Image and Vision Computing 30 (2012) 197–205

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

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

Painful monitoring: Automatic pain monitoring using the UNBC-McMaster shoulder pain expression archive database ☆

Patrick Lucey [a,b,c,*], Jeffrey F. Cohn [b,c], Kenneth M. Prkachin [d], Patricia E. Solomon [e], Sien Chew [f], Iain Matthews [a,c]

[a] Disney Research Pittsburgh, Pittsburgh, PA, United States
[b] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, United States
[c] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, United States
[d] Department of Psychology, University of Northern British Columbia, United States
[e] School of Rehabilitation Sciences, McMaster University, Hamilton, Canada
[f] SAIVT Laboratory, Queensland University of Technology, Brisbane, Australia

ARTICLE INFO

Article history:
Received 1 July 2011
Received in revised form 26 November 2011
Accepted 4 December 2011

Keywords:
Pain
Active Appearance Models (AAMs)
Action Units (AUs)
FACS

ABSTRACT

In intensive care units in hospitals, it has been recently shown that enormous improvements in patient outcomes can be gained from the medical staff periodically monitoring patient pain levels. However, due to the burden/stress that the staff are already under, this type of monitoring has been difficult to sustain so an automatic solution could be an ideal remedy. Using an automatic facial expression system to do this represents an achievable pursuit as pain can be described via a number of facial action units (AUs). To facilitate this work, the "University of Northern British Columbia-McMaster Shoulder Pain Expression Archive Database" was collected which contains video of participant's faces (who were suffering from shoulder pain) while they were performing a series of range-of-motion tests. Each frame of this data was AU coded by certified FACS coders, and self-report and observer measures at the sequence level were taken as well. To promote and facilitate research into pain and augmentcurrent datasets, we have publicly made available a portion of this database, which includes 200 sequences across 25 subjects, containing more than 48,000 coded frames of spontaneous facial expressions with 66-point AAM tracked facial feature landmarks. In addition to describing the data distribution, we give baseline pain and AU detection results on a frame-by-frame basis at the binary-level (i.e. AU vs. no-AU and pain vs. no-pain) using our AAM/SVM system. Another contribution we make is classifying pain intensities at the sequence-level by using facial expressions and 3D head pose changes.

1. Introduction

In Atul Gawande's recent book entitled "The Checklist Manifesto" [1], he notes that tremendous improvement in patient outcomes in intensive care unit (ICU) settings have been achieved through adhering to standardized hygiene and monitoring checklists. One of these items on the checklist is pain monitoring, in which a medical staff member (i.e. nurse or physician) checks on a patient every 4 hours to evaluate whether they are suffering from pain and to make any needed adjustments in pain medication, treatment or diagnosis that may be warranted. However, due to the large workload that hospital staff currently experience, monitoring pain consistently and reliably has been difficult to achieve. As such, having a method to automatically monitor pain could be an ideal solution.

Measuring or monitoring pain is normally conducted via self-report as it is convenient and requires no special skill or staffing. However, self-report measures cannot be used when patients cannot communicate verbally (i.e. unconsciousness, breathing tubes interfering with speech, lacking function speech (infants)), so an observer rating is required where the observer chooses a face on the "faces of pain" scale which resembles the facial expression of the patient [3]. Both these measures have problems though, as they are subjective and do not give a continuous output over time (i.e. is the pain increasing, decreasing, or spiking?).[1] Many researchers have pursued the goal of obtaining a continuous objective measure of pain through analyzes of tissue pathology, neurological "signatures", imaging procedures, testing of muscle strength and so on [4]. These approaches

☆ This paper has been recommended for acceptance by special issue Guest Editors Rainer Stiefelhagen, Marian Stewart Bartlett and Kevin Bowyer.

* Corresponding author at: Disney Research Pittsburgh, Pittsburgh, PA, United States. Tel.: +1 14122986976.
E-mail address: pjlucey@gmail.com (P. Lucey).

[1] See William et al.'s work [2] for full description and analysis of these factors.

0262-8856/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2011.12.003

have been fraught with difficulty because they are often inconsistent with other evidence of pain [4], in addition to being highly invasive and constraining to the patient.

Another potential solution which has been looked at is to use facial expressions. Over the past twenty years, significant efforts have been made in identifying such facial actions [5–7]. Recently, Prkachin and Solomon [7] validated a Facial Action Coding System (FACS) [8] based measure of pain that can be applied on a frame-by-frame basis. A caveat on this manual approach is that it must be performed offline, where manual observations are both timely and costly, which makes clinical use prohibitive. However, such information can be used to train a real-time automatic system which could potentially provide significant advantage in patient care and cost reduction [9–12].

The deployment of an automatic pain monitoring tool represents a slue of applications that would be useful for the community. This is because the detection of psychologically meaningful states from facial behavior alone can be improved by knowing the context (e.g., clinical interview or assessment) and number of outcomes. (say two, i.e. yes/no or pain/no-pain). With these constraints, the application of an automatic facial expression detection system could be very successful. A recent example is of automatic smile detection in digital cameras where Whitehill et al. [13] constrained the goal to detecting only smile or no-smile. Employing a Gabor filter approach, they were able to achieve performance of up to 98% on a challenging dataset consisting of frontal faces spontaneously smiling or not in various environments, although no inferences were made about psychological state (e.g., enjoyment) and no temporal segmenting was required. In addition to pain monitoring, other examples of well-specified problems or contexts in which facial expression detection would be useful include driver fatigue detection, clinical status (e.g., symptomatic or not) and approach/avoidance in consumers (e.g., interested, disgusted or neutral).

The key to the success of these applications is to narrow the context of the target application, so the number of potential outcomes is small ensuring that enough data is available to build robust models so that reliable performance can be gained. However, to do this we need an abundance of data which is representative of the target application. To facilitate this, researchers at the McMaster University and University of Northern British Columbia (UNBC) captured video of patient's faces (who were suffering from shoulder pain) while they were performing a series of active and passive range-of-motion tests to their affected and unaffected limbs on two separate occasions. Each video frame was fully AU coded by certified FACS coders, and both observer and self-report measures at the sequence level were taken as well. To promote and facilitate research into pain as well as facial expression detection, the first portion of this dataset is now available for computer vision and pattern recognition researchers. With their particular needs in mind and through collaboration with CMU and University of Pittsburgh, the UNBC-McMaster Shoulder Pain Expression Archive includes:

1. Temporal spontaneous expressions: 200 video sequences containing spontaneous facial expressions relating to genuine pain,
2. Manual FACS codes: 48,398 FACS coded frames,
3. Self-report and observer ratings: associated pain self-report and observer ratings at the sequence level.
4. Tracked landmarks: 66 point AAM landmarks.

In addition to describing the data distribution, we give baseline pain and AU detection results on a frame-by-frame basis at the binary level (i.e. AU vs. no-AU and pain vs. no-pain). Another major contribution we make is to emulate a human observer by detecting pain intensities at a sequence level. As head pose motion is also indicative of a person in pain, we use the 3D parameters derived from the AAM to detect pain intensity in addition to facial expressions.

2. The UNBC-McMaster shoulder pain expression archive database

A total of 129 participants (63 male, 66 female) who were self-identified as having a problem with shoulder pain were recruited from 3 physiotherapy clinics and by advertisements posted on the campus of the McMaster University. One fourth were students and the rest included people from a wide variety of occupations over different age groups. Diagnosis of the shoulder pain varied, with participants suffering from arthritis, bursitis, tendonitis, subluxation, rotator cuff injuries, impingement syndromes, bone spur, capsulitis and dislocation. Over half of the participants reported use of medication for their pain.

All participants were tested in a laboratory room that included a bed for performing passive range-of-motion tests. After informed consent and information procedures were completed, participants underwent eight standard range-of-motion tests: abduction, flexion, and internal and external rotation of each arm separately [14]. Abduction movements involve lifting the arm forward and up in the sagittal plane. In internal rotation, the arm is bent 90° at the elbow, abducted 90°, and turned internally. External rotation is the same except that the arm is turned externally. Abduction, flexion, and internal and external rotations were performed under active and passive conditions. Active tests differed from the passive tests in being under the control of the patient who was instructed to move the limb as far as possible. Active tests were performed with the patient in a standing position. Passive tests were performed by a physiotherapist who moved the limb until the maximum range was achieved or was asked to stop by the patient. During passive tests, the participant was resting in a supine position on the bed with his or her head supported and stabilized by a pillow. Active tests were performed prior to passive tests because that is the usual sequence in which they are conducted clinically. The order of tests within active and passive conditions was randomized. Tests were performed on both the affected and the unaffected limb to provide a within-subject control.

During both active and passive tests, a Sony digital camera recorded the participants' facial expressions. Camera orientation was initially frontal in the active condition, although change in pose was common. Camera orientation in the passive condition was approximately 70° off frontal, with the face viewed from below. Video sequences approximately ranged from 60 to 700 frames. The audio recordings were not kept for analysis unfortunately.

A card, listing verbal pain descriptors was available to help participants provide verbal ratings of the pain produced on each test. Each card displayed two Likert-type scales [15]. One consisted of words reflecting the sensory intensity of pain. The other consisted of words reflecting the affective-motivational dimension. These scales have been subject to extensive psychophysical analyses, which have established their properties as ratio-scale measures of the respective underlying dimensions. Each scale had 15 items labelled from “A” to “O”. The sensory scale started at “extremely weak” and finished at “extremely intense”; the affective-motivational scale started at “bearable” and finished at “excruciating”. In addition, participants completed a series of 10 cm Visual Analog Scales (VAS), anchored at each end with the words, “No pain” and “Pain as bad as could be”. The three scales were completed by participants after each test. Specifically, after each test, participants rated the maximum pain it had produced using the sensory and affective verbal pain descriptors and the VAS.

Offline, independent observers rated pain intensity (OPI) from the recorded video. Observers had 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 [16]. Correlation between the observer's rating on the OPI and subject's self-reported

Fig. 1. Examples of some of the sequences from the UNBC-McMaster pain shoulder archive: (a) the ratings were OPI = 5, VAS = 9,SEN = 11,AFF = 10, the peak-frame (60) had AU codes of 6c + 9b + 43 which was equal to a PSPI rating of 6 for that frame; (b) the ratingswere OPI = 4, VAS = 6,SEN = 10,AFF = 7, the peak-frame (322) had AU codes of 4a + 6 d + 7 d + 12 d + 43 which was equal to a PSPI rating of 6 for that frame; (c) the ratings were OPI = 3, VAS = 7,SEN = 7,AFF = 7, the peak-frame (352) had AU codes of 4e + 6a + 7e + 9 d + 10 d + 25c + 43 which was equal to a PSPI rating of 14 for that frame; (d) the ratings were OPI = 2, VAS = 6,SEN = 8,AFF = 5, the peak-frame (129) had AU codes of 4b + 6c + 12c + 43 which was equal to a PSPI rating of 6 for that frame.

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 [17] 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. Examples of the active portion of the dataset with the associated self-report measures, along with the FACS codes and frame-by-frame level pain rating (see next subsection) is given in Fig. 1.

2.1. FACS coding

Each frame was extracted from the video and coded using FACS [8]. Each facial action is described in terms of one of 44 individual action units (AUs). Because there is a considerable amount of literature in which FACS has been applied to pain expression, only the actions that have been implicated as possibly related to pain were focused on: brow-lowering (AU4), cheek-raising (AU6), eyelid tightening (AU7), nose wrinkling (AU9), upper-lip raising (AU10), oblique lip raising (AU12), horizontal lip stretch (AU20), lips parting (AU25), jaw dropping (AU26), mouth stretching (AU27) and eye-closure (AU43). With the exception of AU 43, each action was coded on a 5 level intensity dimension (A–E) by one of three coders who were certified FACS coders. Actions were coded on a frame-by-frame basis. All coding was then reviewed by a fourth certified FACS coder.

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. Intercoder percent agreement as calculated by the Ekman–Friesen formula [8] was 95%, which compares favorably with other research in the FACS literature.

2.2. Prkachin and Solomon pain intensity scale

Beginning in 1992, Prkachin [6] found that four actions - brow lowering (AU4), orbital tightening (AU6 and AU7), levator contraction (AU9 and AU10) and eye closure (AU43) - carried the bulk of information about pain. In a recent follow up to this work, Prkachin and Solomon [7] confirmed these four "core" actions contained the majority of pain information. They defined pain as the sum of intensities of brow lowering, orbital tightening, levator contraction and eye closure. The Prkachin and Solomon pain intensity (PSPI) metric is defined as:

$$PSPI = AU4 + max(AU6, AU7) + max(AU9, AU10) + AU43 \quad (1)$$

That is, the sum of intensity of AU4, the maximum intensity value of AU6 and AU7, the maximum intensity value of AU9 and AU10 and AU43 to yield a 16-point scale.[2] For the example in Fig. 1(a), the peak frame here (60) has been coded as AU 6c + 9b + 43. This would result in a PSPI of 3 + 2 + 1 = 6. Similarly in Fig. 1(b), the peak frame has been coded as AU 4a + 6 d + 7 d + 12 d + 43, which equals 1 + 4 + 1 = 6, as AU4 has an intensity of 1, AU6 and AU7 both have intensity of 4 so just the maximum 4 is taken and AU43 has an intensity of 1 (eyes are shut).

The PSPI [7] FACS pain scale is currently the only metric which can define pain on a frame-by-frame basis. All frames in this dataset were coded using the PSPI. For more information on the relative merits of the particular self-report measures and how they relate to PSPI and FACS, please refer to [7].

2.3. Analysis of distributed portion of the pain corpora

From the entire available UNBC-McMaster Pain Shoulder Archive, 200 sequences from 25 different subjects in the active portion of the dataset has been prepared for distribution to the research community. From these 200 sequences there is a total of 48398frames that have been FACS coded and AAM tracked. The inventory of the total number of frames which have been coded from each AU and their intensity is given in Table 1. The number of frames and the associated PSPI score is given in Table 2. From this, it can be seen that 83.6% of the frames had a PSPI score of 0, and 16.4% had frames in which had a PSPI of score ≥ 1.

[2] The intensity of action units (AUs) are scored on a 6-point intensity scale that ranges from 0 (absent) to 5 (maximum intensity). Eye closing (AU43) binary (0 = absent, 1 = present). In FACS terminology, ordinal intensity is denoted by letters rather than numeric weights, i.e., 1 = A, 2 = B, . . . 5 = E.

Table 1
The AU inventory on the UNBC-McMaster shoulder pain archive, where the frequency of each AU and its intensity is given along with the total. Note that for AU43, the only intensity is A (i.e. they eye can only be open or shut).

AU	A	B	C	D	E	Total
4	202	509	225	74	64	1074
6	1776	1663	1327	681	110	5557
7	1362	991	608	305	100	3366
9	93	151	68	76	35	423
10	171	208	63	61	22	525
12	2145	1799	2158	736	49	6887
20	286	282	118	0	20	706
25	767	803	611	138	88	2407
26	431	918	265	478	1	2093
43	2434	–	–	–	–	2434

Table 2
The inventory on the UNBC-McMaster Shoulder Pain Archive according to the Prkachin-Solomon Pain Intensity (PSPI) pain metric, where the percentage of each pain intensity relative to the total number of frames is given (N = 48398).

PSPI	0	1–2	3–4	5–6	7–8	9–10	11–12	13–14	15–16
% total	82.71	10.87	4.57	1.06	0.27	0.20	0.26	0.05	0.01

Examples of this data are given in Fig. 1 and it is apparent that there is some head movement that occurs during these sequences. To quantify how much head movement occurred, we used the 3D parameters from the AAM to estimate the pitch, yaw and roll [18]. The histograms of these parameters are shown in Fig. 2. In terms of pitch, yaw and roll the mean was −0.38, −0.21 and −0.23° and the variance was 23.58, 40.82 and 33.28. However, these parameters differed quite a bit when a person was in no-pain (PSPI = 0) and in pain (PSPI ≥ 1). When the PSPI was equal to 0 the variance in terms of pitch, yaw and roll was 22.69, 37.03 and 29.19. When the PSPI was ≥ 1, the variance increased to 26.72, 55.61 and 48.52 which suggested that head movement coincided with painful facial expression. Overall, close to 90% of all frames in this distribution were within 10 degrees of being fully frontal and over 99% were within 20 degrees from the fully frontal view.

At the sequence level we show the inventory of some self-report and observer measures. Table 3 shows the inventory of the visual analogue scale (VAS) and observer pain intensity (OPI) measures for the 200 sequences. On the left side of the table, it can be said that there is a nice spread of VAS measures from 0–10. With the OPI measures, there is slightly less than half with no observable pain. For the sequence-level experiments, we will be using the OPI ratings so that we can get our automatic system to mimic a human observer. The affective and sensory self-report measures across the 200 sequences will also be available in the distribution.

Table 3
The inventory on the self-report and observer measures of the UNBC-McMaster shoulder pain archive at the sequence level. The self-report Visual Analogue Scale (VAS), ranging from 0 (no-pain) to 10 (extreme pain) and the Observed Pain Intensity (OPI), ranging from 0 (no-pain observed) to 5 (extreme pain observed).

	Frequency										
Scale	0	1	2	3	4	5	6	7	8	9	10
VAS	35	42	24	20	21	11	11	6	18	10	2
OPI	92	25	26	34	16	7					

3. AAM landmarks

In our system, we employ an Active Appearance Model (AAM) based system which uses AAMs to track the face and extract visual features. In the data distribution we include the 66 point AAM landmark points for each image. This section describes how these landmarks were generated.

3.1. Active Appearance Models (AAMs)

Active Appearance Models (AAMs) have been shown to be a good method of aligning a pre-defined linear shape model that also has linear appearance variation, to a previously unseen source image containing the object 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 [19].

The shape **s** of an AAM [19] is described by a 2D triangulated mesh. In particular, the coordinates of the mesh vertices define the shape $s = [x_1, y_1, x_2, y_2, \ldots, x_n, y_n]$, where n is the number of vertices. These vertex locations correspond to a source appearance image, from which the shape was aligned. Since AAMs allow linear shape variation, the shape s can be expressed as a base shape $\mathbf{s}_0$ plus a linear combination of m shape vectors $\mathbf{s}_i$:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} p_i \mathbf{s}_i \tag{2}$$

where the coefficients $\mathbf{p} = (p_1, \ldots, p_m)^T$ are the shape parameters. These shape parameters can typically be divided into rigid similarity parameters $\mathbf{p}_s$ and non-rigid object deformation parameters $\mathbf{p}_o$, such that $\mathbf{p}^T = \left[\mathbf{p}_s^T, \mathbf{p}_o^T\right]$. Similarity parameters are associated with a geometric similarity transform (i.e. translation, rotation and scale). The object-specific parameters, are the residual parameters representing non-rigid geometric variations associated with the determining object shape (e.g., mouth opening, eyes shutting, etc.). Procrustes alignment [19] is employed to estimate the base shape $\mathbf{s}_0$.

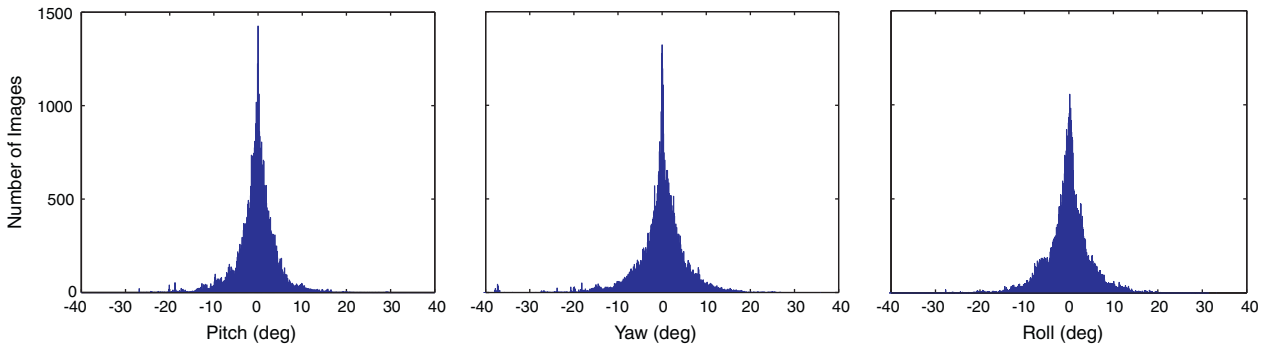

Fig. 2. Histograms of the pitch, yaw and roll taken from the 3D AAM parameters across the UNBC-McMaster shoulder pain expression archive database.

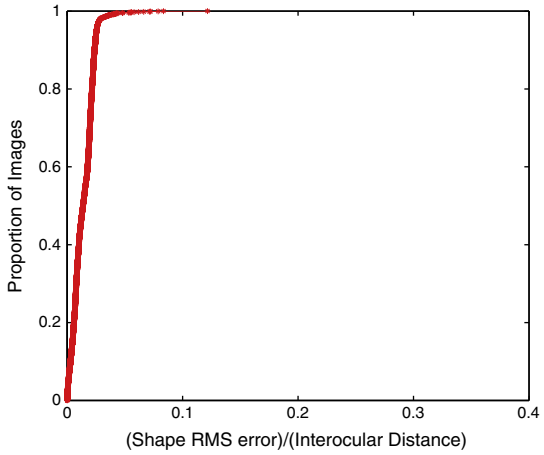

Fig. 3. Fitting curve for the AAM compared against the manual 66 point landmark points. The error ratio of the x axis is the total shape RMS error in terms of pixels after all meshes were similarity normalized divided by the interocular distance of 50 pixels.

Keyframes within each video sequence were manually labelled, while the remaining frames were automatically aligned using a gradient descent AAM fitting algorithm described in [20]. Fig. 4 shows the AAM in action, with the 66 point mesh being fitted to the patient's face in every frame.

3.2. Gaining 3D information from an AAM

From the 2D shape model we can derive the 3D parameters using non-rigid structure from motion. If we have a 2D AAM, a sequence of images $\mathbf{I}^t(\mathbf{u})$ for $t=0,\ldots,N$, and have tracked through the sequence with the AAM, then denote the AAM shape parameters at time t by $\mathbf{p}^t = (p_1^t,\ldots,p_m^t)$. Using Eq. (2) we can compute the 2D AAM shape vectors s^t for each time t:

$$\mathbf{s}^t = \begin{pmatrix} u_1^t & u_2^t & \ldots & u_n^t \\ v_1^t & v_2^t & \ldots & v_n^t \end{pmatrix} \tag{3}$$

A variety of non-rigid structure-from-motion algorithms have been proposed to convert the tracked feature points in Eq. (3) into 3D linear shape models. In this work we stack the 2D AAM shape vectors in all N images into a measurement matrix:

$$\mathbf{W} = \begin{pmatrix} u_1^0 & u_2^0 & \ldots & u_n^0 \\ v_1^0 & v_2^0 & \ldots & v_n^0 \\ \vdots & \vdots & \vdots & \vdots \\ u_1^N & u_2^N & \ldots & u_n^N \\ v_1^N & v_2^N & \ldots & v_n^N \end{pmatrix} \tag{4}$$

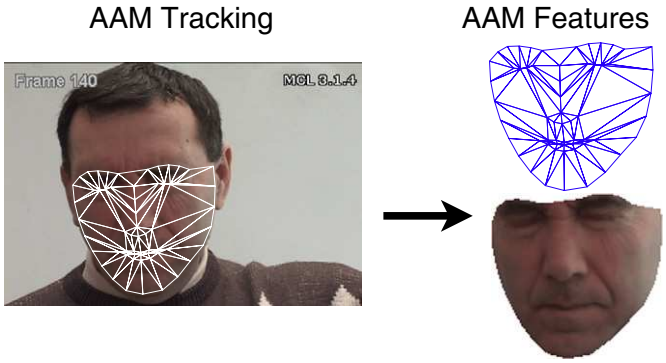

Fig. 4. Once the AAM has tracked a person's face we can derive some feature representations: (top) SPTS — similarity normalized shape and (bottom) CAPP — canonical normalized appearance.

If this data can be explained by a set of 3D linear shape modes, then W can be represented as

$$\mathbf{W} = \begin{pmatrix} \mathbf{P}^0 & p_1^0\mathbf{P}^0 & \ldots & p_{\bar{m}}^0\mathbf{P}^0 \\ \mathbf{P}^1 & p_1^1\mathbf{P}^1 & \ldots & p_{\bar{m}}^1\mathbf{P}^1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{P}^N & p_1^N\mathbf{P}^N & \ldots & p_{\bar{m}}^N\mathbf{P}^N \end{pmatrix} \begin{pmatrix} \bar{\mathbf{s}}_0 \\ \bar{\mathbf{s}}_1 \\ \vdots \\ \bar{\mathbf{s}}_{\bar{m}} \end{pmatrix} \tag{5}$$

which = $\mathbf{MB}$, where $\mathbf{M}$ is a $2(N+1)\times 3(\bar{m}+1)$ scaled projection matrix and $\mathbf{B}$ is a $3(\bar{m}+1)\times n$ shape matrix (setting the number of 3D vertices $\bar{n}$ to equal the number of AAM vertices textitn). Since $\bar{m}$ is the number of 3D shape vectors, it is usually small and the rank of $\mathbf{W}$ is at most $3(\bar{m}+1)$.

We perform a Singular Value Decomposition (SVD) on $\mathbf{W}$ and factorize it into the product of a $2(N+1)\times 3(\bar{m}+1)$ matrix $\tilde{\mathbf{M}}$ and a $3(\bar{m}+1)\times n$ matrix $\tilde{\mathbf{B}}$. This decomposition is not unique, and is only determined up to a linear transformation. Any non-singular $3(\bar{m}+1)\times 3(\bar{m}+1)$ matrix $\mathbf{G}$ and its inverse could be inserted between $\tilde{\mathbf{M}}$ and $\tilde{\mathbf{B}}$ and their product would still equal $\mathbf{W}$. The scaled projection matrix $\mathbf{M}$ and the shape vector matrix $\mathbf{B}$ are then given by:

$$\mathbf{M} = \tilde{\mathbf{M}}\,\mathbf{G}, \textit{and} \qquad \mathbf{B} = \mathbf{G}\tilde{\mathbf{B}} \tag{6}$$

where $\mathbf{G}$ is the corrective matrix. Once $\mathbf{G}$ has been determined, $\mathbf{M}$ and $\mathbf{B}$ can be recovered. So to summarize, given that we have the 2D tracking results, the 3D shape modes can be computed from the 2D AAM shape modes and the 2D AAM tracking results. See [18] for full details.

3.3. AAM accuracy

In checking the AAM alignment accuracy to manually landmarked images, we first similarity normalized all tracked AAM points and manual landmarks to a common mesh size and rotation, with a interocular distance of 50 pixels and aligned to the center of the eye coordinates. We then compared 2584 manually landmarked images against their AAM counterpart. The fitting curve for the AAM is shown in Fig. 3. As can be seen in this curve, nearly all of the AAM landmarks are within0.1 of the error ratio (i.e. pixel RMS (root-mean-square) error divided by the interocular distance). This is negligible given that this corresponds to 2 pixels RMS error of the manual landmarks based on a distance of 50 pixels between the center of the eyes. This highlights the benefit of employing person-specific model such as an AAM, as near perfect alignment can result.

4. Frame-by-frame level experiments

In this section, we describe two binary experiments that we conducted for i) AU (i.e. AU vs. no-AU) and ii) pain (i.e. pain vs. no-pain) detection at a frame-level. We first describe our baseline AAM/SVM system.

4.1. AAM/SVM baseline system

Once we have tracked the patient's face by estimating the shape and appearance AAM parameters, we can use this information to derive features from the face. From the initial work conducted in [9,21,12], we extracted the following features:

- SPTS: the similarity normalized shape, $\mathbf{s}_n$, refers to the 66 vertex points in s_n for both the x- and y- coordinates, resulting in a raw 132 dimensional feature vector. These points are the vertex locations after all the rigid geometric variation (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. (2), that ignores the similarity parameters $\mathbf{p}$.
- CAPP: the canonical normalized appearance $\mathbf{a}_0$ refers to where all the non-rigid shape variation has been normalized with respect to

the base 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. For this study, the resulting 87×93 synthesized grayscale image was used.

Support vector machines (SVMs) were then used to classify individual action units as well as pain. SVMs attempt to find the hyperplane that maximizes the margin between positive and negative observations for a specified class. A linear kernel was used in our experiments due to its ability to generalize well to unseen data in many pattern recognition tasks [22]. LIBSVM was used for the training and testing of SVMs [23].

In all experiments conducted, a leave-one-subject-out strategy (i.e. 25-fold cross validation) was used and each AU and pain detector was trained using positive examples which consisted of the frames that the FACS coder labelled containing that particular AU (regardless of intensity, i.e. A-E) or pain intensity of 1 or more. The negative examples consisted of all the other frames that were not labelled with that particular AU or had a pain intensity of 0.

In order to predict whether or not a video frame contained an AU or pain, the output score from the SVM was used. As there are many more frames with no behavior of interest than frames of interest, the overall agreement between correctly classified framescan skew the results somewhat. As such we used the receiver operating characteristic (ROC) curve, which is a more reliable performance measure. This curve is obtained by plotting the hit-rate (true positives) against the false alarm rate (false positives) as the decision threshold varies. From the ROC curves, we used the area under the ROC curve (A'), to assess the performance. The A' metric ranges from 50 (pure chance) to 100 (ideal classification).[3] An upper-bound on the uncertainty of the A' statistic was obtained using the formula $s = \sqrt{\frac{A'(100-A')}{min\{n_p,n_n\}}}$ where n_p, n_n are the number of positive and negative examples [24,13].

4.2. AU detection results

We conducted detection for ten AUs (4, 6, 7, 9, 10, 12, 20, 25, 26 and 43). The results for the AU detection with respect to the similarity-normalized shape (SPTS), the canonical appearance (CAPP) and the combined (SPTS + CAPP) features are shown in Table 4. In terms of the overall average accuracy of the AU detection, the performance is rather good with combined representation gaining the best overall performance of 81.8, slightly better than CAPP (79.2) and SPTS (78.0).

In terms of individual AU detection, it can be seen that best performance is gained for the strong expressions such as AU6, 10, 12 and 43. Due to the amount of very strong examples in the distribution (i.e. AU intensity is greater than A), it can be seen that robust performance can be gained. For full analysis of AU experiments see [12].

4.3. Pain detection at frame-level

The results for automatically detecting pain are given in Fig. 5, which shows a clearer view of the trend we observed in the AU detection results. For the individual feature sets, SPTS achieved 76.9 area underneath the ROC curve and then the CAPP features yielding the best results with 80.9. When we combine the different feature sets, we again see the benefit of fusing the various representations together showing that there exists complimentary information with the performance increasing to 83.9% (Fig. 6).

The big question that these results raise is that in terms of timing accuracy, what granularity or window of time is required to flag a patient in pain? Can we do this at the frame-level, second-level, minute-level or chunks of minutes? Also, do we need to detect the different intensities of pain (not just a binary pain/no-pain decision). These aspects are looked at in the next section.

[3] In literature, the A' metric varies from 0.5 to 1, but for this work we have multiplied the metric by 100 for improved readability of results.

Table 4
Results showing the area underneath the ROC curve for the similarity-normalized shape (SPTS) and appearance (SAPP) as well as the canonical appearance (CAPP) features. Note the average is a weighted one, depending on the number of positive examples.

AU	SPTS	CAPP	SPTS and CAPP
4	72.5 ± 3.1	60.0 ± 1.5	57.1 ± 1.5
6	80.1 ± 1.7	85.1 ± 0.5	85.4 ± 0.5
7	71.3 ± 0.8	82.6 ± 0.8	80.4 ± 0.7
9	75.1 ± 2.4	84.1 ± 1.6	85.3 ± 1.7
10	87.9 ± 1.7	83.2 ± 1.9	89.2 ± 1.4
12	79.4 ± 0.5	84.6 ± 0.5	85.7 ± 0.4
20	75.7 ± 1.7	61.7 ± 1.9	77.9 ± 1.6
25	78.8 ± 0.9	70.9 ± 1.0	78.0 ± 0.8
26	73.5 ± 1.1	54.7 ± 1.1	71.0 ± 1.0
43	83.1 ± 0.6	86.7 ± 0.7	87.5 ± 0.7
AVG	78.0 ± 0.8	79.2 ± 0.8	81.8 ± 0.8

5. Sequence-level pain detection

A prime motivation behind this paper and our work is to have an automatic facial expression detection system which can replicate the job of a care-giver when monitoring pain of a patient. To see how well we can do that we decided to do a series of sequence-level experiments to determine if our system could estimate the pain intensity like an expert observer could. To do that, we used the OPI sequence-level ratings which range from 0 (no pain) to 5 (strong pain) (see Section 2 for more details).

A leave-one-out-subject data partitioning strategy was employed for SVM training and testing. Positive training instances were taken from the OPI level of interest, and negative training examples were taken from all other OPI levels to train the classifier model for each OPI subject in the dataset. Likewise in testing, the sequences of each and every subject were used for prediction of the OPI class label. A majority vote scheme was used for the prediction, such that the predicted class label fora test frame was assigned by the classifier model which produced the highest probability score. From this, the number of votes obtained by each class (i.e. the OPI intensity level) was used to construct the confusion matrices. To deal with spurious noisy signals, a simple moving-average smoothing filter was applied to the SVM output probability scores. The results for classifying pain at the sequence-level using CAPP features is given in Table 5.

From these results, it can be seen that reasonable classification for OPI(0–1) intensities can be gained, but for other intensities less than

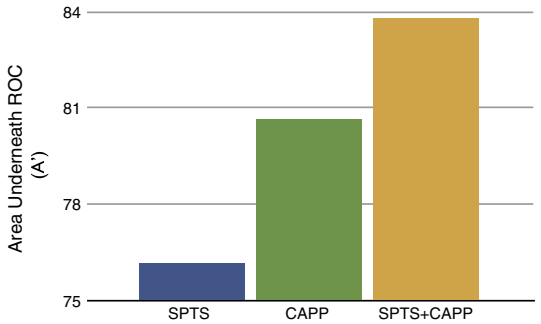

Fig. 5. The performance of the various features for the task of pain detection at the frame-level (yellow = SPTS, green = CAPP). The upper-bound error for all feature sets varied from approximately ± 0.67 to 0.80.

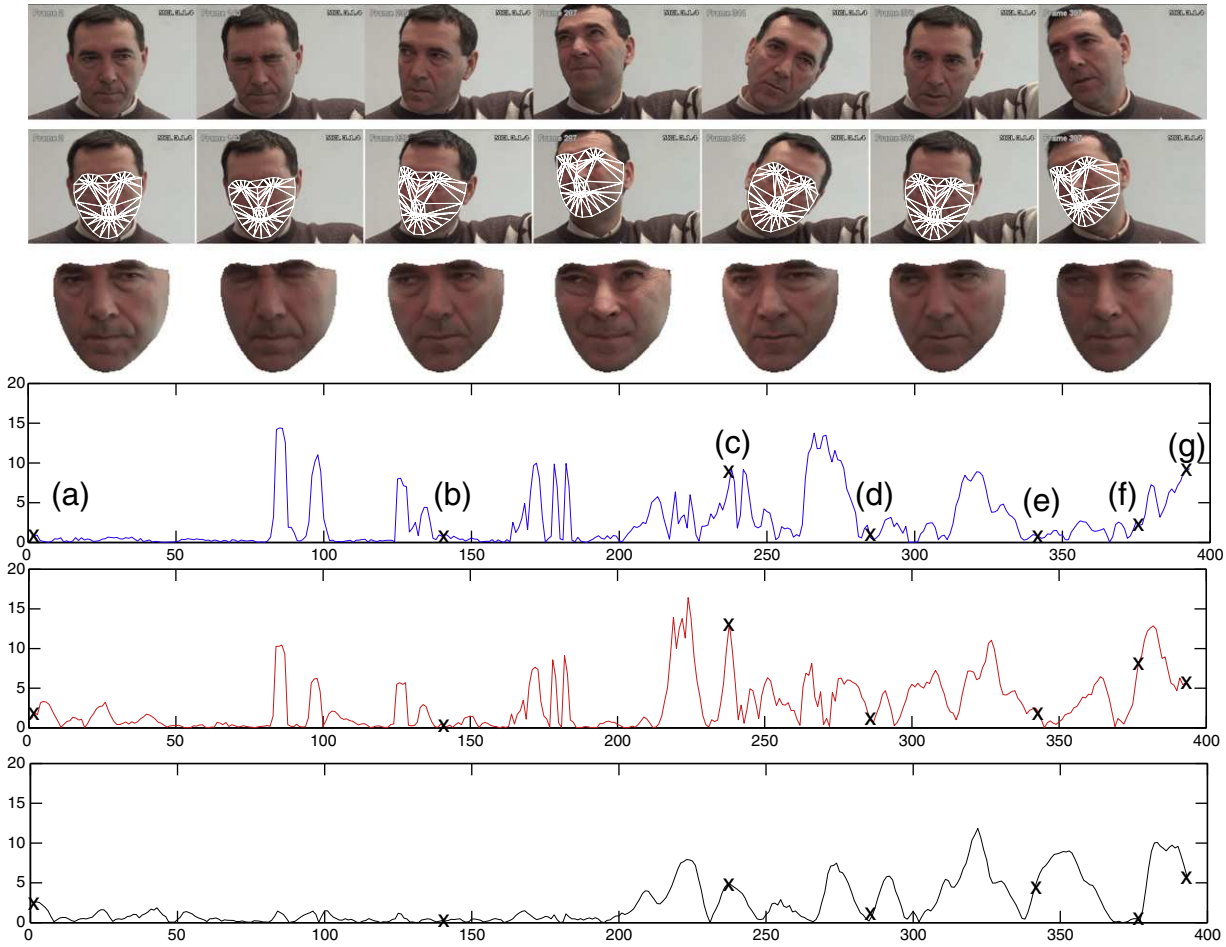

Fig. 6. Top row: image sequence example of a participant in considerable pain (OPI = 5), which includes both considerable facial expression and head movement; 2nd row: even with this large movements the AAM tracks the face very accurately; 3rd row: the normalized appearance features(CAPP) remove the rigid head movement; 4th row: Using the 3D-AAM parameters (see Section 3.2) we plotted the change in pitch motion in degrees over a window of 5 frames; 5th row: change in yaw motion in degrees over a window of 5 degrees; bottom row: change in roll motion in degrees over a window of 5 degrees. (N.B. The "x" on the curves correspond to the frames shown above: (a) frame 2, (b) frame 140, (c) frame 239, (d) frame 287, (e) frame 344, (f) frame 376, (g) frame 397).

desired performance is obtained. In the example given in the top row of Fig. 2, it is obvious that the participant is in considerable pain (i.e. OPI = 5). It is apparent that not only are the facial expressions intense, but there is also considerable head movement. This is a problem associated with our AAM approach, as even though we can track the face (second row) with great accuracy, which is great for facial expression detection, the resultant AAM features (3rd row) nullify the head pose motion information which could be used to classify pain intensities. Using the 3D-AAM parameters (see Section 3.2), it can be seen changes in head pose position coincide when the participant is in pain for pitch (4th row), yaw (5th row) and roll (bottom row).

However, when we analyze the pitch, yaw and roll motion by OPI group (0–5) we see a much different picture. As can be seen in Fig. 7, there is no perceivable correlation between changes in 3D head pose motion and the OPI intensity rating. This shows that there is no common trait across subjects with regards to head pose change and pain (i.e. everyone does not move their head when in pain and it varies from person to person), which suggests that this is individual trait. One possible reason for this is due to the effect of human behavior. It was observed that a distracted patient experiencing little to mid-level pain (OPI2-3) would attempt to display more yaw head movement in an attempt to conceal their perceived embarrassment (from the interviewer). However, when a lot of pain (OPI5) was experienced, the patient's focus turned away their perceived embarrassment to solely onto the sensation of pain experienced. Likewise, this perceived embarrassment could be deemed unnecessary when very little pain was experienced (OPI0). Further work needs to be conducted into this to see if other features can be used to gain better classification.

Table 5
Confusion matrix showing the pain-intensity classification percentage accuracy using the CAPP features (N.B. We coupled OPI ratings to maximize the number of training examples).

	OPI(0–1)	OPI(2–3)	OPI(4–5)
OPI(0–1)	75	21	4
OPI(2–3)	49	38	13
OPI(4–5)	13	40	47

6. Distribution details

The data were collected in the course of a research program devoted to understanding the properties of facial expressions of pain, the processes by which pain expression is perceived and the role of pain expression in clinical assessment of people suffering from pain conditions. Participants provided informed consent for use of their

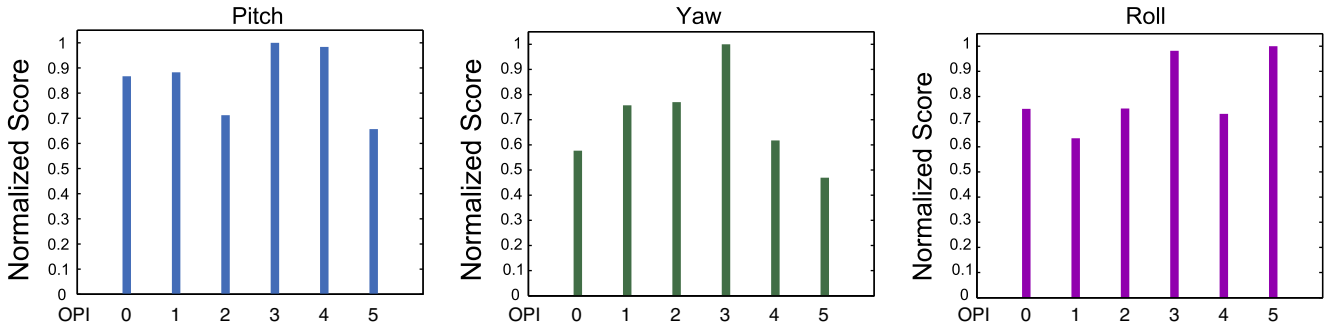

Fig. 7. The 3D head pose motion parameters (i.e. pitch, yaw and roll) when grouped according to OPI group (0–5).

video images for scientific study of the perception of pain including pain detection. Distribution of the database is governed by the terms of their informed consent. Investigators who for scientific purposes are interested in undertaking studies that can be clearly construed as having the potential to advance understanding of the perception of pain expression or contributing to the development of improved techniques for clinical assessment of pain conditions may make application for access to the database. Computer vision studies, which provide a means of modeling human decoding of pain expression, fall into the category of perception of pain expression. Applications should indicate how the proposed work addresses advancement of knowledge in the perception of pain expression or improved clinical assessment. Approved recipients of the data may not redistribute it and agree to the terms of confidentiality restrictions. Use of the database for commercial purposes is strictly prohibited.

If interested in obtaining the database, please sign and return an agreement form available from http://www.pitt.edu/ jeffcohn/PainArchive. Once the signed form has been received, you may expect to receive instructions within 5 business days.

7. Conclusions and future work

In this paper we have described the UNBC-McMaster Shoulder Pain Expression Archive which contains, 1) 200 video sequences containing spontaneous facial expressions; 2) 48,398 FACS coded frames, 3) pain frame-by-frame scores, sequence-level self-report and observer measures; and 4) 66-point AAM landmarks. We have released this data in an effort to address the lack of FACS coded spontaneous expressions available for researchers as well as promoting and facilitating research into the perception of pain. We have included baseline results from our AAM/SVM system for both individual AUs and pain at a frame-by-frame level. We also conducted sequence-level experiments by trying to replicate an expert human observer by classifying pain intensities. As facial expressions do not contain all of the information data, we also explored using the 3D head pose motion information as a predictor of pain. Even though severe and rapid head motion was true for some of the participants, this was not a common trait (i.e. people in no pain also moved their head) so further exploration of this is required.

Pain detection represents a key application in which facial expression recognition could be applied successfully, especially if applied in the context of an heavily constrained situation such as an critical care or intensive care unit where the number of expressions is greatly limited. This is in compared to the situation where a person is mobile and expresses a broad gamut of emotions, where the approach we have take here would be of little use as the painful facial actions are easily confused with other emotions (such as sadness, fear and surprise). For this to occur, a very large dataset which is captured in conditions that are indicative of the behavior to be expected in addition to being accurately coded needs to be collected. Another issue is the requirement of the detection in terms of timing accuracy. In our system presented here, we detect pain at every frame and also at the sequence-level. However, at what level does this need to be accurate at — milliseconds, seconds or minutes? Again this depends on the context in which this system will be used. We plan to look into this area in the future as we get more clinically relevant data.

Acknowledgments

This project was supported in part by CIHR Operating Grant MOP77799 and National Institute of Mental Health grant R01 MH51435. Special mention also goes to Nicole Ridgeway, Zara Ambadar, Nicole Grochowina, Amy Johnson, David Nordstokke, Racquel Kueffner, Shawn Zuratovic and Nathan Unger who provided technical assistance.

References

[1] A. Gawande, The Checklist Manifesto: How to Get Things Right, Metropolitan Books, 2010.
[2] A. Williams, H. Davies, Y. Chadury, Simple pain rating scales hide complex idiosyncratic meanings, Pain 85 (2000) 457–463.
[3] D. Wong, C. Baker, Pain in children: comparison of assessment scales, Pediatr. Nurs. 14 (1988) 9–17.
[4] D. Turk, R. Melzack, The Measurement of Pain and the Assessment of People Experiencing Pain, in: D. Turk, R. Melzack (Eds.), Handbook of Pain Assessment, 2nd ed., Guilford, New York, USA, 2001, pp. 1–11.
[5] K. Craig, K. Prkachin, R. Grunau, The facial expression of pain, Handbook of Pain Assessment, 2nd ed., Guildford, New York, USA, 2001, pp. 153–169.
[6] K. Prkachin, The consistency of facial expressions of pain: a comparison across modalities, Pain 51 (1992) 297–306.
[7] K. Prkachin, P. Solomon, The structure, reliability and validity of pain expression: evidence from patients with shoulder pain, Pain 139 (2008) 267–274.
[8] P. Ekman, W. Friesen, J. Hager, Facial Action Coding System: Research Nexus, Network Research Information, Salt Lake City, UT, USA, 2002.
[9] A. Ashraf, S. Lucey, J. Cohn, T. Chen, Z. Ambadar, K. Prkachin, P. Solomon, B.-J. Theobald, The painful face: pain expression recognition using active appearance models, Proceedings of the 9th international conference on Multimodal interfaces, ACM, Nagoya, Aichi, Japan, 2007, pp. 9–14.
[10] A. Ashraf, S. Lucey, J. Cohn, K.M. Prkachin, P. Solomon, The painful face II — pain expression recognition using active appearance models, Image Vis. Comput. 27 (2009) 1788–1796.
[11] P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, K. Prkachin, Automatically Detecting Pain Using Facial Actions, in: Proceedings of the International Conference on Affective Computing and Intelligent Interaction, pp. 1–8.
[12] P. Lucey, J. Cohn, I. Matthews, S. Lucey, J. Howlett, S. Sridharan, K. Prkachin, Automatically Detecting Pain in Video Through Facial Action Units, IEEE Trans. Syst. Man Cybern. B (2010).
[13] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, J. Movellan, Towards practical smile detection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 2106–2111.
[14] K. Prkachin, S. Mercer, Pain expression in patients with shoulder pathology: validity, coding properties and relation to sickness impact, Pain 39 (1989) 257–265.
[15] M. Heft, R. Gracely, R. Dubner, P. McGrath, A validation model for verbal descriptor scaling of human clinical pain, Pain 9 (1980) 363–373.
[16] A. Anastasi, Psychological Testing, Macmillan, NY, USA, 1982.
[17] J. Cohen, Statistical Power Analysis for the Social Sciences, Lawrence Erlbaum Associates, NJ, USA, 1988.
[18] J. Xiao, S. Baker, I. Matthews, T. Kanade, Real-Time Combined 2D+3D Active Appearance Models, in: Proceedings of the IEEE Conference on Computer Vision and, Pattern Recognition, pp. 535–542.
[19] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 681–685.

[20] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2004) 135–164.
[21] S. Lucey, A. Ashraf, J. Cohn, Investigating spontaneous facial action recognition through AAM representations of the face, in: K. Kurihara (Ed.), Face Recognition Book, Pro Literatur Verlag, 2007.
[22] C. Hsu, C.C. Chang, C.J. Lin, A practical guide to support vector classification, Technical Report, 2005.
[23] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, 2001 Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
[24] C. Cortes, M. Mohri, Confidence Intervals for the Area Under the ROC curve, Advances in Neural Information Processing Systems, 2004.