British Journal of Anaesthesia*, vol. 101, no. 1, pp. 17–24, 2008.

[17] Z. Hammal and J. F. Cohn, "Automatic detection of pain intensity," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 47–52.

[18] Z. Zafar and N. A. Khan, "Pain intensity evaluation through facial action units," in *2014 22nd International Conference on Pattern Recognition*. IEEE, 2014, pp. 4696–4701.

[19] J. O. Egede and M. Valstar, "Cumulative attributes for pain intensity estimation," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 146–153.

[20] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous Pain Intensity Estimation from Facial Expressions," in *Advances in Visual Computing*, 2012, pp. 368–377.

[21] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp. 267–274, 2008.

[22] P. Werner, A. Al-Hamadi, R. Niese, S. Walter, S. Gruss, and H. Traue, "Automatic pain recognition from video and biomedical signals," in *Int Conf Patt Recog*, 2014, pp. 4582–4587.

[23] S. Walter, S. Gruss, H. Traue, P. Werner, A. Al-Hamadi, M. Kächele, F. Schwenker, A. Andrade, and G. Moreira, "Data fusion for automated pain recognition," in *International Conference on Pervasive Computing Technologies for Healthcare*, 2015, pp. 261–264.

[24] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent convolutional neural network regression for continuous pain intensity estimation in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 84–92.

[25] M. Bellantonio, M. A. Haque, P. Rodriguez, K. Nasrollahi, T. Telve, S. Escalera, J. Gonzalez, T. B. Moeslund, P. Rasti, and G. Anbarjafari, "Spatio-temporal pain recognition in cnn-based super-resolved facial images," in *Video Analytics. Face and Facial Expression Recognition and Audience Measurement*. Springer, 2016, pp. 151–162.

[26] J. Egede, M. Valstar, M. T. Torres, and D. Sharkey, "Automatic neonatal pain estimation: An acute pain in neonates database," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 1–7.

[27] J. Egede, M. Valstar, and B. Martinez, "Fusing deep learned and hand-crafted features of appearance, shape, and dynamics for automatic pain estimation," in *Int Conf Automat Face & Gesture Recog*, 2017, pp. 689–696.

[28] D. Ahern, M. Follick, J. Council, N. Laser-Wolston, and H. Litchman, "Comparison of lumbar paravertebral EMG patterns in chronic low back pain patients and non-patient controls," *Pain*, vol. 34, no. 2, pp. 153–160, 1988.

[29] G. Gioftsos and D. Grieve, "The use of artificial neural networks to identify patients with chronic low-back pain conditions from patterns of sit-to-stand manoeuvres," *Clin Biomech*, vol. 11, no. 5, pp. 275–280, 1996.

[30] H. Grip, F. Ohberg, U. Wiklund, Y. Sterner, J. Karlsson, and B. Gerdle, "Classification of Neck Movement Patterns Related to Whiplash-Associated Disorders Using Neural Networks," *IEEE T Inf Technol B*, vol. 7, no. 4, pp. 412–418, 2003.

[31] D. Lai, P. Levinger, R. K. Begg, W. Gilleard, and M. Palaniswami, "Automatic Recognition of Gait Patterns Exhibiting Patellofemoral Pain Syndrome Using a Support Vector Machine Approach," *IEEE T Inf Technol B*, vol. 13, pp. 810–817, 2009.

[32] J. Bishop, M. Szpalski, S. Ananthraman, D. McIntyre, and M. Pope, "Classification of low back pain from dynamic motion characteristics using an artificial neural network," *Spine*, vol. 22, no. 24, pp. 2991–2998, 1997.

[33] J. Rivas, F. Orihuela-espina, L. Sucar, L. Palafox, J. Hernández-franco, and N. Bianchi-berthouze, "Detecting affective states in virtual rehabilitation," in *PervasiveHealth*, 2015.

[34] J. Dickey, M. Pierrynowski, D. Bednar, and S. Yang, "Relationship between pain and vertebral motion in chronic low-back pain subjects," *Clin Biomech*, vol. 17, no. 5, pp. 345–352, 2002.

[35] M. Kachele, P. Thiam, M. Amirian, F. Schwenker, and G. Palm, "Methods for Person-Centered Continuous Pain Intensity Assessment from Bio-Physiological Channels," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 9, p. 1, 2016.

[36] P. Werner, A. Al-hamadi, K. Limbrecht-Ecklundt, S. Walter, S. Gruss, and H. C. Traue, "Automatic Pain Recognition with Facial Activity Descriptors," *IEEE Transaction on Affective Computing*, 2016.

[37] V. Legrain, S. V. Damme, C. Eccleston, K. Davis, D. Seminowicz, and G. Crombez, "A neurocognitive model of attention to pain: Behavioral and neuroimaging evidence," *Pain*, vol. 144, no. 3, pp. 230–232, 2009.

[38] J. J. Rivas, F. Orihuela-Espina, L. Palafox, N. Berthouze, M. d. C. Lara, J. Hernández-Franco, and E. Sucar, "Unobtrusive inference of affective states in virtual rehabilitation from upper limb motions: A feasibility study," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[39] J. Rivas, S. S. L. E. Orihuela-Espina, F., A. Williams, and N. Berthouze, "Automatic recognition of multiple affective states in virtual rehabilitation by exploiting the dependency relationships," in *2018 13th IEEE International Conference on Affective COmputing and Intelligent Interaction (ACII 2019)*. IEEE, 2019, pp. 715–722.

[40] T. Olugbade, A. Singh, N. Bianchi-Berthouze, N. Marquardt, M. Aung, and A. Williams, "How Can Affect Be Detected and Represented in Technological Support for Physical Rehabilitation?" *Transactions on Computer-Human Interaction*, 2019.

[41] F. Keefe and A. Block, "Development of an observation method for assessing pain behavior in chronic low back pain patients," *Behav Ther*, pp. 363–375, 1982.

[42] J. W. S. Vlaeyen and S. J. Linton, "Fear-avoidance and its consequences in chronic musculoskeletal pain: A state of the art," *Pain*, vol. 85, no. 3, pp. 317–332, 2000.

[43] T. A. Olugbade, N. Bianchi-Berthouze, and A. Williams, "The relationship between guarding, pain, and emotion," *PAIN Report*, 2019.

[44] T. Olugbade, N. Berthouze, N. Marquardt, and A. Williams, "Human observer and automatic assessment of movement related self-efficacy in chronic pain: from exercise to functional activity," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[45] M. Aung, N. Bianchi-Berthouze, P. Watson, and A. Williams, "Automatic recognition of fear avoidance behavior in chronic pain physical rehabilitation," *International Conference on Pervasive Computing Technologies for Healthcare*, pp. 158–161, 2014.

[46] C. Wang, T. A. Olugbade, A. Mathur, A. C. De C Williams, N. D. Lane, and N. Bianchi-Berthouze, "Recurrent network based automatic detection of chronic pain protective behavior using mocap and semg data," in *Proceedings of the 23rd International Symposium on Wearable Computers (ISWC)*. ACM, 2019, pp. 225–230.

[47] C. Wang, M. Peng, T. Olugbade, N. Lane, A. Williams, and N. Bianchi-Berthouze, "Learning temporal and bodily attention in protective movement behavior detection," *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 324–330, 2019.

[48] I. Lawrence and K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, pp. 255–268, 1989.

[49] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochimica et Biophysica Acta*, vol. 405, no. 2, pp. 442–451, 1975.

[50] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *International Conference on Automatic Face & Gesture Recognition*. IEEE, 2018, pp. 59–66.

[51] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner *et al.*, "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[52] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint 1409.1556*, 2014.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[54] D. Kollias, P. Tzirakis, M. Nicolaou, A. Papaioannou, G. Zhao, B. Schuller, I. Kotsia, and S. Zafeiriou, "Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond," *Int J Comput Vis*, vol. 127, no. 6-7, pp. 907–929, 2019.

[55] S. Jaiswal, S. Song, and M. Valstar, "Automatic prediction of depression and anxiety from behaviour and personality attributes," in *Int Conf Affect Comput Intell Interact*, 2019, pp. 1–7.

[56] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J Mach Learn Res*, vol. 15, no. 1, pp. 1929–1958, 2014.

[57] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[58] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.