Weakly Supervised Pain Localization using Multiple Instance Learning

Karan Sikka, Abhinav Dhall and Marian Bartlett

Abstract— **Automatic pain recognition from videos is a vital clinical application and, owing to its spontaneous nature, poses interesting challenges to automatic facial expression recognition (AFER) research. Previous pain vs no-pain systems have highlighted two major challenges: (1) ground truth is provided for the sequence, but the presence or absence of the target expression for a given frame is unknown, and (2) the time point and the duration of the pain expression event(s) in each video are unknown. To address these issues we propose a novel framework (referred to as MS-MIL) where each sequence is represented as a bag containing multiple segments, and multiple instance learning (MIL) is employed to handle this weakly labeled data in the form of sequence level ground-truth. These segments are generated via multiple clustering of a sequence or running a multi-scale temporal scanning window, and are represented using a state-of-the-art Bag of Words (BoW) representation. This work extends the idea of detecting facial expressions through 'concept frames' to 'concept segments' and argues through extensive experiments that algorithms like MIL are needed to reap the benefits of such representation.**

The key advantages of our approach are: (1) joint detection and localization of painful frames using only sequence-level ground-truth, (2) incorporation of temporal dynamics by representing the data not as individual frames but as segments, and (3) extraction of multiple segments, which is well suited to signals with uncertain temporal location and duration in the video. Experiments on UNBC-McMaster Shoulder Pain dataset highlight the effectiveness of our approach by achieving promising results on the problem of pain detection in videos.

I. INTRODUCTION

Pain is difficult to examine and considered critical in clinical settings. It is a subjective measure and is often reported by patient self-report, either through clinical interview or visual analog scale (VAS). However these self-report measures have several drawbacks like idiosyncratic use, subjective differences, [1] etc. Hence there has been a considerable research effort to identify and quantify pain [2]. However most of these methods entail manual labeling of facial action units and are time consuming.

Over the years there has been a significant progress in analyzing facial expressions related to emotions using machine learning (ML) and computer vision [3]. Most of those works have focused on posed facial expressions that are obtained under controlled lab settings and differ from spontaneous facial expression in both which muscles are moved and dynamics of the movement [4], [5]. We refer our readers to a survey on automatic facial expression recognition (AFER) by Bartlett et.al [5] that has identified the difficulties faced by AFER on spontaneous expressions. A major challenge of spontaneous expressions is temporal segmentation of the target expressions. Videos may exist in which the target emotion or state was elicited, but the onset, duration, and frequency of facial expressions within the video are unknown.

A significant progress in research on spontaneous expressions has been the introduction of UNBC-McMaster Shoulder Pain dataset [6] that involves subjects experiencing shoulder pain in a clinical setting and this work focuses on pain vs no-pain detection in videos on this dataset. In this dataset each video is labeled for presence or absence of pain, but there is no information about the location or duration of facial expressions within each video. This poses a challenge for training sliding window classifiers and further limits the performance of the standard approach of obtaining fixed length features through averaging and training a classifier. Previous approaches [7], [8] follow a common paradigm of assigning each frame the label of the corresponding video and using them to train a support vector machine (SVM). Pain is detected in a video if the average output score (distance from separating hyperplane) of member frames is above a pre-computed threshold. Such approaches suffer from two major limitations: (1) not all frames in a video have the same label, (2) averaging output scores across all the frames may dampen the signal of interest. This paper proposes to address these challenges by employing multiple instance learning (MIL) [9].

MIL is one of the approaches for handling 'weakly labeled' training data. In such cases the training data only specifies the presence (or absence) of a signal of interest in the data without indicating where it might be present. For instance in the case of pain vs no-pain detection, a sequence label only specifies if a subject is/not in pain without any details regarding locations or duration of pain. Other techniques for tackling weakly labeled data includes part-based models [10] and latent models like pLSA and LDA [11]. Most of these approaches try to identify the signal of interest by inferring the values of some latent variables while minimizing a loss function. MIL was introduced to address the problem of weakly supervised object detection [9] [12]. Compared to other approaches, MIL offers a tractable way to train a discriminative classifier that avoids complex inference procedures. MIL has been successfully employed for face recognition from video [9] and more recently has been proposed for handling labeling noise in video classification [13].

This work was support by NSF grant IIS-0905622.

Karan Sikka and Dr. Marian Bartlett are affiliated with Machine Perception Lab at University of California, San Diego, USA.

Email: {ksikka,mbartlett@ucsd.edu}

Abhinav Dhall is affiliated with Research School of Computer Science, Australian National University, Australia.

Email: {abhinav.dhall@anu.edu.au}

Here we apply MIL to the problem of detecting spontaneous facial expression in video. We combine it with a dynamic extension of concept frames, into a novel framework called multiple-segment multiple instance learning (MS-MIL). Our major contributions are as follows:

1) Inherent challenges in previous approaches for pain detection in videos have been identified and a suitable pipeline has been proposed to address these concerns. The most salient feature of our approach is that it can jointly detect and localize pain by using only sequence level labels.
2) In order to address the demand of pain detection task, we have proposed to represent each video as a bag containing multiple segments which are modeled using MIL. The multiple segment based representation and MIL are able to address the potential problem with spontaneous expressions, like pain, that can have uncertain locations, durations and occurrences.
3) We compare the performance MS-MIL with other approaches for pain detection in videos and also highlight the localization capability of our approach by comparing per frame prediction from our algorithm with ground-truth pain intensities for 4 cases. Results indicated the performance advantages obtained using our approach for pain classification and shows appreciable localization capabilities.

II. Pain vs No-Pain Detection in Videos

Our experiments employ data from the UNBC-McMaster Pain Shoulder Archive that was distributed to the research community in [6], and included 200 sequences from 25 subjects. Each subject was undergoing some kind of shoulder pain and was asked to perform a series of active and passive movements of their affected and unaffected limbs. Active tests were self-initiated shoulder movements and in passive tests the physiotherapist was responsible for the movement. For complete details of the experimental setting we refer the readers to [6].

These sequences were then coded on a number of levels by experts. The coding of interest to this work is the Observer Pain Intensity (OPI) rating that was assigned to each sequence on a level of 0 (no-pain)-5 (strong pain) by an independent observer trained in identification of pain expressions. Following the protocol proposed in [7] [8], labels were binarized into 'pain' and 'no pain' by defining training instances with OPI$\geq$ 3 as the positive class (Pain) and OPI$=0$ as the negative class (No-Pain). Only those subjects were included in our experiments who had a minimum of one trial with an OPI rating of 0 (no pain) and one trial with an OPI rating of either 3, 4 or 5 (pain). Intermediate pain intensities of 1 and 2 were omitted, per the protocol in [7] [8]. This yielded 147 sequences from 23 subjects for our experiments.

III. Related Work and Motivation

One of the first works on automatic pain detection in videos is by Ashraf et.al [8]. Their approach starts by first extracting AAM based features from each frames and using these to cluster the frames in order to create a training data whose size is manageable by SVM. Following this, each of these clustered frames are assigned the label of their corresponding sequence and used to train a linear SVM. Finally during prediction each test-frame is assigned a score based on its distance from separating hyperplane. Then a test-video is predicted as pain if the average score of its member frames exceeds a threshold. Lucey et.al [7] extended this work by highlighting that temporal information is enhanced by compressing the signal in spatial rather than temporal domain. They borrowed ideas from the related field of visual speech recognition and proposed to compress the signal in the spatial rather than temporal domain using the Discrete Cosine Transform (DCT). Lucey et.al [7] used the system in [8] as their baseline system and showed significant improvement in performance using their idea.

The first limitation of earlier work is the ambiguity introduced by weakly labeled data where each member frame is assigned the label of the sequence and such approaches lead to a lower performance compared to the case when ground-truth for each frame is known [8][14]. We address this particular concern by proposing to use MIL that has been designed specifically to handle weakly labeled data.

Secondly, [7] highlighted that incorporating the dynamics of the pain signal is difficult since there is no information about the number of times pain expressions can occur or their location and duration in a sequence. Following this, [7] suggested to add temporal information by appending adjacent frames onto the frame of interest, as input to the SVM [15]. [7] tested this idea of appending adjacent frames in their paper, however they found that their performance degraded. One possible explanation is that SVM classifiers are not well suited to weakly labeled training data and may suffer from mislabels when the data is in this form.

Motivated by the last idea we propose to incorporate temporal dynamics by representing each sequence not as individual frames (as done earlier) but as sets of frames, referred to as 'multiple segments'. The benefits of such a representation are reaped by using MIL, which can efficiently handle data in such form. Since MIL handles data as bags, we can visualize every sequence as a bag containing multiple segments. Multiple segments (MS) has two fold advantages: (1) it allows pain to have random duration and occurrence, and (2) it incorporates temporal information by pooling across multiple frames in a segment.

A third limitation of earlier work is the way in which prediction is done for each sequence using the average decision score of its frames. Such an approach may not be optimal in all situations since averaging operation tends to dampen the signal of interest. The MIL framework employed in this work avoids this limitation based on its inherent property of using $\max$ operation to predict the label of a bag based on the posterior probability of its instances (see Section. IV).

In a recent paper [16] Tax et al. explored the question of whether it is always necessary to fully model the entire

sequence, or whether the presence of specific frames, called 'concept frames', might be sufficient for reliable detection of facial expressions. In their study two different approaches for AFER were investigated: (1) modeling full sequences using approaches like Hidden Markov Models and Conditional Random Fields, and (2) modeling only certain frames, for AU detection in sequences. The author in [16] also suggested that for modeling only particular key frames, algorithms like MIL are required and investigated one such approach. Through extensive experiments the authors showed that for reliable classification, modeling certain key frames is sufficient compared to modeling entire sequence. A limitation of 'concept frames', however, is that they do not incorporate temporal information, which could potentially be exploited by learning algorithms such as MIL (and to some extent SVM [17]).

The present paper takes a leap forward by proposing a dynamic variant of 'concept frames'. Here we extend the idea of 'concept frames' to 'concept segments' consisting of multiple frames. These 'concept segments' can be thought of as localized sub-expressions that contain the expression of interest in a sequence. Hence we state the inherent research direction in this work as: 'Reliable detection of facial expression can be achieved by detection of key localized segments using tailored algorithms like MIL'. It is also worth mentioning the work in [17] where a segment based approach, called k-Seg SVM, was explored. Structured-SVM was employed to detect temporal events (AU segments in their case). Our work differs from this work in several respects, most notably that [17] is a completely supervised algorithm requiring location information in the training data, whereas the approach presented here operates on weakly labeled data.

IV. MIL

The general machine learning paradigm involves finding a classification function that minimizes a loss function $\mathcal{L}(D, h(x))$ over training data provided as N samples along with their corresponding labels, $D = \{x_i, y_i\}_{i=1}^{N}$, where $x_i \in X$ and $y_i \in Y$. On the other hand, the MIL paradigm is designed to handle problems involving training data in the form of bags, $B = \{X_i, y_i\}_{i=1}^{N}$, where $X_i = \{x_{ij}\}_{j=1}^{N_i}$, $y_i \in Y$ and N_i is the number of instances in X_i. Since this work deals with binary classification problem, we shall use the output space $Y \in \{-1, 1\}$. Such problems occur frequently in computer vision since it is easier to obtain a group label for the data compared to individual labels and such labels can also suffer from annotator bias and noise [13]. Recently several works have adopted MIL to address these concerns in domains like handling label noise in video classification [13], face recognition in videos with subtitles [18], object localization [12], etc.

As shown in Fig. 1, the MIL framework defines two kinds of bags, positive and negative, in a fashion similar to positive and negative instances in traditional ML. A bag is positive if it contains at least one positive instance, while a negative bag contains no positive instances.

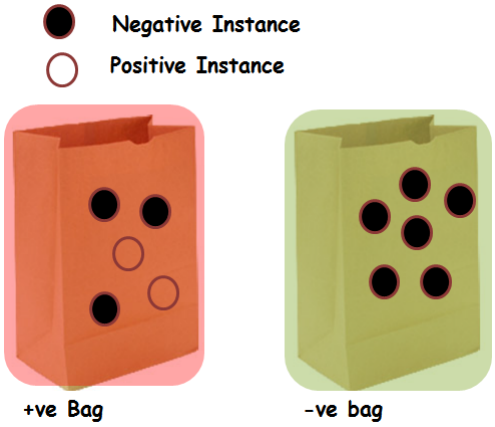


Fig. 1: Figure showing positive and negative bags used in MIL. A positive bag contains one positive instance and negative contains only negative instance.

We employed the Multiple Instance Learning based on boosting (MilBoost) algorithm proposed by Viola et.al [9] for this work. In next sections we shall give a brief overview of MilBoost algorithm which is based on Friedman's gradient boosting framework [19].

A. MilBoost

MilBoost combines the gradient boosting framework with the idea of MIL to handle training data as bags. The i^{th} bag is denoted by X_i and j^{th} instance inside it is represented as x_{ij}. The posterior probabilities over bags and instances are defined as:

$$p_i = Pr(y_i = 1/X_i) \tag{1}$$

$$p_{ij} = Pr(y_i = 1/x_{ij}) \tag{2}$$

We shall be using the original formulation defined in [9] for the loss function given by negative log-likelihood:

$$\mathcal{L} = -\sum_{i}^{N} t_i \log p_i + (1 - t_i) \log (1 - p_i) \tag{3}$$

where $t_i = 1$ if $y_i = 1$ and $t_i = 0$ if $y_i = -1$. This is the same loss function used in methods like logistic regression.

This formulation for loss function seems intuitive since the only information available about a MIL dataset is label information for each bag (y_i). We lack any information about the probabilities (or labels) of individual instances (p_{ij}), which can also be seen as latent variables, whose values are inferred during the boosting process.

Since a positive bag contains atleast one positive instance, the probability of bag being positive (p_i) is defined in terms of individual instances as:

$$p_i = \max_j(p_{ij}) \tag{4}$$

Since the max function is not differentiable, number of differentiable approximations to the max function have been proposed for MilBoost [9], [18], [20]. In this work we shall refer to these approximations as soft-max functions $g(p_{ij})$.

The most common choice of soft-max function is noisy-or (NOR).

A major disadvantage with NOR is that it deviates from the $\max$ function as size of the bag increases which we shall refer to as 'bagsize-bias'. To realize this shortcoming we consider a toy example which consists of two bags B_1 and B_2 of sizes of 3 and 5. The instance probabilities for these bags are given by $B_1 = [.15\ .15\ 2]$ and $B_2 = [.15\ .15\ .15\ 2]$. As is evident the $\max$ for both cases is 2 however the NOR formulation yields maximas as $.42$ and $.50$ respectively. This observation clearly highlights the bagsize-bias associated with NOR. Such a problem is critical for cases where bag sizes might differ across training examples and ours is one such case since number of frames per sequence vary from $60 - 600$. We have investigated the effect of choosing other soft-max functions in this work by providing a comparison of NOR with Generalized mean (GM) [20].

Similar to other boosting frameworks, MilBoost involves constructing a strong classifier $H_T(x_{ij})$ by iteratively combining many weak classifiers $h_t(x_{ij})$ that belong to a certain family of function denoted by $\mathcal{H}$. The equation is given as

$$H_T(x_{ij}) = \sum_{t=1}^{T} \alpha_t h_t(x_{ij}) \tag{5}$$

Here H_T denotes the classifier constructed at T^{th} iteration. The classifier score for each instance x_{ij} is given by $H_T(x_{ij})$. These raw scores are mapped into probabilities p_{ij} by a sigmoid function ($p_{ij} = \sigma(H_T(x_{ij}))$. In order to add another weak classifier to H_T, gradient boosting framework proposes to compute weights for each instance x_{ij}, denoted by w_{ij}, by taking the gradient of the loss function $\mathcal{L}$ w.r.t $H_T(x_{ij})$ i.e. the classifier score at each instance at T^{th} iteration. This is given as

$$w_{ij} = -\frac{\partial L}{\partial H_T(x_{ij})} \tag{6}$$

Estimating weights w_{ij} is a tractable procedure once we have a differentiable soft-max function [20]. The next step involves finding a new weak learner ($h_{T+1} \in \mathcal{H}$) that has the highest correlation with the weights w_{ij}. This is necessary since w_{ij} could define any arbitrary direction and thus it is intended to find the best approximation to w_{ij} in the space of $\mathcal{H}$

$$h_{T+1} = \arg\max_{h} \sum_{ij}^{N} w_{ij} h(x_{ij}) \tag{7}$$

This work employs binary decision stumps as weak learner, that assigns a threshold to one of the feature dimensions and are common choice in boosting frameworks [9]. Thus $\mathcal{H}$ refers to the class of decision stumps. In this setting it is easy to realize that our choice of $h_{T+1}(x_{ij})$ should be the one that follows the sign of w_{ij} for instances with highest weights. We refer the readers to Borris et.al [20], who have provided a simple mathematical formulation on how Eq. 7 can be transformed into:

$$h_{T+1} = \arg\min_{h} \sum_{ij} [h(x_{ij}) \neq sgn(w_{ij})] w'_{ij} \tag{8}$$

where $[.]$ is the Iverson bracket, $w'_{ij} = \frac{|w_{ij}|}{\sum_{ij}|w_{ij}|}$ and $sgn(l)$ is the signum function. This Equation is formulation for any learning algorithm that can learns binary labels over weighted samples. Hence we can find a function $h_{T+1}(x_{ij})$ at $(T+1)^{th}$ iteration that has the highest correlation with w_{ij} by using training algorithm for binary decision stumps.

V. Multiple Instance Learning based on Multiple Segments (MS-MIL)

A. Overview

Each sequence S_i is represented as a bag containing many segments or sub-sequences $\{s_{ij}\}_{j=1}^{N_i}$, where N_i is the number of segments in sequence S_i. Temporal consistency is maintained inside a segment s_{ij} by restricting it to contain only contiguous frames (see Section. V-C), $s_{ij} = \{f_i^k, f_i^{k+1}, ..., f_i^{N_{ij}-k+1}\}$, where f_i^k represents the k^{th} frame in sequence S_i and N_{ij} is the number of frames in subsequence s_{ij}. The only information we know about a sequence is if it has pain i.e. $y_i = 1$ or no-pain i.e. $y_i = -1$. We shall give a brief overview of the entire algorithm here.

Representation: Feature extraction process for a frame shall be denoted by a mapping $\phi_{Fr} : R^{m*n} \rightarrow R^d$ that maps frames in image space to a $d-$dimensional vector space. While feature representation for a subsequence (or segment) is represented as a mapping $\phi_S : \mathcal{S} \rightarrow R^d$ that transform a subsequence in space S to a $d-$dimensional vector space. We shall use these notations in this overview and have described them in detail in next two Subsections.

Training: Training data in the form of bags is trained using MilBoost framework described in Section. IV-A. This process yields a classifier $H_T : R^d \rightarrow R$. The number of iterations for MilBoost have been set to 100 in our experiments.

Prediction: Suppose we have a test sequence $S_i = \{s_{i1}, ..., s_{iN_i}\}$. Each subsequence s_{ij} is assigned a posterior probability p_{ij} using the trained classifier H_T and sigmoid function σ as:

$$p_{ij} = \sigma(H_T(\phi_S(s_{ij}))) \tag{9}$$

The posterior probability of test sequence S_i is predicted as a function of the probabilities of its instances using a soft-max function $g(l)$ as:

$$p_i = g(p_{ij}) \tag{10}$$

Avoiding Local-Minima: MilBoost algorithms can often converge to local minima. This problem is more critical for pain detection since theoretically the algorithm can converge even after learning a single instance of pain expression in a sequence as the loss function is defined over bags. In such cases the learned function won't be able to generalize well over unseen data. Hence we draw parallel ideas from bagging predictors proposed by Brieman [21], which proposes to combine multiple versions of a predictor to get an aggregated prediction. They showed improvement in results for predictors that are unstable/get caught up in multiple local minima. Since the problem formulation is very similar to ours, we also ran MilBoost over multiple initializations and

bootstrapped data. The final predictions for each segment were obtained by averaging the predictions p_{ij} made from multiple MilBoost classifiers. Using this approach we found an improvement in predictions and moreover this procedure allowed us to report results that would be reproducible. Based on our experiments we opted to run MilBoost 25 times (any large enough number will work fine).

Pain Localization: The prediction process estimates the posterior probability of each segment s_{ij} in S_i. Each frame in a sequence is assigned a posterior probability by first identifying the segment it belongs to. Following this, the frame is assigned a score based on its proximity from the center of that segment. In this work a hamming window is employed to assign a smoothly varying score to different frames in a segment. Since a frame could belong to multiple segments, it is assigned the maximum score from all these segments. For instance, the probability of frame f_i^k in pain is predicted using the following:

$$p_{f_i^k} = p(y = 1/f_i^k) = \max_j(w(s_{ij}) * p_{ij} | f_i^k \in s_{ij}) \quad (11)$$

where $w(s_{ij})$ is the hamming window function centered at segment s_{ij}. Thus our algorithm not only yields the probability for a sequence but also the probability for each frame that can be used to localize painful frames in a video (see Section VIII).

B. Bag of Words based Representation (BoW)

Recently computer vision has witnessed significant research in BoW models and their extensions, and as a result they have been applied across multiple domains. Sikka et al. [22] presents a survey of different BoW Architectures for AFER. They identified many advantages of BoW based approaches over previous approaches to AFER based on Gabor wavelets, local binary patterns, etc. and have proposed a state-of the-art feature pipeline through experimental analysis.

We employed the system proposed in [22] for representation, and built a spatial pyramid of level 4 on top of highly discriminative multi-scale dense SIFT (MSDF) features, which are encoded using LLC encoding followed by max-pooling. As proposed, we also employed a separate dataset ($CK+$ [23]) for building a codebook (size 200 in this case) for encoding features. It is important to note that this strategy highlights the fact that feature extraction process is completely independent of the dataset. Our experiments yielded that MSDF features at two scales are sufficient for this problem and hence extracted MSDF features with window sizes of 4 and 8 and strides of 2 pixels. As mentioned in the Overview section, the feature extraction operation using BoW is denoted as a mapping ϕ_{Fr}.

C. Multiple Segments

A segment is defined as a subset of original sequence that contains only contiguous frames. Moreover in the current framework a sequence is allowed to contain overlapping segments. As highlighted in Section III, the motivation behind multiple segments is that it allows random onset of pain expression, incorporates dynamic information, and can be efficiently handled by the MIL framework. It is assumed that for a sequence labeled as pain, at least one of the segments will contain a painful expression, and such a positive segment is referred to as a 'concept segment'.

Construction: We propose two ways to generate these multiple segments. A naive procedure is to run overlapping temporal scanning windows at multiple scales (referred to as Sc-wind) across the sequence and represent each subset of frames as a segment. This idea is motivated by the traditional approach in computer vision of running multi-scale scanning windows prior to a detection task. This idea has been exploited in previous work on weakly-supervised object localization [10] [9]. A parallel approach of generating multiple segments was explored in [12]. Here an image was segmented into many clusters using the idea of multiple stable segmentation. Each segmentation is obtained by varying the parameters of normalized cuts (referred to as Ncuts). We explored an analogous approach by clustering the frames in a sequence using Ncuts. Since we wanted to restrict a segment to contain only contiguous frames, the weight matrix was defined to incorporate the similarity between the time index of two frames along with their feature similarity. This weight matrix is used as an input to Ncuts. Each element of this weight matrix $W_i(r, s)$ defines the similarity between frames f_i^r and f_i^s of sequence S_i:

$$W(r,s) = \exp\Big(-|\frac{\phi_{F_r}(f_i^r) - \phi_{F_r}(f_i^s)}{\sigma_f}|^2\Big) + .. \quad (12)$$

$$.. \exp(-|\frac{t_r - t_s}{\sigma_t}|^2) \quad (13)$$

where t_r refers to time indexes of frame f_i^r.

Once the segments are constructed using either of the two approaches, it is important to represent them as fixed-length vectors while also maintaining temporal information. [7] have highlighted that an elegant way of doing this is to append adjacent frames onto the frame of interest. We employed this idea along with max feature pooling, proposed for AFER in [22], for feature extraction. This process is represented as a mapping $\phi_S : \mathcal{S} \rightarrow R^d$ that maps a segment $s_{ij} = \{f_i^k, f_i^{k+1}, ..., f_i^{N_{ij}-k+1}\}$ belonging to set $\mathcal{S}$ to a d-dimensional vector space and can be shown as:

$$\phi_S(s_{ij}) = max_k(\phi_{Fr}(f_i^k) | f_i^k \in s_{ij}) \quad (14)$$

Recently several works [24] [25] have highlighted the performance advantages of the max pooling operation compared to standard pooling operations like averaging.

VI. Experiments

The details of UNBC-McMaster pain dataset which was used for our experiments can be found in Section II. Experiments were conducted in a leave-one-subject-out cross-validation strategy. Thus there is no overlap between subjects in the training and testing data. For reporting the results, we follow the strategy employed in [7] [8], where they reported total classification rate, which refers to the percentage of

correctly classified sequences, computed at Equal Error Rate in the Receiver Operation Curve.

A. MIL vs traditional machine learning approaches

We have argued the aptness of MIL to handle sequences represented as multiple segments compared to traditional ML algorithms. This argument has been validated by using the same multiple segment representation but replacing MIL with linear SVM. All the segments in the training data are assigned the label of sequence and used to train this SVM. This strategy, if not same, is in spirit similar to that employed in previous works ([7] ([8]). Finally during prediction a combining rule is used to assign each sequence a decision score based on the score of its member segments [16]. We have explored two common combining rules, namely maxima (similar to MIL and used in [16]) and average ([7] [8]) and the corresponding SVMs are referred to as MS-SVM$_{max}$ and MS-SVM$_{avg}$.

We have also explored two approaches for generating multiple segments, namely via multi-scale temporal scanning windows and multiple clustering. This is accomplished by considering four scenarios as shown in Table. I by using different settings for the two methods for generating multiple segments. This will also allow us to study the effect of varying number of segments. Since number of frames are different across videos, we determine the number of clusters for Ncuts by fixing minimum number of elements (frames) in a cluster. The values of other parameters are kept constant for all experiments ($\sigma_t = 100$ and $\sigma_f = 10k$). The two parameters for scanning window based method are: (1) window size, and (2) overlap between two windows. The overlap is set as 50% of the window size in all cases. We have also studied the effect of two different soft-max functions on our algorithm. This is necessary since soft-max functions show deviation from $\max$ function with different number of bags.

Name	Method	Min frames	Window size
S_1	Ncut	30	-
S_2	Ncut	$30, 40, 50$	-
S_3	Sc-wind	-	31
S_4	Sc-wind	-	$31, 41, 51$

TABLE I: Methods compared for generating multiple clusters. Each S_i refers to a different setting.

Setting	MS-MIL		ML-SVM$_{max}$	ML-SVM$_{avg}$
	NOR	GM		
S_1	83.7	81.52	75.51	68.71
S_2	82.61	82.99	72.79	69.39
S_3	82.61	82.61	76.19	68.71
S_4	81.52	**83.7**	74.83	70.75

TABLE II: Comparison of MS-MIL with traditional machine learning approaches across four different settings.

B. MS-MIL vs Other Pain Detection Algorithms

We have also provided a comparison between MS-MIL and previous state of the art algorithms ([8] [7] for Pain Detection as shown in Table. III. Although it is not possible to directly compare the results owing to different number of subjects and samples, we would like to highlight that our experiments have been conducted with a larger number of samples (147 vs 142 and 84 in [7] and [8] respectively). Secondly there isn't much difference between the number of samples used in [7] and ours ($5/147$ samples), in which case the results could be comparable to some extent.

Method	Accuracy (at EER)	#subjects	#samples
MS-MIL	**83.7**	23	147
Lucey et.al [7]	80.99	20	142
Ashraf et.al [8] (shown in [7])	68.31	20	142
Ashraf et.al [8]	81.21	21	84
ML-SVM$_{avg}$	70.75	23	147
ML-SVM$_{max}$	76.19	23	147

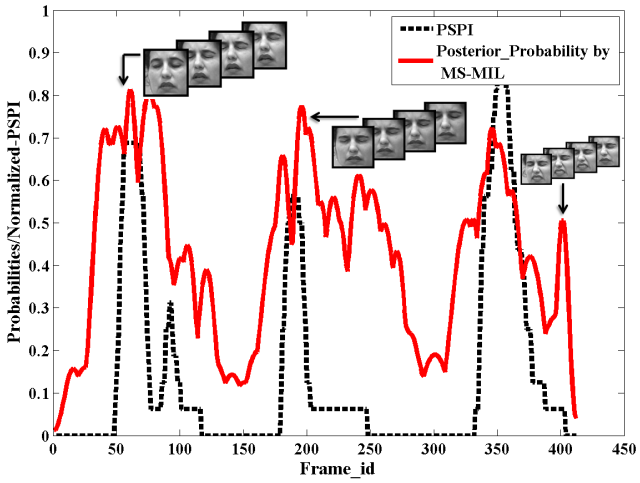
TABLE III: with different algorithms for pain detection in videos.

VII. Results and Discussion

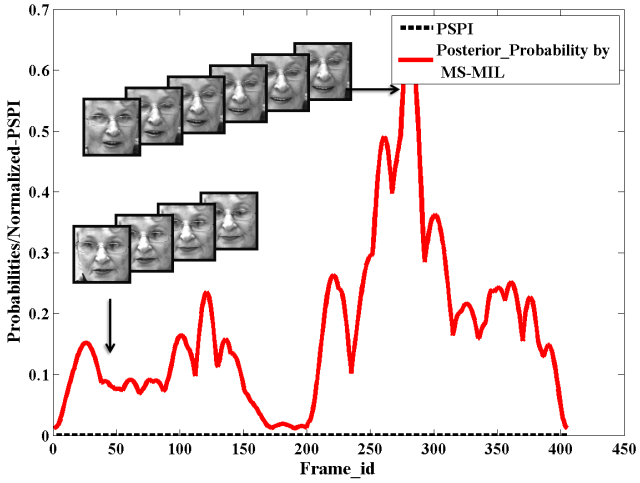
Table. III shows that MS-MIL achieves a higher performance compared to its counterparts. MS-MIL shows more than 7% improvement over both ML-SVM$_{avg}$ and ML-SVM$_{max}$. Although we have stated that a direct comparison with previous algorithms is not possible, some inferences could still be made. Firstly the results of [8] (as implemented by Lucey et al. in [7]) and [7], whose experimental settings are very close to ours, shows an accuracy of 68.31% and 80.99% respectively, compared to 83.7% performance of MS-MIL. Thus it could be argued that MS-MIL shows significant performance improvement over [8] and is comparable to (or better) than [7].

Our argument that multiple segments are efficiently handled by MIL is validated by the comparison of MS-MIL with SVM based methods in Table. II. Here MS-MIL performs better than both ML-SVM$_{avg}$ and ML-SVM$_{max}$ in all of the settings considered. A trend can be observed in moving from lower (S_1 and S_3) to higher number (S_2 and S_4) of segments for different algorithms for generating multiple segments. MS-MIL based on NOR soft-max function shows a decrease in performance, while for MS-MIL employing GM soft-max function the performance improves. This observation falls in line with the argument presented in Section IV-A that NOR soft-max approximation to the $\max$ function deviates as the size of the bags increases. Based on this observation, GM seems to be a better choice for the soft-max function. The results also indicate that ML-SVM$_{max}$ outperforms ML-SVM$_{avg}$ for all cases since the averaging operation is known to dampen the signal of interest.

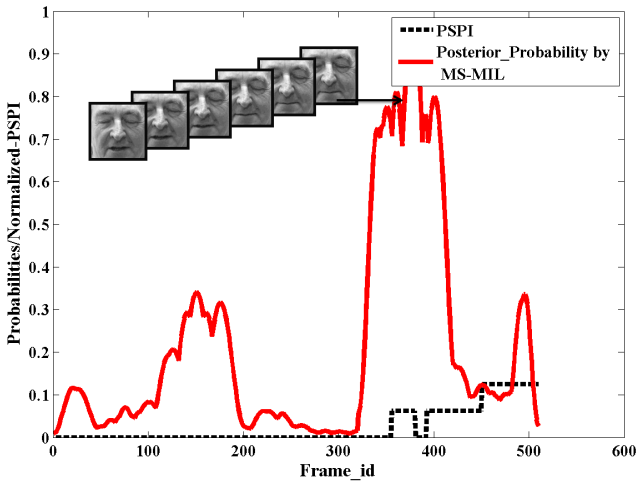
The best result of 83.7% was obtained with MS-MIL under the setting $S4$ and GM soft-max function. Although a similar result was obtained with setting $S1$ and NOR soft-max function, it could have resulted from bag-size bias associated with NOR function. Hence we propose setting S_4 with GM soft-max function for pain detection and for performing pain localization in Section. VIII.



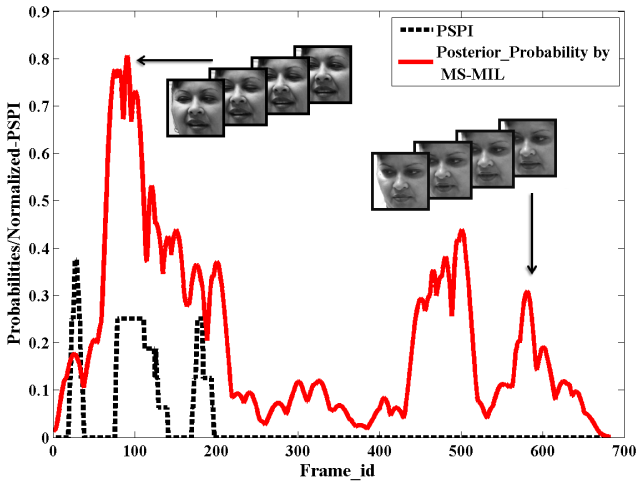
(a) Case1: Pain expression with multiple occurences



(b) Case2: Case where PSPI index showed no pain but MS-MIL could correctly localize pain.



(c) Case3: Pain expression with multiple occurences



(d) Case4: MS-MIL is able to correctly identify an ambigious segment that looks similar to a segment expressing pain. (e.g. probabilty $< .5$ is assigned to segment around frame-id 600).

Fig. 2: **Pain Localization**: Example showing the performance of our algorithm for pain localization vs ground-truth frame labels (PSPI).

VIII. Localization of Pain

We show 4 cases in Fig. 2 to highlight the ability of our algorithm to localize pain. This visualization compares the posterior probability per frame predicted by MS-MIL against frame level pain intensities (PSPI index) provided as a form of ground-truth labels with this dataset. The Prkachin and Solomon pain intensity index (PSPI) is a metric that combines intensities of 4 Action units (AUs) from the Facial Action Coding System (FACS) [26]. Each frame in the dataset was FACS coded by experts, and the PSPI was computed for each frame. The PSPI does not always agree with observer ratings of pain.

We remind readers that MS-MIL uses only sequence level labels to generate these predictions. In order to facilitate direct comparison between probabilities and the PSPI index on the same vertical scale, PSPI index was normalized in the range of $[0, 1]$ by dividing by maximum PSPI score of 16 ([6]).

These visualizations support our claims that MS-MIL is capable of joint classification and localization of pain. It is evident from Fig. 2a and Fig. 2c that our algorithm is able to identify multiple occurrences of pain. Secondly the posterior probabilities predicted by MS-MIL seem to correlate well with the PSPI index. We wanted to draw attention to one major advantage of learning pain expression through sequence level labels. Fig. 2b shows a case of a pain sequence whose PSPI ground-truth score was zero across all frames but the observer rated the facial expression as showing pain (OPI= 3). Our algorithm, which was trained on observer ratings (at the sequence level), was able to localize pain in this case. One explanation for zeros in the PSPI index is that it is based on intensities of 4 AUs which might not be able to account for all possible ways pain can be expressed, or perceived as pain by observers. We would also like to

highlight a case where MS-MIL correctly assigned a lower probability to a segment that gave an impression of a subject expressing pain. One such case is shown in Fig. 2d, where MS-MIL assigns a probability $< .5$ to the segment around frame 600 where the subject was talking

Videos for all 4 cases shown in Fig. 2 are available for download [1] as supplementary material in quicktime's .mov format[2].

IX. Conclusion

This paper proposed a novel approach to the problem of detecting spontaneous expressions of pain in videos, based on multiple instance learning (MIL). We presented a novel framework called multiple-segment multiple instance learning (MS-MIL) which incorporates with MIL a dynamic extension of concept frames.

The paper first highlighted some limitations of previous approaches and how they motivated the design of proposed algorithm. Next, a brief overview of multiple instance learning was presented, followed by the description of the proposed approach, MS-MIL. We then showed the performance advantages of representing each sequence as multiple segments, and how multiple instance learning efficiently handles such representations compared to traditional machine learning approaches. Different methods for extracting multiple segments and soft-max functions for MIL were also compared. We tested our algorithm on the UNBC McMaster Shoulder Pain archive, and obtained a significant improvement in results over algorithms based on traditional machine learning.

From our experiments it is evident that pain detection in videos is a challenging problem owing to the variability associated with how pain can be expressed by different subjects at different times and scenarios. The present algorithm is able to do an appreciable job of not only detecting pain, but also identifying the temporal location of pain expressions within the video clip. The most salient contribution of this work is that pain localization is achieved without any human intervention and employing only sequence level labels.

References

[1] R. Cornelius, *The science of emotion: Research and tradition in the psychology of emotions.* Prentice-Hall, Inc, 1996.

[2] D. Turk and R. Melzack, *Handbook of pain assessment*. The Guilford Press, 2010.

[3] S. Z. Li, A. K. Jain, Y.-L. Tian, T. Kanade, and J. F. Cohn, "Facial expression analysis," in *Handbook of Face Recognition*. Springer New York, 2005, pp. 247–275.

[4] K. Craig, S. Hyde, and C. Patrick, "Genuine, suppressed and faked facial behavior during exacerbation of chronic low back pain," *Pain*, vol. 46, no. 2, pp. 161–171, 1991.

[5] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, 2006.

[6] P. Lucey, J. Cohn, K. Prkachin, P. Solomon, and I. 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*. IEEE, 2011, pp. 57–64.

[1] `http://www.youtube.com/playlist?list=PL-EecoNSKE159u-sa0laeE6oOPmmt2CqJ`

[2] QuickTime player might be required to play the video. It can downloaded from `http://www.apple.com/quicktime/download/`

[7] P. Lucey, J. Howlett, J. Cohn, S. Lucey, S. Sridharan, and Z. Ambadar, "Improving pain recognition through better utilisation of temporal information," in *In the Proceedings of the International Conference on Auditory-Visual Speech Processing*, 2008.

[8] A. Ashraf, S. Lucey, J. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, and B. Theobald, "The painful face: pain expression recognition using active appearance models," in *Proceedings of the 9th international conference on Multimodal interfaces*. ACM, 2007, pp. 9–14.

[9] P. Viola, J. Platt, and C. Zhang, "Multiple instance boosting for object detection," *Advances in neural information processing systems*, vol. 18, p. 1417, 2006.

[10] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 9, pp. 1627–1645, 2010.

[11] G. Wang, Y. Zhang, and L. Fei-Fei, "Using dependent regions for object categorization in a generative framework," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1597–1604.

[12] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie, "Weakly supervised object localization with stable segmentations," in *Proceedings of the 10th European Conference on Computer Vision: Part I*. Springer-Verlag, 2008, pp. 193–207.

[13] T. Leung, Y. Song, and J. Zhang, "Handling label noise in video classification via multiple instance learning," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2056–2063.

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

[15] G. Potamianos, C. Neti, G. Iyengar, A. Senior, and A. Verma, "A cascade visual front end for speaker independent automatic speechreading," *International Journal of Speech Technology*, vol. 4, no. 3, pp. 193–208, 2001.

[16] D. Tax, E. Hendriks, M. Valstar, and M. Pantic, "The detection of concept frames using clustering multi-instance learning," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, aug. 2010, pp. 2917 –2920.

[17] T. Simon, M. H. Nguyen, F. De la Torre, and J. Cohn, "Action unit detection with segment-based svms," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[18] P. Wohlhart, M. Köstinger, P. Roth, and H. Bischof, "Multiple instance boosting for face recognition in videos," *Pattern Recognition*, pp. 132–141, 2011.

[19] J. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, pp. 1189–1232, 2001.

[20] B. Babenko, P. Dollár, Z. Tu, S. Belongie *et al.*, "Simultaneous learning and alignment: Multi-instance and multi-pose learning," in *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[21] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[22] K. Sikka, T. Wu, J. Susskind, and M. Bartlett, "Exploring Bag of Words Architectures in the Facial Expression Domain," in *ECCV Workshop: What's in the face*, 2012.

[23] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[24] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 2. IEEE, 2005, pp. 994–1000.

[25] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

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