Proceedings of Machine Learning Research 116:141–154, 2020 Machine Learning for Health (ML4H) at NeurIPS 2019

Pain Evaluation in Video using Extended Multitask Learning from Multidimensional Measurements

Xiaojing Xu XIX068@UCSD.EDU
Department of Electrical and Computer Engineering
University of California San Diego, CA, USA

Jeannie S. Huang JSHUANG@UCSD.EDU
Rady Childrens Hospital
Department of Pediatrics
University of California San Diego, CA, USA

Virginia R. de Sa DESA@UCSD.EDU
Department of Cognitive Science and
Halıcıoğlu Data Science Institute,
University of California San Diego, CA, USA

Editors: Adrian V. Dalca, Matthew B.A. McDermott, Emily Alsentzer, Samuel G. Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones

Abstract

Previous work on automated pain detection from facial expressions has primarily focused on frame-level pain metrics based on specific facial muscle activations, such as Prkachin and Solomon Pain Intensity (PSPI). However, the current gold standard pain metric is the patient's self-reported visual analog scale (VAS) level which is a video-level measure. In this work, we propose a multitask multidimensional-pain model to directly predict VAS from video. Our model consists of three stages: (1) a VGGFace neural network model trained to predict frame-level PSPI, where multitask learning is applied, i.e. individual facial action units are predicted together with PSPI, to improve the learning of PSPI; (2) a fully connected neural network to estimate sequence-level pain scores from frame-level PSPI predictions, where again we use multitask learning to learn multidimensional pain scales instead of VAS alone; and (3) an optimal linear combination of the multidimensional pain predictions to obtain a final estimation of VAS. We show on the UNBC-McMaster Shoulder Pain dataset that our multitask multidimensional-pain method achieves state-of-the-art performance with a mean absolute error (MAE) of 1.95 and an intraclass correlation coefficient (ICC) of 0.43. While still not as good as trained human observer predictions provided with the dataset, when we average our estimates with those human estimates, our model improves their MAE from 1.76 to 1.58. Trained on the UNBC-McMaster dataset and applied directly with no further training or fine-tuning on a separate dataset of facial videos recorded during post-appendectomy physical exams, our model also outperforms previous work by 6% on the Area under the ROC curve metric (AUC).

1. Introduction

Reading facial expressions is one of the most useful ways that humans perceive pain in others (Fordyce, 1976). Accurate measurement of the pain severity, however, is difficult

even for trained professionals. The current clinical gold standard and most widely employed method of assessing clinical pain is patient self-report (Zamzmi et al., 2016). However, this method is vulnerable to social and self-presentation biases and requires substantial cognitive, linguistic, and social competencies (Zamzmi et al., 2016; Sikka et al., 2015; Aung et al., 2016). The goal of an automated facial pain recognition model is to generate a pain level based on facial videos that predicts the patient's self-reported visual analog scale (VAS) pain level. The model should be able to generalize to new patients, for example those with communication disabilities.

Pain is multidimensional. Major dimensions of pain include physiological, sensory, affective, cognitive, behavioral, and sociocultural (McGuire, 1992) aspects. Self-reported VAS reports the subjective nature of pain but other multidimensional assessments have also been useful (Ramelet et al., 2007; Ahles et al., 1983; Clark et al., 2002). In this paper, we analyze the relationship between several pain measurements and their predictions from a machine learning model, and propose a novel method to learn a pain score as a combination of several dimensions of pain to better approximate the patient's reported VAS level.

A natural way to predict a pain score using video is to use a 3D CNN. However, this is not feasible in clinical pain detection because (1) clinical pain datasets are usually too small to train a deep model and (2) the length of the video is not fixed. By contrast, there exist a lot of models designed and trained for face analysis tasks in images, and we only need to fine-tune such a model to apply to pain data frames. This paper proposes an efficient three-stage model to estimate pain in video. In the first stage, we use deep neural networks pre-trained on other face datasets to predict frame-level pain features such as Prkachin and Solomon Pain Intensity (PSPI) (Prkachin and Solomon, 2008) scores directly from raw images. We then extract statistics from the output of the first stage, and send them into a neural network to get the sequence-level multidimensional pain scales. Further, we find an optimal linear combination of these pain scales to estimate VAS. We also use multitask learning in each of the first two stages, and show that both helped improve the final VAS estimation. We show on the UNBC-McMaster Shoulder Pain dataset (Lucey et al., 2011) and a post-surgery child pain dataset (Xu et al., 2018a; Hawley et al., 2019) that the proposed extended multitask-learning multidimensional-pain approach outperforms current state-of-the-art methods on pain intensity estimation in video.

1.1. Contributions

- We propose a three-stage model to evaluate the current gold standard pain metric VAS in video from video frames directly
- We show that multitask learning of pain-related ratings improves the learning of target pain ratings
- We propose to learn VAS as a combination of several dimensions of pain which are learned through multitask learning, and show that this extended multitask learning method performs significantly better than predicting VAS directly
- Our model beats the current state-of-the-art performance on two datasets
- Our model when combined with human estimates improves human estimation of pain.

1.2. Related Work

Two types of pain metrics are considered in pain studies (Ashraf et al., 2009). In facial video pain recognition, frame-level pain metrics are calculated from the intensity of objective facial action units (AUs), such as PSPI in individual video frames. Sequence-level pain metrics are rated by observers or subjects themselves for a collection of frames or video.

Most research on automatic pain detection using facial expression has focused on frame-level pain metrics. Early studies have primarily involved two steps: extracting features from facial images, and then using machine learning models to predict pain levels. Ashraf et al. (2009) and Lucey et al. (2011) used AAM (Active Appearance Model)-based features and SVM (Support Vector Machine) to detect pain. Monwar and Rezaei (2006) extracted location and shape features of the face and used a neural network to recognize pain expressions. Rudovic et al. (2013) proposed the heteroscedastic Conditional Ordinal Random Field to change the variance in the ordinal probit model to adapt to the pain expressiveness level specific to each subject. Recently, deep learning has been increasingly used to assess pain directly from raw pixels. Wang et al. (2017) fine-tuned a face verification network. Zamzmi et al. (2018) combined deep features from pre-trained VGGFace with traditional features for neonates' pain facial expression detection. There is also work considering spatiotemporal information when estimating pain in a single frame. Rodriguez et al. (2017) linked CNNs to a Long Short-Term Memory Networks (LSTM) model. Tavakolian and Hadid (2018) used 3D CNNs to capture a wide range of spatiotemporal variations of the faces. Other work has attempted to detect peak pain intensity of the entire video using multiple-instance learning (Sikka et al., 2013; Ruiz et al., 2016).

None of the above methods estimate a sequence-level self-reported pain measure, but pain is a subjective experience and self-rating with measures such as 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. (2015) and Xu et al. (2018a) detected postoperative pain in children using AUs extracted by iMotions (imotions.com). Liu et al. (2017) proposed a two-stage method to first train a NN (Neural Network) model at frame level using sequence-level VAS as label and AAM landmarks as inputs, and then obtained video VAS score from frame-level predictions using a Gaussian process regression model. Martinez et al. (2017) used a bidirectional LSTM model to predict PSPI of each video frame using AAM landmarks and then applied personalized HCRFs (Hidden Conditional Random Fields) to predict VAS using the PSPI sequences.

Our model can be decomposed to frame-level and sequence-level predictions in a similar way to the two stages in Liu et al. (2017); Martinez et al. (2017); Xu et al. (2018a) but our model takes raw images as inputs in Stage 1, which involves the use of deep learning and transfer learning, and doesn't require AAM landmarks or AUs on test data which require expensive trained human annotation of key frames and automated landmark/AU detector and tracking algorithms.

2. Method

We developed our model based on the widely used UNBC-McMaster Shoulder Pain dataset (Lucey et al., 2011). It includes facial videos of participants suffering from shoulder pain while performing a series of active and passive range-of-motion tests to their affected and unaffected

limbs on two separate occasions. The dataset has 25 subjects, 200 videos and 48,398 frames of size 320 x 240 pixels in total with two types of labels: frame-level labels and sequence-level labels. Frame-level labels include 66 AAM landmarks, 11 facial action unit (AU) intensities and 1 PSPI score. Both of the previous papers predicting VAS using this dataset (Martinez et al., 2017; Liu et al., 2017) used AAM landmarks as input features but in this work we only used images as inputs. We also used AUs and PSPIs as outputs during training.

AUs are defined by the FACS (Facial Action Coding System) (Ekman and Friesen, 1976) to code nearly all anatomically possible facial expressions. In this dataset, the 11 AUs are brow lowering (AU4), cheek raising (AU6), eyelid tightening (AU7), nose wrinkling (AU9), upper-lip raising (AU10), oblique lip raising (AU12), horizontal lip stretch (AU20), lips parting (AU25), jaw dropping (AU26), mouth stretching (AU27) and eye closure (AU43) coded by trained human FACS coders. These AUs are among those believed to be related to pain expression (Lucey et al., 2011).

PSPI (Prkachin and Solomon, 2008) is a pain evaluation metric computed from a specific set of pain-related AU intensities:

$$PSPI = AU4 + max(AU6, AU7) + max(AU9, AU10) + AU43$$

AU intensities are integers ranging from 0-5 (weakest trace to maximum intensity possible), except for AU43 which is only scored with values of 0 or 1, so PSPI rating is also an integer and ranges from 0-16 (with larger values reflecting more pain).

Sequence-level labels include the gold standard self-rating VAS pain score ranging from 0-10, as well as three other pain ratings: OPR (Observers Pain Rating) 0-5, AFF (Affective-motivational scale) 0-16 and SEN (Sensory Scale) 0-16. The properties of AFF and SEN are discussed in Gracely et al. (1978); Heft et al. (1980). OPR is an estimation of pain intensity given by independent trained observers from the recorded video. The observers are shown to have high inter-observer reliability (Lucey et al., 2011).

With the help of the labels described above, our goal is to train a model that predicts VAS from the video or image sequence directly. We chose the hyper-parameters of the neural networks based on training/validation learning curves and validation performance. The learning rates were selected using grid search at logarithmic intervals. The number of epochs and early stopping criterion were decided by observing the learning curves. The choice of optimizer doesn't affect the validation performance much so we chose it based on previous work (Parkhi et al., 2015; Xu et al., 2018a) and experience.

2.1. Stage 1: PSPI Estimation in Facial Images

Our first stage predicts frame-level PSPI score from RGB images. We built our model based on the VGGFace model (Parkhi et al., 2015). The architecture was designed and pre-trained to classify 2622 individuals, and we simply replaced the last layer with our own linear fully-connected regression layer. During training, we updated all parameters in the neural network, but we used different initial learning rates (1e-4 for the last layer and 1e-5 for other layers). We used Adam optimization and weight decay of 5e-4. We applied batch-weighted (Sellami and Hwang, 2019) Mean Squared Error (MSE) loss, where the weight of a sample in loss is inversely proportional to the proportion of its label (which is PSPI score here) in the current batch, to overcome the class imbalance problem. We used a batch

size of 32, max epochs of 50, and early stopping when the validation loss hadn't decreased for 20 epochs.

We did the same image pre-processing as done in the VGGFace model (Parkhi et al., 2015). Specifically, we used the cascade DPM Face Detector (Wolf et al., 2011; Mathias et al., 2014) to detect the face and then extended the bounding box by a factor of 0.1 and resized the cropped image to 224×224. We normalized the images with the mean and standard deviations per RGB channel over the pre-training data.

2.2. Stage 2: VAS Estimation in Facial Videos using Sequence of Predictions

After we obtained PSPI predictions of each frame in Stage 1, we extracted 9 statistics (mean, max, min, standard deviation, 95th, 85th, 75th, 50th, 25th percentiles) over all frames of a video to form a video feature vector which was fed to a fully connected neural network with one 18-unit hidden layer to predict VAS in a linear output unit using batch-weighted MSE loss similar to Stage 1. We used Adam with initial learning rate 1e-2. Batch size was set to 32 and max epochs to 200; early stopping, when the validation loss hadn't decreased for 20 epochs, was used.

Combining Stage 1 and 2 we obtained our baseline model which predicts VAS score from video. This is illustrated in Figure 1 by solid blocks. Stage 1 and 2 were trained separately due to memory capacity limitations of our GPU.

2.3. Multitask Learning

The UNBC-McMaster Shoulder Pain dataset contains other pain metrics besides VAS and PSPI at both the frame and sequence level. At the frame level, there are 11 manually coded FACS AUs. At the sequence level, besides VAS, three other pain ratings are available. We hypothesized that a multitask network (Caruana, 1997) learning these metrics with the same hidden layer/representations as those learning to predict PSPI and VAS may be better able to learn PSPI and VAS.

For example, in the first stage, PSPI is a non-linear combination (due to the max operation) of 6 AUs. The same PSPI score could be due to many different combinations of AU scores and underlying facial expressions. Thus there is a noisy many to one mapping between facial muscle activations and PSPI scores. Learning individual AU activations is a simpler mapping, and a network that performs well on the underlying AU representations should be able to compute PSPI.

Similarly, in the second stage, OPR, AFF, and SEN are related to the VAS pain score. OPR in particular is more directly related to the video than VAS. The OPR scores resulted from trained human observers estimating pain using only the video. On the other hand, a person's self-reported VAS may not be fully reflected in their video; if the person is particularly expressive or stoic, their VAS score may be more or less related to the video features.

Our proposed multitask architecture is illustrated in Figure 1. In Stage 1, instead of only one output estimating PSPI, we concatenated several AU values and the PSPI score to form a multitask vector output. During training, we scaled the labels into the same range to make sure all elements contribute equally to the loss. AU labels are even more sparse than PSPI labels, so we only used the 9 AUs (AU4, 6, 7, 10, 12, 20, 23, 26, 43) labeled in more than

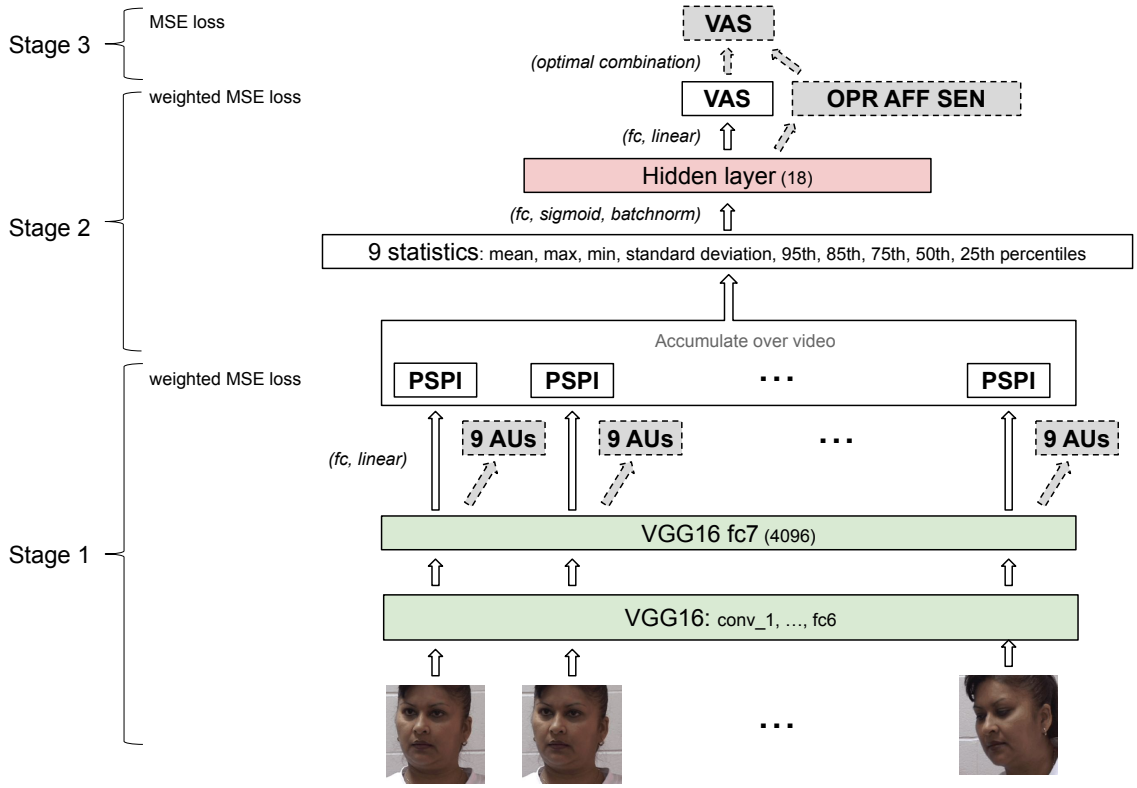


Figure 1: The proposed three-stage structure. The baseline model is represented by solid blocks, and shaded blocks with dashed outlines show added parts in multitask learning and ensemble learning with multidimensional pain scales. During training, Stage 1 is first trained with batches of frames and used to predict a sequence of PSPI scores. Then Stage 2 is learnt using batches of video features obtained from PSPI sequences. The network can't be trained end-to-end due to limited GPU memory.

500 frames out of the 48398 frames in the dataset. For a similar reason, we weighted the loss function using PSPI score distribution and only looked at PSPI for validation loss for early stopping.

In Stage 2, similarly, we used a 4-dimensional vector representing the four pain ratings instead of a single value representing VAS as output. The losses are weighted based on the distribution of VAS scores, and the validation loss is the mean MSE of the 4 outputs.

2.4. Stage 3: Ensemble Learning of Multidimensional Pain Measurement

On the UNBC-McMaster dataset, each of the 4 sequence-level scores (VAS, OPR, AFF, and SEN) can be seen as pain estimates that focus on different aspects of pain. For example, VAS reflects how much pain the patient perceives and relies on the patient's personal understanding of pain, whilst OPR is based on third-party observation of facial expressions, and will be

influenced by how much "pain expression" the patient shows on his/her face and how good the observer is at reading facial expressions of pain. They also have different properties. For example, OPR may be more consistent across subjects when scored by the same observer. As OPR entirely depends on facial video it should be more easily learned from facial video than VAS in the same way that AUs should be more learnable from video than a non-linear function of them. At the same time, OPR may be limited as a measure of actual pain as it is only able to reflect pain revealed by facial expressions and will be biased if the subject hides it, but that will also be a limitation of our system and any computer vision system unless it incorporates features from other sensors (Xu et al., 2018b).

OPR, AFF, and SEN are all highly correlated with VAS and can be considered as predictions of VAS. In fact, after scaling the outputs of Stage 2 to the same range as VAS, all 4 outputs do a reasonable job at estimating VAS. That is, we have 4 "experts" each with its own prediction of pain level. This suggests an ensemble averaging method to reduce variance at no cost to bias (Hashem, 1997). This corresponds to Stage 3 in Figure 1. The optimal linear combination of experts to form a least mean squared error estimation of the target score was discussed in Hashem (1997). Below we briefly discuss the derivation of our ensemble model weights.

Consider each data point $(\boldsymbol{x}, y)$ as an observation of random variables $(\boldsymbol{X}, Y)$ from an unknown multivariate distribution over $\mathbb{R}^9 \times \mathbb{R}$. And $f_i : \mathbb{R}^9 \to \mathbb{R}$ $(i = 1, 2, 3, 4)$ maps Stage 2 inputs to a real number, each corresponding to one of the 4 scores.

We learn the final prediction of VAS as a weighted sum of the four experts. If each expert is f_i, then the overall model $\tilde{f}$ can be defined as:

$$\tilde{f}(\boldsymbol{x}) = \sum_{i=1}^{4} \alpha_i f_i(\boldsymbol{x})$$

where we apply the constraint $\sum_{i=1}^{4} \alpha_i = 1$ (and $\alpha_0 = 0$) as suggested by Clemen (1986); Trenkler and Liski (1986); Hashem (1997).

The MSE loss of the final model is:

$$\text{MSE}(\tilde{f}(\boldsymbol{X})) = E[(\tilde{f}(\boldsymbol{X}) - Y)^2] = E[(\sum_{i=1}^{4} \alpha_i f_i(\boldsymbol{X}) - Y)^2] = E[(\sum_{i=1}^{4} \alpha_i (f_i(\boldsymbol{X}) - Y))^2]$$

Our goal is to minimize MSE subject to $\sum_{i=1}^{4} \alpha_i = 1$. The Lagrangian expression is:

$$L(\mathbf{X}, \lambda) = \text{MSE}(\tilde{f}(\mathbf{X})) - \lambda(\sum_{i=1}^{4} \alpha_i - 1) \tag{1}$$

where λ is the Lagrange multiplier.

Differentiating Equation (1) with respect to α_k:

$$\frac{\partial L(\mathbf{X}, \lambda)}{\partial \alpha_k} = E[2 \sum_{i=1}^{4} \alpha_i (f_i(\mathbf{X}) - Y)(f_k(\mathbf{X}) - Y)] - \lambda = 2 \sum_{i=1}^{4} \alpha_i E[(f_i(\mathbf{X}) - Y)(f_k(\mathbf{X}) - Y)] - \lambda$$

and setting the derivative to 0 gives us the optimal $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \alpha_3, \alpha_4]^T$ as:

$$\boldsymbol{\alpha} = \frac{\Omega^{-1}\mathbf{1}}{\mathbf{1}^T \Omega^{-1} \mathbf{1}} \tag{2}$$

where $\Omega = [\omega_{ij}] = E[(f_i(\boldsymbol{X}) - Y)(f_j(\boldsymbol{X}) - Y)]$

3. Experiments

On the UNBC-McMaster dataset, we performed 5-fold cross validation with each fold consisting of 5 subjects. We used the same training/test splits for the three stages in each iteration.

Our results are reported in Table 1 and Table 2. We report Mean Absolute Error (MAE), Mean Squared Error (MSE), Intraclass Correlation Coefficient (ICC) and Pearson Correlation Coefficient (PCC) averaged over five cycles of 5-fold cross validation (along with the standard deviation over the 5 runs). To ensure reproducibility, we used the same set of random seeds to make sure all models are trained and tested on the same data and have the same initial states. In Table 1, the predictions of the final extended multitask learning model are given in the last row, and the previous rows are discussed in the next subsections and reflect ablation analyses to examine the effect of different components of our model. In Table 2, we compare our model to previously published research.

We run all our experiments on a single GPU (NVIDIA GeForce RTX 2080); it takes about 4 hours to train a three-stage model using 4 folds of the UNBC-McMaster data.

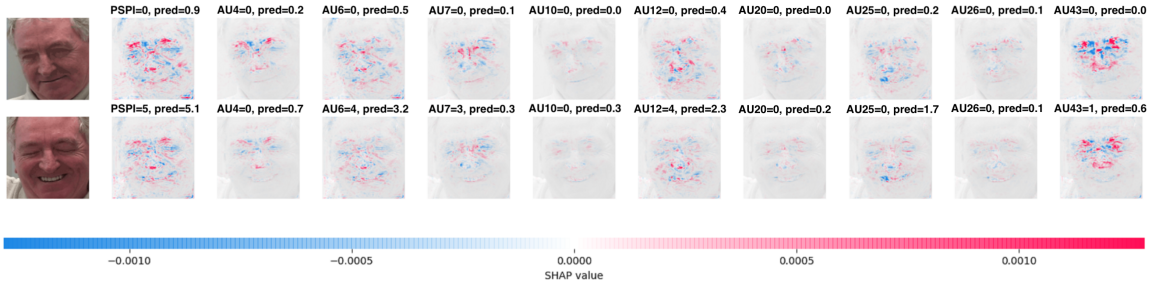
3.1. Stage 1: PSPI Estimation using Multitask Learning

We obtained an MAE of 0.84 ± 0.06 in PSPI estimation in Stage 1. We could likely improve the performance by using temporal information, but as our final goal is to estimate VAS using the PSPI predictions, we decided to focus on that aspect and looked at the final performance on VAS directly.

In Table 1, the second row compared to the first row shows the benefit of multitask learning in Stage 1. It should be noticed that in multitask learning although the AUs are learned in Stage 1, they are not fed into the next stage, but having them as targets during Stage 1 training helps the net learn more informative PSPI predictions for predicting VAS.

To better understand this aspect, we plot the contributions of image pixels to the outputs using SHAP introduced by Lundberg and Lee (2017) in Figure 2. SHAP is a framework that interprets complex models by assigning each feature an importance value for a particular prediction. Pixels with larger absolute values of SHAP (darker color on image) reflect a greater influence on the output. In the 3rd column of Figure 2 (a), the PSPI output of the MTL model has captured more meaningful pixels on the face compared to the baseline model shown in the 2nd column of Figure 2 (a). We can see clearer shapes of the eyebrows and the mouth, and even the nasolabial folds in the 3rd column relative to the 2nd. Figure 2 (b) shows that many of these areas are relevant for the prediction of several AUs, such as eyebrows in AU4, eye area in AU7 and AU43, corners of the lip and nasolabial furrows in AU12, mouth in AU25, etc. Note this is true even though some of the AUs are not well learned because of a lack of training data (AU10 and AU20 both are present in less than 1000 frames).

(*a*) Baseline (2nd col) vs MTL (3rd col)



(*b*) MTL PSPI (2nd col) and AUs (remaining cols)

Figure 2: Contributions in Stage 1 of pixels for two frames are explained in the figures above. (a) compares PSPI predictions and pixel contributions for the baseline model (that outputs only PSPI) with the MTL model (that outputs 9 AUs in addition to PSPI). (b) shows the processing of the PSPI as well as the 9 AUs in the MTL model. The first column in both (a) and (b) show two frames from different videos (top row for PSPI=0, bottom row for PSPI=5). The rest of each row reflect processing of each of these frames. The second column in (a) shows the contributions of pixels to the baseline Stage 1 model using SHAP. The third column in (a) shows the contributions of pixels to the MTL model predicting both PSPI and 9 AUs. In (b) this column is reproduced next to plots of the SHAP contributions towards predicting the AU outputs. Positive SHAP value means a positive contribution of a pixel to the corresponding output, and negative SHAP value means negative contribution. For example, in (b), when predicting AU4, the model focuses on the area around eyes and eyebrows, especially the inner portion of the eyebrows and the area between them, which is consistent with the description of AU4.

3.2. Stage 2 and 3: VAS Estimation using Extended Multitask Learning of Multidimensional Pain Scales

Using the PSPI estimations from Stage 1, we trained a neural network to predict VAS. Again multitask learning was used and the 4 pain scales shared the same hidden layer. In Table 1, we show in row 3 compared to row 1 (as well as row 4 compared to row 2) that multitask learning in Stage 2 improves VAS prediction slightly.

We first observed the performance of each of the 4 outputs from Stage 2, shown in Figure 3. Interestingly, the best approximation of a metric is not always given by its corresponding

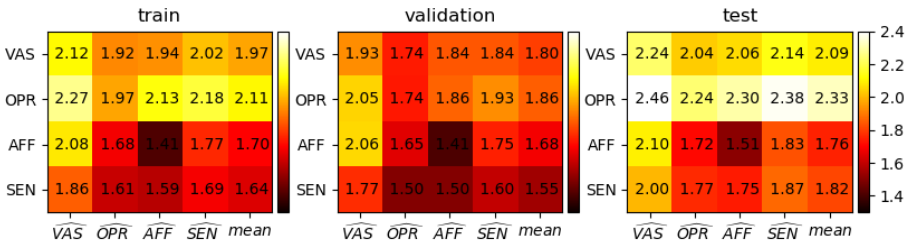


Figure 3: Average MAE matrices on training, validation and test data. The y-axis gives the true label, and the x-axis the prediction (or the mean of the 4 predictions). Each entry is the mean absolute difference between the two variables.

Table 1: Sequence-level VAS Prediction using Frame-level Predictions from Stage 1

Model	MAE	MSE	ICC	PCC
Baseline	2.34 ± 0.09	7.27 ± 0.51	0.34 ± 0.04	0.50 ± 0.04
MTL (AUs)	2.23 ± 0.08	6.76 ± 0.37	0.37 ± 0.02	0.52 ± 0.03
MTL (pain scores)	2.30 ± 0.06	7.06 ± 0.31	0.37 ± 0.04	0.52 ± 0.02
MTL (AUs + pain scores)	2.20 ± 0.06	6.53 ± 0.30	0.37 ± 0.03	0.54 ± 0.02
MTL (AUs) + Ensemble	1.97 ± 0.04	6.10 ± 0.26	0.40 ± 0.03	0.52 ± 0.03
MTL (AUs + pain scores) + Ensemble	1.95 ± 0.06	5.98 ± 0.22	0.43 ± 0.02	0.54 ± 0.03

trained output. The OPR output does a better job in estimating OPR than other outputs. The same is true for AFF. However, the OPR output gives a better estimate of VAS than the VAS output.

One way to solve this problem is to consider them as four different estimators of VAS and learn an ensemble model on top of them. This third stage of our model has been discussed in Section 2.4 and experimental results are shown in the last row in Table 1. The optimal weights were found on training and validation data, and the ensemble outperforms each of the 4 outputs on the test data.

In order to show that multitask learning in Stage 2 is also helpful, we also trained 4 separate networks for the 4 scores with no shared parameters. Then we combined these scores using the same method for learning an ensemble model and obtained a final prediction of VAS. The performance (body row 5 in Table 1) is slightly worse than using multitask learning at this stage.

It should also be noted that learning the weights in Stage 2 and 3 together through back propagation didn't give as much improvement as in the extended multitask learning where we learned to predict multiple pain dimensions first and combine them afterward.

Overall, the multitask and ensemble contributions to our model (bottom row Table 1) improve the performance by 17% over the baseline model (top row Table 1).

3.3. Comparison with Other Work

Our results are compared with previous work estimating VAS using the UNBC-McMaster dataset or a child pain dataset in Table 2. The child pain dataset contains facial video from

children aged 10 to 15 who had undergone medically necessary laparoscopic appendectomy. Details of this dataset can be found in Xu et al. (2018a). Without any retraining or fine-tuning, we tested the model trained on the UNBC-McMaster dataset with 134 videos of 70 subjects from the child pain dataset. Our model performs significantly better than previous work. Our 95% Confidence Interval of MAE on the shoulder pain dataset is 1.95 ± 0.0526 and that of ICC is 0.43 ± 0.0175. If we assume equal variability for the previous state-of-the-art (Liu et al., 2017) (not provided in the paper), our MAE is significantly lower ($p = 0.0002$). Similarly, our AUC on the child pain data is significantly higher than that in Xu et al. (2018a) with $p = 0.00001$.

Table 2: Comparison with Other Work

Model	Dataset	MAE	ICC	AUC
pRNN-HCRF (p=1) (Martinez et al., 2017)	UNBC	2.47 ± 0.18	0.36 ± 0.08	-
pRNN-HCRF (p=2) (Martinez et al., 2017)	UNBC	2.46 ± 0.23	0.34 ± 0.04	-
DeepFaceLIFT (Liu et al., 2017)	UNBC	2.18	0.35	-
Extended MTL (Our Model)	UNBC	$\mathbf{1.95 \pm 0.06}$	$\mathbf{0.43 \pm 0.02}$	-
TransferLearning (Xu et al., 2018a)	Child	-	-	0.72 ± 0.02
Extended MTL (Our Model)	Child	2.22 ± 0.10	0.33 ± 0.05	$\mathbf{0.76 \pm 0.01}$

4. Discussion and Conclusion

We designed an automatic system which takes in facial videos and outputs a pain score. It can either be applied to new videos directly without any human annotation (as we did for the child pain dataset), or retrained with new data following our method for possible performance improvement. The new dataset doesn't have to be annotated with the same labels. For example, other pain dimensions/action units can also be helpful using the extended multitask learning framework.

We can compare our result to trained human raters (the provided OPR value). When the given OPR is used to estimate VAS, the MAE is 1.76. The machine learning model (MAE 1.95) is not quite as good as expert human observers for now, but it is close and cheaper, more consistent, and available 24 hours a day. More usefully, in situations where an observer's rating is available, averaging our model output with the real OPR gives better performance than real OPR (HUMAN) alone. With this simple averaging, the MAE is reduced to 1.58. A Wilcoxon signed-rank test revealed that the absolute errors are significantly lower for the average of OPR and our output than OPR alone ($p = 4.06e\text{-}7$). **Thus our system can help expert humans estimate VAS better.**

To summarize, we propose a three-stage model to predict VAS in facial videos directly, and propose a method using multitask learning, multidimensional pain measurement and ensemble learning to effectively improve the performance of the model. Our approach achieves state-of-the-art performance on the UNBC-McMaster Shoulder Pain dataset and a child pain dataset. Our approach can also be easily transferred to other healthcare tasks and general machine learning application tasks. The three-stage structure can be used when end-to-end training is not feasible with limited computational memory and data. Breaking the problem into separate stages allows the use of transfer learning from pre-trained models

and well-studied structures and simplifies the learning and interpretation of the system. Our idea of extended multitask learning and the usage of multidimensional measurements can be applied to other healthcare data which are noisy, high-dimensional, and limited in the number of samples.

Acknowledgments

This work was supported by National Institutes of Health National Institute of Nursing Research grant R01 NR013500 and NSF IIS 1528214.

References

Tim A Ahles, Edward B Blanchard, and John C Ruckdeschel. The multidimensional nature of cancer-related pain. *Pain*, 17(3):277–288, 1983.

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.

Min SH Aung, Sebastian Kaltwang, Bernardino Romera-Paredes, Brais Martinez, Aneesha Singh, Matteo Cella, Michel Valstar, Hongying Meng, Andrew Kemp, Moshen Shafizadeh, et al. The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset. *IEEE transactions on affective computing*, 7(4): 435–451, 2016.

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

W Crawford Clark, Joseph C Yang, Siu-Lun Tsui, Kwok-Fu Ng, and Susanne Bennett Clark. Unidimensional pain rating scales: a multidimensional affect and pain survey (maps) analysis of what they really measure. *Pain*, 98(3):241–247, 2002.

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

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

Wilbert Evans Fordyce. *Behavioral methods for chronic pain and illness*, volume 1. Mosby St. Louis, 1976.

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

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

Kara Hawley, Jeannie S Huang, Matthew Goodwin, Damaris Diaz, Virginia R de Sa, Kathryn A Birnie, Christine T Chambers, and Kenneth D Craig. Youth and parent appraisals of participation in a study of spontaneous and induced pediatric clinical pain. *Ethics & Behavior*, 29(4):259–273, 2019.

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.

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.

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.

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.

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.

Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *European conference on computer vision*, pages 720–735. Springer, 2014.

Deborah B McGuire. Comprehensive and multidimensional assessment and measurement of pain. *Journal of pain and symptom management*, 7(5):312–319, 1992.

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.

O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

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.

Anne-Sylvie Ramelet, Nancy Rees, Susan McDonald, Max Bulsara, and Huda Huijer Abu-Saad. Development and preliminary psychometric testing of the multidimensional assessment of pain scale: Maps. *Pediatric Anesthesia*, 17(4):333–340, 2007.

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.

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.

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.

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.

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.

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.

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.

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

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.

Lior Wolf, Tal Hassner, and Itay Maoz. *Face recognition in unconstrained videos with matched background similarity*. IEEE, 2011.

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 *Joint Workshop on Artificial Intelligence in Health*, pages 10–21. CEUR-WS, 2018a.

Xiaojing Xu, Büsra Tugçe Susam, Hooman Nezamfar, Damaris Diaz, Kenneth D Craig, Matthew S Goodwin, Murat Akcakaya, Jeannie S Huang, and Virginia R de Sa. Towards automated pain detection in children using facial and electrodermal activity. In *Joint Workshop on AI in Health*, pages 208–211. CEUR-WS, 2018b.

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.

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