Exploring Multidimensional Measurements for Pain Evaluation using Facial Action Units

Xiaojing Xu[1] and Virginia R. de Sa[2]
[1] Department of Electrical and Computer Engineering, UC San Diego, CA, USA
[2] Department of Cognitive Science and Halıcıoğlu Data Science Institute, UC San Diego, CA, USA

Abstract— **Although pain is widely recognized to be a multidimensional experience, it is typically measured by unidimensional patient self-reported visual analog scale (VAS). However, self-reported pain is subjective, difficult to interpret and sometimes impossible to obtain. Machine learning models have been developed to automatically recognize pain at both the frame level and sequence (or video) level. Many methods use or learn facial action units (AUs) defined by the Facial Action Coding System (FACS) for describing facial expressions with muscle movement. In this paper, we analyze the relationship between sequence-level multidimensional pain measurements and frame-level AUs and an AU derived pain-related measure, the Prkachin and Solomon Pain Intensity (PSPI). We study methods that learn sequence-level metrics from frame-level metrics. Specifically, we explore an extended multitask learning model to predict VAS from human-labeled AUs with the help of other sequence-level pain measurements during training. This model consists of two parts: a multitask learning neural network model to predict multidimensional pain scores, and an ensemble learning model to linearly combine the multidimensional pain scores to best approximate VAS. Starting from human-labeled AUs, the model achieves a mean absolute error (MAE) on VAS of 1.73. It outperforms provided human sequence-level estimates which have an MAE of 1.76. Combining our machine learning model with the human estimates gives the best performance of MAE on VAS of 1.48.**

I. Introduction

The current gold standard of estimating clinical pain is patient self-report given by visual analog scale (VAS), despite its known limitations [1], [2]. One of these limitations is that it is difficult to obtain in populations with verbal or neurological disabilities [2]. Automated pain recognition models have been developed to solve this problem using various nonverbal signals such as facial expressions, head/body movement and physiological signals [3], [4], [5], [6]. Research has shown that facial expressions can provide sensitive and reliable information about pain across the life span [7], [8], from infants [9] to elderly patients [10], [11].

Two types of pain metrics are usually considered in pain studies: frame-by-frame metrics and sequence-level metrics. One prominent example of frame-level metrics, are the muscle-based facial action units (AUs) defined by the Facial Action Coding System (FACS) [12]; they have been widely used as a consistent and reliable way to represent facial expressions including pain [13] expression. The names of some of the pain-related AUs can be found in Table I.

We gratefully acknowledge the support of NSF IIS 1817226, IBM Research AI, and NVIDIA Corporation for the donation of the Titan V GPU used for this research

Another frame-level metric, built on top of the AUs, is the Prkachin and Solomon Pain Intensity (PSPI) [14]. It defines a single number that measures pain as a combination of AU intensities:

$$\text{PSPI} = \text{AU4} + \max(\text{AU6,AU7}) + \max(\text{AU9,AU10}) + \text{AU43}$$

Most research on automatic pain detection from facial expression has focused on predicting frame-level PSPI scores. A widely used 2-step framework is to first extract low-dimensional relevant non-rigid geometric or appearance features from raw pixels and then learn a classification or regression model [15], [16], [17], [18]. Otherwise, deep learning can be used to learn from raw pixels directly [19], [20]. In addition to these "static approaches" that extract features from single frames, it is also useful to learn dynamic features when data is available in the form of video sequences [21], [22]. Multiple-instance learning has been used to learn frame-level scores using sequence-level labels in a weakly supervised manner [23], [24].

Automated detection of facial AUs has also been well studied, and PSPI ratings can be calculated directly from AU estimates. Many approaches of AU detection focused on finding regions of interest [25], [26], [27], [28]. Jaiswal et al. and Chu et al. combined CNN and LSTM, and Kumawat proposed a 3D convolutional layer called Local Binary Volume layer, to learn temporal information [29], [30], [31]. Baltrušaitis et al. studied the benefit of person-specific neutral expression normalisation and multiple datasets for generic model training, and presented a pipeline that detects AUs in real-time [32]. Tang et al. and Romero et al. fine-tuned VGG models pretrained on face datasets to detect AUs under different facial views [33], [34]

In contrast, to the automated work above, sequence-level pain metrics are more often used in clinics, and the understanding and interpretation of pain in the literature is mostly based on sequence level assessments, rated by observers or by self-report. The sequence-level self-rated VAS is still the most commonly used pain score in clinical settings. Only a few papers have addressed the problem of estimating VAS score in facial videos. Sikka et al. [35] and Xu et al. [36] detected pain in children after surgery using AUs extracted by iMotions (imotions.com). Liu et al., Martinez et al., and Xu et al. used a two-stage method to first train a model to predict pain scores at the frame level, and then predicted video VAS score using these frame-level predictions [37], [38], [39] although only [39] started from raw pixels.

Although sequence-level metrics are considered to have more clinical relevance, frame-level pain recognition has been studied more thoroughly and there exist software packages and toolkits such as iMotions (imotions.com) and OpenFace [40] to automatically detect AUs. There are many reasons for this. First, it is difficult to obtain a large number of sequence-level samples. A pain dataset with each video lasting less than 1 minute can have three orders of magnitude more frames than videos. Second, machine learning models on videos require significantly more space and time to train. This problem is not unique to pain; there are many deep neural networks trained on facial images, but there is no publically available model trained on facial videos, so it is hard to leverage prior work when working with videos. Currently most sequence-level models use frame-level models as building blocks [35], [37], [38], [36], [39], and the problem of learning sequence-level metrics is usually broken down into two parts: learning frame-level metrics and learning sequence-level metrics based on the frame-level metrics. Since there has been a lot of research addressing the first part (learning frame-level metrics), in this work, we focus on whether and how well we can solve the second part. In order to not be dependent on the quality of model solving the first part, we study the second problem for human coded frame-level AUs and PSPIs (which are usually used as ground truth in AU and PSPI estimation models). We do this through a two-stage model similar to the last two stages in the extended multitask learning model which is the current state-of-the-art for estimating VAS [39]. In the first stage, we send statistics of AUs and PSPI over frames of each video as inputs to a neural network to get a sequence-level VAS prediction, and use multitask learning to improve the VAS prediction while obtaining multidimensional pain scales. Then, as in [39], we extend the multitask learning framework by finding an optimal linear combination of these pain scales to further improve VAS prediction. We show on the UNBC-McMaster Shoulder Pain dataset [16] that this method outperforms human video-level labels, and can be further improved when combined with those human ratings.

The contributions of this paper are as follows:

- We analyze the relationship between multidimensional pain measurements and their predictions from a machine learning model
- We study the relationship between sequence-level and frame-level pain metrics, and build an extended multitask learning model to estimate sequence-level pain scores using human-coded frame-level features
- We explore ways of utilizing human-coded AUs and multidimensional pain ratings to improve VAS prediction, and study the contribution of each component of the multitask-ensemble multidimensional-pain model
- Our model serves as a baseline of how well one can predict VAS using human-coded AUs
- Our model can be combined with automated AU/PSPI detection systems to achieve end-to-end VAS prediction and provides an upper-bound on expected performance.

II. Method

This paper studies the widely used UNBC-McMaster Shoulder Pain dataset [16]. It contains videos of patient faces (who were suffering from shoulder pain) while they were performing a series of active and passive range-of-motion tests to their affected and unaffected limbs on two separate occasions. The dataset includes 25 subjects, 200 videos and 48,398 frames.

TABLE I: AU Description

AU4	brow lowering	AU12	oblique lip raising
AU6	cheek raising	AU20	horizontal lip stretch
AU7	eyelid tightening	AU25	lips parting
AU9	Nose wrinkling	AU26	jaw dropping
AU10	upper lip raising	AU43	eye closure

The dataset provides 11 facial action unit (AU) intensities coded each frame by certified FACS coders, and 1 PSPI score calculated from the AUs. AUs are defined by FACS (Facial Action Coding System) [12] to code movements of individual facial muscles. In this work, we work with the 9 AUs (AU4, 6, 7, 10, 12, 20, 25, 26 and 43) present in more than 500 frames.

In addition to the frame-level features, the dataset also provides 4 sequence-level labels: VAS (Visual Analog Scale) 0-10, OPR (Observers Pain Rating) 0-5, AFF (Affective-motivational scale) 0-15 and SEN (Sensory Scale) 0-15. OPR is the human observers' rating of pain level of the video. The other three measures are provided by the patients themselves. The sensory scale consists of a numeric scaling associated with the following words of increasing SEN scale: extremely weak, faint, very weak, weak, very mild, mild, slightly moderate, moderate, barely strong, clear cut, slightly intense, strong, intense, very intense, extremely intense. The affective-motivational (AFF) scale uses the following affect-based words: slightly unpleasant, slightly annoying, annoying, unpleasant, slightly distressing, slightly miserable, very annoying, distressing, very unpleasant, miserable, very distressing, slightly intolerable, very miserable, intolerable, very intolerable [41], [42].

With the features and labels described above, our goal is to train a model that predicts VAS using AU and PSPI intensities. Our model structure and hyper-parameters follow that of stage 2 and 3 of the model proposed in [39].

A. VAS Estimation in Facial Videos using AU Sequences

For each video, we form a 10-D feature vector by taking the maximum rating over all frames for each of the 9 AUs and 1 PSPI to form a 10 dimensional feature vector of the video that is input to a fully connected neural network with one 20 unit hidden layer to predict VAS in a linear output layer using batch-weighted MSE loss [43]. We used the Adam optimizer, initial learning rate of 1e-2, batch size of 32, max number of epochs of 200, and used early stopping when the validation loss hadn't decreased for 20 epochs.

B. Multitask Learning

As mentioned in [39], the three other sequence level pain ratings are very related to the VAS pain score which motivates a multitask learning (MTL) approach [44] that leverages "the domain-specific information contained in the training signals of related tasks" [44]. OPR may be especially useful as it should be fully constrained by information in the video (unlike VAS that may reflect strong pain but masked facial expression). The multitask architecture is straight forward. We use 4 scores instead of a single VAS as outputs of the neural network. The labels are normalized into the same range so that all elements contribute equally to the loss during training. The losses are weighted based on the distribution of VAS scores, and the validation loss is the mean MSE of the 4 outputs.

C. Ensemble Learning of Multidimensional Pain Scores

Each of the four sequence-level scores (VAS, OPR, AFF, and SEN) reports on different aspects of pain. VAS reflects the patient's overall rating of their perceived pain. AFF and SEN are designed to try to separate out affective vs sensory aspects of pain and are also reported by the patient. OPR, on the other hand, is scored by an external observer and is only based on the facial video so may be a more predictable function of the video for training a machine learning system. If humans are considered the gold standard at facial pain recognition, then OPR could be considered an approximate upper bound for a machine-learning facial video system.

OPR, AFF, and SEN are all highly correlated with VAS (see Figure 1 LEFT) and can be considered as predictions of VAS. After scaling their outputs to the same range as VAS, they all do a reasonable job at estimating VAS and can be considered as four different "experts" (Fig. 1 RIGHT). Ensemble averaging can be used to compute the optimal linear combination of experts to reduce variance of the estimator [45].

As in [39], the final prediction of VAS is learned as a weighted sum of the four experts. If each expert outputs f_i, then the overall model $\tilde{f}$ is defined as:

$$\tilde{f} = \sum_{i=1}^{4} \alpha_i f_i$$

We solve the optimization problem minimizing MSE of the final prediction $\tilde{f}$ subject to $\sum_{i=1}^{4} \alpha_i = 1$ [46], [47], [45], [39]. The optimal $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]^T$ is:

$$\boldsymbol{\alpha} = \frac{\Omega^{-1}\mathbf{1}}{\mathbf{1}^T\Omega^{-1}\mathbf{1}}$$

where $\Omega = [\omega_{ij}] = [E[(f_i - VAS)(f_j - VAS)]]$ and VAS gives the true VAS labels. The ensemble weights an expert more if it is more accurate in estimating VAS.

III. Experimental Analysis

On the UNBC-McMaster dataset, we performed 5-fold cross validation with each fold consisting of 5 subjects. To prevent overfitting, we used the same training/test splits for the two stages in each iteration. One of the 4 training folds is randomly selected as the validation set for neural network training. After 5 iterations we evaluate the models using Mean Absolute Error (MAE), Mean Squared Error (MSE), Intraclass Correlation Coefficient (ICC) and Pearson Correlation Coefficient (PCC) on all test data. ICC is useful when MAE scores are deceptively low. For example, for the current dataset, if the model outputs the average VAS for all samples, the MAE will be 2.44, but the ICC will be approximately zero. So we want low MAE with high ICC.

For all models in this paper, we performed the above 5-fold cross validation 5 times, and report mean and standard deviation over 5 experiments. All experiments were run on a single NVIDIA Titan V GPU.

A. Relationship between Sequence-level Metrics in the Data

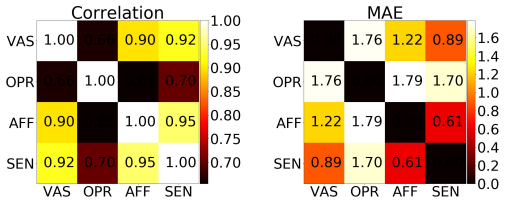

Fig. 1: Correlation (left) and MAE (right) between each pair of the 4 sequence-level true scores. The scores have been scaled to the same range 0-10.

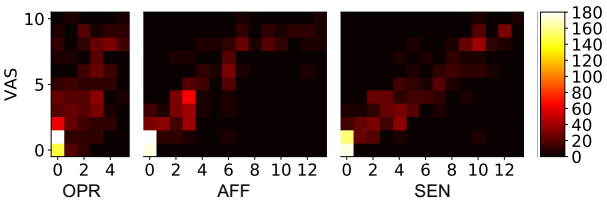

Fig. 2: 2D histogram of sequence-level score pairs.

The relationship between the 4 sequence-level scores in the UNBC-McMaster dataset is shown in Fig. 1. We can see from the heatmap on the left that VAS, AFF and SEN are highly correlated, and OPR is also correlated with these 3 self-rated scores but not as much. The right side of the figure shows how well (in terms of MAE) each of the multimodal pain measures predicts the others (after appropriate rescaling). For example OPR (human ratings) predicts VAS with an MAE of 1.76.

Figure 2 shows the joint distributions of VAS with OPR, AFF and SEN plotted as 2D histograms. It can be seen that although VAS is linearly correlated with the three other scores, they are not strictly proportional.

B. Relating Sequence- and Frame-level Metrics in the Data

Fig. 3, shows the correlation between the frame-level and sequence-level pain scores. We see again the high correlations between the sequence-level measures and some correlation between the frame-level measures. Of the sequence measures, OPR generally has a higher correlation with

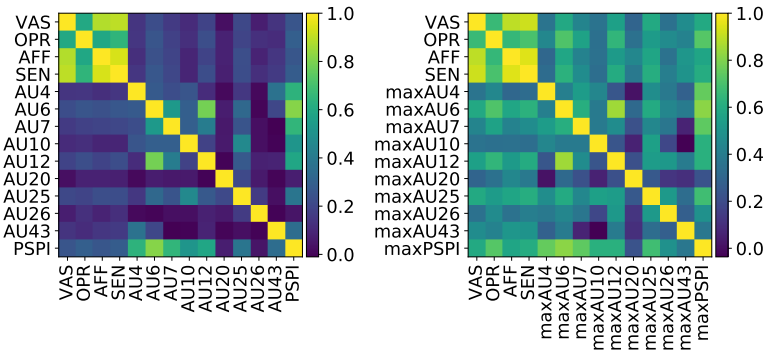

Fig. 3: The correlation between 4 sequence-level scores (VAS, OPR, AFF, SEN) and 10 frame-level scores (9 AUs and PSPI) in the data. On the left is the correlation at the frame level, where the VAS for a frame is the VAS of the video it belongs to. On the right is the correlation at the sequence level, where the maximum AU/PSPI for a video is taken.

the AUs and PSPI. This shows the potential of predicting sequence-level pain ratings from frame-level measurements.

C. Multidimensional Pain Prediction using Neural Networks

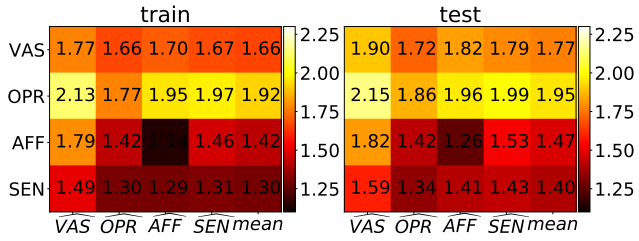

Fig. 4: Average MAE on training and test data. The y axis gives the true label, and x axis the predictions. Each entry is the mean absolute difference between the two variables. All the labels and predictions have been mapped to the range 0-10 before calculation, but MAEs in different rows are not strictly comparable because OPR only takes 6 values while AFF and SEN can take 16.

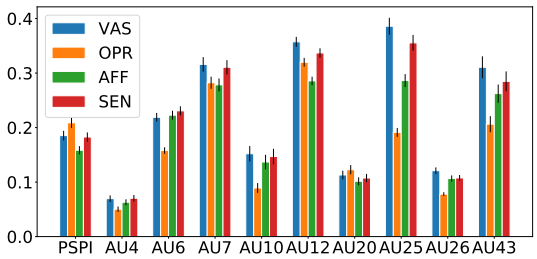

Fig. 5: Contributions of each of the AUs to the neural network outputs that use max of AUs and PSPI as input. The heights of the bars represent feature importance measured as the mean absolute shap values. Error bars show the standard deviation of the mean absolute shap values.

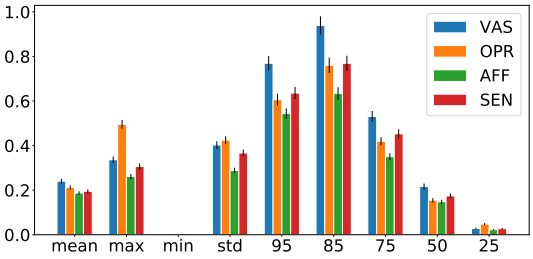

Fig. 6: Contributions of each of the 9 PSPI statistics to the neural network outputs that use 9 PSPI stats as input. The heights of the bars represent feature importance measured as the mean absolute shap values. Error bars show the standard deviation of the mean absolute shap values.

Fig. 4, presents, as heatmaps, the MAEs of the multitask neural network with 4 outputs (prior to ensembling) corresponding to the 4 sequence-level pain ratings. For example, diagonal elements show the MAEs of each output predicting the corresponding metric, and the second element in the first row shows the MAE of using the OPR output to predict VAS. Interestingly, the best MAE in predicting a metric is not always given by its corresponding output, e.g. the OPR output predicts VAS better than the VAS output, and the OPR output works better in SEN prediction than the SEN output. Actually, the OPR output works well when used to estimate all the metrics despite being trained to only estimate OPR, the metric with the lowest correlation with the other pain scores. This may be because OPR is more consistent across subjects and is based purely on video features. As a result, OPR may be learned more easily from facial features such as AUs and PSPI, and serve as a better pain metric when tested across subjects.

AU Importance. We use the shap framework [48] to calculate the contribution of each of the AUs to the four output scores, and plot the importance values in Fig. 5. The bar graph shows, for example, that AU7, 12, 25 and 43 are very useful in pain prediction, except that AU25 is much less important when predicting OPR than predicting the self-ratings. OPR uses PSPI more while not using as much the

individual AUs compared to the self-report measures of VAS, AFF and SEN. Interestingly, while AU4 is considered to be among the "core expressions of pain" and contributes to PSPI score [49], [14], [50], it is not a very important feature in this model on this dataset.

There is a fair amount of consistency between Fig. 5 and Fig. 3. For example, PSPI has higher correlation with OPR than the 3 self-rated scores, and also higher importance for predicting OPR. AU25 and 43 are less important for OPR and also less correlated with OPR than the other 3 pain scores.

Benefit of Multitask Learning. We explore the benefit of multitask learning in the neural network in Table II row 1-2. The first row shows the VAS prediction performance without multitask learning, i.e. when the neural network only has one output predicting VAS. The second row corresponds to the multitask learning model, where the performance is evaluated only with the output trained to predict VAS. Learning the three other scores from a shared hidden layer, together with VAS helps the model's VAS output to better predict VAS.

Different Input Features. When extracting sequence-level features for a video from a sequence of frame-level features, we simply take the maximum of the AU/PSPI sequence as in [15], but it is also common to use other statistics such as standard deviation, minimum, mean, etc. [35], [37], [36], [39]. To explore how different choices of input features work, we extracted 9 statistics (mean, max, min, standard deviation, 95th, 85th, 75th, 50th, 25th percentiles) from the PSPI and AU sequences to form a length-90 (9 stats $\times$ (9 AU + 1 PSPI)) feature vector. The performance using 90 features is not as good as using 10 maxima (row 4-6 compared to row 1-3 in Table II). The reason may be that 90 dimensional inputs is too large for our model. To address this, we also tried using 9 statistics of PSPI only following [37], [39] since PSPI is defined to represent pain and contains the most comprehensive information about pain expressions. The results are shown in the last three rows in Table II. Using 9 statistics of PSPI works fine, but still not as good as using 10 maxima of PSPI and AUs. The shap importance values for this model are plotted in Fig. 6. Min and 25 percentile are two inputs that are not very useful for this model.

D. Optimal Linear Combination of Multidimensional Pain

While multitask learning results in improved training of VAS prediction through joint learning of all 4 measures, ensembling the 4 predicted outputs discussed in Section II-C results in significantly ($p < 0.0001$) better performance as shown in row 3 in Table II.

E. Contributions of Different Components: Multitask Learning, Ensemble Learning and Multidimensional Pain

In this section, we perform ablation studies to explore the relative contributions of different components of the extended multitask learning model.

In order to see whether multitask learning helps, we trained models with separate hidden layer for each of the four sequence-level outputs i.e. with the same inputs and outputs but without multitask learning/hidden layer sharing. The performance ("4 scores") is not as good as using multitask learning ("4 scores MTL") (see Table II and Fig. 7).

In order to compare the importance of ensemble learning to that of multi-task learning, we trained a model with the same structure as our best model, i.e. with 4 neural network outputs and ensemble learning on top of them, but instead of using 4 different pain scores as labels for NN outputs, we trained each of the 4 outputs with identical VAS labels (but different initial conditions). This allows the model to start from 4 different initial states and explore different areas of the weight space with different final predictions. The ensemble model will then find the best way to linearly combine these predictions to obtain a new random variable as the prediction of VAS. The results show that this simple ensemble model also performs better than a single network predicting only VAS but slightly worse than the best model predicting 4 different pain scores, as plotted in Fig. 7 "VAS $\times 4$ MTL". From these results we conclude that ensembling is most helpful for the excellent performance, but that using multidimensional pain scores is also helpful.

We also trained a version of the network with 4 VAS outputs where each output had its own (unshared) hidden layer. ("VAS $\times 4$" in Fig. 7). This model performed slightly worse. This is likely because the multitask learning model has less parameters and so learns faster with less overfitting.

Lastly, since ensemble learning contributes significantly to the performance, we considered a model with extra copies of outputs to provide more "expert" predictions to ensemble. We considered 4 copies of the 4 different sequence-level scores, and separately, 16 copies of VAS, to make 16 output NNs. This didn't further improve the performance.

To summarize, with the same inputs, the model with ensemble learning on multidimensional pain predictions yields the best performance. This corresponds to the third row in Table II for each input type, as well as the first (blue) bar in Fig. 7 in each group.

F. Comparison with Humans and Other Work

We compare our model with humans in Table III. The human ratings are given by the OPR scores in the dataset. Our extended multitask learning model using AU features and multidimensional pain outputs beats the MAE of those humans. Moreover, when averaging our prediction with the human predictions, the performance can be further improved. This implies that learning pain as a function of individual AUs may be a more accurate and systematic way than learning pain from the whole face.

We also compare our model using true AUs with [39] that has a model with similar structure but uses AUs predicted automatically from the output of a deep convolutional network. Our results significantly outperform [39] demonstrating the potential of an end-to-end VAS prediction model if the AU prediction stage is improved.

IV. Discussion and Conclusion

We explored a model that predicts VAS using facial actions units, and beats human observers on the UNBC-McMaster Shoulder Pain dataset. When a human observer is

TABLE II: Sequence-level VAS Prediction using Frame-level Labels

NN Input	NN Output	Ensemble Learning	MAE	MSE	ICC	PCC
PSPI+AU max	VAS	-	1.94 ± 0.05	5.25 ± 0.18	0.57 ± 0.02	0.64 ± 0.02
PSPI+AU max	4 scores MTL	-	1.90 ± 0.04	4.98 ± 0.10	0.59 ± 0.01	0.67 ± 0.01
PSPI+AU max	4 scores MTL	Ensemble	$\mathbf{1.73 \pm 0.03}$	$\mathbf{4.61 \pm 0.19}$	$\mathbf{0.61 \pm 0.02}$	$\mathbf{0.67 \pm 0.02}$
PSPI+AU stats	VAS	-	2.02 ± 0.05	5.83 ± 0.14	0.51 ± 0.04	0.58 ± 0.02
PSPI+AU stats	4 scores MTL	-	1.94 ± 0.05	5.39 ± 0.22	0.56 ± 0.02	0.61 ± 0.02
PSPI+AU stats	4 scores MTL	Ensemble	1.81 ± 0.04	5.04 ± 0.17	0.58 ± 0.01	0.63 ± 0.01
PSPI stats	VAS	-	2.07 ± 0.05	5.81 ± 0.23	0.52 ± 0.04	0.63 ± 0.03
PSPI stats	4 scores MTL	-	2.03 ± 0.05	5.58 ± 0.23	0.53 ± 0.03	0.65 ± 0.02
PSPI stats	4 scores MTL	Ensemble	1.76 ± 0.03	4.81 ± 0.18	0.59 ± 0.02	0.65 ± 0.02

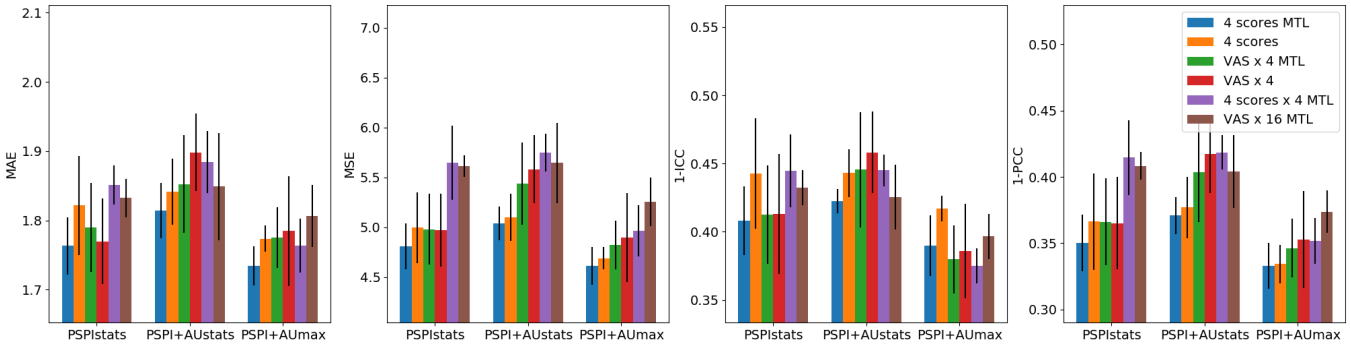

Fig. 7: Bar graphs showing the MAE, MSE, 1-ICC, 1-PCC (we plot 1-ICC and 1-PCC instead of ICC and PCC so that for all sub-figures shorter bars mean better performance) of the following models predicting VAS using 3 different combinations (PSPI, PSPI+AU, PSPI+AU max) of frame-level labels: (1) 4 scores MTL. Predicting 4 scores using multitask learning. (2) 4 scores. Predicting 4 scores using 4 separate models. (3) VAS $\times$ 4 MTL. Predicting 4 VAS using multitask learning. (4) VAS $\times$ 4 predicting 4 VAS using 4 separate models. (5) 4 scores $\times$ 4 MTL. Predicting 4 copies of 4 scores using multitask learning. (6) VAS $\times$ 16 MTL. Predicting 16 copies of VAS using multitask learning.

TABLE III: Comparison with Humans and Other Work

	MAE	MSE	ICC	PCC
EMTL with true AU (this paper)	1.73 ± 0.03	4.61 ± 0.19	0.61 ± 0.02	0.67 ± 0.02
EMTL from pixels [39]	1.95 ± 0.06	5.90 ± 0.23	0.43 ± 0.03	0.55 ± 0.03
Human (OPR)	1.76	6.26	0.66	0.66
Average of EMTL (with true AU) and Human	1.48 ± 0.02	4.22 ± 0.10	0.70 ± 0.01	0.71 ± 0.01

available, the performance can be largely improved simply by averaging our prediction and the human prediction. While the human observer in the UMBC-McMaster dataset is not necessarily the same human that labeled the AUs, it would be interesting to explore whether this method of using human-labeled AUs can beat the same observer at VAS prediction.

We studied ablations of the Extended Multitask Learning Model. The approaches using multitask learning, multidimensional pain measurement and ensemble learning can be used in similar healthcare datasets and tasks. Our model can be combined with existing frame-level pain estimation models such as AU or PSPI extractors to easily form a video-level metric prediction model. In this case, the performance shown in this paper provides an upper bound on the accuracy that can be achieved when using automatically estimated AUs instead of manually labeled AUs. It also provides a baseline for estimating sequence-level pain ratings such as VAS using widely-used frame-level pain related measurements such as AUs and PSPI.

REFERENCES

[1] Kenneth D Craig. The facial expression of pain better than a thousand words? *APS Journal*, 1(3):153–162, 1992.

[2] Ghada Zamzmi, Chih-Yun Pai, Dmitry Goldgof, Rangachar Kasturi, Yu Sun, and Terri Ashmeade. Machine-based multimodal pain assessment tool for infants: a review. *preprint arXiv:1607.00331*, 2016.

[3] Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O Andrade, and Gustavo Moreira da Silva. The biovid heat pain database data for the advancement and systematic validation of an automated pain recognition system. In *2013 IEEE international conference on cybernetics (CYBCO)*, pages 128–131. IEEE, 2013.

[4] Temitayo A Olugbade, Nadia Bianchi-Berthouze, Nicolai Marquardt, and Amanda C Williams. Pain level recognition using kinematics and muscle activity for physical rehabilitation in chronic pain. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 243–249. IEEE, 2015.

[5] Eun-Hye Jang, Byoung-Jun Park, Mi-Sook Park, Sang-Hyeob Kim, and Jin-Hun Sohn. Analysis of physiological signals for recognition of boredom, pain, and surprise emotions. *Journal of physiological anthropology*, 34(1):25, 2015.

[6] Evan David Campbell, Angkoon Phinyomark, and Erik Justin Scheme. Feature extraction and selection for pain recognition using peripheral physiological signals. *Frontiers in neuroscience*, 13:437, 2019.

[7] Jean-Francois Payen, Olivier Bru, Jean-Luc Bosson, Anna Lagrasta, Eric Novel, Isabelle Deschaux, Pierre Lavagne, and Claude Jacquot.

Assessing pain in critically ill sedated patients by using a behavioral pain scale. *Critical care medicine*, 29(12):2258–2263, 2001.

[8] Amanda C de C Williams. Facial expression of pain: an evolutionary account. *Behavioral and brain sciences*, 25(4):439–455, 2002.

[9] Ruth VE Grunau and Kenneth D Craig. Pain expression in neonates: facial action and cry. *Pain*, 28(3):395–410, 1987.

[10] Paolo L Manfredi, Brenda Breuer, Diane E Meier, and Leslie Libow. Pain assessment in elderly patients with severe dementia. *Journal of Pain and Symptom Management*, 25(1):48–52, 2003.

[11] Thomas Hadjistavropoulos, Keela Herr, Kenneth M Prkachin, Kenneth D Craig, Stephen J Gibson, Albert Lukas, and Jonathan H Smith. Pain assessment in elderly adults with dementia. *The Lancet Neurology*, 13(12):1216–1227, 2014.

[12] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[13] Zhanli Chen, Rashid Ansari, and Diana Wilkie. Automated pain detection from facial expressions using facs: A review. *arXiv preprint arXiv:1811.07988*, 2018.

[14] Kenneth M Prkachin and Patricia E Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008.

[15] Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F Cohn, Tsuhan Chen, Zara Ambadar, Kenneth M Prkachin, and Patricia E Solomon. The painful face–pain expression recognition using active appearance models. *Image and vision computing*, 27(12):1788–1796, 2009.

[16] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, and Iain Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Face and Gesture 2011*, pages 57–64. IEEE, 2011.

[17] Md Maruf Monwar and Siamak Rezaei. Pain recognition using artificial neural network. In *Signal Processing and Information Technology, 2006 IEEE International Symposium on*, pages 28–33. IEEE, 2006.

[18] Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Automatic pain intensity estimation with heteroscedastic conditional ordinal random fields. In *International Symposium on Visual Computing*, pages 234–243. Springer, 2013.

[19] Feng Wang, Xiang Xiang, Chang Liu, Trac D Tran, Austin Reiter, Gregory D Hager, Harry Quon, Jian Cheng, and Alan L Yuille. Regularizing face verification nets for pain intensity regression. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1087–1091. IEEE, 2017.

[20] Ghada Zamzmi, Dmitry Goldgof, Rangachar Kasturi, and Yu Sun. Neonatal pain expression recognition using transfer learning. *arXiv preprint arXiv:1807.01631*, 2018.

[21] Pau Rodriguez, Guillem Cucurull, Jordi Gonzàlez, Josep M Gonfaus, Kamal Nasrollahi, Thomas B Moeslund, and F Xavier Roca. Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE transactions on cybernetics*, 2017.

[22] Mohammad Tavakolian and Abdenour Hadid. Deep spatiotemporal representation of the face for automatic pain intensity estimation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 350–354. IEEE, 2018.

[23] Karan Sikka, Abhinav Dhall, and Marian Bartlett. Weakly supervised pain localization using multiple instance learning. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8. IEEE, 2013.

[24] Adria Ruiz, Ognjen Rudovic, Xavier Binefa, and Maja Pantic. Multi-instance dynamic ordinal random fields for weakly-supervised pain intensity estimation. In *Asian Conference on Computer Vision*, pages 171–186. Springer, 2016.

[25] SL Happy and Aurobinda Routray. Automatic facial expression recognition using features of salient facial patches. *IEEE transactions on Affective Computing*, 6(1):1–12, 2014.

[26] Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3391–3399, 2016.

[27] Wei Li, Farnaz Abtahi, and Zhigang Zhu. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1841–1850, 2017.

[28] Wei Li, Farnaz Abtahi, Zhigang Zhu, and Lijun Yin. Eac-net: Deep nets with enhancing and cropping for facial action unit detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(11):2583–2596, 2018.

[29] Shashank Jaiswal and Michel Valstar. Deep learning the dynamic appearance and shape of facial action units. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–8. IEEE, 2016.

[30] Wen-Sheng Chu, Fernando De la Torre, and Jeffrey F Cohn. Learning spatial and temporal cues for multi-label facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 25–32. IEEE, 2017.

[31] Sudhakar Kumawat, Manisha Verma, and Shanmuganathan Raman. Lbvcnn: Local binary volume convolutional neural network for facial expression recognition from image sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[32] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 6, pages 1–6. IEEE, 2015.

[33] Chuangao Tang, Wenming Zheng, Jingwei Yan, Qiang Li, Yang Li, Tong Zhang, and Zhen Cui. View-independent facial action unit detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 878–882. IEEE, 2017.

[34] Andrés Romero, Juan León, and Pablo Arbeláez. Multi-view dynamic facial action unit detection. *Image and Vision Computing*, 2018.

[35] Karan Sikka, Alex A Ahmed, Damaris Diaz, Matthew S Goodwin, Kenneth D Craig, Marian S Bartlett, and Jeannie S Huang. Automated assessment of children's postoperative pain using computer vision. *Pediatrics*, 136(1):e124–e131, 2015.

[36] Xiaojing Xu, Kenneth D. Craig, Damaris Diaz, Matthew S. Goodwin, Murat Akcakaya, Büşra Tuğçe Susam, Jeannie S. Huang, and Virginia R. de Sa. Automated pain detection in facial videos of children using human-assisted transfer learning. In *Artificial Intelligence in Health. AIH 2018. Lecture Notes in Computer Science*, volume 11326, pages 162–180. Springer International Publishing, Cham, 2019.

[37] Dianbo Liu, Fengjiao Peng, Andrew Shea, Rosalind Picard, et al. Deepfacelift: interpretable personalized models for automatic estimation of self-reported pain. *arXiv preprint arXiv:1708.04670*, 2017.

[38] Lopez Martinez, Daniel Rosalind Picard, et al. Personalized automatic estimation of self-reported pain intensity from facial expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 70–79, 2017.

[39] Xiaojing Xu, Jeannie S. Huang, and Virginia R. de Sa. Pain evaluation in video using extended multitask learning from multidimensional measurements. In *Machine Learning for Health ML4H at NeurIPS 2019*, Proceedings of Machine Learning Research. PMLR, 2019.

[40] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science, 2016.

[41] Richard H Gracely, Patricia McGrath, and Ronald Dubner. Ratio scales of sensory and affective verbal pain descriptors. *Pain*, 5(1):5–18, 1978.

[42] Marc W Heft, Richard H Gracely, Ronald Dubner, and Patricia A McGrath. A validation model for verbal descriptor scaling of human clinical pain. *Pain*, 9(3):363–373, 1980.

[43] Ali Sellami and Heasoo Hwang. A robust deep convolutional neural network with batch-weighted loss for heartbeat classification. *Expert Systems with Applications*, 122:75–84, 2019.

[44] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[45] Sherif Hashem. Optimal linear combinations of neural networks. *Neural networks*, 10(4):599–614, 1997.

[46] Robert T Clemen. Linear constraints and the efficiency of combined forecasts. *Journal of Forecasting*, 5(1):31–38, 1986.

[47] G Trenkler and EP Liski. Linear constraints and the efficiency of combined forecasts. *Journal of Forecasting*, 5(3):197–202, 1986.

[48] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[49] Kenneth M Prkachin. The consistency of facial expressions of pain: a comparison across modalities. *Pain*, 51(3):297–306, 1992.

[50] Kenneth M Prkachin. Assessing pain by facial expression: facial expression as nexus. *Pain Res Manag.*, 14(1):53–58, 2009.