PAINFUL DATA: The UNBC-McMaster Shoulder Pain Expression Archive Database

Patrick Lucey, Jeffrey F. Cohn, Kenneth M. Prkachin, Patricia E. Solomon and Iain Matthews

Abstract— A major factor hindering the deployment of a fully functional automatic facial expression detection system is the lack of representative data. A solution to this is to narrow the context of the target application, so enough data is available to build robust models so high performance can be gained. Automatic pain detection from a patient's face represents one such application. To facilitate this work, researchers at McMaster University and University of Northern British Columbia captured video of participant'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 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. This database is called the UNBC-McMaster Shoulder Pain Expression Archive Database. To promote and facilitate research into pain and augment current datasets, we have publicly made available a portion of this database which includes: 1) 200 video sequences containing spontaneous facial expressions, 2) 48,398 FACS coded frames, 3) associated pain frame-by-frame scores and sequence-level self-report and observer measures, and 4) 66-point AAM landmarks. This paper documents this data distribution in addition to describing baseline results of our AAM/SVM system. This data will be available for distribution in March 2011.

I. INTRODUCTION

Automated facial expression detection has made great strides over the past decade [1], [2], [3]. From initial focus on posed images recorded from a frontal view, current efforts extend to posed data recorded from multiple views [4], [5], 3D imaging [6], and, increasingly, non-posed facial behavior in real-world settings in which non-frontal pose and head motion are common [7], [8]. With respect to the latter, lack of well-annotated, ecologically valid, representative data has been a significant limitation.

In many real-world applications, the goal is to recognize or infer intention or other psychological states rather than facial actions alone. For this purpose, both narrowing the number of facial actions of interest and paying attention to context may be critical to the success of an automated system. For example, if one were to apply an automated facial analysis system to detect hostile intent in airports, a very large number of facial actions could occur and the shear number of permutations that could explain a person's facial expression would be far too great to have any confidence in detecting intention. The range of facial action variability in this setting is considerable. Hundreds of facial action units in combinations could occur and their signal value highly diverse. A target person could be running late, afraid of flying, upset at leaving a loved one, or agitated over visa problems or missed connections, among other states. Facial expression analysis would have to be combined with other modalities, including speech, in an iterative process to have confidence in intention detection.

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). With these constraints, the application of an automatic facial expression detection system could be very successful. A recent example is that of automatic smile detection in digital cameras where Whitehill et al. [8] 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. 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).

A. Painful Motivation

An application that would be of great benefit is that of pain/no-pain detection [9], [10], [11], [12]. In Atul Gawande's recent book entitled "The Checklist Manifesto" [13], he notes that massive improvements in patient outcomes in intensive care unit (ICU) settings have been achieved through adhering to standardized hygiene and monitoring per a priori checklists. One of these is pain monitoring, in which a nurse checks on a patient every 4 hours or so to evaluate whether they are suffering from pain and to make any needed adjustments in pain medication that may be warranted. Pain monitoring especially beyond the ICU has been hard to implement due to competing demands on nursing staff. Automatic monitoring could be an ideal solution.

Fig. 1. Examples of some of the sequences from the UNBC-McMaster Pain Shoulder Archive: (a) the sequence-level ratings were Observer Rated Pain Intensity (OPI) = 5, Visual Analog Scale (VAS) = 9, Sensory Scale (SEN) = 11, Affective-Motivational Scale (AFF) = 10, the peak-frame (60) had AU codes of 6c + 9b +43 which was equal to a Prkachin and Solomon Pain Intensity (PSPI) rating of 6 for that frame; (b) the ratings were OPI = 4, VAS = 6,SEN = 10,AFF = 7, the peak-frame (322) had AU codes of 4a + 6d + 7d+ 12d + 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 + 9d + 10d + 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.

Outside of the ICU, most pain assessment is via self-report. Self-reported pain is convenient and requires no special skills or staffing, but has several limitations. Self-report is subjective, lacks specific timing information about whether pain is increasing, decreasing, or spiking, and cannot be used when patients are impaired. Breathing tubes interfere with speech, consciousness may be transient, and patients may yet to have achieved functional speech (e.g., infants). Over the past twenty years, significant efforts have been made in identifying such facial actions [14], [15], [16]. Recently, Prkachin and Solomon [16] validated a Facial Action Coding System (FACS) [17] based measure of pain that can be applied on a frame-by-frame basis. A caveat on this 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.

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.

This paper documents this database and describes baseline results and protocol based on our AAM/SVM system.

II. 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 others were from the community and included a wide variety of occupations. 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
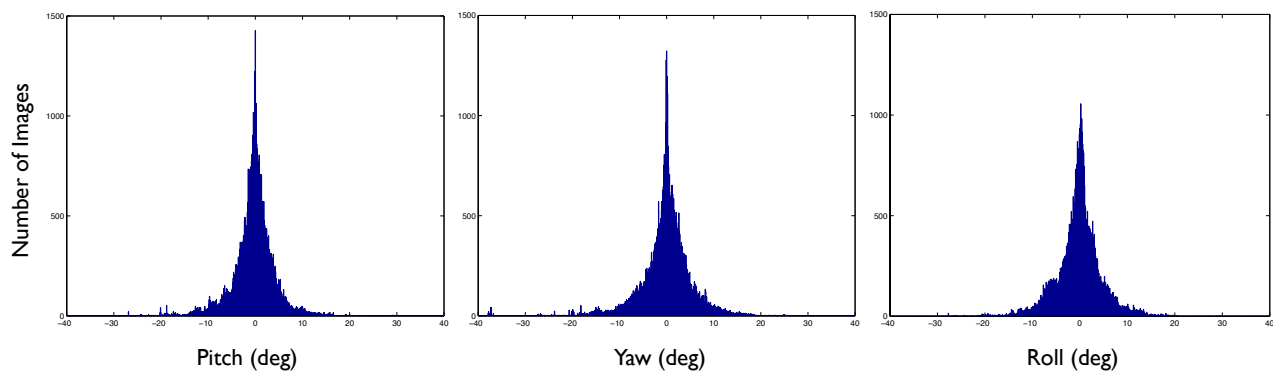
Fig. 2. Histograms of the pitch, yaw and roll taken from the 3D AAM parameters across the UNBC-McMaster Shoulder Pain Expression Archive database.

rotation of each arm separately [18]. Abduction movements involve lifting the arm forward and up in the sagittal plane. In internal rotation, the arm is bent 90 degrees at the elbow, abducted 90 degrees, 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, two Sony digital cameras recorded 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 degrees off frontal, with the face viewed from below.

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 [19]. 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 (SEN) started at "extremely weak" and finished at "extremely intense"; the affective-motivational scale (AFF) started at "bearable" and finished at "excruciating". In addition, participants completed a series of 10cm 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 [20]. Correlation between the observers rating on the OPI and subjects 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 [21] 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 Figure 1.

A. FACS coding

Each test was extracted from the video and coded using FACS [17]. 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 focussed 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 [17] was 95%, which compares favorably with other research in the FACS literature.

B. Prkachin and Solomon Pain Intensity Scale

Beginning in 1992, Prkachin [15] 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 [16] 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:

$$\text{Pain} = \text{AU4} + (\text{AU6 or AU7}) + (\text{AU9 or AU10}) + \text{AU43} \quad (1)$$

That is, the sum of AU4, AU6 or AU7 (whichever is higher in intensity), AU9 or AU10 (whichever is higher in intensity) and AU43 to yield a 16-point scale[1]. For the example in Figure 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 Figure 1(b), the peak frame has been coded as AU 4a + 6d + 7d+ 12d + 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 [16] 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 [16].

C. 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 48398 frames 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 I. The number of frames and the associated PSPI score is given in Table II. 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.

Examples of this data are given in Figure 1 and it is apparent that there is some head movement that occurs during these sequences. To quantify how much head movement

[1]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.

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 II

The inventory on the UNBC-McMaster Shoulder Pain Archive according to the Prkachin-Solomon Pain Intensity (PSPI) pain metric, where the frequency of each pain intensity is given.

PSPI Score	Frequency
0	40029
1-2	5260
3-4	2214
5-6	512
7-8	132
9-10	99
11-12	124
13-14	23
15-16	5

occurred, we used the 3D parameters from the AAM to estimate the pitch, yaw and roll [22]. The histograms of these parameters are shown in Figure 2. In terms of pitch, yaw and roll the mean was -0.38, -0.21 and -0.23 degrees 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$\geq$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 III 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

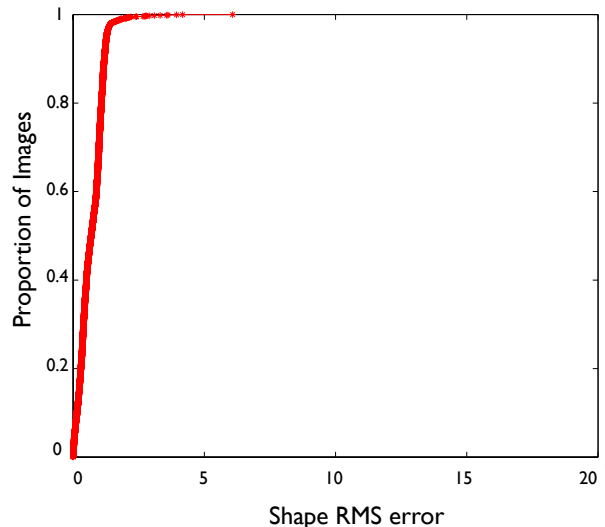| VAS	Frequency	OPI	Frequency
0	35	0	92
1	42	1	25
2	24	2	26
3	20	3	34
4	21	4	16
5	11	5	7
6	11	Total	200
7	6		
8	18		
9	10		
10	2		
Total	**200**		

Fig. 3. Fitting curve for the AAM compared against the manual 68 point landmark points. The Shape RMS error refers to the total shape RMS error in terms of pixels after all meshes were similarity normalized with an inter-occular distance of 50 pixels.

ratings so as we can have our automatic system to mimic that of a human observer. The affective and sensory self-report measures across the 200 sequences will also be available in the distribution.

III. 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.

A. 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 [23].

The shape $\mathbf{s}$ of an AAM [23] 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 $\mathbf{p}^T = [\mathbf{p}_s^T, \mathbf{p}_o^T]$. 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 determing object shape (e.g., mouth opening, eyes shutting, etc.). Procrustes alignment [23] is employed to estimate the base shape $\mathbf{s}_0$.

Keyframes within each video sequence were manually labelled, while the remaining frames were automatically aligned using a gradient descent AAM fitting algorithm described in [24]. Figure 4 shows the AAM in action, with the 68 point mesh being fitted to the patient's face in every frame. From the 2D shape model we can derive the 3D parameters using non-rigid structure from motion. See [22] for full details.

B. 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 an inter-occular distance of 50 pixels and aligned to the centre of the eye coordinates. We then compared 2584 manually landmarked images against their AAM counterpart. The fitting curve for the AAM is shown in Figure 3. As can be seen in this curve, nearly all of the AAM landmarks are within 2 pixels RMS error of the manual landmarks, which is negligible when one considers that this is 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.

IV. EXPERIMENTS

In this section, we describe two experiments that we conducted for i) AU and ii) pain detection at a frame-level. We first describe our baseline AAM/SVM system.
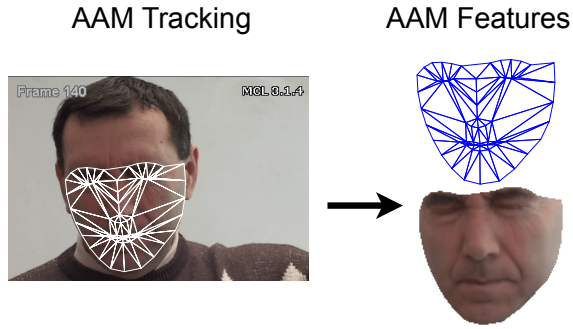
AAM Tracking   **AAM Features**

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.

A. 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], [25], [12], we extracted the following features:

- **SPTS:** The *similarity normalized shape or points*, $\mathbf{s}_n$, 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 Equation 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 [26]. LIBSVM was used for the training and testing of SVMs [27].

In all experiments conducted, a leave-one-subject-out strategy 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

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

there are many more frames with no behavior of interest than frames of interest, the overall agreement between correctly classified frames can skew the results somewhat. As such we used the receiver-operator 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)[2]. 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 [28], [8].

B. 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 IV. 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].

C. Pain Detection at Frame-level

The results for automatically detecting pain are given in Figure 5, which shows a clearer view of the trend we observed in the AU detection results. For the individual

[2]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
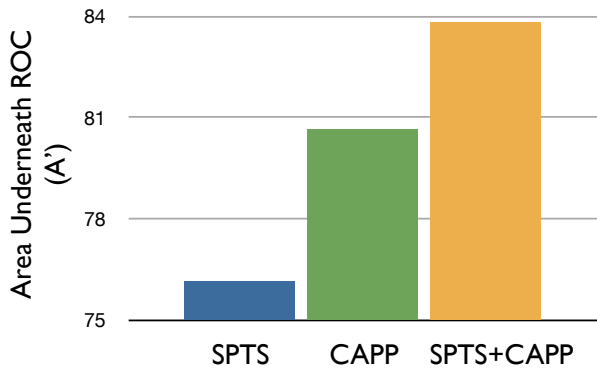
Fig. 5. The performance of the various features for the task of pain detection at the frame-level (yellow = SPTS, green = CAPP). The upper-bound error for all feature sets varied from approximately ±0.67 to 0.80.

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%.

V. DISTRIBUTION DETAILS

The data was 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 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.

This data will be available from March 2011. 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.

VI. 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 also included baseline results from our AAM/SVM system.

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 ICU ward 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 taken 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. However, at what level does this need to be accurate at - milliseconds, seconds or minutes? Again this is depends on the context in which this system will be used. A more likely scenario would be to detect pain as an event or at a sequence level (i.e. if a person was in pain over a window of 1 to 2 mins). We plan to look into this area in the future.

VII. ACKNOWLEDGMENTS

REFERENCES

[1] Y. Tian, J. Cohn, and T. Kanade, "Facial expression analysis," in *The handbook of emotion elicitation and assessment*, S. Li and A. Jain, Eds. New York, NY, USA: Springer, pp. 247–276.

[2] Y. Tong, W. Liao, and Q. Ji, "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1683–1699, 2007.

[3] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A Survey of Affect Recognition Methods: Audio, Visual and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[4] R. Gross, I. Matthews, S. Baker, and T. Kanade, "The CMU Multiple Pose, Illumination, and Expression (MultiPIE)," Robotics Institute, Carnegie Mellon University, Tech. Rep., 2007.

[5] M. Valstar and M. Pantic, "Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database," in *Proceedings of the 3rd International Workshop on EMOTION: Corpora for Research on Emotion and Affect*, 2010.

[6] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A High-resolution 3D Dynamic Facial Expression Database ," in *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 2008.

[7] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic Recognition of Facial Actions in Spontaneous Expressions," *Journal of Multimedia*, 2006.

[8] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, "Towards Practical Smile Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2106–2111, 2009.

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

[10] A. Ashraf, S. Lucey, J. Cohn, K. M. Prkachin, and P. Solomon, "The Painful Face II– Pain Expression Recognition using Active Appearance Models," *Image and Vision Computing*, vol. 27, no. 12, pp. 1788–1796, 2009.

[11] P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, and K. Prkachin, "Automatically Detecting Pain Using Facial Actions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2009, pp. 1–8.

[12] P. Lucey, J. Cohn, I. Matthews, S. Lucey, J. Howlett, S. Sridharan, and K. Prkachin, "Automatically Detecting Pain in Video Through Facial Action Units," *IEEE Transactions on Systems Man and Cybernetics, Part B*, 2010.

[13] A. Gawande, *The Checklist Manifesto: How to Get Things Right*. Metropolitan Books, 2010.

[14] K. Craig, K. Prkachin, and R. Grunau, "The facial expression of pain," in *Handbook of pain assessment*.

[15] K. Prkachin, "The consistency of facial expressions of pain: a comparison across modalities," *Pain*, vol. 51, pp. 297–306, 1992.

[16] K. Prkachin and P. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, pp. 267–274, 2008.

[17] P. Ekman, W. Friesen, and J. Hager, *Facial Action Coding System: Research Nexus*. Salt Lake City, UT, USA: Network Research Information, 2002.

[18] K. Prkachin and S. Mercer, "Pain expression in patients with shoulder pathology: validity, coding properties and relation to sickness impact," *Pain*, vol. 39, pp. 257–265, 1989.

[19] M. Heft, R. Gracely, R. Dubner, and P. McGrath, "A validation model for verbal descriptor scaling of human clinical pain," *Pain*, vol. 9, pp. 363–373, 1980.

[20] A. Anastasi, *Psychological Testing*. Macmillan, NY, USA, 1982.

[21] J. Cohen, *Statistical Power Analysis for the Social Sciences*. Lawrence Erlbaum Associates, NJ, USA, 1988.

[22] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-Time Combined 2D+3D Active Appearance Models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004, pp. 535–542.

[23] T. Cootes, G. Edwards, and C. Taylor, "Active Appearance Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[24] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[25] S. Lucey, A. Ashraf, and J. Cohn, "Investigating spontaneous facial action recognition through aam representations of the face," in *Face Recognition Book*, K. Kurihara, Ed. Pro Literatur Verlag, 2007.

[26] C. Hsu, C. C. Chang, and C. J. Lin, "A practical guide to support vector classification," Tech. Rep., 2005.

[27] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[28] C. Cortes and M. Mohri, "Confidence Intervals for the Area Under the ROC curve," *Advances in Neural Information Processing Systems*, 2004.