Image and Vision Computing 32 (2014) 659–670

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Classification and weakly supervised pain localization using multiple segment representation☆

Karan Sikka [a],*, Abhinav Dhall [b], Marian Stewart Bartlett [a]

[a] University of California San Diego, 9450 Gilman Drive # 0440, La Jolla, CA 92093, USA
[b] Australian National University, CSIT, Building 108, North Road, Canberra 2601, Australia

ARTICLE INFO

Article history:
Received 2 June 2013
Received in revised form 10 December 2013
Accepted 21 February 2014
Available online 4 March 2014

Keywords:
Emotion classification
Action classification
Pain
Temporal segmentation
Bag of Words
Weakly supervised learning
Boosting
Bagging

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 such as 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. Extensive experiments on UNBC-McMaster Shoulder Pain dataset highlight the effectiveness of the approach by achieving competitive results on both tasks of pain classification and localization in videos. We also empirically evaluate the contributions of different components of MS-MIL. The paper also includes the visualization of discriminative facial patches, important for pain detection, as discovered by our algorithm and relates them to Action Units that have been associated with pain expression. We conclude the paper by demonstrating that MS-MIL yields a significant improvement on another spontaneous facial expression dataset, the FEEDTUM dataset.

1. Introduction

Pain is one of the most challenging problems in medicine and biology and has substantial eco-social costs associated with it [9]. It has been estimated that there might be more than 30 million people in USA with chronic or recurrent pain [34]. Also nearly half of Americans seeking treatment from a physician report pain as their primary symptom. The United States Bureau of the Census estimated the total cost for chronic pain to exceed $150 billion annually in year 1995–96 [9,34]. Thus there has been a significant research effort in improving pain management over the years.

Identifying pain among patients is considered critical in clinical settings since it is used for regulating medications, long-term monitoring, and gauging the effectiveness of a treatment. Pain assessment in most cases involves patient self-report, obtained through either clinical interview or visual analog scale (VAS) [9]. For the latter case the nurse asks the patient to mark his pain on a linear scale with ratings from 0 to 10, denoting no-pain to unbearable-pain. The fact that VAS is easy to use and returns a numerical rating of pain has made VAS the most prevalent pain assessment tool. However VAS suffers from a number of drawbacks such as subjective differences, and patient idiosyncrasies. Therefore it cannot be used for unconscious or verbally-impaired patients [6] and may suffer from high individual bias. These drawbacks have led to a considerable research effort to identify and quantify objective pain indicators using human facial expression [33]. However most of these methods entail manual labeling of facial Action Units or evaluations by highly trained observers, which in most cases is time consuming and unfit for real-time applications.

Over the years there has been a significant progress in analyzing facial expressions related to emotions using machine learning (ML) and computer vision [19]. Most of this work has focused on posed facial

☆ This paper has been recommended for acceptance by Qiang Ji, Ph.D.
* Corresponding author.
E-mail addresses: ksikka@ucsd.edu (K. Sikka), abhinav.dhall@anu.edu.au (A. Dhall), mbartlett@ucsd.edu (M.S. Bartlett).

http://dx.doi.org/10.1016/j.imavis.2014.02.008
0262-8856/© 2014 Elsevier B.V. All rights reserved.

expressions that are obtained under controlled laboratory settings and differ from spontaneous facial expression in a number of ways [4,7]. We refer our readers to a survey on automatic facial expression recognition (AFER) by Bartlettet al. [4] 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 contribution to in research on spontaneous expressions was the introduction of UNBC-McMaster Shoulder Pain dataset [21] that involves subjects experiencing shoulder pain in a clinical setting. This dataset was provided with two levels of annotations for measuring pain — (1) per-frame pain ratings based on a formula applied to Action Unit (AU) annotations, and (2) per-video pain ratings as measured by experts (see Section 5.1). This work utilizes the per-video pain ratings for training a binary pain classification system. Pain localization is then evaluated using the per-frame pain ratings based on AU labels, which are more costly to obtain. Thus our setting is such that each video is labeled for the presence or absence of pain, but there is no information about the location or duration of facial expressions within each video. This setting is referred to as weakly labeled data and 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 [2,22] 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 and (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) framework [35].

MIL is an approach 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 the time point or duration of pain. Other techniques for tackling weakly labeled data include part-based models [11] and latent models such as pLSA and LDA [37]. 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 [13,35]. 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 [35] and more recently has been proposed for handling labeling noise in video classification [18].

This work focuses at detecting spontaneous pain expression in video when given only sequence level ground-truths. The phrase *detection* is used throughout the paper to denote the joint tasks of pain classification and localization in time. Explicitly, classification refers to predicting the absence/presence of pain in a video, while localization refers to predicting pain/no-pain at the frame level. The novelty of this work lies in combining MIL 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 drawbacks in previous approaches for pain detection in videos are identified and a pipeline has been proposed to address these concerns. The most salient feature of our approach is that it can jointly classify and localize pain by using only sequence level labels (Section 2).
2. For addressing the demands of the pain detection task, we propose 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 spontaneous expressions, such as pain, that can have uncertain locations, durations and occurrences (Section 4).
3. The performance of MS-MIL is compared on the detection task with other competitive algorithms. We also perform systematic evaluation to highlight the contribution of multiple segment representation and MIL, in MS-MIL, separately. These results indicate the advantage of using the MS-MIL approach along with some interesting insights (Section 6).

The problem of detecting pain through facial expressions in general includes many challenges and this work is trying to focus on a particular aspect of the problem. Other challenges in objective pain measurement include differences between acute and chronic pain, as well as differences in personality including pain catastrophizing, which may affect the intensity of pain expression. We are undertaking a separate study to begin to address some of these factors [14].

2. Related work and motivation

The first computer vision work on automatic pain detection in videos on the UNBC-McMaster Pain dataset was by Ashraf et al. [2]. Their approach started by first extracting AAM based features from each frames and using these to cluster the frames in order to create a training data with size that is manageable by a SVM. Following this, each of these clustered frames was assigned with the label of their corresponding sequence and used to train a linear SVM. Finally during prediction each test-frame was assigned a score based on its distance from separating hyperplane. Then a test-video was predicted to be in pain if the average score of its member frames exceeded a threshold. Lucey et al. [22] extended this work by borrowing 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. [22] used the system in [2] as their baseline system and showed significant improvement in performance using their idea.

Previous works didn't address the ambiguity introduced by weakly labeled data, and each member frame was assigned the label of the sequence. Such approaches lead to a lower performance compared to the case when ground-truth for each frame is known [1,2]. We address this particular concern by proposing to use MIL (in-place of SVM) which has been designed specifically to handle weakly labeled data.

Secondly, [22] 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, [22] suggested to add temporal information by appending adjacent frames onto the frame of interest, as input to the SVM [25]. [22] 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) have twofold advantages: (1) it allows pain expression to have random duration and occurrence, and (2) it incorporates temporal information by pooling across multiple frames in a segment. Thirdly, the earlier work performed prediction for each sequence using the average decision score of its frames. Such an approach