Image and Vision Computing 27 (2009) 1788–1796

Contents lists available at ScienceDirect

Image and Vision Computing

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

The painful face – Pain expression recognition using active appearance models

Ahmed Bilal Ashraf, Simon Lucey, Jeffrey F. Cohn *, Tsuhan Chen, Zara Ambadar, Kenneth M. Prkachin, Patricia E. Solomon

University of Pittsburgh, Psychology, 3137 Sennott Square, 210 S. Bouquet St., Pittsburgh, PA 15260, USA

ARTICLE INFO

Article history:
Received 4 May 2008
Received in revised form 4 February 2009
Accepted 3 May 2009

Keywords:
Active appearance models
Support vector machines
Pain
Facial expression
Automatic facial image analysis
FACS

ABSTRACT

Pain is typically assessed by patient self-report. Self-reported pain, however, is difficult to interpret and may be impaired or in some circumstances (i.e., young children and the severely ill) not even possible. To circumvent these problems behavioral scientists have identified reliable and valid facial indicators of pain. Hitherto, these methods have required manual measurement by highly skilled human observers. In this paper we explore an approach for automatically recognizing acute pain without the need for human observers. Specifically, our study was restricted to automatically detecting pain in adult patients with rotator cuff injuries. The system employed video input of the patients as they moved their affected and unaffected shoulder. Two types of ground truth were considered. Sequence-level ground truth consisted of Likert-type ratings by skilled observers. Frame-level ground truth was calculated from presence/absence and intensity of facial actions previously associated with pain. Active appearance models (AAM) were used to decouple shape and appearance in the digitized face images. Support vector machines (SVM) were compared for several representations from the AAM and of ground truth of varying granularity. We explored two questions pertinent to the construction, design and development of automatic pain detection systems. First, at what level (i.e., sequence- or frame-level) should datasets be labeled in order to obtain satisfactory automatic pain detection performance? Second, how important is it, at both levels of labeling, that we non-rigidly register the face?

1. Introduction

Pain is difficult to assess and manage. Pain is fundamentally subjective and is typically measured by patient self-report, either through clinical interview or visual analog scale (VAS). Using the VAS, patients indicate the intensity of their pain by marking a line on a horizontal scale, anchored at each end with words such as "no pain" and "the worst pain imaginable". This and similar techniques are popular because they are convenient, simple, satisfy a need to attach a number to the experience of pain, and often yield data that confirm expectations. Self-report measures, however, have several limitations [9,16]. These include idiosyncratic use, inconsistent metric properties across scale dimensions, reactivity to suggestion, efforts at impression management or deception, and differences between clinicians' and sufferers' conceptualization of pain [11]. Moreover, self-report measures cannot be used with young children, with individuals with certain types of neurological impairment and dementia, with many patients in postoperative care or transient states of consciousness, and those with severe disorders requiring assisted breathing, among other conditions.

Significant efforts have been made to identify reliable and valid facial indicators of pain [10]. These methods require manual labeling of facial action units or other observational measurements by highly trained observers [6,14]. Most must be performed offline, which makes them ill-suited for real-time applications in clinical settings. In the past several years, significant progress has been made in machine learning to automatically recognize facial expressions related to emotion [20,21]. While much of this effort has used simulated emotion with little or no head motion, several systems have reported success in facial action recognition in real-world facial behavior, such as people lying or telling the truth, watching movie clips intended to elicit emotion, or engaging in social interaction [5,7,22]. In real-world applications and especially in patients experiencing acute pain, out-of-plane head motion and rapid changes in head motion and expression are particularly challenging. Extending the approach of [18], we applied machine learning to the task of automatic pain detection in a real-world clinical setting involving patients undergoing assessment for pain.

In this paper we will attempt to gain insights into two questions pertinent to automatic pain recognition: (i) how should we be labeling datasets for learning to automatically detect pain?, and (ii) is there an inherent benefit in non-rigidly registering the face and decoupling the face into shape and appearance components when recognizing pain?

* Corresponding author. Tel.: +1 412 624 8825.
E-mail addresses: bilal@cmu.edu (A.B. Ashraf), slucey@ieee.org (S. Lucey), jeffcohn@cs.cmu.edu (J.F. Cohn), tsuhan@cmu.edu (T. Chen), ambadar@pitt.edu (Z. Ambadar), kmprk@unbc.ca (K.M. Prkachin), solomon@mcmaster.ca (P.E. Solomon).

0262-8856/
doi:10.1016/j.imavis.2009.05.007

1.1. How should we register the face?

Arguably, the current state-of-the-art system for recognizing expression (specifically for AUs) is a system reported by Bartlett et al. [1,4,5], which first detects the fully frontal face using a Viola and Jones face detector [25], and then rigidly registers the face in 2D using a similarly designed eye detector. Visual features are then extracted using Gabor filters which are selected via an AdaBoost feature selection process. The final training is performed using a support vector machine (SVM). As noted above, this system was adapted recently [2] and applied to the task of detecting "genuine" versus "faked" pain. Tong et al. [33] also reported good AU detection performance with their system which uses a dynamic Bayesian Network (DBN) to account for the temporal nature of the signal as well as the relationship with other AUs. Pantic and Rothkrantz [26] used a rule-based method for AU recognition. Pantic and Patras [27] investigated the problem of posed AU recognition on profile images.

A possible limitation of these approaches is that they employ a rigid rather than non-rigid registration of the face. We refer to non-rigid registration as any shape variation of an object that cannot be modeled by a 2D rigid warp (i.e. translation, scale and rotation). Non-rigid registration of the face may be beneficial from two perspectives. One is to normalize unwanted variations in the face due to out-of-plane head motion. In many real-world settings, such as clinical pain assessment, out-of-plane head motion may be common. The other possible advantage is to enable decoupling of the shape and appearance components of the face, which may be more perceptually important than rigidly registered pixels. We found in previous work [18,3] that this type of alternative representation based on the non-rigid registration of the face was useful for expression recognition in a deception-interview paradigm. In that work we employed an active appearance model [8,19] (AAM) to derive a number of alternative representations based on a non-rigid registration of the face. In the current paper we extend that work. We explore whether such representations are helpful for classifiers to learn "pain"/"no-pain" in clinical pain assessment, and we compare the relative efficacy of sequence- and frame-level labeling.

1.2. How should we label data for learning pain?

With near unlimited access to facial video from a multitude of sources (e.g., movies, Internet, digital TV, laboratories, home video, etc.) and the low cost of digital video storage, the recording of large facial video datasets suitable for learning to detect expression is becoming less of an issue. However, video datasets are essentially useless unless we have some type of labels (i.e., "pain"/"no-pain") to go with them during learning and testing.

In automatic facial expression recognition applications, the de facto standard is to label image data at the frame level (i.e., assigning a label to each image frame in a video sequence). The rationale for this type of labeling stems from the excellent work that has been conducted with respect to facial action unit (AUs) detection. AUs are the smallest visibly discriminable changes in facial expression. Within the FACS (Facial Action Coding System: [13,14]) framework, 44 distinctive AUs are defined. Even though this represents a rather small lexicon in terms of individual building blocks, over 7000 different AU combinations have been observed [15]. From these frame-by-frame AU labels, it has been demonstrated that good frame-by-frame labels of "pain"/"no-pain" can be inferred by the absence and presence of specific AUs (i.e., brow lowering, orbit tightening, levator contraction and eye closing) [10,31].

The cost and effort, however, associated with doing such frame-by-frame labeling by human experts can be extremely large, which is a rate limiter in making labeled data available for learning and testing. If systems could use more coarsely labeled image data, larger datasets could be labeled without increasing labor costs. In this paper we present a modest study to investigate the ability of an automatic system for "pain"/"no-pain" detection trained from sequence- rather than frame-level labels. In sequence-level labeling one label is given to all the frames in the video sequence (i.e., pain present or not present), rather than labels for every frame in the sequence. We compare the performance of pain/no-pain detectors trained from both frame- and sequence-level labels. This work differs considerably from our own previous work in the area [3] in which only sequence-level labels for learning/evaluation were considered. To our knowledge no previous study has compared algorithms trained in both ways.

One other study of automatic pain detection can be found in [2]. Littlewort and colleagues pursued an approach based on their previous work to AU recognition [4,5]. Their interest was specifically in the detection of "genuine" versus "faked" pain. Genuine pain was elicited by having naïve subjects submerge their arm in ice water. In the faked-pain condition, the same subjects simulated pain prior to the ice-water condition. To discriminate between conditions, the authors rigidly registered the face and extracted a vector of confidence scores corresponding to different AU recognizers at each frame. These AU recognizers were learnt from frame-based labels of AU and the corresponding facial image data. Based on these scores the authors studied which AU outputs contained information about genuine versus faked-pain conditions. A secondary SVM was then learnt to differentiate the binary pain conditions based on the vector of AU output-scores. Thus, frame-level labels were used to classify pain- and no-pain conditions, or in our terminology pain- and no-pain sequences.

To summarize, previous work in pain and related expression detection has used rigid representation of face appearance and frame-level labels to train classifiers. We investigated both rigid and non-rigid registration of appearance and shape and compared use of both frame- and sequence-level labels. In addition, previous work in pain detection is limited to sequence-level detection. We report results for both sequence- and frame-level detection.

2. Image and meta data

2.1. Image data

Image data for our experiments was obtained from the UNBC-McMaster shoulder pain expression archive. One hundred twenty-nine subjects with rotator-cuff injury (63 male, 66 female) were video-recorded in "active" and "passive" conditions. In the active condition, subjects initiated shoulder rotation on their own; in passive, a physiotherapist was responsible for the movement. Camera angle for active tests was approximately frontal to start; camera angle for passive tests was approximately 70 deg to start. Out-of-plane head motion in both conditions was common. Images were captured at a resolution of 320×240 pixels. The face area spanned an average of approximately 140×200 (28,000) pixels. For comparability with previous literature, in which initial camera orientation has typically varied from frontal to about 15 deg, we focused on the active condition in the experiments reported below. Sample pain sequences are shown in Fig. 1.

2.2. Meta data

Pain was measured at the sequence- and frame-level.

2.2.1. Sequence-level measures of pain

Pain ratings were collected using subject and observer report. Subjects completed a 10-cm Visual Analog Scale (VAS) after each

Fig. 1. Examples of temporally subsampled sequences. (a) and (c) illustrate pain and (b) and (d) no pain.

movement to indicate their level of subjective pain. The VAS was presented on paper, with anchors of "no pain" and "worst pain imaginable". Subsequently, observed pain intensity (OPI) rating was rated from video by an independent observer with considerable training in the identification of pain expression. Observer ratings were performed on a 6-point Likert-type scale that ranged from 0 (no pain) to 5 (strong pain).

To assess inter-observer reliability of the OPI pain ratings, 210 randomly selected trials were independently rated by a second rater. The Pearson correlation between the observers' OPI was 0.80, $p < 0.001$, which represents high inter-observer reliability [28]. Correlation between the observer's rating on the OPI and subject's self-reported pain on the VAS was 0.74, $p < 0.001$ for the trials used in the current study. A value of 0.70 is considered a large effect [29] and is commonly taken as indicating high concurrent validity. Thus, the inter-method correlation found here suggests moderate to high concurrent validity for pain intensity.

2.2.2. Frame-level measures of pain

In addition to pain ratings for each sequence, facial actions associated with pain were annotated for each video frame using FACS [14]. Each action was coded on a 6-level intensity dimension (0 = absent, 1 = trace . . . 5 = maximum). Because there is considerable literature in which FACS has been applied to pain expression [10,24,30,31], we restricted our attention to those actions that have been implicated in previous studies as possibly related to pain (see [24] for complete list).

To assess inter-observer agreement, 1738 frames selected from one affected-side trial and one unaffected-side trial of 20 participants were randomly sampled and independently coded. Inter-coder percent agreement as calculated by the Ekman–Friesen formula [14] was 95%, which compares favourably with other research in the FACS literature. Following previous literature in the psychology of pain, a composite pain score was calculated for each frame, representing the accumulated intensity scores of four facial actions: brow lowering, orbit tightening, levator contraction and eye closing (see [24] for construction of this index). For the sequences evaluated in these experiments, pain scores ranged from 0 to 12.

2.2.3. Subject selection

Subjects were included if they had a minimum of one trial with an OPI rating of 0 (i.e. no pain) and one trial with an OPI rating of 3, 4, or 5 (defined as pain). To maximize experimental variance and minimize error variance [31,32] movements with intermediate ratings of 1 or 2 were omitted. Forty-four subjects had both pain- and without-pain rated movements. Of these subjects, 23 were excluded for technical errors (8), maximum head rotation greater than about 70 deg (1), and glasses (7) or facial hair (7). The final sample consisted of 21 subjects with 69 movements, 27 with pain and 42 without pain.

3. Active appearance models

In machine learning, the choice of representation is known to influence recognition performance [12]. Active appearance models (AAMs) provide a compact statistical representation of the shape and appearance variation of the face as measured in 2D images. This representation decouples the shape and appearance of a face image. Given a pre-defined linear shape model with linear appearance variation, AAMs align the shape model to an unseen image containing the face and facial expression of interest. In general, AAMs fit their shape and appearance components through a gradient descent search, although other optimization methods have been employed with similar results [8]. In our implementation, keyframes within each video sequence were manually labeled, while the remaining frames were automatically aligned using a gradient-descent AAM fit described in [19,23].

3.1. AAM derived representations

The *shape* $\mathbf{s}$ of an AAM [8] is described by a 2D triangulated mesh. In particular, the coordinates of the mesh vertices define the shape $\mathbf{s}$ (see row 1, column (a), of Fig. 2 for examples of this mesh). These vertex locations correspond to a source appearance image, from which the shape is aligned (see row 2, column (a), of Fig. 2). Since AAMs allow linear shape variation, the shape $\mathbf{s}$ can be expressed as a base shape $\mathbf{s}_0$ plus a linear combination of m shape vectors $\mathbf{s}_i$:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} p_i \mathbf{s}_i \tag{1}$$

where the coefficients $\mathbf{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$ are the shape parameters. These shape parameters are typically divided into similarity parameters $\mathbf{p}_s$ and object-specific parameters $\mathbf{p}_o$, such that $\mathbf{p}^T = [\mathbf{p}_s^T, \mathbf{p}_o^T]$.

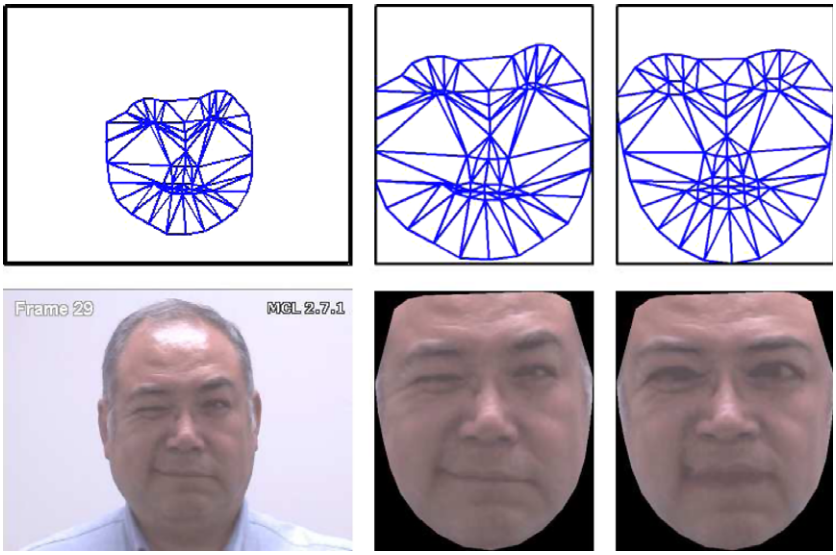


Fig. 2. Example of AAM derived representations (a) top row: input shape ($\mathbf{s}$), bottom row: input image, (b) top row: similarity normalized shape ($\mathbf{s}_n$), bottom row: similarity normalized appearance ($\mathbf{a}_n$), (c) top row: base shape ($\mathbf{s}_0$),0 bottom row: shape normalized appearance ($\mathbf{a}_0$).

We shall refer to $\mathbf{p}_s$ and $\mathbf{p}_o$ herein as the *rigid* and *non-rigid* shape vectors of the face, respectively. Rigid parameters are associated with the geometric similarity transform (i.e., translation, rotation and scale). Non-rigid parameters are associated with residual shape variations such as mouth opening, eyes shutting, etc. Procrustes alignment [8] is employed to estimate the base shape $\mathbf{s}_0$. Once we have estimated the base shape and shape parameters, we can normalize for various variables to achieve different representations as outlined in the following subsections.

3.1.1. Rigid normalized shape, $\mathbf{s}_n$

As the name suggests, this representation gives the vertex locations after all rigid geometric variation (i.e., translation, rotation and scale), relative to the base shape, has been removed. The similarity normalized shape $\mathbf{s}_n$ can be obtained by synthesizing a shape instance of $\mathbf{s}$, using Eq. (1), that ignores the similarity parameters of $\mathbf{p}$. An example of this similarity normalized mesh can be seen in row 1, column (b), of Fig. 2.

3.1.2. Rigid normalized appearance, $\mathbf{a}_n$

This representation contains appearance from which rigid geometric variation has been removed. Once we have rigid normalized shape $\mathbf{s}_n$, as computed in Section 3.1.1, the rigid normalized appearance $\mathbf{a}_n$ can be produced by warping the pixels in the source image with respect to the required translation, rotation, and scale (see row 2, column (b), of Fig. 2). This representation is similar to those employed in methods like [4,5] where the face is geometrically normalized with respect to the eye coordinates (i.e., translation, rotation and scale).

3.1.3. Non-rigid normalized appearance, $\mathbf{a}_0$

In this representation we can obtain the appearance of the face from which the non-rigid geometric variation has been normalized with respect to the base face shape $\mathbf{s}_0$. This is accomplished by applying a piece-wise affine warp on each triangle patch appearance in the source image so that it aligns with the base face shape. We shall refer to this representation as the face's *canonical appearance* (see row 2, column (c), of Fig. 2 for an example of this canonical appearance image) $\mathbf{a}_0$.

If we can remove all shape variation from an appearance, we'll get a representation that can be called as shape normalized appearance, $\mathbf{a}_0$.$\mathbf{a}_0$ can be synthesized in a similar fashion as $\mathbf{a}_n$ was computed in Section 3.1.2, but instead ensuring that the appearance contained within $\mathbf{s}$ now aligns with the base shape $\mathbf{s}_0$.

3.2. Features

Based on the AAM derived representations in Section 3.1 we define three types of features:

- **S-PTS:** *similarity normalized shape* $\mathbf{s}_n$ representation (see Eq. (1)) of the face and its facial features. There are 68 vertex points in $\mathbf{s}_n$ for both x and y coordinates, resulting in a raw 136 dimensional feature vector.
- **S-APP:** *similarity normalized appearance* $\mathbf{a}_n$ representation. Due to the number of pixels in $\mathbf{a}_n$ varying from image to image, we apply a mask based on $\mathbf{s}_0$ so that the same number of pixels (approximately 27,000) are in $\mathbf{a}_n$ for each image.
- **C-APP:** *canonical appearance* $\mathbf{a}_0$ representation where all shape variation has been removed from the source appearance except the base shape $\mathbf{s}_0$. This results in an approximately 27,000 dimensional raw feature vector based on the pixel values within $\mathbf{s}_0$.

The naming convention **S-PTS**, **S-APP**, and **C-APP** will be employed throughout the rest of this paper.

One might reasonably ask, why should **C-APP** be used as a feature as most of the expression information has been removed through the removal of the non-rigid geometrical variation? Inspecting Fig. 2 one can see an example of why **C-APP** might be useful. The subject is tightly closing his right eye. Even after the application of the non-rigid normalization procedure one can see there are noticeable visual artifacts (e.g., wrinkles) left that could be considered important in recognizing the presence/absence of pain. These appearance features may be critical in distinguishing between similar action units. Eye closure (AU 43), for instance, results primarily from relaxation of the levator palpebrae superioris muscle, which in itself produces no wrinkling. The wrinkling shown in Fig. 2 is produced by contraction of the orbicularis oculi (AU 6). The joint occurrence of these two actions, AU 6+43, is a reliable indicator of pain [10,31]. If AU 6 were ignored, pain detection would be less reliable. For any individual facial action, shape or appearance may be more or less important [6]. Thus, the value of appearance features will vary for different facial actions.

In the AAM, appearance can be represented as either **S-APP** or **C-APP**. They differ with respect to representation (rigid vs. non-rigid alignment, respectively) and whether shape and appearance are coupled (**S-APP**) or decoupled (**C-APP**). In training a classifier, the joint **C-APP** and **S-PTS** feature could perhaps offer improved performance over **S-APP** as it can treat the shape and appearance representations separately and linearly (unlike **S-APP**).

4. SVM classifiers

Support vector machines (SVMs) have proven useful in many pattern recognition tasks including face and facial action recognition. Because they are binary classifiers, they are well suited to the task of "pain" vs. "no-pain" classification. SVMs attempt to find the hyper-plane that maximizes the margin between positive and negative observations for a specified class. A linear SVM classification decision is made for an unlabeled test observation $\mathbf{x}^*$ by,

$$\mathbf{w}^T\mathbf{x}^* \underset{\text{false}}{\overset{\text{true}}{\gtrless}} b \tag{2}$$

where $\mathbf{w}$ is the vector normal to the separating hyperplane and b is the bias. Both $\mathbf{w}$ and b are estimated so that they minimize the structural risk of a train-set, thus avoiding the possibility of overfitting the training data. Typically, $\mathbf{w}$ is not defined explicitly, but through a linear sum of support vectors. As a result SVMs offer additional appeal as they allow for the employment of non-linear combination functions through the use of kernel functions, such as the *radial basis function* (RBF) and *polynomial* and *sigmoid* kernels. A linear kernel was used in our experiments due to its ability to generalize well to unseen data in many pattern recognition tasks [17]. Please refer to [17] for additional information on SVM estimation and kernel selection.

5. Experiments

5.1. Pain model learning

To ascertain the utility of various AAM representations, different classifiers were trained by using features of Section 3.2 in the following combinations:

S-PTS : *similarity normalized shape* $\mathbf{s}_n$
S-APP : *similarity normalized appearance* $\mathbf{a}_n$
C-APP + S-PTS : *canonical appearance* $\mathbf{a}_0$ combined with the similarity normalized shape $\mathbf{s}_n$.

To check for subject generalization, a leave-one-subject-out strategy was employed for cross validation. Thus, there was no overlap of subjects between the training and testing set. The number of training frames from all the video sequences was prohibitively large to train an SVM, as the training time complexity for a SVM is $O(m^3)$, where m is the number of training examples. In order to make the step of model learning practical, while making the best use of training data, each video sequence was first clustered into a preset number of clusters. Standard k-means clustering was employed, with k set to a value that reduces the training set to a manageable size. The value of k was chosen to be a function of the sequence length, such that the shortest sequence in the dataset had at least 20 clusters. Clustering was used only in the learning phase. Testing was carried out without clustering as described in the following sections.

Linear SVM training models were learned by iteratively leaving one subject out, which gives rise to N number of models, where N is the number of subjects. SVMs were trained at both the *sequence-* and *frame-levels*. At the *sequence-level*, a frame was labeled as pain if the sequence in which it occurred met criteria by the OPI (see Section 2.2.3). At the *frame-level*, following [10], a frame was labeled pain if its FACS-based pain intensity was equal to 1 or higher.

5.2. How important is registration?

At the *sequence-level*, each sequence was classified as pain present or pain absent. Pain present was indicated if the observer rating was 3 or greater. Pain absent was indicated if observer rating was 0. Learning was performed on clustered video frames; testing was carried out on individual frames. The output for every frame was a score proportional to the distance of the test-observation from the separating hyperplane. The predicted pain scores for individual frames across all the test sequences ranged from −2.35 to 3.21. The output scores for a sample sequence are shown in Fig. 3. For the specific sequence shown in Fig. 3, the predicted scores ranged from 0.48 to 1.13. The score values track the pain expression, with a peak response corresponding to frame 29 shown in Fig. 3.

To predict whether a sequence was labeled as "pain" the output scores of individual frames were summed together to give a cumulative score (normalized for the duration of the sequence) for the entire sequence,

$$D_{\text{sequence}} = \frac{1}{T}\sum_{i=1}^{T} d_i \tag{3}$$

where d_i is the output score for the i^{th} frame and T is the total number of frames in the sequence.

Having computed the sequence-level cumulative score in Eq. (3), we seek a decision rule of the form:

$$D_{\text{sequence}} \underset{\text{nopain}}{\overset{\text{pain}}{\gtrless}} \text{Threshold} \tag{4}$$

By varying the threshold in the decision rule of Eq. (4) one can generate the Receiver Operating Characteristic (ROC) of the classifier, which is a plot of the relation between the false acceptance rate and the hit rate. The false acceptance rate represents the proportion of no-pain video sequences that are predicted as pain containing sequences. The hit rate represents the detection of true pain. Often, a detection system is gauged in terms of the Equal Error Rate (EER). The EER is determined by finding the threshold at which the two errors, the false acceptance rate, and the false rejection rate, are equal.

In Fig. 4, we present the ROC curves for each of the representations discussed in Section 3.2. The EER point is indicated by a cross on the respective curves. The best results (EER = 15.7%) are for canonical appearance combined with similarity normalized shape (C-APP + S-PTS). This result is consistent with our previous work [18], in which we used AAMs for facial action unit recognition.

The similarity normalized appearance features (S-APP) performed at close-to-chance levels despite the fact that this representation can be fully derived from canonical appearance and similarity normalized shape.

5.3. How should we label data for learning pain?

A limitation of the approach described in Section 5.2 is that the ground truth was considered only at the video sequence level. In any given sequence the number of individual frames actually showing pain could be quite few. A coarse level of ground truth is common in clinical settings. We were fortunate, however, to have frame-level ground truth available as well, in the form of FACS annotated action units for each video frame. Following [24], as described in Section 2.2.2, a composite pain score was calculated for each frame. Composite pain scores ranged from 0 to 12.

Following [24], for the binary ground truth labels, we considered a pain score greater than zero to represent pain, and a score

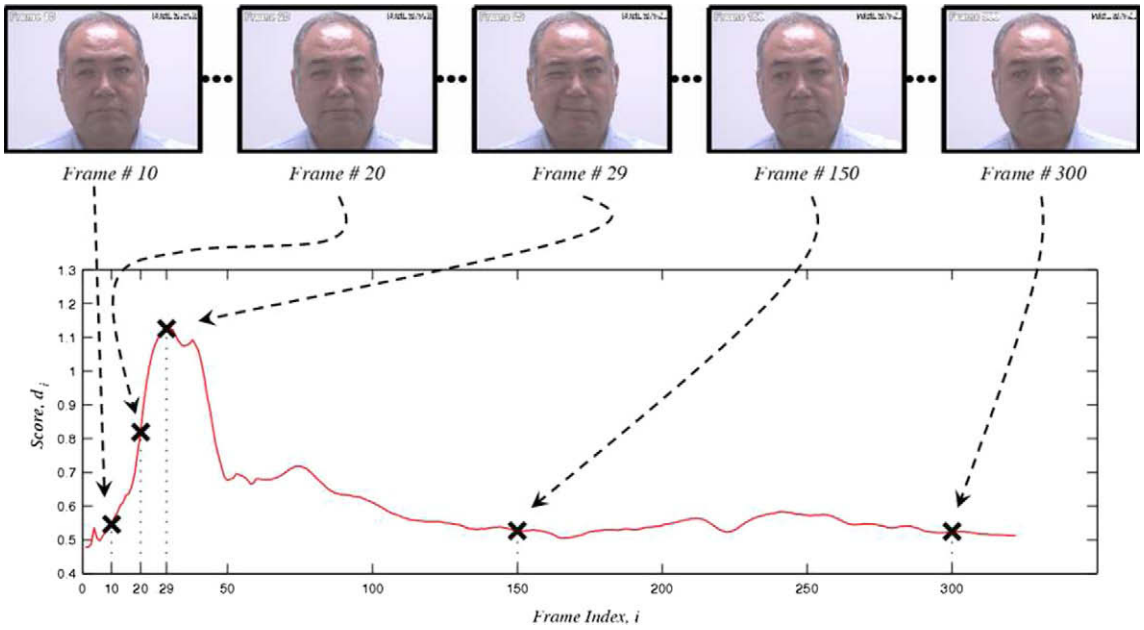


Fig. 3. Example of video sequence prediction. The *x*-axis in the above plot represents the frame index in the video, while the *y*-axis represents the predicted pain score. The dotted arrows show the correspondence between the image frames (top-row) and their predicted pain scores. For instance, Frame 29 in the top row shows an intense pain and corresponds to the peak in the plot.

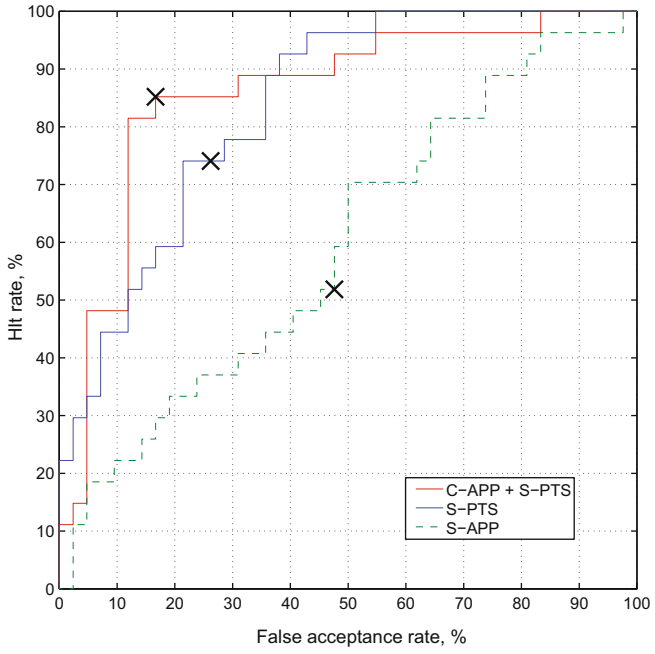


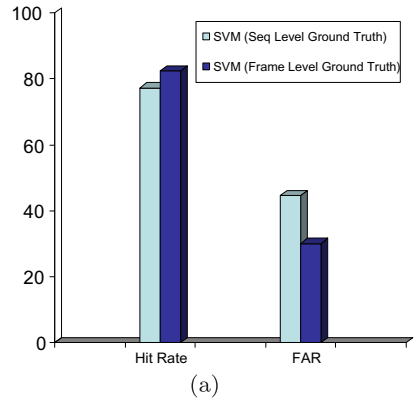
Fig. 4. Sequence-level pain detection results for experiments performed in Section 5.2, showing the ROC for classifiers based on three different representations. The crosses indicate the EER point. The best results (EER: 15.7%) are achieved by using a combination of canonical-appearance and similarity normalized shape (C-APP + S-PTS).

of zero to represent no-pain. For model learning, the previous clustering strategy was altered. Instead of clustering the video sequences as a whole, positive and negative frames were clustered individually prior to inputting them into the SVM. As before, clustering was not performed in the testing phase. As a comparison, we present results for frame-level prediction using an SVM trained on sequence level labels. In both cases, results are for S-PTS + C-APP features and leave-one-out cross-validation. They differ only in whether sequence-level or frame-level labels were provided to the SVM. The SVM based on frame-level ground truth improved the frame-level hit rate from 77.9% to 82.4%, and reduced False Acceptance Rate (FAR) by about a third, from 44% to 30.1% (see Fig. 5).

In Fig. 6 we show an example of how the respective SVM outputs compare with one another for a representative subject. The SVM trained on sequence level ground truth has consistently higher output in regions in which pain is absent. The SVM trained on frame level ground truth gives a lower score for the portion of the video sequence in which pain is absent. Previously, many no-pain frames that were part of pain video sequences were all forced to have a ground-truth label of 'pain'. This suggests why the previous SVM model has much higher FAR and lower correlation with frame-level ground truth. The present scheme precisely addresses the issue by employing frame-level ground truth and thus leads to better performance. The range of predicted pain scores for SVMs trained on frame-level ground truth was −2.45 to 3.29 across all the video sequences, while the range for the video sequence shown in Fig. 6 was −0.33 to 0.78.

Across all subjects, the improvement in performance should not come as a surprise, as the frame-level approach trains the classifier directly for the task at hand (i.e., frame-level detection). Whereas the sequence-level SVM was trained for the indirect task of sequence classification. More interestingly, the classifier trained with coarser (sequence-level) labels performs significantly better than "random chance" when tested on individual frames. In Fig. 7 we present the ROC curve for frame-level pain detection for classifiers trained with different ground-truth granularity and the ROC of a random classifier (i.e., applying an unbiased coin–toss to each frame). As one can see the ROC of the sequence-trained classifier lies significantly above that of the "random chance" classifier.

This result is especially interesting from a machine learning perspective. Hitherto, a fundamental barrier in learning and evaluating pain recognition systems is the significant cost and time associated with frame-based labeling. An interesting question for future research could be posed if one used the same labeling time and resources at the sequence-level. For same level of effort, one



(a)

Ground truth \ Prediction	Pain	No pain
Pain	2792	792
No pain	5363	6809

$$Hit\ rate = \frac{2792}{2792+392} = 77\%$$

$$\mathcal{FAR} = \frac{5363}{5363+6809} = 44\%$$

Ground truth \ Prediction	Pain	No pain
Pain	2957	632
No pain	3671	8501

$$Hit\ rate = \frac{2957}{2957+632} = 82\%$$

$$\mathcal{FAR} = \frac{3671}{3671+8501} = 30\%$$

(c)

Fig. 5. Frame-level performance based on experiments performed in Section 5.3. (a) Hit rate and false-acceptance rates for SVMs trained using different ground-truth granularity. Training SVMs by using frame-level ground truth improved performance. Frame-level hit rate increased from 77.9% to 82.4%, and frame-level false acceptance rate (FAR) decreased from 44% to 30.1%. (b) Confusion matrix for sequence-trained SVM. (c) Confusion matrix for frame-trained SVM.

could ground-truth a much larger sequence-level dataset, in comparison with frame-level labeling, and as a result employ that larger dataset during learning. One is then left with the question of which system would do better. Is it advantageous to have large sequence-level labeled datasets or smaller frame-level labeled datasets? Or more interestingly, what learning methods could be developed to leverage a hybrid of the two?

Because different kinds of expressions involve activities of different facial muscles we wished to visualize what regions of the face contribute towards effective pain detection. To accomplish this we formed an intensity image from the weighted combination of the learned support vectors for pain and no pain classes using their support weights (Fig. 8). For pain, the brighter regions represent more contribution, while for no pain, the darker regions represent less contribution. These plots highlight that regions around the eyes, eyebrows, and lips contribute significantly towards pain vs. no pain detection. These are same regions identified in previous literature as indicative of pain by observers.

6. Discussion

In this paper we explored various face representations derived from AAMs for detecting pain from the face. We explored two important questions with respect to automatic pain detection. First, how should one represent the face given that a non-rigid registration of the face is available? Second, at what level (i.e., sequence- or frame-based) should one label datasets for learning an automatic pain detector?

With respect to the first question we demonstrated that considerable benefit can be attained from non-rigid rather than rigid registrations of the face. In particular, we demonstrated that decoupling a face into separate non-rigid shape and appearance components offers significant performance improvement over those that just normalize for rigid variation in the appearance (e.g., just locating the eyes and then normalizing for translation,

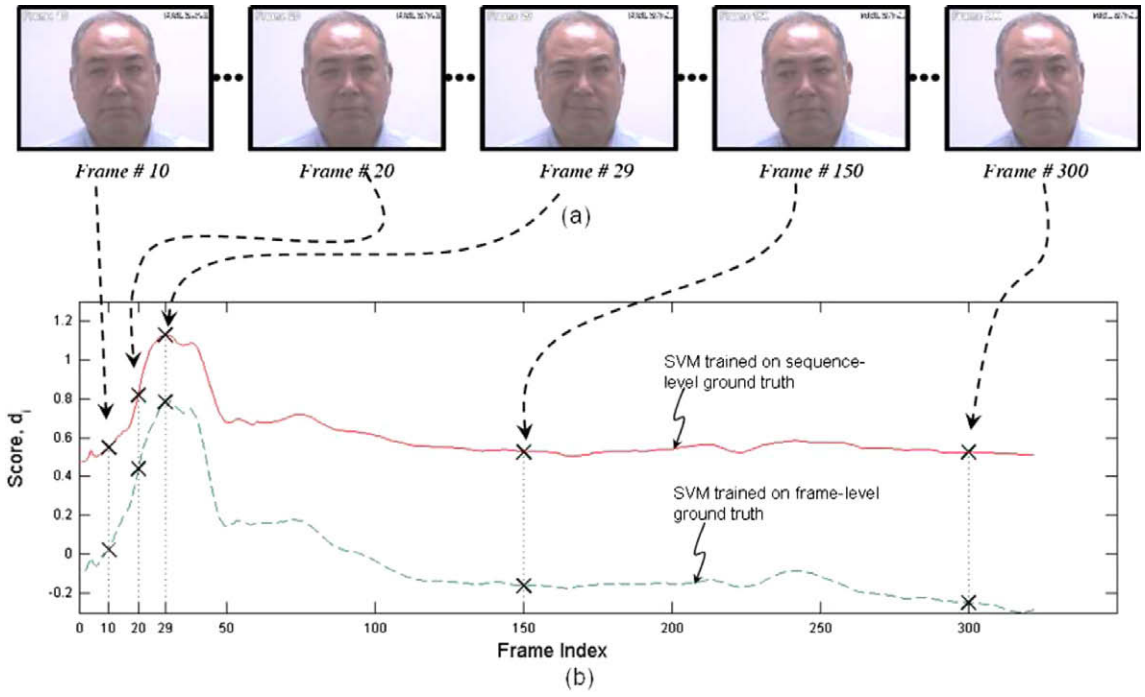


Fig. 6. Comparison between the SVM scores for sequence-level ground truth and frame-level ground truth. (a) Sample frames from a pain-video sequence with their frame indices, (b) scores for individual frames for the two SVM training strategies. Points corresponding to the frames shown in (a) are highlighted as crossed. Output of SVM trained on frame-level groundtruth remains lower for frames without pain, and hence leads to a lower false acceptance rate.

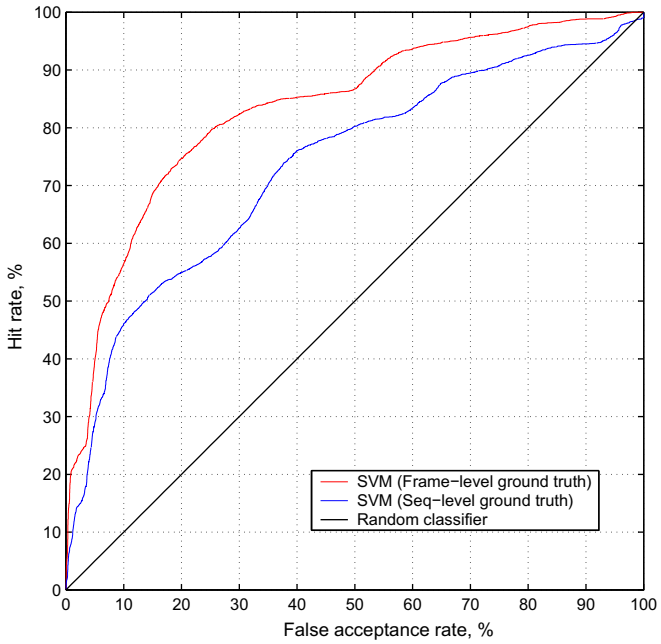


Fig. 7. Comparison of ROCs for SVMs trained on sequence- and frame-level labels. To demonstrate the efficacy of the sequence-level trained SVM on the frame-level detection task the ROC for a "random-chance" classifier is also included. One can see that although the sequence-level SVM behaves worse than the frame-level SVM it is significantly better than random-chance demonstrating that coarse-level labeling strategies are effective and useful in automatic pain recognition tasks.

Fig. 8. Weighted combination of support vectors to visualize contribution of different face regions for pain recognition. (a) For pain, (b) For no pain. For pain, the brighter regions represent more weightage. For no pain, the darker regions represent more weightage.

rotation and scale). This result is significant as most leading techniques for action unit [4,5] and pain [2] detection tasks are employing rigid rather than non-rigid registrations of the face.

We did not explore differences among possible appearance features. Relative strengths and weaknesses among various appearance features is an active area of ongoing research (see, for instance, [33]). Our findings have implications for work on this topic. Previous studies with Gabor filter responses, for instance, use rigid registration [2,33]. While rigid registration may be adequate for some applications (e.g., posed behavior or spontaneous behavior with little out-of-plane head motion), for others it appears not. We found that rigid registration of appearance had little information value in video from clinical pain assessments. Out-of-plane head motion was common in this context. Non-rigid registration of appearance greatly improved classifier performance. Our findings suggest that type of registration (rigid vs. non-rigid) may influence the information value and robustness of appearance features. When evaluating features, it is essential to consider the issues of out-of-plane rotation and types of registration.

We also did not consider the relative advantages of first- versus second-order classifiers. That is, is it better to detect pain directly or to detect action units first and then use the resulting action unit outputs to detect pain (or other expression of interest). This is an important topic in its own respect. Littlewort [2], for instance, first detected action units and then used the (predicted) action units in a classifier to detect pain. In the current study and in our own previous work [3] we detected pain directly from shape and appearance features without going through action unit detection first. Research on this topic is just beginning. Most previous studies in expression detection or recognition have been limited to posed behavior and descriptions of facial expression (e.g., action units or emotion-specific expressions, such as happy or sad). The field is just now beginning to address the more challenging question of detecting subjective states, such as clinical or induced pain. Our concern with second-order classifiers is that they are vulnerable to error at the initial step of action unit detection. Human observers have difficulty achieving high levels of reliability [6]; and classifiers trained on human-observer labeled data will be affected by that source of error variance. Alternatively, to the extent that specific facial actions are revealing [34,35], second-order classifiers may have an advantage. We are pursuing these questions in our current research.

Our results for the second question demonstrate that unsurprisingly, frame-level labels in learning are best for frame-level detection of pain. However, sequence-level trained classifiers do substantially better than chance even though they are being evaluated on a task they have not been directly trained for. This result raises the interesting question over how researchers in the automatic pain detection community should be using their resources when labeling future datasets. Should we still be labeling at the frame-level, ensuring that the datasets we learn from are modestly sized. Or, should we be employing hybrid labeling strategies where we label some portions at the frame- and some portions at the sequence-level allowing for learning from much larger datasets. The answer to these questions shall be the topic of our continuing research.

In summary, in a study of clinical pain detection, we found that the combination of non-rigidly registered appearance and similarity normalized shape maximized pain detection at both the sequence and frame levels. By contrast, rigidly registered appearance was of little value in sequence- or frame-level pain detection. With respect to granularity of training data, for frame-level pain detection, use of frame-level labels resulted in hit rate of 82% and false positive rate of 30%; the corresponding rates for sequence-level labels were 77% and 44%, respectively. These findings have implications for pain detection and machine learning more generally. Because sequence-level labeling affords collection of larger data sets, future work might consider hybrid strategies that combine sequence- and frame-level labels to further improve pain and expression detection. The current findings in clinical pain subjects suggest the feasibility of automatic pain detection in medical settings.

Acknowledgements

This research was supported by CIHR Grant MOP 77799 and NIMH Grant MH 51435.

References

[1] M. Bartlett, G. Littlewort, C. Lainscesk, I. Fasel, J. Movellan, Machine learning methods for fully automatic recognition of facial expressions and facial actions, in: IEEE International Conference on Systems, Man and Cybernetics, October 2004, pp. 592–597.
[2] G. Littlewort, M. Bartlett, K. Lee, Faces of pain – automated measurement of spontaneous facial expressions of genuine and posed pain, in: Proceedings of

the 9th International Conference on Multimodal Interfaces (ICMI), 2007, pp. 15–21.
[3] A.B. Ashraf, S. Lucey, J. Cohn., T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, The painful face – pain expression recognition using active appearance models, in: Proceedings of the 9th International Conference on Multimodal Interfaces (ICMI), 2007, pp. 9–14.
[4] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscesk, I. Fasel, J. Movellan, Recognizing facial expression: machine learning and application to spontaneous behavior, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2 (2005) 568–573. June.
[5] M.S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, J. Movellan, Fully automatic facial action recognition in spontaneous behavior, in: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, 2006, pp. 223–228.
[6] J.F. Cohn, Z. Ambadar, P. Ekman, Observer-based measurement of facial expression with the facial action coding system, in: J.A. Coan, J.B. Allen (Eds.), The Handbook of Emotion Elicitation and Assessment. Oxford University Press Series in Affective Science, Oxford University Press, New York, NY, 2007, pp. 203–221.
[7] J.F. Cohn, K.L. Schmidt, The timing of facial motion in posed and spontaneous smiles, International Journal of Wavelets Multiresolution and Information Processing 2 (2004) 1–12.
[8] T. Cootes, G. Edwards, C. Taylor, Active appearance models, PAMI 23 (6) (2001) 81–685.
[9] R.R. Cornelius, The Science of Emotion, Prentice Hall, Upper Saddler River, New Jersey, 1996.
[10] K.D. Craig, K.M. Prkachin, R.V.E. Grunau, The facial expression of pain, in: D.C. Turk (Ed.), Handbook of Pain Assessment, 2nd ed., Guilford, New York, 2001.
[11] A.C.d.C. Williams, H.T.O. Davies, Y. Chadury, Simple pain rating scales hide complex idiosyncratic meanings, Pain 85 457–463.
[12] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, second ed., John Wiley and Sons, Inc., New York, NY, USA, 2001.
[13] P. Ekman, W.V. Friesen, Facial Action Coding System, Consulting Psychologists Press, Palo Alto, CA, 1978.
[14] P. Ekman, W.V. Friesen, J.C. Hager, Facial action coding system: Research Nexus, Network Research Information, Salt Lake City, UT 2002.
[15] K.R. Scherer, Methods of research on vocal communication: paradigms and parameters, in: Scherer, P. Ekman (Eds.), Handbook of Methods in Non Verbal Behavior Research, Cambridge University Press, 1982.
[16] T. Hadjistavropoulos, K.D. Craig, Social influences and the communication of pain, in: Pain: Psychological Perspectives, Erbaulm, Newyork, 2004, pp. 87–112.
[17] C.W. Hsu, C.C. Chang, C.J. Lin, A practical guide to support vector classification, Technical Report, 2005.
[18] S. Lucey, A.B. Ashraf, J. Cohn, Investigating spontaneous facial action recognition through AAM representations of the face, in: K. Kurihara (Ed.), Face Recognition Book, Pro Literature Verlag, Mammendorf, Germany, 2007. April.
[19] I. Matthews, S. Baker, Active appearance models revisited, IJCV 60 (2) (2004) 135–164.
[20] M. Pantic, A. Pentland, A. Niholt, T.S. Huang, Human computing and machine understanding of human behavior: a survey, in: Proceedings of the ACM International Conference on Multimodal Interfaces, 2006.
[21] Y. Tian, J.F. Cohn, T. Kanade, Facial expression analysis, in: S.Z. Li, A.K. Jain (Eds.), Handbook of Face Recognition, Springer, New York, NY, 2005, pp. 247–276.
[22] M.F. Valstar, M. Pantic, Z. Ambadar, J.F. Cohn, Spontaneous vs. posed facial behavior: automatic analysis of brow actions, in: Proceedings of the ACM International Conference on Multimodal Interfaces, November 2006, pp. 162–170.
[23] J. Xiao, S. Baker, I. Matthews, T. Kanade, 2d vs. 3d deformable face models: representational power, construction, and real-time fitting, International Journal of Computer Vision 75 (2) (2007) 93–113.
[24] K. M Prkachin, The consistency of facial expressions of pain: a comparison across modalities, Pain 51 (1992) 297–306.
[25] P. Viola, M.J. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) (2004) 137–154.
[26] M. Pantic, L.J.M. Rothkrantz, Facial action recognition for facial expression analysis from static face images, IEEE Transactions on Systems, Man, and Cybernetics 34 (3) (2004) 1449–1461.
[27] M. Pantic, I. Patras, Detecting facial actions and their temporal segments in nearly frontal-view face image sequences, in: IEEE Conference on Systems, Man and Cybernetics (SMC), October 2005, pp. 3358–3363.
[28] A. Anastasi, Psychological Testing, fifth ed., Macmillan, NY, USA, 1982.
[29] J. Cohen, Statistical Power Analysis for the Social Sciences, Lawrence Erlbaum Associates, Hillsdale, NJ, USA, 1988.
[30] K. Prkachin, S. Berzinzs, R.S. Mercer, Encoding and decoding of pain expressions: a judgment study, Pain 58 (1994) 253–259.
[31] K. Prkachin, P. Solomon, The structure, reliability and validity of pain expression: evidence from patients with shoulder pain, Pain 139 (2008) 267–274.
[32] F.N. Kerlinger, Foundations of Behavioral Research: Educational, Psychological and Sociological Inquiry, Holt, Rinehart and Winston, NY, USA, 1973.
[33] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 683–1699.
[34] C. Darwin, The expression of the Emotions in Man and Animals, third ed., Oxford University, New York, 1872/1998.
[35] P. Ekman, E. Rosenberg, What the Face Reveals, second ed., Oxford, New York, 2005.