Personalized Pain Detection in Facial Video with Uncertainty Estimation

Xiaojing Xu[1] and Virginia R. de Sa[2]

***Abstract*— Pain is a personal, subjective experience, and the current gold standard to evaluate pain is the Visual Analog Scale (VAS), which is self-reported at the video level. One problem with the current automated pain detection systems is that the learned model doesn't generalize well to unseen subjects. In this work, we propose to improve pain detection in facial videos using individual models and uncertainty estimation. For a new test video, we jointly consider which individual models generalize well generally, and which individual models are more similar/accurate to this test video, in order to choose the optimal combination of individual models and get the best performance on new test videos. We show on the UNBC-McMaster Shoulder Pain Dataset that our method significantly improves the previous state-of-the-art performance.**

I. Introduction

Two types of pain metrics are considered in pain studies [1]. In facial video pain recognition, frame-level pain metrics are calculated from the intensity of objective facial muscle movements called facial action units (AUs) defined by the Facial Action Coding System (FACS). Relevant pain related AU descriptions are given in Figure 1. A commonly used combination of some pain-related action units is called the Prkachin and Solomon Pain Intensity(PSPI) [2] : PSPI=AU4+max(AU6,AU7)+max(AU9,AU10)+AU43). Sequence-level pain metrics are overall pain levels rated by observers or the subjects themselves.

The current gold standard to evaluate pain is the sequence-level self-rated 0-10 Visual Analog Scale (VAS). Automated pain evaluation systems developed to help detect pain [3]–[7] can usually be broken down into two stages: Stage 1 predicts the PSPI score in each frame, and Stage 2 learns VAS using predicted PSPI scores in a video. This work follows the same two stage approach to predict VAS.

Pain is a personal, subjective experience, and VAS is a noisy label that differs in its relationship to facial expression across subjects. This makes automated pain estimation difficult when generalizing to subjects not in the training dataset. To address this issue, Martinez et al. introduced a facial expressiveness score, unique for each person, but their method requires labeled data for new subjects [8]. Liu et al. personalized the estimation of self-reported pain via a set of hand-crafted personal features including age, gender and complexion [6]. The labeling of these personal features is easier, but still the model can't automatically generalize to unseen subjects. There is also work tackling pain personalization in images instead of videos [9], [10].

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

Fig. 1: Pain-related AU descriptions.

In this work, we propose a systematic way to model the noise and bias in VAS in different subjects, and design a pain estimation model that can be optimized for new subjects using uncertainty estimation.

A. Uncertainty in Machine Learning Models

Uncertainty can be generally categorized into two types: epistemic or aleatory [11], [12]. Epistemic uncertainty can be reduced given enough data, while aleatoric uncertainty captures noise that is inherent in the observations.

In a supervised learning problem, suppose data points (x_i, y_i) are related via a model $y_i = f(x_i) + \varepsilon_i$, where f is the true function that maps data input to output, and ε_i is the noise inherent in the observations with zero mean and variance σ_i^2. A machine learning model seeks to find a function $\hat{f}(x; D)$ that approximates the true $f(x)$, using training data $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$. Using mean squared error to evaluate the approximation, the expected squared error between $\hat{f}(x; D)$ and y on new observation (x, y) is:

$$E[(y - \hat{f}(x; D))^2] = E[(f(x) + \varepsilon - \hat{f}(x; D))^2] \quad (1)$$

$$= \sigma^2 + E[(f(x) - \hat{f}(x; D))^2] \quad (2)$$

Eq (2) follows from (1) because ε is independent of $\hat{f}$.

σ^2 is often called the irreducible error. It is a property of the data, not the model, so it captures the aleatoric uncertainty. The second term doesn't exactly capture epistemic uncertainty because $\hat{f}$ is only one deterministic model, but it is correlated to epistemic uncertainty, evaluating how much the solution $\hat{f}$ over D varies from the true solution f assuming infinite data.

Neural networks have been used to estimate the input dependent $f(x)$ as well as the variance $\sigma^2(x)$ of the prediction $\hat{f}(x)$ [13], [14]. In this work, we make the same assumption that the noise ε is input/subject dependent and can be predicted using a machine learning model.

Work supported by NSF IIS 1528214, and 1817226, IBM Research AI, NVIDIA Corporation (Titan V), NIH R01 NR013500,the Sanford Institute for Empathy and Compassion Center for Empathy and Technology and the UC San Diego Division of Social Sciences.

[1] X. Xu is with the Department of Electrical and Computer Engineering, UC San Diego, CA, USA. xix068@ucsd.edu

[2]V. R. de Sa is with the Department of Cognitive Science and the Halıcıoğlu Data Science Institute, UC San Diego, CA, USA. desa@ucsd.edu

B. Contributions

- We learn personalized individual models to predict the current gold standard pain metric VAS in video from video frames directly.
- We learn PSPI and VAS as a combination of the output of individual models to improve the generalizability of the pain prediction model.
- We learn the uncertainty of VAS prediction of each individual model, and improve the VAS prediction on new test subjects by adjusting ensemble weights based on the uncertainty of individual predictions
- Our models beat the current state-of-the-art performance on the UNBC-McMaster dataset.

II. Methods

Our model uses the Extended Multi-Task Learning (EMTL) model described in [4] as the baseline structure. The original EMTL model is trained using all training subjects together; in this work we train individual models on data from individual training subjects, and explore ways to combine these individual models so that the ensemble prediction is optimal for samples from test subjects.

A. Optimal Linear Combination of Individual Models

We previously [4] proposed an optimal linear combination of multidimensional pain estimations (VAS, OPR, SEN, and AFF) to obtain an improved prediction of VAS. This method works well in aggregating different aspects of pain to produce a better estimation but doesn't consider the subject-dependent aspect of pain, i.e. different patients experience pain differently and express their pain in different facial expressions.

We address this problem by training personalized models: instead of training one model with all training subjects, we train several models each using video from one subject.

Consider each data point $(\boldsymbol{x}, y)$ as an observation of random variables $(\boldsymbol{X}, Y)$, and denote the model for subject s as $\hat{f}_s$. We learn the final prediction of VAS as a weighted sum of the predictions $\hat{f}_s(x)$. The overall model $\tilde{f}$ can be represented as:

$$\tilde{f}(\boldsymbol{x}) = \sum_s \alpha_s \hat{f}_s(\boldsymbol{x}) = \boldsymbol{\alpha}^T \hat{\boldsymbol{f}}(\boldsymbol{x}) \tag{3}$$

The solution to minimizing the MSE of the the final model $E[(\tilde{f}(\boldsymbol{X}) - Y)^2]$ subject to $\sum_s \alpha_s = 1$ can be obtained using constrained optimization with solution [4]:

$$\hat{\boldsymbol{\alpha}} = \frac{\Omega^{-1}\mathbf{1}}{\mathbf{1}^T\Omega^{-1}\mathbf{1}} \tag{4}$$

where $\Omega = E\left[(Y - \hat{\boldsymbol{f}}(\boldsymbol{X}))(Y - \hat{\boldsymbol{f}}(\boldsymbol{X}))^T\right]$.

What this means is that, if a subject generalizes to others better than another subject, then the weight of the first subject should be larger than the weight of the second subject in the ensemble model $\tilde{f}$. The optimal linear combination accounts for the covariance between the different estimators and is optimal for minimizing mean squared error.

B. Ensemble using Predicted Variance

The optimal linear combination(OLC) model in section II-A only aims to reduce epistemic uncertainty, and helps the model generalize to data in the same distribution.

However, the data distribution is different for different subjects, and this is captured in the first term, σ^2, in equation (2). In this section we propose to learn the variances of individual model predictions to account for both aleatoric and epistemic uncertainty. In practice, we learn $\hat{\sigma}_s^2(\boldsymbol{x})$ to approximate $(y - \hat{f}_s(\boldsymbol{x}))^2$. This is not the variance exactly, but equals the variance of the label noise if $\hat{f}_s = f_s$.

The original MSE loss is only dependent on predicted means $\hat{f}_s(\boldsymbol{x})$, and assumes the same σ^2 for all data points. This is not true, especially across subjects, because both $\boldsymbol{x}$, the facial expression of pain, and y, the self-rated pain level VAS, are quite different across subjects. In other words, for different subjects, $(\boldsymbol{x}, y)$ data are in different domains. Our variance prediction model is able to predict such uncertainty due to domain shift and use this in determining parameters for the ensemble model. For example, if a video is quite similar to training subject 1, and completely different from training subject 2, then the pain score prediction from the model trained on subject 1 should have smaller σ^2 (and hence a higher ensemble weighting) for this sample than the model trained on subject 2, meaning this sample is out-of-distribution for the subject 2 model and prediction from the subject 1 model is more trustworthy.

The OLC model in section II-A can't do this because the optimal weight in equation (4) is only dependent on training samples. We bring in $\hat{\sigma}_s^2(\boldsymbol{x})$ which also depends on the test input $\boldsymbol{x}$ to predict the best weighting in the ensemble model for individual test samples.

We propose a new loss function which applies Tikhonov regularization to integrate predicted variance in personalized models:

$$Loss(\tilde{f}) = (y - \tilde{f})^2 + \beta\tilde{\sigma^2} \tag{5}$$

where

$$\tilde{f} = \sum_s a_s \hat{f}_s = \boldsymbol{a^T}\hat{\boldsymbol{f}}, \text{and } \tilde{\sigma}^2 = var(\tilde{\varepsilon})$$

$\boldsymbol{a} = [a_s]$ is the weight vector, $\hat{\boldsymbol{f}} = [\hat{f}_s(\boldsymbol{x})]$ is the input vector, and

$$\Sigma = diag(\hat{\sigma}_s^2(\boldsymbol{x})) \tag{6}$$

is a diagonal matrix where the learned variances are on the diagonal.

So the loss (5) can be expressed as

$$Loss(\tilde{f}) = \boldsymbol{a^T}(\Omega + \beta\Sigma)\boldsymbol{a} \tag{7}$$

The first term is the MSE of the final prediction. It finds individual models that generalize well on the whole data distribution, and the MSE matrix Ω is the same Ω in equation (4), learned on the training data. The second term on the other hand looks for models performing better especially for the current video, and is different for each sample. At test

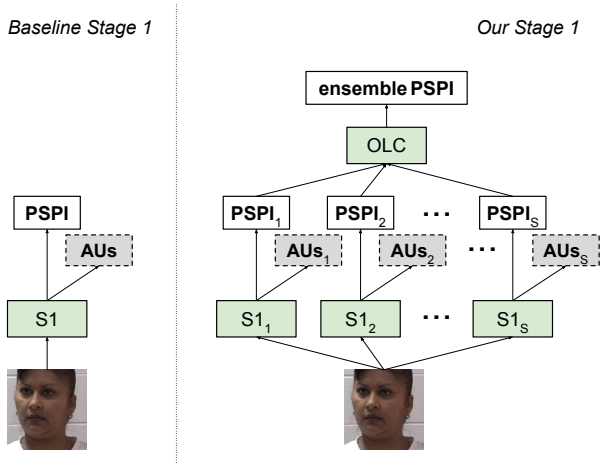


Fig. 2: Stage 1 model structure. $S1$ and $S1_s$'s have a similar structure to VGG16 [15]. They are trained to predict PSPI and AUs. In our model $S1_s$ is trained with subject s, and OLC parameters are learned to combine predictions from individual models to get a better ensemble PSPI prediction.

time, the ensemble model will calculate the optimal weights $\boldsymbol{a}$ for the loss above using the same method as in section II-A, using Ω learned from training data and Σ arising from the variance prediction model (6).

The optimal weight vector is determined by the following equation:

$$\hat{\boldsymbol{a}} = \frac{(\Omega + \beta\Sigma)^{-1}\mathbf{1}}{\mathbf{1}^T(\Omega + \beta\Sigma)^{-1}\mathbf{1}} \tag{8}$$

where the optimal $\hat{\boldsymbol{a}}$ is dependent on the input as Σ is.

C. Ensemble using Predicted Error

In the analysis above we ignored the correlation between the errors in individual model outputs. It may not be true that the errors are independent, so in this section we generalize the method above to consider correlations between errors in different personalized model predictions.

Instead of learning $\hat{\sigma}_s^2(\boldsymbol{x})$, we use the same neural network structure as $\hat{f}_s(\boldsymbol{x})$ to predict $\hat{\varepsilon}_s(\boldsymbol{x})$ which approximates $y - \hat{f}_s(\boldsymbol{x})$. This allows us to calculate the covariance matrix

$$\Sigma = [\sigma_{ij}] = [\varepsilon_i \varepsilon_j] \tag{9}$$

of the multivariate prediction. To take covariance in prediction noise into consideration, we replace Σ in section II-B (6) by (9).

III. Experiments

A. Dataset

We developed our model based on the widely used UNBC-McMaster Shoulder Pain dataset [16] which was collected in accordance with the ethical standards and approval of the Institutional review board. It includes facial videos of participants suffering from shoulder pain while performing a series of active and passive range-of-motion tests to their

Algorithm 1: Pain Estimation Model Training

Data: $D = (\boldsymbol{X}, Y)$, D_s = data from subject s
Result: Model to predict $y \in Y$ given $\boldsymbol{x} \in \boldsymbol{X}$

```
/* Train Stage-1 personalized models   */
for s in training subjects do
    train S1_s using D_s
    for x in D do
        get predictions S1_s(x)
    end
end
/* Ensemble learning on Stage1 predictions  */
```

Over D_{train}, learn a_s to minimize the MSE of $S1(\boldsymbol{x}) = \sum_s a_s S1_s(\boldsymbol{x})$

```
for x in D do
    get predictions S1(x)
end
/* Train Stage-2 personalized models   */
for s in training subjects do
    train S2_s using D_s
    for x in D do
        get predictions S2_s(S1(x))
    end
end
/* Train Stage-2b variance/error prediction
   models                                */
for s in training subjects do
    train S2b_s using D_train
    for x in D do
        get predictions S2b_s(S1(x))
    end
end
/* Ensemble learning on Stage-2 predictions
   using Stage-2b uncertainty estimations  */
```

Over $D_{training}$, learn error matrix to minimize the input dependent loss

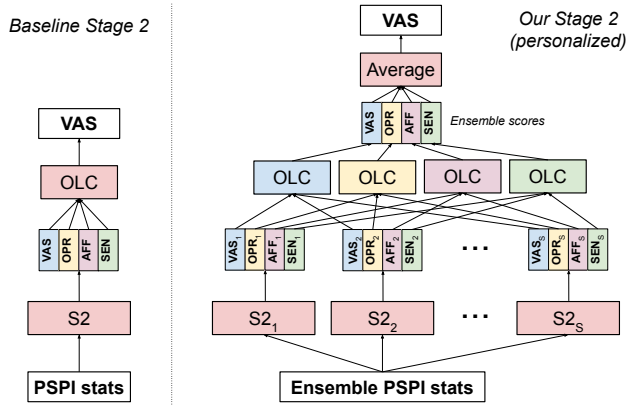


Fig. 3: Stage 2 model structure using individual models. The baseline model uses OLC to combine four pain scores, and we use OLC to combine individual models, and then average the four scores to get the final estimation of VAS. $S2$ and $S2_s$ are fully-connected neural networks with one hidden layer [4].

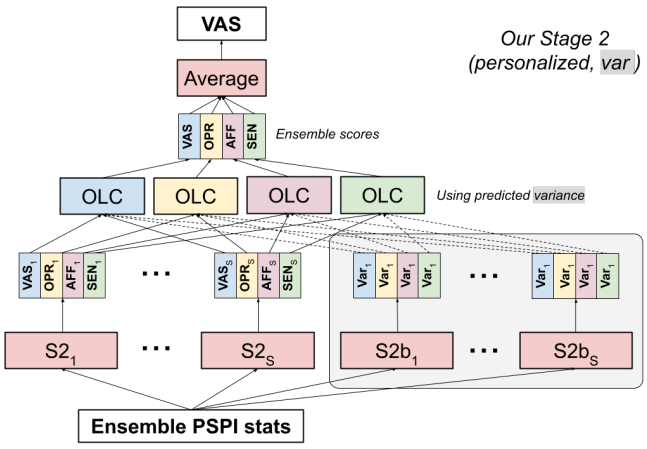


Fig. 4: Stage 2 model structure using individual models and uncertainty estimation. $S2b_s$ models have the same structure as $S2_s$ and learn to predict $(y_s - S2_s(\boldsymbol{x}))^2$ after the $S2_s$ models have been trained. This diagram uses variance predictions as an example. For error prediction models, we simply replace all the "Var" with "Error" in this figure.

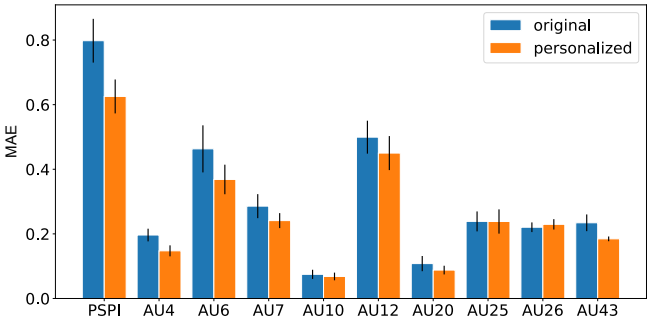


Fig. 5: Stage 1 Performance. Error bars are standard deviations of the MAE metrics over 5 experiments.

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.

We run all our experiments on a single GPU (NVIDIA Titan V). It takes about 2 hours to train each individual $S1$ model on approximately 2000 frames.

The dataset has two types of labels: frame-level labels and sequence-level labels. Frame-level labels include 66 AAM landmarks, 11 facial action unit (AU) [17] intensities and 1 PSPI [2] score. In the first stage of our model, we train individual models to predict PSPI as well as AUs.

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 - An estimate of the VAS given by a human observer of the video) 0-5, AFF (Affective-motivational scale) 0-15 and SEN (Sensory Scale) 0-15. The AFF and SEN measures are designed to separate the emotional and sensory aspects of pain. Their properties are discussed in more detail in [18], [19].

B. Algorithm, Model Training and Evaluation

Our model uses the EMTL model described in [4] as the baseline structure where Stage 1 fine-tunes a VGGFace network with the last layer replaced by a regression layer to predict frame-lavel PSPI and AUs from video frames, and Stage 2 uses a fully connected neural network to estimate sequence-level pain scores using 9 statistics of predicted PSPI from Stage 1. The difference between our model and the EMTL model is shown in Figures 2, 3 and 4.

The training algorithm of our pain estimation model is shown in Algorithm 1. Implementation details such as image pre-processing and optimization methods are the same as [4].

Following [4], we performed 5-fold cross validation with each fold consisting of 5 subjects. We used the same training/test splits for all stages in each iteration. One of the 4 training folds is randomly selected as the validation set during neural network training. After 5 iterations, we concatenated all the test samples and calculated the Mean Absolute Error (MAE), Mean Squared Error (MSE), Intra-class Correlation Coefficient (ICC) and Pearson Correlation Coefficient (PCC).

For all models, we report mean and standard deviation of MAE, MSE, ICC and PCC over 5 separate runs of 5-fold cross validation.

C. Frame-level Pain using Individual Models

For the first stage, we train an individual VGGFace model for each subject. We didn't train from scratch but instead trained a Stage 1 model using all training subjects, and then fine-tuned it for 10 epochs using data from each subject to get the individual model for this subject. The model structure is shown in Figure 2.

The performance of individual models on their own subject's data is good. The training accuracy of individual models are higher on their own data than the training accuracy of the model trained on all subjects. But the test accuracy on subjects not used for training is lower for individual models. This is as expected because individual models can learn personalized distributions better, but won't work so well when used for other subjects.

We apply the optimal linear combination to individually trained models in section II-A to Stage 1, and show better performance on PSPI prediction (Table I) and most AU predictions (Figure 5). We didn't use variance of S1 predictions based on inputs because the validation error of learning $\hat{\sigma}^2$ or $\hat{\varepsilon}$ didn't decrease while training.

D. Sequence-level Pain using Individual Models

After getting predictions of PSPI, we train individual Stage 2 models to predict VAS. The model structure is shown in Figure 3. The VAS prediction performance of the models is shown in Table II.

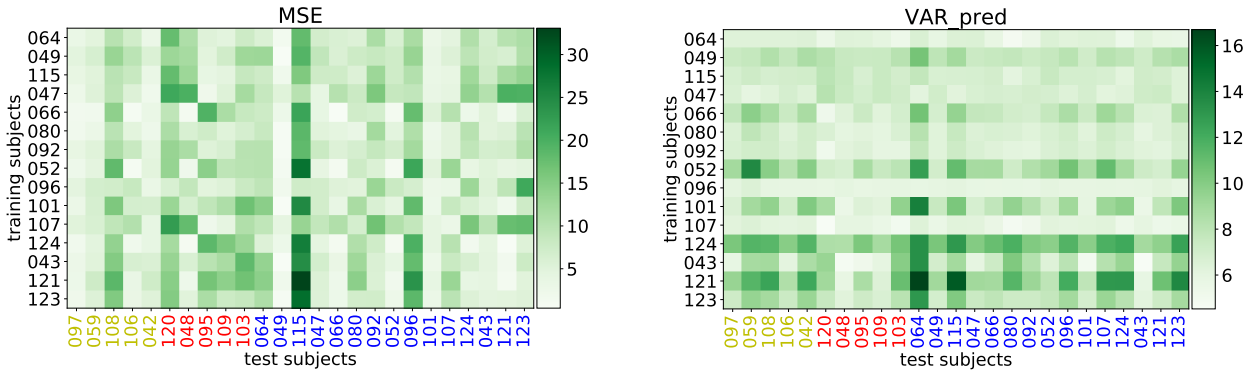
The first row is the original model proposed in [4], and is the previous state-of-the-art. The second row uses personalized models for Stage 1, as described in section III-C, and Stage 2 remains the same except using PSPI predictions learned with optimal linear combination on individual

Stage 1 Model	MAE	MSE	ICC	PCC
Baseline	0.80 ± 0.07	1.53 ± 0.14	0.47 ± 0.04	0.49 ± 0.04
Personalized model	$\mathbf{0.63 \pm 0.05}$	$\mathbf{1.28 \pm 0.11}$	$\mathbf{0.45 \pm 0.05}$	$\mathbf{0.50 \pm 0.05}$

TABLE I: Frame-level PSPI Prediction

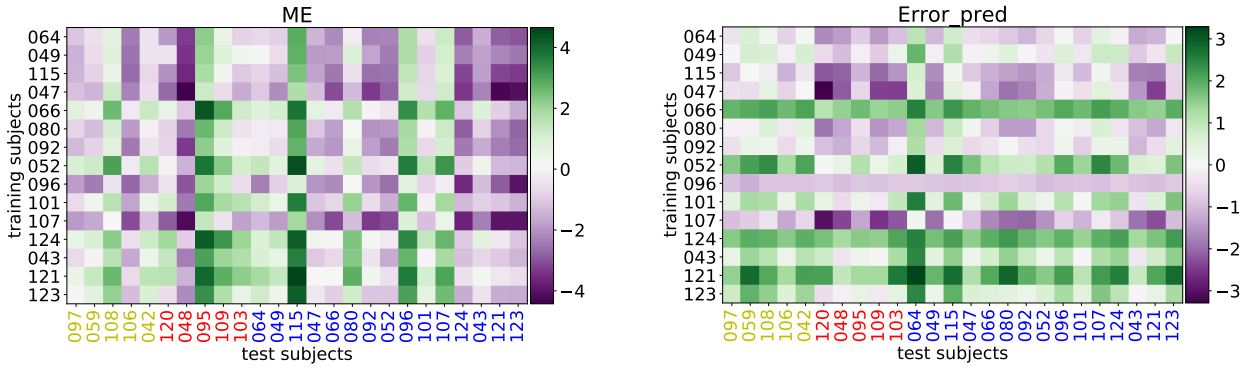
Stage 1 Model	Stage 2 Model	MAE	MSE	ICC	PCC
Baseline [4]	Baseline [4]	1.95 ± 0.06	5.90 ± 0.23	0.43 ± 0.03	0.55 ± 0.03
Personalized [Eq.(4)]	Baseline	1.95 ± 0.07	5.66 ± 0.37	0.46 ± 0.03	0.57 ± 0.04
Personalized	Personalized [Eq.(4)]	$\mathbf{1.88 \pm 0.07}$	5.70 ± 0.47	$\mathbf{0.50 \pm 0.04}$	0.57 ± 0.04
Personalized	Personalized, reg-variance [Eq.(8) with Eq.(6)]	$\mathbf{1.88 \pm 0.07}$	$\mathbf{5.58 \pm 0.37}$	$\mathbf{0.49 \pm 0.04}$	$\mathbf{0.59 \pm 0.04}$
Personalized	Personalized, reg-error [Eq.(8) with Eq.(9)]	$\mathbf{1.88 \pm 0.07}$	$\mathbf{5.57 \pm 0.37}$	$\mathbf{0.50 \pm 0.04}$	$\mathbf{0.59 \pm 0.04}$

TABLE II: Sequence-level VAS Prediction



(a) $(y - \hat{f}_s)^2$ (b) $\hat{\sigma}_s^2$

Fig. 6: Personalized Model MSE of VAS on Individuals. Actual (a) and Predicted (b)



(a) $y - \hat{f}_s$ (b) $\hat{\varepsilon}_s$

Fig. 7: Personalized Model Mean Error of VAS on Individuals. Actual (a) and Predicted (b)

predictions. The performance is better than the first row, showing that learning models tuned to individual faces and combining the outputs with OLC at Stage 1 helps both PSPI prediction and VAS prediction.

The third row uses PSPI predictions based on OLC, as well as individual models in Stage 2 and OLC on top of individual VAS predictions. The performance is further improved. For Stage 2 individual models, each model is trained from scratch on one training subject. This shows that, even without uncertainty estimation, learning individual models and running ensemble learning on top of the individual predictions can improve the performance of the model on unseen test subjects significantly.

In Figure 6(a) we take one fold in one iteration as an example, and plot the MAE of each individual model on each test subject. We can see that although clearly some subjects are generally good as training or test subjects, there are significant differences across subjects, e.g. subject 049 is easy to predict as a test subject, but its performance using the training subject 066 is not as good as subject 106 which is not performing as well using other training subjects. For some of the test subjects, such as subjects 048 and 121, the MAE varies a lot across training models.

It is also not true that a training subject always performs

the best on itself. We don't see a clear diagonal pattern in the square on the right side where the test subjects are in the same order as the training subjects. For example, subject 115 performs better on models trained with subject 047 and 096 than the model trained on itself.

E. Sequence-level Pain using Individual Models and Uncertainty Estimation

In this work, we use the same structure as the Stage 2 sequence-level prediction models to predict the error or variance of the predictions, and the final model is shown in Figure 4. For each personalized model $\hat{f}_s(\boldsymbol{x})$, we train additional models to predict $(y - \hat{f}_s(\boldsymbol{x}))^2$ (var model), or $y - \hat{f}_s(\boldsymbol{x})$ (err model) using all training subjects. These models are denoted as $\hat{\sigma}_s^2(\boldsymbol{x})$ and $\hat{\varepsilon}_s(\boldsymbol{x})$ respectively, and we refer to them as variance predictions and error predictions.

Figure 6 plots the average squared error $(y - \hat{f}_s)^2$ and the average predicted squared error $\hat{\sigma}_s^2$ of each individual model on each test subject. For a test subject, we'd like the variance prediction models $\hat{\sigma}_s^2$ to be able to predict, from the training data, which training models will be more reliable on test data, and they successfully recognize such differences. For example, test subject 115 picks out training subjects 107 and 096 and 047 as having low mean squared error predictions, and would weigh them more using our input-dependent ensemble methods.

Similarly, Figure 7 plots the average (predicted) error. The error prediction models $\hat{\varepsilon}_s$ not only learn the reliability of different individual models $\hat{f}_s$, but also their bias, e.g. they learn that subject 066 is more stoic in facial expression of pain than his VAS score, whereas subject 107 tends to rate his VAS lower than shown in his facial expression.

We show performance using personalization with uncertainty estimations in the last two rows in Table II (regularization using variance/error). For the regularization methods, in practice with enough data, β can be fit using cross-validation. In this work we simply fix $\beta = 1/|D_{training}|$.

The variance over cross-validation folds is large, resulting in relatively large standard deviations in the performance metrics. However, as the train-test cross-validation splits were identical across all models, more sensitive pairwise tests can be used to test for significant improvements in performance with the personalized approaches. Wilcoxon signed-rank one-sided tests indicated that MAE for each full personalized method is lower than the baseline method ($p < 0.0001$)). MSE for both data-dependent uncertainty estimation methods (last two rows in Table II) are lower than our personalized model without uncertainty estimation (row 3 in Table II) with $p < 0.05$.

IV. Conclusion

The relationship between perceived pain and facial expression of that pain is different for different people. In this work we addressed this issue by learning data-dependent personalized models. Personalization is performed at stage one acting on video frames and also at stage two predicting VAS from statistics of the PSPI measure. Uncertainty estimation is used at the second stage to adjust ensemble weights to improve performance on new subjects. Our method improves upon the non-personalized model on the UNBC-McMaster Shoulder Pain dataset and achieves the state-of-the-art performance.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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, 2018.

[6] 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.

[7] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic analysis of facial actions: A survey. *IEEE Trans on Affective Computing*, 2017.

[8] 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.

[9] Sia Rezaei, Abhishek Moturu, Shun Zhao, Kenneth M Prkachin, Thomas Hadjistavropoulos, and Babak Taati. Unobtrusive pain monitoring in older adults with dementia using pairwise and contrastive training. *IEEE Journal of Biomedical and Health Informatics*, 2020.

[10] Ognjen Rudovic, Nicolas Tobis, Sebastian Kaltwang, Björn Schuller, Daniel Rueckert, Jeffrey F Cohn, and Rosalind W Picard. Personalized federated deep learning for pain estimation from face images. *arXiv preprint arXiv:2101.04800*, 2021.

[11] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[12] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.

[13] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. IEEE, 1994.

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017.

[15] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

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

[19] 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.