Learning Pain from Action Unit Combinations: A Weakly Supervised Approach via Multiple Instance Learning

Zhanli Chen
Department of Electrical and
Computer Engineering
University of Illinois at Chicago
Email: zchen35@uic.edu

Rashid Ansari
Department of Electrical and
Computer Engineering
University of Illinois at Chicago
Email: ransari@uic.edu

Diana J. Wilkie
College of Nursing
University of Florida
Email: diwilkie@ufl.edu

Abstract—**Facial pain expression is an important modality for assessing pain, especially when a patient's verbal ability to communicate is impaired. A set of eight facial muscle-based action units (AUs), which are defined by the Facial Action Coding System (FACS), have been widely studied and are highly reliable means for detecting pain through facial expressions. Unfortunately, using FACS is a very time consuming task that makes its clinical use prohibitive. An automated facial expression recognition system (AFER) reliably detecting pain-related AUs would be highly beneficial for efficient and practical pain monitoring. On the other hand, automated pain detection under clinical settings is based on spontaneous facial expressions with limited knowledge about ground truth and can be viewed as a weakly supervised problem, which limits the application of general AFER system that trained on well labeled data. Existing pain oriented AFER research either focus on the individual pain-related AU recognition or bypassing the AU detection procedure by training a binary pain classifier from pain intensity data. In this paper, we decouple pain detection into two consecutive tasks: the AFER based AU labeling at video frame level and a probabilistic measure of pain at sequence level from AU combination scores, which naturally imitates the strategies of human coders in clinical settings. Our work is distinguished in the following aspects, 1) State of the art AFER tools Emotient is applied on pain oriented data sets for single AU labeling. 2) Two different data structures are proposed to encode AU combinations from single AU scores, which forms low-dimensional feature vectors for the learning framework. 3) Two weakly supervised learning frameworks namely multiple instance learning (MIL)[22] and multiple clustered instance learning (MCIL)[24] are employed corresponding to each feature structure to learn pain from video sequences. The results shows a 87% pain recognition accuracy with 0.94 AUC (Area Under Curve) on UNBC-McMaster Shoulder Pain Expression dataset. Tests on long videos in Wilkie's lung cancer patient video dataset suggests the potential value of the proposed system for pain monitoring task under clinical settings.**

Index Terms—**FACS, Action Unit Combinations, Pain, MIL.**

I. Introduction

Assessing pain is a difficult but important task in clinical settings, which practically relies on self-report by patients through simple subjective measures like visual analog scale(VAS). Research has shown that facial expressions can provide reliable measures of pain across human lifespan [6] and there is also good consistency of facial expressions corresponding to pain stimuli. The Facial Action Coding System is widely used in pain analysis, because it provides an objective assessment to score and recognize Action Units (AUs), which represent the muscular activity that produces momentary changes in facial appearance [8]. Several studies[7], [17] using FACS have identified a collection of core Action Units, which are specific to pain and that occur singly or in combination as summarized in Table 1. These results are also confirmed in the study of facial expressions of pain suffered by cancer patients [23]. The facial expression coding using FACS is generally performed offline by trained experts on the video of a patient. A long video is typically divided into multiple subsequences with fixed length and the Action Units are coded at each time step (i.e. each video frame) within the video subsequence. Pain is assessed across the entire sequence based on the occurrence and frequency of pain-related AUs. However the Action Unit coding via human observations is very time consuming, which makes its real-time clinical use prohibitive [13],[1]. Therefore, the development of an automated FACS based pain detection would be a significant and efficient innovation for clinical practice.

TABLE I
Action Unit Definition and Pain-Related AU combinations

AU	Description	Pain-Related Combinations
4	eye brow lower	6/7
6	cheek raiser	20
7	eye lid tightener	4+6/7/43
9	nose wrinkler	4+9/10
10	upper lip raiser	4+26
20	lip stretcher	9/10+26
26	jaw drop	
43	eyes closed	

Over the years, progress of computer vision and machine learning (CVML) techniques has led to significant development of automated facial expression recognition(AFER) systems. Comparing with recognition of basic emotion problems in early days, the focus on spontaneous facial expressions context has extended the application of AFER in solving many practical problems. However, its application to pain analysis is very limited. One major challenge is the difficulty in establishing a comprehensive dataset with sufficient pain-related expressions. Existing video datasets containing pain-related facial expressions are developed under professional projects with no public access, most of which are small in size thus lack sufficient diversity to train a robust automated system in general. One milestone facilitates the research on spontaneous expressions recognition has been the introduction of UNBC-McMaster Shoulder Pain Archieve, which is the only publicly available comprehensive pain oriented facial expression dataset. This dataset records subjects experiencing shoulder pain in a clinical setting with complete labels on both frame level (AUs) and sequence level (OPI VAS etc), and it has been widely employed by FACS based AFER research as a standard dataset for performance evaluation. On the other hand, shoulder pain involved in UNBC-McMaster dataset is acute pain, there is few research suggests whether this data is beneficial to study other type of pain, for example, chronic pain caused by cancer.

Most existing research on FACS based automated AU recognition focuses on detection of single AUs. Pain-related AUs could occur in conjunction with other AUs to form combinations irrelevant to pain, therefore measurement only based on occurrence of individual AUs is not sufficient for pain identification. While the ground truth of facial expressions and action units are available on frame level, the ground truth about pain is typically available on sequence level only via self-report, which is also known as 'weakly labeled'. In the attempts of automated pain analysis, early approaches [13][1] employ an average paradigm by assigning the sequence label to each frame and training a support vector machine (SVM) on the frame level label. Pain is decided to be detected from a video if the average output score of the frames exceeds a threshold. However, pain-related frames could be a small portion in a long video thus averaging output score could attenuate the signal of interest. Recent research[20] suggests that video based pain detection can be formulate as a weakly supervised learning problem and multiple instance learning (MIL) is a preferable machine learning tool to handle this problem. They trained a binary pain classifier directly via the high-dimensional features extracted from video frames without going through the AU coding procedure. Although encouraging results are reported from experiments on UNBC-McMaster dataset, this setting may cause performance degradation for trans-dataset application, due to the undesirable disturbance from person-specific features and demographic variations encoded in the the high-dimensional features.

In a common procedure, FACS certified coders will first perform AU coding for every frame and then infer sequence level pain label from the occurrence and frequency of corresponding AU combinations. In general, action units coding is highly correlated with the appearance of facial expressions, and the reliability of pain detection strongly depends on the accuracy of AU coding. However, pain is a subjective measure and suffering from pain is not necessarily accompanied by facial expressions. Facial expressions of pain are more likely appear when pain intensity level is high or during the transition of pain intensity to a higher level. Due to the sparsity of pain, videos captured under clinical settings usually lacks sufficient positive samples to train a binary pain classifier directly. On the contrary, state-of-the-art AFER systems could have been trained on millions of online images[16], which possesses sufficient compatibility to all kinds of video datasets including pain oriented ones. These observations lead us to design the entire automated pain detection framework in a decoupled structure: a robust generic AFER system for frame level AU encoding followed by a MIL framework trained on simple low-dimensional features formed by the confidence scores of AU combinations. As our major contribution, the decoupled framework alleviates the difficulty in training a dataset specific pain classifier on high-dimensional features with limited positive samples, and may potentially lead to a generic pain detector for analyzing multiple types of pain. Our second contribution is to perform pain detection based on AU combinations rather than single AUs. We propose two distinct low-dimensional feature vectors for AU combination representation: the first structure encodes all AU combinations of interest into a single vector and analyzed by the MIL framework; the second structure uses a sparse representation to encode each AU combination separately and analyzed by the multiple clustereded instance learning (MCIL) framework, which is an extension of MIL. To our best knowledge, this is the first work to apply MCIL in facial expression related research. Our third contribution is to perform a preliminary study on chronic pain by employing the Wilkie's Lung Caner Patient dataset.

The rest of the paper is structured as follows, Section 2 reviews literature on automated pain detection and analysis. Section 3 gives brief overview of two pain oriented datasets involved in this work: the UNBC-McMaster dataset t(acute shoulder pain) and Wilkie's dataset(chronic cancer pain). In Section 4 we presents the pain detection framework in a decoupled structure, with focus on the feature representation based on AU combinations and corresponding learning tools(MIL and MCIL). In Section 5 we demonstrates the advantages of the proposed framework based on the testing results on both datasets in Section 3. Finally the conclusion is given in Section 6.

II. Related Work

In the past decades, significant progress of computer vision and machine learning techniques has boosted the development of AFER system. With increasing demand of facial expression based applications, the highlights of AFER research have shifted from posed expressions obtained under controlled setting to spontaneous expressions evoked under natural settings. More details can be found under the surveys[3][5][25]. A robust AFER system is capable of handling disturbances from demographic difference, environment variation and rigid motions, and is applicable to different dataset without being retrained. Emotient [12] and Affectiva [16] are two examples of state of the art AFER systems which are commercially available as platforms that facilitate the development of facial expression based applications. However, general AU scores from such system are more appropriate as the intermediate results and further customization is required for advanced applications including pain analysis. On the other hand, there has been very limited efforts on exploring machine learning based pain interpretation from facial expressions. While popular CVML tools are widely used in most AFER research, distinct learning methods and measure metrics are developed under individual automated pain detection research. Ashraf et al. [1] studied the UNBC-McMaster dataset and proposed three feature types that are extracted from the Active Appearance Model (AAM) to train SVM pain classifiers and used an averaging scheme to generate sequence level labels. In their follow up research[15][14], the same set of features are used to train a binary classifier for each single pain-related AU on frame level, and a sequence level pain intensity classifier using the OPI labels. Chen *et al*[4] employed a simple rule based method to model temporal dynamics of AUs to study pain of patients suffering from lung cancer in the Wilkie's dataset. Sikka *et al* [19] employed a CVML-based model to assess pediatric postoperative pain on a video dataset of neurotypical youth. 14 single AUs are extracted under 3 statistics to form a 42-dimensional descriptor for each pain event which serves as the input to logistic regression models of both binary pain classification and pain intensity estimation. Sikka *et al* [20] modeled video sequences from UNBC-McMaster dataset with the Bag of Words representation and applied multiple segment multiple instance learning (MS-MIL) for jointly detecting and localizing painful frames using sequence-level ground truth (OPI). There are three problems observed in previous research : (1) Action Unit and pain recognition are treated as separate problem which are handled by AU classier and pain classifier respectively, without given sufficient attention to their casual relationship; (2) Pain analysis is more focused on single AU detection rather than corresponding AU combination; (3) The difficulty in establishing pain oriented video datasets with public accessibility, and insufficient effort on developing advanced comprehensive automated analysis tools to fulfill clinical need.

III. Datasets

The UNBC-McMaster Shoulder Pain Expression Archive Dataset contains 200 videos sequences captured from patients suffering from shoulder pain and spontaneous facial expressions are triggered by moving their affected and unaffected limbs. All frames are FACS coded by certified coders for 10 single pain-related AUs and the frame-level pain score is rated by the Prkachin and Solomon Pain Intensity (PSPI). Sequence level pain label is given by self-reported Visual Analog Scale (VAS) and observer-rated pain intensity (OPI). In addition, 66 point facial landmarks from the Active Appearance Model(AAM) are also provided for each frame to facilitate the development of user customized AFER system. It is the only publicly available pain oriented facial expression video dataset and the spontaneous facial expressions are evoked solely by acute pain.

One major distinguishing feature of this study is that we conduct research on the unique dataset created by D. Wilkie [23], contains videos of 43 patients suffering from lung cancer. The patients were required to repeat a standard set of randomly ordered instructions for action such as sit, stand up, walk, and recline, in a 10-minute video with a camera focused on the face area to record their facial expressions. Each video was partitioned into 30 equal durations of 20 seconds per subsequence. The subsequences were reviewed and scored for 9 AUs occurring in combination by three trained human FACS coders independently, and the results were entered in a scoresheet that served as ground truth. Pain was scored in one video subsequence if at least two coders agree with each other on a set of specific AU combinations listed on the scoresheet. The intensity of pain for the entire video was measured on the total number of subsequences that were associated with pain label. Due to the illumination issue and video quality degradation, 1164 out of the total 1290 video subsequences are processable by Emotient and about 600 subsequences are suitable for pain analysis.

In this paper, we use UNBC-McMaster dataset to train a MIL based pain detector and then test it on Wilkie's dataset [23] with chronic cancer pain. We hypothesize that the pain detector could capture the prevalent feature of pain synthesized from the low-dimensional feature vectors of AU combinations.

IV. Pain Detection Framework

Pain detection is one of the applications that are based on spontaneous facial expression recognition. It will be very beneficial in developing pain detection system if we take advantage of existing progress on spontaneous AFER research. Recent state-of-the-art generic AFER systems [9][16] have proved to be robust to context variations, especially illumination and rigid motions, and have shown good performance on new dataset with different demographics. This is a key motivation for us to adopt commercialized Emotient system. On the other hand, the generic system is not optimized for the application of pain recognitionso that deterministic decisions made directly from the AU scores with a preset threshold could cause a high false alarm rate. Therefore we investigate the second layer of machine learning framework to refine the single AU scores and give a more reliable prediction of pain.

The proposed automated pain detection system is composed of two independent machine learning system, an (Automated Facial Expression Recognition) AFER system that computes frame level confidence scores for single AUs and a Multiple Instance Learning (MIL) system that performs sequence-level pain prediction based on contributions from a set of AU combinations. The entire framework is shown in Fig.1., where both systems are trained on the UNBC-McMaster dataset with respective labels.

AU coding relies on observable facial muscular movements or facial expressions, whereas pain is more like a latent

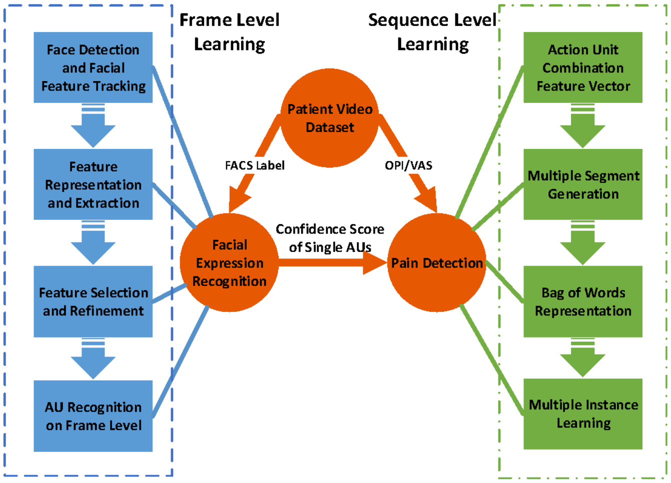
Frame Level Learning
Sequence Level Learning
Face Detection and Facial Feature Tracking
Feature Representation and Extraction
Feature Selection and Refinement
AU Recognition on Frame Level
Patient Video Dataset
FACS Label
OPI/VAS
Facial Expression Recognition
Confidence Score of Single AUs
Pain Detection
Action Unit Combination Feature Vector
Multiple Segment Generation
Bag of Words Representation
Multiple Instance Learning

Fig. 1. The Decoupled Pain Detection Framework

variable which does not necessarily accompanied by facial expressions, especially in the chronic case. Hence learning pain from AU scores in an indirect manner could be more justifiable and efficient than directly identifying pain high dimensional facial features. MIL is a well-suited approach to handle the 'weakly labelled' pain data represented by a bag of word (BOW) structure. In addition, the decoupled framework performs pain analysis in low-dimensional AU score space, which facilitates data fusing from different pain-oriented video datasets and potentially helps to develop a commercial robust generic automated pain analysis system in future research.

A. Automated Facial Expression Recognition

An AFER system typically can be described with four key blocks consisting of face detection, feature representation, feature selection and classification, as shown in Fig.1. The first block identifies face area with rectangular box in every video frames. The second block aligns the detected face areas and employs various descriptors to extract features from the facial images. The third block is responsible to select features most relevant to the non rigid motions caused by facial expressions, and dimension reduction techniques are also necessary to compress the feature vector size to be maneuverable by the classifier. The fourth block contains a set of one versus all classifiers that trained on the refined feature vectors for each AU of interest, and the output could be either binary decisions or soft scores that reflects probability or confidence about the targets. Existing research [2][11][15][10] on spontaneous facial expression recognition shows AFER systems are highly customizable, more blocks could be added to this framework to boost performance depending on the application context.

One example of such an AFER system is the computer expression recogntion toolbox (CERT)[12], which is the core of system we are using. In the system setup, face detection is based on an extentsion of classic Viola-Jones approach. Ten facial feature points are tracked using GentleBoost and the detected face area is aligned to a canonical template patch through an affine warp estimated from the feature positions. A Gabor filter bank is then applied to extract features in 8 directions and 9 spatial frequencies and the filter outputs are concatenated into a single feature vector. The feature vector is then fed into separate linear support vector machines(SVM) for individual AU recognition.

We use Emotient to track and label a set of AUs $\{4, 6, 7, 9, 10, 20, 26, 43\}$, which is commonly used in most pain-oriented research. The processing results are represented by the flow of Evidence numbers, where an Evidence number ranging between -2 to 2 is assigned to each AU on every frame. The Evidence output for an expression channel represents the odds in logarithm (base 10) scale of a target AU being present. For example: Evidence = 2 ($10^2 = 100$) means the observed expression is 100 times more likely to be categorized by an expert human coder as target AU and Evidence = 0 means the chances that the expression to be categorized by an expert human coder as target AU or not are equal. The Evidence scores can be conveniently transformed to Probability measurements by the following equation:

$$Probability = \frac{1}{1 + 10^{-Evidence}} \tag{1}$$

The probability measurements derived from the Evidence scores are used as the features of pain analysis framework. The evidence/probability score profile of every AU can be viewed as a $1 - D$ time domain signal.

B. Action Unit Combination Encoding

*1) **Compact Structure Vs. Clustered Structure**:* Practically, Action Units 6/7 are the most frequently observed pain-related AUs in FACS coding. Multiple AU combinations could also be activated in a video segment for pain evalutation. Furthermore, the Evidence number produced by AFER suggests uncertainty about AU coding, where higher uncertainty could frequently be associated to spontaneous facial expressions due to their low intensity. Hence when we design feature vectors based on AU combinations, it worth to have a comprehensive consideration in terms of frequency, individual contribution and confidence of measurement for each AU combination. Therefore, in the task of designing feature vectors for pain analysis, we not only consider the activation of individual AU combination but also take into account the correlation among activation of multiple AU combinations. Two different feature vector structures based on AU combination scores, which referred to as compact and clustered respectively, are proposed as follows,

Compact Structure: Let $A = \{4, 6, 7, 9, 10, 20, 26, 43\}$ be the set of single pain-related AUs. The AU combination feature vector for frame i of a video sequence S is an 11 dimensional column vector $v_i = \{6, 7, 20, 4+6, 4+7, 4+43, 4+9, 4+10, 4+26, 9+26, 10+26\}$, where each entry of the feature vector is the probability estimate of corresponding AU or AU combination. The probability estimate of $AU(i+j)$ depends on the smaller probability between AUi and AUj so that $P_{AU_{ij}} = min(P_{AU_i}, P_{AU_j}), \forall i, j \in A$. As a result, the pain information about frame i is conveyed by the probability

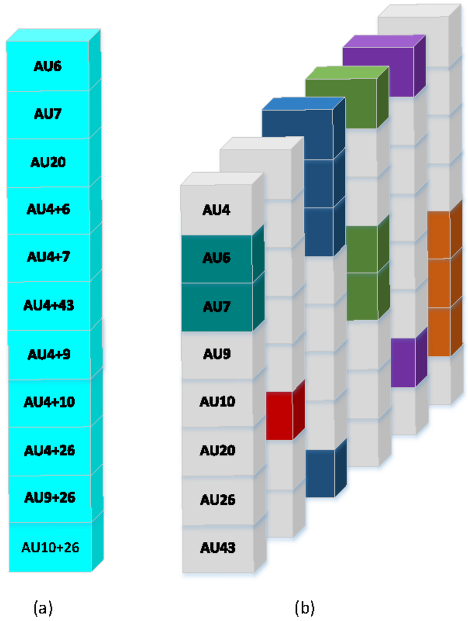

Fig. 2. AU combination structure: (a)Compact Structure, (b)Clustered Structure

measurement of all pain-related AU combinations that are compressed in a single low-dimensional vector, as shown in Fig.3(a)

Clustered Structure: We follows similar coding strategies in Wilkie's dataset to group the 11 pain-related AU combinations into clusters according to two criteria, 1)there is common AU shared by the combinations in the cluster and 2) the AU combinations within a cluster are actuated in adjacent area on the face. Six clusters are formed in this way, including $\{6/7\}, \{20\}, \{4{+}6/7/43\}, \{4{+}9/10\}, \{4{+}26\}$ and $\{9/10{+}26\}$. The AU combination feature of the video frame i under clustered structure is composed of a 8×6 matrix, where the column j is highlighted by all the single AUs involved in the combinations belongs to cluster $j, j \in \{1, 2, 3, 4, 5, 6\}$. The non-zero entry for column j is the probability measurement of the highlighted single AU in cluster j and all the rest entries of the feature matrix are set to zero, which results in a sparse representation of features, as shown in Fig.3(b).

*2) **Bag of Words representation**:* Nowadays, self-report is still the golden rule for pain evaluation in patient care. A pain label is commonly available for video segment without more accurate information about pain occurance per frame. Such a problem occurs frequently in computer vision since it is easier to obtain group labels for the data compare to individual labels, and is known as 'weakly supervised' problem. On the other hand, although temporal dynamics of spontaneous facial expressions have good intra-dataset consistency, it could vary significantly under clinical settings depending on the pain a patient is suffering. While AUs evoked by acute pain occur in less then a second, those evoked by chronic cancer pain could last for minutes. Hence conventional temporal modeling with fixed moving window or a preset duration parameter[13][11] is not sufficient to handle practical applications. To address the challenges from weakly labeled data and complicated temporal dynamics, we employ the bag of word (BOW) representation as suggested in[21][20].

A video sequence S_i can be represented as a bag that contains a number of segments generated from S_i. The bag is defined as $\{s_{ij}\}_{j=1}^{N_i}$ where s_{ij} is the jth segment in the bag $s.t.$ $s_{ij} = \{f_i^k, f_i^{k+1}, \dots, f_i^{N_{ij}-k+1}\}$. s_{ij} contains only contiguous frames and N_{ij} is the total number of frames in S_{ij} taken from S_i and f_i^k is the kth frame in S_i. The bags are then associated with the label of sequence S_i as $B = \{S_i, y_i\}_{i=1}^{N}$, $y_i \in \{-1, 1\}$ which defines two kinds of bags, positive and negative. A positive bag contains at least one positive instance, while a negative bag contains no positive instances. Specifying this representation for the pain detection problem, a positive bag refers to a video sequence containing pain-related facial expressions, and a negative bag refers to a video sequence not containing any pain-related facial expression. Practically, pain-related AU temporal segments only occupy a small portion of the entire video sequence. The sparsity of positive training samples fits well in the context of BOW structure, which is another motivation to adopt this type of data structure. It takes three steps to generate a BOW representation from the feature space.

Feature Extraction at Frame Level: Define a mapping $\phi_{Fr} : R^{m \times n} \to R^d$ as the feature extraction process on frame level that maps frames in image space to a d dimensional feature vector (in the case of clustered structure, the mapping is defined as $\phi_{Fr} : R^{m \times n} \to R^D$, where R^D refers to a 8×6 sparse feature matrix). Feature vector are typically with high dimensions in existing unifying framework. However, R^d is simply the low dimensional AU combination feature vector space in the proposed decoupled pain detection framework.

Multiple Segment Generation: The instances in a bag are video segments containing consecutive frames generated from the sequence. The bound of each segment can be generated conveniently in two ways. A typical way is to run overlapping temporal scanning windows at multiple scales known as Sc-Wind. A parallel way is clustering the frames in a sequence using normalized cuts (Ncuts). Each element of the weight matrix of Ncut algorithm is obtained by a similarity measure between frames f_i^u and f_i^v in sequence i is measured by

$$\begin{aligned} W(u, v) = &\exp\left(-\left|\frac{\phi_{Fr}(f_i^u) - \phi_{Fr}(f_i^v)}{\sigma_f}\right|^2\right) \\ &+ \exp\left(-\left|\frac{t_u - t_v}{\sigma_t}\right|^2\right) \end{aligned} \tag{2}$$

where t_u refers to frame index of f_i^u, and σ_f and σ_t are constants selected for feature domain and time domain respectively. Details of Ncuts are provided in[18] [20].

Feature Representation at Segement Level: The feature representation of video segment is denoted by the mapping $\phi_S : S \to R^d$ that transform video segment in sequence space S to a d dimensional feature vector. This mapping is specified by a max-pooling strategy from the feature representation of all the frames in the segment as:

$$\phi_S(s_{ij}) = \max_k(\phi_{Fr}(f_i^k) \,|\, f_i^k \in s_{ij}) \tag{3}$$

The instance in a bag is now represented by a single feature vector with the same dimension d as the frame-level feature vector. After associating the pain label to the bag, a multiple instance learning(MIL) framework can be trained on the BOW data for automated pain detection.

C. Multiple Instance Learning

The general idea for solving machine learning problems is to establish a classifier and optimize it with respect to a loss function. Viola *et al.*[22] first solved the MIL problem with a boosting framework, which is known as MILboost, and discussed its application in object detection from images. In this section, we give a brief overview of MILboost and how it can be customized for pain detection. The decision of presence of pain is based on the probability of bags been positive. The posterior probabilities of bags and instances are defined as:

$$p_i = \mathcal{P}(y_i = 1|S_i) \tag{4}$$

$$p_{ij} = \mathcal{P}(y_i = 1|s_{ij}) \tag{5}$$

The only available ground truth is the label of the bag, and all the instances in a bag carry the same label as the bag.

A classifier $H_T : R^d \to R$ is trained on the feature vectors of instances, and a posterior probability is assigned to each instance based on the classifier output, $s.t.$

$$p_{ij} = \sigma(H_T(\phi_s(s_{ij}))) \tag{6}$$

where $\sigma()$ is a sigmoid function $s.t.$

$$\sigma(x) = \frac{1}{1+\exp(-x)}, \, \forall x \in R \tag{7}$$

The loss function is defined by negtive log-likelihood, which is the same as used in logistic regression problem:

$$\mathcal{L} = -\sum_i^N (r_i * p_i + (1-r_i)(1-p_i)) \tag{8}$$

where $r_i = 1$ if $y_i = 1$ and $r_i = 0$ if $y_i = -1$. Since a positive bag contains at least one positive instance, the probability of bag to be positive p_i depends on the probability of the instance that is most likely to be classified as positive, i.e.

$$p_i = \max_j(p_{ij}) \tag{9}$$

The MILboost uses the boosting procedure to construct a strong classifier $H_T(s_{ij})$ by iteratively combining a set of weak classifiers $h_t(s_{ij})$ as,

$$H_T(s_{ij}) = \sum_{t=1}^T \alpha_t h_t(s_{ij}) \tag{10}$$

where H_T denotes the classifier constructed in the T^{th} iteration and all weak classifiers are the same type of learners that are generated from space $\mathcal{H}$. Note $H_T(s_{ij})$ and $h_t(s_{ij})$ are simplified notations of $H_T(\phi_s(s_{ij}))$ and $h_t(\phi_s(s_{ij}))$ respectively, and similar notation will be used in the following derivations.

The boosting algorithm updates the weight of instances at the end of each iteration by taking the gradient of the loss function $\mathcal{L}$ $w.r.t$ $H_T(s_{ij})$

$$\omega_{ij} = -\frac{\partial \mathcal{L}}{\partial H_T(s_{ij})} = -\frac{\partial \mathcal{L}}{\partial p_i}\frac{\partial p_i}{\partial p_{ij}}\frac{\partial p_{ij}}{\partial H_T(s_{ij})} \tag{11}$$

The instance weights are then normalized as $\omega'_{ij} = \frac{|\omega_{ij}|}{\sum_{ij}|\omega_{ij}|}$. Misclassified instances will be assigned higher weights and a weak classifier $h_{T+1}(s_{ij})$ is trained on the reweighed data and added to H_{T+1} in the $(T+1)^{th}$ iteration

$$h_{T+1} = \arg\max_{h \in \mathcal{H}} \sum_i^N \sum_j^{N_i} \omega'_{ij} h(s_{ij}) \tag{12}$$

where N is the total number of bags and N_i is the number of instances in the ith bag. However, the $\max$ function is not differentiable, a soft-max function has to be used as an approximation, which is noted as

$$p_i = g_j(p_{ij}) \approx \max_j(p_{ij}) \tag{13}$$

Among all the options for soft-max function, the generalized mean(GM) is a preferable model as suggested by past research[20]. For the instances$\{s_{ij}\}_{j=1}^{N_i}$ in the bag of S_i, the GM approximation is given by:

$$g_{GMj}(p_{ij}) = (\frac{1}{N_i}\sum_j p_{ij}^u)^{\frac{1}{u}} \tag{14}$$

where u is the parameter controls sharpness and accuracy in GM model $s.t.$ $g_{GM}(p_{ij}) \to \max(p_{ij})$ as $u \to \infty$. Now the gradient of the GM soft-max is given by:

$$\frac{\partial p_i}{\partial p_{ij}} = p_i \frac{p_{ij}^{(u-1)}}{\sum_{s=1}^{N_i} p_{is}^u} \tag{15}$$

D. Multiple Clustered Instance Learning

In the compact feature settings, scores of all the AU combinations are encoded in one feature vector, which can be conveniently handled by the original MIL framework. However, it may be desirable to distinguish the contribution of individual AU combinations for more precise analysis on different type of pain. Practically, the clustered representation is a more natural way used by human coders and a positive decision on any of the clusters is sufficiently to identify pain. The multiple clustered instance learning (MCIL) proposed by Xu *et al.*[24] is an extension of existing MIL that provides patch-level clustering of 4 subclasses of cancer tissues, and facilitates both image-level classification and pixel level segmentation (cancer vs. non-cancer). Based on the similarity of the problems, we employ MCIL to handle the clustered feature structure for pain recognition.

MCIL assumes there are K clusters in a positive bag and existing a hidden variable $y_{ij}^k \in \{-1, 1\}$ that denotes whether the instance s_{ij} belongs to the kth cluster. A instance could be considered as a positive instance if it belongs to one of the K clusters and a bag is labeled as positive bag only if it contains at least one positive instance. The goal of MCIL is to learn one boosting classifier $H_T^k(s_{ij}^k)$ for each of the K cluster. In our clustered data representation settings, each video frame is encoded by a 8×6 matrix. If we treat each column vector as an independent instance, all the instances in one bag will form six clusters naturally. A six-cluster MCIL learner can be trained and the overall decision is on the cluster classifier that gives maximum output:

$$H_T(s_{ij}) = \max_k (H_T^k(s_{ij})) \tag{16}$$

Similar to the core of MIL, the posterior probability of bag i is given by,

$$p_i = \max_j \max_k (p_{ij}^k) \tag{17}$$

and the max function is approximated by soft-max function.

$$p_i = g_j(p_{ij}) = g_j(g_k(p_{ij}^k)) = g_{jk}(p_{ij}^k) \tag{18}$$

Note that the order of soft-max functions is interchangeable, i.e.

$$g_j(g_k(p_{ij}^k)) = g_{jk}(p_{ij}^k) = g_k(g_j(p_{ij}^k)) \tag{19}$$

Taking GM model as an example, the proof goes as follows,

$$\begin{aligned} g_k(g_j(p_{ij}^k)) &= (\frac{1}{K}\sum_k (p_i^k)^u)^{\frac{1}{u}} \\ &= (\frac{1}{K}\sum_k ((\frac{1}{N_i}\sum_j (p_{ij}^k)^u)^{\frac{1}{u}})^u)^{\frac{1}{u}} \\ &= (\frac{1}{KN_i}\sum_{k,j} (p_{ij}^k)^u)^{\frac{1}{u}} = g_j(g_k(p_{ij}^k)) \end{aligned} \tag{20}$$

Finding the strong classifier of a cluster follows standard boosting procedure and all the classifiers are trained on the same set of BOW instances. However, the weight of instances are updated respectively as per cluster,

$$\omega_{ij}^k = -\frac{\partial \mathcal{L}}{\partial H_T^k(s_{ij})} = -\frac{\partial \mathcal{L}}{\partial p_i}\frac{\partial p_i}{\partial p_{ij}^k}\frac{\partial p_{ij}^k}{\partial H_T^k(s_{ij})} \tag{21}$$

The partial derivative of $\frac{\partial p_i}{\partial p_{ij}^k}$ for the GM model is given by,

$$\frac{\partial p_i}{\partial p_{ij}^k} = p_i \frac{(p_{ij}^k)^{(u-1)}}{\sum_{s=1}^{N_i}\sum_{t=1}^{K}(p_{is}^t)^u} \tag{22}$$

For the rest two items in the partial derivative of the weight update expression, $\frac{\partial \mathcal{L}}{\partial p_i}$ is the same as in MILboost and $\frac{\partial p_{ij}^k}{\partial H_T^k(s_{ij})} = p_{ij}^k(1 - p_{ij}^k)$, which is the derivative $w.r.t$ a sigmoid function.

V. Experiment

The pain detection system is first tested on the UNBC-McMaster dataset, where video sequences with OPI ratings≥ 3 are treated as positive samples and those with OPI$= 0$ are treated as negative samples. This yielded the same set of 147 sequences from 23 subjects as in [20]. Video sequences are processed by Emotient and the output of single AU Evidence score dataflows are encoded with the compact and clustered AU combination structures separately. These two type of BOW features are used to train MIL (refered to as Compact-MIL) and MCIL (referred to as Clustered-MCIL) learners respectively. Instances in a bag are generated by two temporal aggregation methods, Sc-wind and Ncuts. We set the multiple scaling window size at $30, 40, 50$ for Sc-wind. The size of segment in a cluster is limited between 21 and 81 for Ncuts, and $\sigma_f = 0.1$, $\sigma_t = 30$ in computing the frame correlation matrix. The GM soft-max function is adopted for the approximation of $\max$ in MIL training. The performance of MIL and MCIL learners are evaluated by accuracy and area under curve (AUC). We use the work in [20],referred to as MS-MIL, for comparison and the results based on a 10-fold validation are summarized in Table 2.

TABLE II
Comparison of the decoupled framework with MS-MIL[20]

Framework	Sc-wind		Ncut	
	Accuracy(%)	AUC	Accuracy(%)	AUC
MS-MIL	83.7		82.99	
Compact-MIL	83.96	0.875	85.04	0.9
Cluster-MCIL	85.18	0.92	**86.84**	**0.94**

The best performance is achieved by the Clustered-MCIL framework in conjunction with Ncuts. The improvement on classifier accuracy may be attributed to the advantage of the decoupled structure. AU coding is a 'hard' problem involving learning a mapping from a very high-dimensional pixel space to the low dimensional AU space under complicated environment. However, due to the fact that AUs are evoked by certain facial muscular movement, reliable AUs coding can be achieved by a robust AFER system if trained on sufficient data. On the other hand, pain is a subjective measure and can be viewed as a latent variable. It will be easier to synthesis similarity from purified low-dimensional features under the impact of problem uncertainty. In addition, the improvement on AUC between Clustered-MCIL and Compact-MIL could be attributed to the feature sparsity from the clustered representation, which not only increases the margin on features but also follows more naturally to human coder's decisions.

Next, we employ the Clustered-MCIL settings with GM approximation to train a pain detector on UNBC-McMaster dataset and test it on video sequences from selected patients in Wilkie's dataset. Each testing sequence is divided into 30 subsequences and each subsequence has a duration of for 20 seconds. Three human coders perform FACS coding on each subsequence and pain is identified if any pain-related AU combination is detected by at least two coders in the original research. However, since the human coding does not reveal pain intensity information and coders do not always agree to each other, this ground truth is more suitable for qualitative analysis. 393 subsequences from 27 patients are selected for this experiment, where at least 50% of each subsequence

TABLE III

COMPARISON OF MACHINE PREDICTION WITH HUMAN CODER DECISION ON SUBSEQUENCES IN WILKIE'S DATASET[23]

Subsequences Label	Pain (2+ coders scored)	Pain (1 coder scored)	No Pain (3 coders agreed)
Human Coder	82	121	190
Automated System	68	83	169
Consistency Rate	82.9%	68.6%	88.9%

are analyzable by Emotient. A subsequence is considered as positive (pain) if AU combinations are coded by at least two coders (more crediable) or only one coder (less credible). A subsequence is negative (no pain) if no AU combination is scored by any coder. As a result, 82 subsequences are identified as positive by at least two coders, 121 subsequences are identified as negative by only one coder, and the rest 190 subsequences are identified as negative samples. The automated system is used as an independent coder and check the consistency between machine prediction and human coder decisions and the results are summarized in Table 3. Note the system has no prior knowledge about the testing dataset. In general, We observe the decisions of automated system are highly correlated with that of majority human coders on both pain and no pain videos. Additionally, a 68.6% consistency rate is observed between the system and the only coder, however this is more likely due to the ambiguity on low credible videos rather than to the accuracy degradation. As a result, the system shows its potential in pain assessment for long videos under clinical settings and we will conduct advanced tests with focus on patient monitoring in future research.

VI. CONCLUSION

This paper proposed and investivatedthe performance of an automated pain detection via spontaneous facial expressions in the context of clinical applications. The proposed framework imitates the decision of FACS certified coder by following the procedures of *Facial Expression →AU→AU Combination→Pain*. Due to the difficulty in constructing pain oriented video dataset, we proposed a decoupled structure for automated pain detection task: 1) the AU coding from facial expressions takes full advantage of the AFER development, 2)pain detection is based on simply low-dimensional features and handled by MIL as a weakly supervised problem. The proposed system not only demonstrates improvement on existing state of the art work, but also shows adaptiveness on trans-dataset learning and long video analysis. In future work, we will conduct comprehensive test on new pain oriented videos datasets and investigate practical methodologies that facilitates clinical pain analysis using our system.

ACKNOWLEDGMENT

This research and publication was made possible by Grant Number P30 NR010680-S1 from the National Institutes of Health, National Institute of Nursing Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Institute of Nursing Research.

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] Marian Stewart Bartlett, Gwen Littlewort, Mark G Frank, Claudia Lainscsek, Ian R Fasel, and Javier R Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of multimedia*, 1(6):22–35, 2006.

[3] Konstantinos Bousmalis, Marc Mehu, and Maja Pantic. Towards the automatic detection of spontaneous agreement and disagreement based on nonverbal behaviour: A survey of related cues, databases, and tools. *Image and Vision Computing*, 31(2):203–221, 2013.

[4] Zhanli Chen, Rashid Ansari, and Diana J Wilkie. Automated detection of pain from facial expressions: a rule-based approach using aam. In *SPIE Medical Imaging*, pages 83143O–83143O. International Society for Optics and Photonics, 2012.

[5] Ciprian Adrian Corneanu, Marc Oliu Simón, Jeffrey F Cohn, and Sergio Escalera Guerrero. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications. *IEEE transactions on pattern analysis and machine intelligence*, 38(8):1548–1568, 2016.

[6] KENNETH D Craig, Kenneth M Prkachin, and Ruth VE Grunau. The facial expression of pain. *Handbook of pain assessment*, 2:257–276, 1992.

[7] Kathleen S Deyo, Kenneth M Prkachin, and Susan R Mercer. Development of sensitivity to facial expression of pain. *Pain*, 107(1):16–21, 2004.

[8] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[9] iMotions A/S. imotions biometric research platform 6.0. 2016.

[10] Bihan Jiang, Michel F Valstar, and Maja Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321. IEEE, 2011.

[11] Sander Koelstra, Maja Pantic, and Ioannis Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE transactions on pattern analysis and machine intelligence*, 32(11):1940–1954, 2010.

[12] Gwen Littlewort, Jacob Whitehill, Tingfan Wu, Ian Fasel, Mark Frank, Javier Movellan, and Marian Bartlett. The computer expression recognition toolbox (cert). In *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 298–305. IEEE, 2011.

[13] Patrick Lucey, Jeffrey Cohn, Simon Lucey, Sridha Sridharan, and Kenneth M Prkachin. Automatically detecting action units from faces of pain: Comparing shape and appearance features. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 12–18. IEEE, 2009.

[14] Patrick Lucey, Jeffrey F Cohn, Kenneth M Prkachin, Patricia E Solomon, Sien Chew, and Iain Matthews. Painful monitoring: Automatic pain monitoring using the unbc-mcmaster shoulder pain expression archive database. *Image and Vision Computing*, 30(3):197–205, 2012.

[15] 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 *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 57–64. IEEE, 2011.

[16] Daniel McDuff, Rana Kaliouby, Thibaud Senechal, May Amr, Jeffrey Cohn, and Rosalind Picard. Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 881–888, 2013.

[17] Kenneth M Prkachin. Assessing pain by facial expression: facial expression as nexus. *Pain Research and Management*, 14(1):53–58, 2009.

[18] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

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

[20] Karan Sikka, Abhinav Dhall, and Marian Stewart Bartlett. Classification and weakly supervised pain localization using multiple segment representation. *Image and vision computing*, 32(10):659–670, 2014.

[21] Karan Sikka, Tingfan Wu, Josh Susskind, and Marian Bartlett. Exploring bag of words architectures in the facial expression domain. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 250–259. Springer, 2012.

[22] Paul Viola, John C Platt, Cha Zhang, et al. Multiple instance boosting for object detection. In *NIPS*, volume 2, page 5, 2005.

[23] Diana J Wilkie. Facial expressions of pain in lung cancer. *Analgesia*, 1(2):91–99, 1995.

[24] Yan Xu, Jun-Yan Zhu, I Eric, Chao Chang, Maode Lai, and Zhuowen Tu. Weakly supervised histopathology cancer image segmentation and classification. *Medical image analysis*, 18(3):591–604, 2014.

[25] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31(1):39–58, 2009.