A new framework with multiple tasks for detecting and locating pain events in video

Junkai Chen, Zheru Chi, Hong Fu

Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hung Hom, Kowloon, 999077, Hong Kong
Shenzhen Research Institute of the Hong Kong Polytechnic University, 18 Yue Xing Road, Shenzhen, 518052, China
Department of Computer Science, Chu Hai College of Higher Education, 80 Castle Peak Road, Castle Peak Bay, Tuen Mun, 999077, Hong Kong

ABSTRACT

Automatically detecting and locating pain events in video is an important task in medical assessment. It is a challenging problem in facial expression analysis due to spontaneous faces, head movements and pose variations. In this paper, we explore the role of facial information at various time scales (frame, segment and sequence) and propose a new framework for pain event detection and locating in video. We introduce a feature-level fusion method for pain event detection and a multiple-task fusion method for locating pain events, respectively. Both spatial and spatial–temporal features are utilized in our study. At first, we employ the histogram of oriented gradients (HOG) of fiducial points (P-HOG) to extract spatial features from each video frame and train an SVM as a frame-based pain event detector. Secondly, HOG from Three Orthogonal Planes (named as HOG-TOP) is used to characterize the dynamic textures of a video segment, a segment-based classifier (SVM) is then trained for segment-level detection. We further apply a max pooling strategy to obtain the global P-HOG and HOG-TOP to represent the whole video sequence and a multiple kernel fusion is employed to find an optimal feature-level fusion. An SVM with multiple kernels is trained to perform sequence-level (pain event) detection. Finally, an effective probabilistic fusion method is proposed to integrate the detection results of the three different tasks (frame-level, segment-level and sequence-level detection) to locate pain events in video. Extensive experiments conducted on the UNBC-McMaster Shoulder Pain database show that our proposed method outperforms other state-of-the-art methods both in pain event detection and locating in video. Our sequence-level event detection method has also been applied to facial expression recognition in video with good results.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Pain monitoring and measurement is an important task in medical assessment. Pain diagnosis can be used to identify many surgical diseases, like shoulder frozen, arthritis and ligament injury etc. (Lucey et al., 2012). A great challenge is how to effectively assess and measure the pain, since pain is a kind of subjective feeling. A widely used technique to evaluate pain is patient’s self-reporting, which is convenient and simple, and does not rely on any special skills. It is typically measured by either through a clinical interview or by using a visual analog scale (VAS) (Lynch et al., 2011). With the VAS, a patient is asked to mark his pain on a linear scale with a range from 0 (“no pain”) to 10 (“the worst pain”). It has become a very popular method due to its simplicity. However, this method has several limitations including idiosyncratic use, subjective variations etc. (Williams et al., 2000). Therefore, it is not available for some important populations like young children, people who are deaf-and-dumb, and patients who require assisted breathing. Some researchers attempt to acquire a continuous objective measure of pain through analyzing tissue pathology, neurological “signatures”, imaging procedures, testing of muscle strength and so on (Turk and Melzack, 2001). These efforts have been fraught with difficulty because they are often inconsistent with other evidence of pain, in addition to being highly invasive and constraining to the patient (Turk and Melzack, 2001). It is, therefore, necessary to find a reliable and effective method to automatically detect pain events.

Event detection in video is a challenging problem which has received great attention in the computer vision community. Conventional methods heavily rely on carefully hand-designed features. Recently, deep learning models have demonstrated impressive abilities for many computer vision tasks. We have witnessed good progress achieved by deep learning models in general event detection (Gan et al., 2015) and human actions detection (Simonyan and Zisserman, 2014; Wang et al., 2015). Facial event detection (Simon et al., 2010; Ding et al., 2013) is also a kind of event detection task.
General event and human action detection mainly deal with objects, scenes and human activities. Facial event detection concentrates on analyzing facial muscle activities, which provide valuable cues to infer human emotions or intentions.

A potential solution for pain event detection is to analyze facial expression. Facial expression is a powerful nonverbal way for human beings to transmit messages and reveal emotions. Since pain is a kind of emotion or affection, facial expression provides important cues for pain analysis. With the advancement of techniques for facial expression recognition, recently, automatic pain detection through analyzing facial expressions has become an evolving research area and has attracted a growing interest from affective computing community.

Analyzing facial expressions in video provides a promising way for pain detection and measurement. Pain can be revealed by facial expression and we know that facial expression is caused by facial muscle activities, a kind of dynamic process. Video sequences can record this process and provide useful cues for facial event detection. Pain analysis in video has two problems to be solved: pain event detection and pain event locating. The first problem is to detect whether there exists any pain event in a video sequence; the second problem is to locate the pain events in a video sequence.

To handle these two problems, some previous works focused on frame-based methods (Lucey et al., 2011a, 2012; Ashraf et al., 2009) have been reported. These methods classified each frame as positive or negative sample by solving a two-class classification problem. In this paper, we propose a new framework with multiple tasks to jointly tackle the two problems. Considering that information with various time scales (frame, segment and sequence) can make different contributions, we propose to combine three different tasks, that is, frame, segment and sequence detection, to effectively detect and locate pain events in video. In our work, HOG of fiducial points (P-HOG) is used to characterize spatial features from video frames and an SVM classifier is trained for the frame-level paint detection. In order to further exploit spatial–temporal information among contiguous frames, we propose segment-level detection to assist the event detection. HOG from Three Orthogonal Planes (HOG-TOP) is utilized to characterize dynamic textures of video segments. We train an SVM classifier as a segment-based pain event detector. We further apply the max pooling strategy to obtain global P-HOG and HOG-TOP to represent the whole video sequence and employ multiple kernel fusion to optimally combine the two types of global features. An SVM with multiple kernels is trained to perform the sequence-level (pain event) detection. At last, an effective probabilistic fusion method is proposed to integrate the detection results of the three different tasks (frame, segment and sequence detection) to locate pain events in video. By integrating three different tasks, our method provides a more robust and precise detection of pain events in video than the other previously reported techniques which usually focus on one of these tasks only.

The contributions of our study are highlighted as follows:

1. To the best of our knowledge, this is the first time to integrate three different tasks for joint pain event detection and locating. Most previously reported methods emphasize the features or classifiers employed for one task only. And we show that the three tasks can complement with each other and improve the performance.
2. We introduce a feature-level fusion method for pain event detection and show that both the spatial features (P-HOG) and dynamic textures (HOG-TOP) contribute to pain event detection and multiple kernel fusion can enhance the discriminative power of the SVM classifier.
3. We propose an effective probabilistic fusion method to integrate the three tasks for pain event locating and show that integrating three different tasks can achieve a better performance than applying any individual task alone.

The rest of this paper is organized as follows. We introduce some related works in Section 2. Our methodology is presented in Section 3. We describe our experiments and report experimental results with discussions in Section 4. Finally, the paper is concluded in Section 5.

2. Related work and motivations

2.1. Facial expression analysis

Facial expression recognition has been an active research field for many years. It can be widely used in many fields such as lie detection, human computer interaction and affective analysis (Jaimes and Sebe, 2007). An early work carried out by Ekman etc. (Ekman and Friesen, 1978) defined six universal facial expressions: happiness, anger, disgust, fear, sadness and surprise. For the past few years, a number of approaches have been proposed to recognize these six expressions. In general, these approaches can be divided into two mainstreams: appearance based methods (Gu et al., 2012; Shan et al., 2009) and geometry based methods (Maaej et al., 2011; Saeed et al., 2012). An appearance based method mainly extracts the texture feature from a face image. On the other hand, a geometry based method primarily characterizes the shape of a face image with some fiducial points. A more detailed discussion on the two methods can be found in Sariyandi et al. (2015).

Many previous works focused on posed facial expression recognition (Yu and Bhanu, 2006; Wang and Yin, 2007; Moore and Bowden, 2011; Yang et al., 2011). Recently, spontaneous facial expression recognition has attracted a growing attention, because it is close to the real world situation and can reveal the true intentions or emotions of human beings. Compared with posed facial expression recognition, spontaneous facial expression recognition is more challenging due to head movements and poses variations. Some progress has been made on this problem (Dhall et al., 2012; Whitehill et al., 2009; Pfister et al., 2011).

2.2. Pain analysis in video

The research works mentioned above mainly focused on the recognition of the six universal expressions, anger, disgust, fear, happiness, sadness and surprise. Pain is also considered as a facial expression. Pain analysis has attracted great interest recently. A significant contribution to the research on pain analysis was the introduction of the UNBC-McMaster Shoulder Pain dataset (Lucey et al., 2011a), which recorded videos of faces of adult subjects with shoulder injuries. All the videos in the dataset were provided with two levels of annotation for measuring pain: frame level and sequence level. For frame annotation, it followed the description proposed in Prkachin and Solomon (2008) which defined pain as the sum of the intensities of certain facial action units including brow lowering, orbital tightening and eye closure and then employed Facial Action Coding System (FACS) (Ekman et al., 2002; Essa and Pentland, 1997) to code each frame.

Pain detection can be regarded as a spontaneous facial expression recognition problem. An early research on automatic pain recognition was done by Ashraf et al. by developing a “pain-no pain” detection system (Ashraf et al., 2009). In their framework, Active Appearance Models (AAMs) were used to decouple shape and appearance in face images. An SVM was trained for classification. Both frame-level detection and sequence-level detection results were reported.

Following this idea, different algorithms have been proposed to tackle the problem. Lucey et al. (Lucey et al., 2008) pointed
out that temporal information plays a vital role in recognition and showed that by compressing the spatial signal instead of the temporal signal, a better pain recognition performance could be achieved. In addition, some researchers realized that the relationship between the facial expressions and facial action units defined in Facial Action Coding System (FACS) could be used to detect pain. In (Prkachin, 1992), pain was defined as a combination of several action units. Following this idea, automatically detecting pains in video via facial action units was reported in Lucey et al. (2011a,b, 2012). Hammal et al. found that previous works paid attention mainly to the detection of pain occurrences in each video frame and ignored the pain intensity (Hammal and Cohn, 2012). They proposed a method based on Log-Normal filters and SVMs for four-level pain intensity estimation. Sebastian et al. also worked on pain intensity estimation (Kaltwang et al., 2012). They extracted shape and appearance features from face images. Relevance Vector Regression (RVR) was trained to predict different intensity levels.

2.3. Limitations of current techniques

The works mentioned above mostly focus on frame-level detection, i.e. detecting the occurrence or intensity of pain in each video frame. When a video sequence is long, it is computationally very time consuming. A better approach is to apply a global feature vector to characterize the whole video sequence to have a more compact representation. In addition, performing the sequence-level detection can reduce the number of training instances and therefore the training would be more efficient. Some works reported in Wang et al. (2012) and Sikka et al. (2014) have demonstrated the power of this approach. The method reported in (Wang et al., 2012) is based on Bag-of-Word (BoW) architecture and is composed of three steps: feature extraction, feature encoding and normalization. The method transforms a video sequence into a feature vector of fixed length and an SVM is trained to perform the classification. In (Sikka et al., 2014), a method called MS-ML is proposed to jointly detect and locate pain events in video. In their framework, each video sequence is represented as a bag of multiple segments, and Multiple Instance Learning (MIL) is employed to deal with this weakly labeled data in the form of sequence level ground truth. The studies reported in (Wang et al., 2012; Sikka et al., 2014) can be enhanced by considering also spatial–temporal information. In fact, spatial–temporal information plays a vital role for video-based facial expression analysis. However, few work utilized spatial–temporal information for pain event detection and location.

Most previous works surveyed above focused on using one type of information only. However, it is our belief that information at various time scales (frame, segment and sequence) plays different roles. All these information can complement with each other. So we have developed a framework which combines three different tasks (frame, segment and sequence detection) to effectively detect and locate pain events in video. In our study, we show that feature-level fusion is efficient for pain event detection and combining three tasks (frame, segment and sequence detection) for locating pain events is more robust than carrying out any individual task alone.

3. Pain event detection and locating

In this section, we introduce our proposed framework with multiple tasks for joint pain event detection and locating in video. Unlike some previous methods which focus on one particular task, our method combines three different tasks: (1) a frame-level detection task that predicts pain presence/absence based on information extracted from a single frame; (2) a segment-level detection task that classifies pain segments of contiguous frames; (3) a sequence-level detection task that detects pain events in video. The sequence-level detection is used to detect whether there exist pain events in video. (4) Finally, a multi-task fusion method is proposed to integrate the three tasks to locate pain events in video. Fig. 1 shows the pipeline of our proposed framework. Next we describe each task in more detail.

3.1. Frame-level detection

Frame-level detection tries to detect the pain presence/absence of each frame, which is a binary classification problem. We train an SVM with spatial features extracted from video frames for this task. For facial expression analysis, there are two major types of features considered: appearance and geometric features. Geometric features often use facial fiducial points to describe the face shape while appearance features mainly characterize the textures of faces. In our study, both geometric information (facial fiducial points) and appearance information (HOG) are utilized. We extract HOG from the neighborhoods of facial fiducial points (P-HOG) to characterize the spatial feature of each video frame and train an SVM to perform the classification.

Facial expressions are caused by facial muscle movements, especially the muscles around the mouth, nose and eyes. We can extract textures from the interesting regions directly. A face image is first tracked with Active Appearance Models (AAM) (Matthews and Baker, 2004), and some facial landmarks are labeled on the face, such as the blue points shown in Fig. 2. We ignore the facial landmarks around the face outline and only consider the landmarks around the mouth, nose and eyes. We can draw a neighborhood each centered in each landmark (a $16 \times 16$ local window) and extract the appearance features from each neighborhood. Here we employ HOG to characterize texture information. The HOG features from each neighborhood are concatenated to represent the whole appearance feature of each frame.

In our experiment, each video frame is tracked with 66 facial landmarks. We use 49 fiducial points which cover the brows, eyes, nose and mouth, as shown in Fig. 2. A $16 \times 16$ local patch centered on each fiducial point is cropped and thus we can obtain 49 local patches from each frame. HOG is further used to encode each local patch. The vector length of HOG extracted from a local patch with the default setting is 36 (Dalal and Triggs, 2005). The global feature including the HOG from 49 local patches is a vector with a length of $36 \times 49 = 1764$.

3.2. Segment-level detection

Some previous studies have pointed out that facial expression is a dynamic and contiguous process (Scherer and Ekman, 1982; Koelstra et al., 2010). It means that in a video sequence, pain frames and no-pain frames are clustered by themselves. Fig. 3 shows the ground truth frame labels in a video sequence. We can see that there are three pain events in this video sequence, although the duration of each pain event is different, the pain frames of each pain event are contiguous and the no-pain frames between pain events are also contiguous. It inspires us to consider segment-level detection. Each long video sequence can be partitioned into many non-overlap segments with each segment containing a set of contiguous frames. We can locate the clustered pain/no-pain frames by classifying each segment.

In our work, we apply the HOG from Three Orthogonal Planes (HOG-TOP) proposed in (Chen et al., 2014) to extract the dynamic textures from segments. HOG-TOP is an extension of the 2-D HOG. Given a video sequence, each pixel in the sequence has a 3-D neighborhood which lies in the three orthogonal planes ($X$–$Y$, $X$–$T$ and $Y$–$T$). A $3 \times 3$ Sobel mask is first applied to compute the
gradients along the X, Y and T axes, respectively. The gradient orientations are defined as 
\[ \theta_{XY} = \tan^{-1}(G_Y/G_X), \quad \theta_{XT} = \tan^{-1}(G_T/G_X), \quad \theta_{YT} = \tan^{-1}(G_T/G_Y), \]
where \( G_X, G_Y, G_T \) are the gradients along the X, Y and T directions, respectively. These angles are quantized into \( K \) (\( K \) is set 9 in our study) orientation bins with a range of 0° – 180°. Then a histogram is generated in each plane and the three histograms generated are concatenated to form a global descriptor to characterize spatial-temporal information of a segment.

A block-based method is applied in our study, as shown in Fig. 4. We divide the video segment into a number of block volumes and extract the HOG-TOP features from each block volume.
The HOG-TOP features of all the block volume are concatenated to represent the whole video segment. In our experiments, the face image is first cropped from the original image and resized to $64 \times 64$. We partition the face image into $8 \times 8$ blocks with each block with size of $8 \times 8$. The number of bins is set to 9 with an angle range of $0^\circ - 180^\circ$. At each block, we can obtain HOG-TOP with a dimension of $3 \times 9 = 27$. We concatenate the HOG-TOP features of the $8 \times 8$ blocks into a long feature vector with a dimension of $3 \times 9 \times 8 = 1728$.

After feature extraction, we train an SVM for segment-level detection. In our experiment, a segment contains at least one pain frame is labeled as a positive instance (pain segment) and a segment which contains only no-pain frames is labeled as a negative instance (no-pain segment).

### 3.3. Sequence-level detection

In order to determine whether there exist pain events in a video sequence, we propose a sequence-level detection method which is based on feature-level fusion. We first adopt the max-pooling strategy to transform the feature image (P-HOG) and segment feature (HOG-TOP) to a global P-HOG and global HOG-TOP, respectively. After that, multiple kernel fusion is applied to find an optimal combination of the two sets of features. Finally, an SVM with multiple kernels is trained to perform the classification.

Given a long video sequence with $N$ frames segmented into $M$ segments, after feature extraction, we can obtain $N$ P-HOG features and $M$ HOG-TOP features. Here, we take the max-pooling strategy to get the global P-HOG and global HOG-TOP. Suppose that we have a set of features $S = \{ x_i \in \mathbb{R}^D | i = 1, 2, \ldots, N \}$, and $x_i$ is a $D$-dimension feature vector, i.e. $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$. Here we denote the final feature vector after the max pooling as $F$. Then the elements in $F$ should satisfy

$$F_j = \max_{i=1,2,\ldots,N} x_{ij}, \quad j = 1, 2, \ldots, D$$

(1)

where $F_j$ is the $j$th element of $F$ and $x_{ij}$ is the $j$th element of $x_i$. From Eq. (1), we can see that each element in $F$ is the maximum value of all the corresponding elements in the feature set $S$, where $S$ is a $N \times D$ matrix, with $N$ being the number of feature vectors and $D$ the dimensionality of the feature vector. Then $F$ contains the maximum value of each column in $S$. Max pooling transforms a set of feature vectors to a global feature vector, which is used to represent the whole video sequence.

Once we have acquired the global features, the next thing is to train a classifier to perform the detection. SVM is a widely used classification model. However, a traditional SVM with a single kernel is not efficient to handle the training problem of multiple features. Recently, multiple kernel fusion has attracted a growing attention. Previous works have showed that features with multiple kernels can enhance the discriminative power of the SVMs (Chen et al., 2014; Sikka et al., 2013).

In our study, we have two types of features: global P-HOG and global HOG-TOP. We apply multiple kernel fusion to find an optimal fusion of the two types of features. Given a training set with labeled samples: $S = \{ (x_i, y_i) \mid x_i \in \mathbb{R}^D, y_i \in \{-1, 1\} \}^N_{i=1}$. The dual formulation of the traditional single kernel SVM optimization problem is then given by

$$\max_{\alpha} \left[ \sum_{i=1}^{N} y_i \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j K_{ij} \right]$$

subject to

$$\sum_{i=1}^{N} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C$$

(2)

where $K_{ij}$ is the kernel matrix, and $K_{ij} = k(x_i, x_j)$, where $k(\cdot, \cdot)$ is the kernel function and $x_i, x_j$ are the feature vectors.

We adopt the formulation proposed in Rakotomamonjy et al. (2008) in which the kernel is actually a convex combination of several basis kernels. We define the global P-HOG as $x$ and global HOG-TOP as $z$, then we have

$$K_{ij} = \beta_1 k_1(x_i, x_j) + \beta_2 k_2(z_i, z_j)$$

with $\beta_1, \beta_2 > 0, \beta_1 + \beta_2 = 1$.

The basis kernels can be linear kernels, radial basis function (RBF) kernels and polynomial kernels, etc. In our study, we use a linear kernel for each type of features and adopt the grid search with LIBSVM (Chang and Lin, 2011) to learn the kernel weights $\beta_1, \beta_2$ and coefficients $\alpha$.

Given a test sample which contains global P-HOG $x$ and global HOG-TOP $z$, the label $y$ can be predicted by

$$y = \text{sgn} \left( \sum_{i=1}^{N} \alpha_i (\beta_1 k_1(x_i, x) + \beta_2 k_2(z_i, z)) + b \right)$$

(3)

### 3.4. Probabilistic fusion of three tasks

Our framework aims to effectively detect and locate pain events in video. We have illustrated how to apply a sequence-level detection method which incorporated multiple features (global P-HOG and HOG-TOP) and multiple kernel fusion for pain event detection. For pain event locating, we propose a probabilistic fusion method to integrate three different tasks (frame, segment and sequence detection) to achieve this goal. Fig. 5 shows the semantic diagram of combining three tasks for pain event locating. Frame-level detection can detect the pain presence/absence of each individual frame. Segment-level detection can catch the clustered pain/no-pain frames in a video sequence. And sequence-level detection can eliminate some false positives caused by frame and segment detection in a true negative video sequence. We will show that the combined multiple-task outperforms any individual task, and each task plays a different but vital role for pain event locating.

The classifiers trained for the three tasks in our work are all SVMs. We know that the output of an SVM is a decision value obtained by a linear function, $\delta = \mathbf{w} \cdot \mathbf{x} + b$. We apply the sigmoid function to transform the decision value $\delta$ to a probabil-
ity \( p = \frac{1}{1 + e^{-x}} \). Note that for the frame-level detection, each frame has a probability; while for the segment-level detection, the frames contained in the same segment share the same probability.

We denote the probability of a frame which is predicted as a pain frame by \( p_f \) and it can be acquired from the frame-level detection. \( p_s \) is the probability of a segment which is predicted as a positive segment, meaning that the segment contains at least one pain frame. \( p_s \) can be obtained from the segment-level detection. We define the following rule to fuse the frame probability and segment probability:

\[
p = \begin{cases} 
    p_f & |p_f - \tau| > |p_s - \tau| \\
    p_s & |p_f - \tau| < |p_s - \tau|
\end{cases}
\]

(5)

where \( \tau \) is a threshold and it is set to 0.5 in our experiment. \( |p_f - \tau| \) is the distance between the frame probability and the threshold, while \( |p_s - \tau| \) is the distance between the segment probability and the threshold. The distance shows the confidence of a prediction. The greater the distance is, the higher confidence of the prediction is. However, combining the frame-level and segment-level detection might still not be able to achieve a reliable prediction. For instance, considering a true negative video sequence, meaning there are no pain frames in it, if we only integrate the frame-level detection and the segment-level detection, there still exist some false positives. In order to eliminate these false positives, we can bring in the segment-level detection. Suppose that the prediction of sequence-level detection is \( y = 1 \) (when the sequence is predicted as a positive instance) and \( y = 0 \) (when the sequence is predicted as a negative instance), the fused frame probability defined in Eq. (5) is multiplied by the sequence prediction \( y \). Then, the final frame probability which combines the three detections is defined as

\[
p = \begin{cases} 
    p_f \cdot y & |p_f - \tau| > |p_s - \tau| \\
    p_s \cdot y & |p_f - \tau| < |p_s - \tau|
\end{cases}
\]

(6)

Compared with Eq. (5), Eq. (6) brings in the sequence prediction which can wipe out the false positives caused by the frame-level and segment-level detection in a true negative video sequence.

4. Experimental results and discussions

4.1. The database

In order to evaluate our method, we conduct experiments on the UNBC-McMaster shoulder pain dataset (Lucey et al., 2011b). This dataset includes 200 sequences from 25 subjects. Each subject was suffering from some kind of shoulder pain and was requested to make some passive or active movements. Active tests were performed with the patient in a standing position. Passive tests were performed with the help of a physiotherapist. More details about the dataset can be found in Lucey et al. (2011b).

The facial expressions recorded in the dataset are spontaneous with head movements. Each video frame provided 66 facial landmarks tracked by using the active appearance model. The dataset provides two kinds of labels: frame-level label and sequence-level label. The frame-level label is called Prkachin and Solomon pain intensity (PSPI) metric which is a sum of certain facial action units intensities from Facial Action Coding System (FACS) (Essa and Pentland, 1997). The PSPI (with a range from 0 to 16) can describe the pain intensity of each frame. In our study, we focus on detecting pain presence/absence on each frame. A frame with PSPI greater than 1 is considered as a pain frame and otherwise the frame is a no-pain frame. This database also provides the sequence-level label called Observer Pain Intensity (OPI) rating that categorizes each sequence into one of the six intensities from 0 (no-pain) to 5 (strong pain). Following the protocol proposed in (Lucey et al., 2008; Sikka et al., 2014), all the video sequences are classified into “pain” and “no pain” in our study. When OPI \( \geq 3 \), the sequence is a positive instance (pain) and when OPI = 0, the sequence is a negative instance (no-pain). The video sequences of pain intensities of 1 and 2 are removed. We consider both of two level labels and used 139 video sequences with 50 positive instances and 89 negative instances for our experiments. Table 1 shows the descriptions of positive video sequences and negative video sequences. We can see that a positive video sequence has its OPI of greater than or equal to 3 and there at least exist some pain frames in the video sequence. On the other hand, a negative sequence contains no-pain frames with an OPI equal to 0.

4.2. Pain event detection

We first apply our sequence-level method for pain event detection. As mentioned above, we employ two types of features: P-HOG and HOG-TOP. We compute HOG features on the facial landmarks of each frame provided by the dataset. In order to compute HOG-TOP, we first utilize the facial landmarks to crop the face region from each frame and then resize the face region to 64 \( \times 64 \). And then we split the sequences into a number of non-overlapping fixed length segments and apply the HOG-TOP feature descriptor proposed in (Chen et al., 2014) to extract the dynamic textures from each segment. The max pooling is used to find the global features from the feature set.

In our experiment, we take the leave-one-subject-out cross validation strategy. The video instances from one subject are used for testing and the video instances from the other subjects are used for training. In each try, there is no overlapping between the subjects in the training and test data. Since there are 25 subjects, we carry out 25 cross validation experiments. We follow the strategy employed in (Lucey et al., 2008; Sikka et al., 2014) and use the overall classification rate for performance evaluation.

We first apply P-HOG and HOG-TOP individually. An SVM with a linear kernel is trained for the classification. The overall classification accuracy acquired by using P-HOG is 87.1%. For HOG-TOP, we set different segment lengths to explore the performance of HOG-TOP under different scales. Table 2 shows the overall classification rates obtained by applying HOG-TOP at different scales. The S1, S2 and S3 are the scale settings with the segment length indicating the number of frames in each segment.

We further apply multiple kernel fusion to combine the P-HOG and HOG-TOP features and train an SVM with multiple kernels to perform the classification. We compare the results with P-HOG and HOG-TOP features used individually. Since P-HOG is the frame-level feature, the segment length does not affect P-HOG; the performance of P-HOG is the same in the three settings. From Table 2, we can see that the hybrid feature outperforms the individual features, with an improvement of around 5%, meaning that multiple kernel fusion can enhance the discriminative power of an SVM.

We further compare our method with the other methods. The results are shown in Table 3. The best performance in our method is S2 with a classification rate of 91.4%. The improvement is significant compared with the methods reported in Ashraf et al. (2009) (68.3%) and Lucey et al. (2008) (81.0%). The methods reported in Ashraf et al. (2009) and Lucey et al. (2008) are using frame-based detection. Experimental results shows that a sequence based detection method can achieve a better performance than a frame based detection method. In addition, the algorithms reported in Wang et al. (2012) employed BoW to encode each frame and adopted the max pooling strategy to obtain a global feature from all the frames of a video sequence. And an SVM is trained for classification. The performance (81.5%) is comparable with HOG-TOP applied individually in our method, but is not as good as when the hybrid feature is used in our method. It shows that multiple features are more ef-
effective to handle this problem. Note that Multiple Instance Learning (MIL) was applied in Sikka et al. (2014). Our best performance (91.4%) is better than that reported in Sikka et al. (2014) (83.7%), with an improvement of about 8%. Although the classification accuracy in our method is affected by the segment length, as shown in Table 3, we have shown that even the lowest accuracy (90.6%) in our method is still better than the other methods.

### 4.3. Locating pain events

Locating pain events focus on predicting the pain presence/absence at the frame level. Like pain event detection, we also take the leave-one-subject-out cross validation strategy. In our experiment, we apply two evaluation metrics as employed in Sikka et al. (2014), classification accuracy and maximum F1-score. The F1-score is defined as \( F1 = 2 \cdot \frac{P \cdot R}{P + R} \), known to give a trade-off between recall \( R \) and precision \( P \). This dataset provides a PSPI with a range of 0–16 to indicate the pain intensity for each frame. In this work, we attempt to detect pain/no-pain frames. We first transform the PSPI to a binary label. In our experiment, a frame is labeled as a pain frame when PSPI \( \geq 2 \); otherwise, it is a no-pain frame.

For frame-level detection, we extracted P-HOG from each frame and trained an SVM to perform the binary classification. The locating accuracy is 76.5% and the maximum F1-score is 0.452.

For segment-level detection, we applied HOG-TOP to extract the dynamic textures from each segment and an SVM was trained to perform the classification. We set different segment lengths to explore the performance of segment-level detection at different scales. Fig. 6 shows the locating accuracy and the max-F1 score of segment-level detection under different scales. Here S1, S2 and S3 denote the segment lengths in different settings. We can find that there is no much difference on the performance under different scales. Compared with the frame-level detection, the segment-level detection achieves a higher locating accuracy while obtaining a lower F1-score.

We also provide the performance of pain event locating based on the sequence-level detection. In this case, all the frames contained in a video sequence share the same prediction result. The locating accuracy and the maximum F1-score are 56.9% and 0.358, respectively. Since there are a small amount of pain frames in a positive video sequence, when all the frames share the same prediction output, there are too many false positives which make the sequence-level detection is not as good as the frame-level detection.

From the experimental results of three individual detections, we can see that information with various time scales (frame, segment and sequence) can make different contributions. Each piece of information is vital, which inspires us to combine three detection methods to enhance the pain event locating performance.

We combine the three detection methods (frame, segment and sequence detection) as defined in Eq. (6). Experimental results are shown in Fig. 7. The segment length is only meaningful to segment-level detection. It is obvious to find that the combined method outperforms individual detection methods. We can also see that a combination of three detection methods performs better than a combination of two detection methods (frame and segment detection), especially for the maximum F1-score, with an improvement of about 3%. Bringing in the sequence-level detection can eliminate some false positives in a negative video sequence. Fig. 8 shows the pain event locating results in a positive video sequence and a negative video sequence, respectively. We can see that three different tasks can complement with one another and integrating the three tasks can achieve a better performance. Fig. 8(a) shows that some false positives generated by the frame-level detection method can be wiped out when we bring in the segment-level detection method. Fig. 8(b) illustrates that the sequence-level detection method can eliminate some false positives generated by the frame-level detection method on a negative video sequence.

We also compare our method with the other methods. Since we aim to perform joint pain event detection and locating, there are fewer previous methods trying to tackle both tasks except for the study reported in Sikka et al. (2014) in which Sikka et al. proposed a method called MS-MIL to jointly detect and locating pain events in video. They also designed an MS-SVM\(_{\text{max}}\) method for the comparison purpose. We compare our method with MS-MIL and MS-SVM\(_{\text{max}}\) in Table 4. The detection accuracy measures the performance of video based pain event detection as discussed in Section 4.2. Locating accuracy and the maximum F1-score demonstrates the performance of pain event locating. We can find that the locating accuracy in our method is much higher than MS-SVM\(_{\text{max}}\) and MS-MIL reported in Sikka et al. (2014), with an improvement.
of about 13% and 10%, respectively. The maximum F1-score is also higher than MS-\(\text{SVM}_{\text{max}}\) and slightly better than MS-MIL. Although the number of samples in our experiments is slightly less than that in Sikka et al. (2014), it is reasonable to conclude that our method compares favorably with MS-\(\text{SVM}_{\text{max}}\) and MS-MIL for pain event detection. For pain event locating, our method can achieve a comparable maximum F1-score as MS-MIL while obtaining a much higher locating accuracy.

### 4.4. The effect of segment length

A challenging problem for us is to set the length of a segment when we compute HOG-TOP. In our experiments, we explored the effect of the segment length for pain event detection and locating. Since there is no overlap between the segments, segment length is the only parameter we need to consider. We set different segment lengths with a range from 10 to 30 and then evaluate the performance of pain event detection and locating. Fig. 9 shows the detection accuracy obtained by three different feature sets under different segment scales. P-HOG is a frame feature which is not affected by segment length. The performance of P-HOG is therefore the same under six different segment lengths. On the other hand, HOG-TOP is a kind of dynamic texture feature to characterize spatial–temporal information of an image sequence. The segment length affects the detection performance of HOG-TOP significantly. We can see that the highest accuracy is 83.4% when the segment length is set to 20. The performance falls to 76.9% when
4.6. Experiments on facial expression recognition

Although our framework is proposed for pain event detection in video, it can also be used to recognize other facial expressions in video. We conducted the experiments on the GEMEP-FER2011 database (Valstar et al., 2011) to evaluate the effectiveness of our sequence-level detection model. There are 289 video sequences in the database (155 for training and 134 for testing). The training set includes seven subjects with three to five instances of each emotion per subject. Each video sequence is categorized into one of the following five emotions: anger, fear, joy, relief and sadness. There is no information about the onset, peak and duration of the facial expressions in this database, which is different from the UNBC-McMaster shoulder pain database. Only the sequence-level labels of training set are public available. We take the leave-one-subject-out cross validation strategy on the training set to evaluate our framework for facial expression recognition in video.

We apply the Mixtures of Parts (MoPs) model (Zhu and Ramanan, 2012) to locate facial fiducial points first and extract P-HOG from each frame. Max pooling is used to transform the P-HOG of each frame to a global P-HOG feature. Most video sequences in this database contain dozens of frames, it is not necessary to divide the video sequence into segments. We extract HOG-TOP from the whole video sequence directly. We also compare our method with the LBP-TOP (Zhao and Pietikäinen, 2007). Experimental results are shown in Table 5. The hybrid feature is an optimal combination of global P-HOG and HOG-TOP. From Table 5, we can see that the hybrid feature outperforms P-HOG and HOG-TOP applied alone, indicating that spatial information (P-HOG) and spatial-temporal information (HOG-TOP) are complement with each other to some extent. Our model can take advantages of both spatial and spatial-temporal information. Experimental results show that our sequence-level detection method is also robust to deal with facial expression recognition in video.

4.7. Discussions

Experimental results reported in Section 4.2 show the effectiveness of our feature-level fusion method in pain event detection. P-HOG can represent spatial appearance of facial expressions. HOG-TOP can characterize the dynamic appearance changes caused by facial activities. In fact, these two types of features capture information from different time scales. Pain, as a kind of facial expression, actually is a dynamic process. These two kinds of information can be used to build a more useful model to characterize this dynamic process. Moreover, in order to take advantages of two types of features, multiple kernel fusion is used to integrate the two features optimally. Multiple features with an optimal combination can achieve a better performance than individual feature.

Experimental results reported in Section 4.3 show that a multiple-task fusion based on three different tasks is more robust to deal with pain event locating than any individual task. In the dataset we used, a negative video sequence contains no-pain frames only. Even in a positive video sequence, the number of pain frames accounts for a small part of the whole sequence. It means that there are much more negative instances than positive instances. This is an unbalance training problem. The frame-level and segment-level detection tend to make a large number of false positives. But when we combine the two detection methods, they can complement with each other and achieve a better performance. It suggests that the pieces of information from different time scales are complementary and they can be used to reduce false positives. We also observe that the sequence-level detection can help in eliminating some false positives made by frame-
level and segment-level detection in negative video sequences. The maximum F1-score can be improved slightly when we bring in sequence-level detection, as shown in Fig. 7.

5. Conclusion

In this paper, we propose a novel method for joint pain event detection and locating in video. We develop a novel framework which combines three detection tasks, frame-level, segment-level and sequence-level detection, to handle (1) pain event detection which determines whether there exist pain events in a video (a pain/no-pain classification problem) and (2) pain event locating (predicting pain presence/absence in each frame).

For pain event detection, we propose a feature-level fusion method. Both the static attributes of video frames and dynamic attributes of sequences are explored in this work. We apply P-HOG to extract spatial features from video frames to represent the static attributes. HOG-TOP is used to characterize the dynamic textures of video segments. We then employ max pooling to form a global P-HOG and global HOG-TOP to characterize the whole video sequence. Multiple kernel fusion is used to find an optimal combination of these two types of global features. Finally, an SVM with multiple kernels is trained to detect whether there exist any pain events in video.

For pain event locating, we propose a multiple-task fusion method. We first build three sequential tasks namely frame-level detection, segment-level detection and sequence-level detection. For the frame-level detection, we train an SVM with P-HOG features to predict pain presence/absence on each frame. Noting that pain frames or no-pain frames are contiguous in a video sequence, we also detect pain segments of contiguous frames. An SVM is trained with HOG-TOP features extracted from video segments to perform the segment-level detection. At last, we couple the two detection methods with the sequence-level detection to locate pain events.
events in video. Our sequence-level event detection method has also been applied to facial expression recognition in video with good results.

To the best of our knowledge, our work is the first to combine three detection methods at different time scales (frame, segment and sequence) for joint pain event detection and locating. Our approach utilizes information from different time scales. Our method outperforms the other state-of-the-art methods on the public UNBC-McMaster Shoulder Pain dataset.

Acknowledgments

The work reported in this paper was supported by a research Grant from National Natural Science Foundation of China (Project code: 61472423). Junkai Chen would like to acknowledge a postgraduate scholarship from the Hong Kong Polytechnic University.

References


