Automatically Detecting Pain in Video Through Facial Action Units

Patrick Lucey, *Member, IEEE,* Jeffrey F. Cohn, *Associate Member, IEEE,* Iain Matthews, *Member, IEEE,* Simon Lucey, *Member, IEEE,* Sridha Sridharan, *Senior Member, IEEE,* Jessica Howlett, *Student Member, IEEE,* and Kenneth M. Prkachin

***Abstract*—In a clinical setting, pain is reported through either self-report or via an observer. Such measures are: 1) subjective, and 2) give no specific timing information. However, coding pain as a series of facial action units (AUs) can avoid these issues as it can be used to gain an objective measure of pain on a frame-by-frame basis. Using video data from patients with shoulder injuries, in this paper we describe an Active Appearance Model (AAM) based system that can automatically detect the frames in video in which a patient is in pain. This pain dataset highlights the many challenges associated with spontaneous emotion detection, especially that of expression and head movement due to the patient's reaction to pain. In this paper, we show that the AAM can deal with these movements and can achieve significant improvements in both AU and pain detection performance compared to the current-state-of-the-art approaches which utilize similarity-normalized appearance features only.**

***Index Terms*—Emotion, Facial Action Units (AUs), Facial Action Coding System (FACS), Active Appearance Models (AAMs), Support Vector Machines (SVMs), Pain.**

I. Introduction

Reliably assessing and managing pain in a clinical setting is difficult. Patient self-report has become the most widely used technique to measure pain because it is convenient, does not require advanced technology or special skills. It is typically evaluated either through clinical interview or by using a visual analog scale (VAS). With the VAS, the intensity of pain is indicated by marking a line on a horizontal scale, anchored at each end with words such as "no pain" and "the worst pain imaginable".

While useful, self-report measures have significant limitations [1], [2]. These include inconsistent metric properties across scale dimensions, reactivity to suggestion, efforts at impression management or deception, and differences between clinician's and sufferers' conceptualization of pain [3]. Moreover, self-report cannot be used in important populations, such as young children, patients who have limited abilities to communicate, the mentally impaired, and patients who require assisted breathing. In these situations, an observer rating is required where the observer chooses a face on the "faces of pain" scale which best resembles the facial expression of the patient [4]. This is highly impractical and inefficient if the observer is required for long periods of time which could be the case for a patient in an intensive care unit (ICU).

(a) (b)

Fig. 1. In this paper, we develop a system which can detect the frames in a video sequence in which the patient is in either a state of (a) "pain" or (b) "no-pain". When a patient is in pain, it often coincides with facial expression change as well as head motion which can be seen above.

In addition to self-report and observer measures being highly subjective, these measures do not give a continuous output over time, as the only output measured coincides when the patient is at their emotional apex (e.g. highest pain intensity). They do not provide information on the patient's emotional state other than these peak periods. In an effort to address these shortcomings, 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 [5]. These approaches have been fraught with difficulty because they are often inconsistent with other evidence of pain [5], in addition to being highly invasive and constraining to the patient.

Another potential solution is to code pain using facial actions, which is analogous to the "faces of pain" approach. Over the past two decades, significant efforts have been made in identifying such facial actions [6], [7], [8]. Recently, Prkachin and Solomon [8] developed a Facial Action Coding System (FACS) [9] based measure of pain which can be gained

Manuscript was received 28 August 2009; revised 10 April 2010; accepted 1 September 2010.

P. Lucey, and J.F. Cohn are with the Department of Psychology, University of Pittsburgh/Robotics Institute, Carnegie Mellon University, Pittsburgh, PA. Email: {plucey@pitt.edu, jeffcohn@cs.cmu.edu}

I. Matthews is with Disney Research Pittsburgh and is adjunct at the Robotics Institute Carnegie Mellon University, Pittsburgh, PA; S. Lucey is with the Commonwealth Science and Industrial Research Organization (CSIRO), Australia; J. Howlett and S. Sridharan are with the SAIVT lab at the Queensland University of Technology in Brisbane, Australia; K.M. Prkachin is with the Department of Psychology at the University of Northern British Columbia. Email: {iainm@disneyresearch.com, simon.lucey@csiro.au, j.howlett@qut.edu.au, s.sridharan@qut.edu.au, kmprk@unbc.ca}

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

at each time step (i.e. each video frame), which is the only such available measure. 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.

In this paper, we describe an Active Appearance Model (AAM) based computer vision system which can automatically detect pain based on facial expressions coded using FACS. We demonstrate its use on the UNBC-McMaster Shoulder Pain Archive which contains patients with rotator-cuff injuries, eliciting spontaneous facial expressions associated with pain which are not posed or feigned. These facial actions vary in duration and intensity and often coincides with abrupt changes in head position as shown in Figure 1. Using an AAM approach, we show that both shape (i.e. contour) and appearance (i.e. texture) are both vital for gaining accurate detection performance. We also highlight the difficulties associated with detecting spontaneous data such as pain, where there is a lot of head motion. This is a particular problem for systems which use similarity-normalized appearance features (i.e. normalized for translation, rotation and scale), as some parts of the face may not be visible, inhibiting accurate detection.

The key contributions of this paper are:

1) We describe a system which can automatically detect pain from a patient's face using an AAM approach on a frame-by-frame basis (Section V).
2) We show that using the common similarity-normalized appearance features on spontaneous data is problematic due to the major facial expressions and head motion and using an AAM approach can yield significant improvement (Section IV & V).
3) We show that fusing all AAM representations together (i.e. similarity normalized shape, appearance and canonical normalized appearance (synthesized)) using linear logistical regression (LLR), improves both AU and pain detection performance (Section IV & V).

A. Related Work

There have been many recent attempts to detect emotions directly from the face, mostly using FACS [9]. Comprehensive reviews can be found in [10], [11], [12]. These attempts relate mostly to posed data as spontaneous (i.e. real) emotions are subtle and do not occur frequently, which makes this pursuit timely and costly. Pain however, is one spontaneous emotion that can be captured on cue as it can be elicited. This can be achieved ethically through physical movement of a limb or joint which is painful, or through a device such as a cold pressor.

Collecting data via a cold pressor, Littlewort et al. [13] used their AU detector which consisted of Gabor Filters, Adaboost and SVMs, to differentiate between genuine and fake pain. In this work no actual pain/no-pain detection was performed as differentiation between genuine, fake and baseline sequences was done via analyzing the various detected AUs. The classification for this work was done at the sequence level and a cold pressor was used to elicit real pain on the subjects. All images were then similarity-normalized by first coarsely locating the face using a Viola-Jones type of approach, then locating the eyes which were used to scale, rotate and crop the image according to a predefined inter-ocular distance.

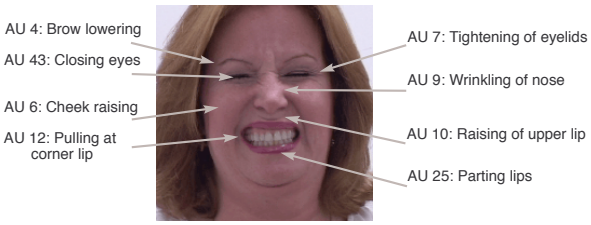


Fig. 2. An example of facial actions associated when a person in is pain. In this example, AU's 4, 6, 7, 9, 10, 12, 25 and 43.

In terms of pain/no-pain detection, Ashraf et al. [14] used the UNBC-McMaster Shoulder Pain Archive which contains data with patient's moving both their injured and uninjured shoulders, to classify video sequences as pain/no-pain. Ashraf et al. [15] then extended this work to the frame-level to see how much benefit would be gained at labeling at the frame-level over the sequence-level. Even though they found that it was advantageous to have the pain data labeled at the frame-level, they proposed that this benefit would be largely diminished when encountering large amounts of training data. In these works, all images were registered using an AAM.

Other than pain, there have been a few other relevant works published on detecting spontaneous facial expressions and emotions. The first one is based on the RUFACS dataset [16], which consists of 34 subjects participating in a interview ($\approx$ 2mins) where they are being asked to take a position on a particular issue (either truthfully or not). This dataset contains a lot of head motion and subtle facial actions, which is indicative of natural human behavior. Due to these challenges, Bartlett et al. [16] found that the performance of their AU detection system greatly diminished compared to the posed scenario[1]. All images in this work were similarity-normalized as per the Littlewort et al. [13] system described above.

More recently, Whitehill et al. [18] published their work on robust smile detection across all environments, motivated for the use in digital cameras. For this work, they collected the GENKI dataset, which contains over 63,000 static images from the internet, which were all frontal. Again, all images were similarity normalized as previously described, however, this work had a greatly improved eye detector which improved the registration of the images. Even with this improved image registration and little head motion, the authors blamed the loss in alignment accuracy decreased smile detection performance up to 5%.

Even though these above works all acknowledge the importance of registration of input images for spontaneous facial expression and emotion detection, none have quantified to the extent in which this effects overall performance. In this paper, we do such an analysis for both AU and pain detection, which

[1]In terms of area under the ROC curve (A'), the mean AU detection rate for RUFACS dataset was 71.0 compared to 92.6 for the posed data of the Cohn-Kanade database [17]

Fig. 3. Examples from the UNBC-McMaster database, showing the instances of pain and also of head pose variation during the sequence.

will be very important if spontaneous expression detection systems such as pain detectors are used in commercial applications in the future.

II. Facial Expressions of Pain

A. Defining Pain via Facial Action Units

Much is known about how humans facially express pain from studies in behavioral science [6], [7], [8]. Most of these studies encode pain from the movement of facial muscles into a series of AUs, based on FACS. An example of the facial actions of a person in pain is shown in Figure 2.

In 1992, Prkachin [7] conducted a study on facial expressions and 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 [8] 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 scale is defined as:

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

That is, the sum of AU4, AU6 or AU7 (whichever is higher), AU9 or AU10 (whichever is higher) and AU43 to yield a 16-point scale[2]. Frames that have an intensity of 1 and higher are defined as pain. For the example in Figure 2, which has been coded as AU4B + AU6E + AU7E + AU9E + AU10D + AU12D+ AU25E + AU43A, the resulting pain intensity would be $2+5+5+1 = 13$. This is because AU4 has an intensity of 2, AU6 and AU7 are both of intensity 5 so just the maximum is taken, AU9 is of intensity 5 and AU10 is of intensity 4 so again the maximum is taken which is 5, and AU43 is of intensity 1 (eyes are shut).

The Prkachin and Solomon [8] FACS pain scale is currently the only metric which can define pain on a frame-by-frame basis. All frames that were used in this study were coded via this metric.

TABLE I
Mean and variance of the pitch, yaw and roll parameters of the pain data relative to the pain metric. N is the number of frames analyzed.

	N	Pitch		Yaw		Roll	
		μ	σ^2	μ	σ^2	μ	σ^2
Pain $= 0$	40461	-0.25	22.69	-0.29	37.03	-0.01	29.16
Pain > 0	7937	-0.93	26.72	0.12	55.61	-1.12	48.52

B. UNBC-McMaster Shoulder Pain Expression Archive Database

The UNBC-McMaster Shoulder Pain Expression Archive database was used for this work. It contains video of the faces of adult subjects (129 subjects - 63 male, 66 female) with rotator cuff and other shoulder injuries. Subjects were recorded during movement of their affected and unaffected shoulder during active and passive conditions. In the active condition, subjects initiated shoulder rotation on their own. In the passive condition, a physiotherapist was responsible for the movement. In the experiments conducted in this paper, only the active condition was used. Within the active condition, tests were performed on both the affected and the unaffected shoulder to provide within subject control. The camera angle for these tests were approximately frontal. However, moderate head motion was common. Video of each trial was rated offline by a FACS certified 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 [9] was 95%, which compares favorably with other research in the FACS literature. For more information on the database, please refer to [8].

From the database, we used 203 sequences from 25 different subjects. Overall, there were 48,398 frames of data analyzed and all of these frames were used in our experiments. Out of this data, according to the pain metric given in the previous subsection, 83.6% of the frames had a pain score of 0, and 16.4% had frames in which had a person in pain (pain score ≥ 1). Examples of this data are given in Figure 3. Clearly, considerable head movement occurs during the sequence. To

[2]Action units 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.

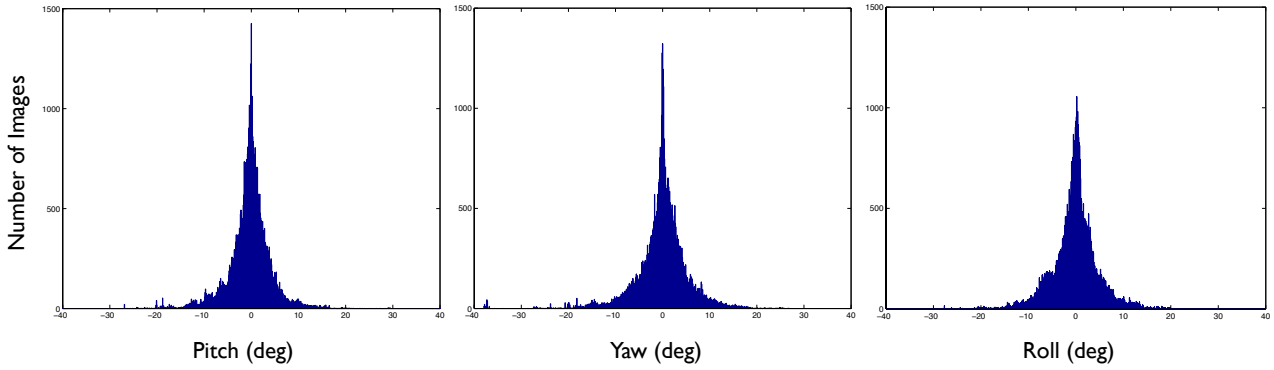


Fig. 4. Histograms of the pitch, yaw and roll taken from the 3D AAM parameters across the UNBC-McMaster Shoulder Pain Expression Archive database.

TABLE II
Proportion of frames in which the patient was less than 5, 10, 15 and 20 degrees, as well as greater than 20 degrees from frontal in terms of pitch, yaw and roll (absolute degrees). $N = 40461$ for pain score =0 and $N = 7937$ for pain score ≥ 1).

	Pitch		**Yaw**		**Roll**	
	Pain	No-Pain	Pain	No-Pain	Pain	No-Pain
<5 deg	81.0%	81.3%	82.2%	75.4%	78.7%	62.6%
<10 deg	95.7%	95.8%	93.6%	93.8%	95.2%	88.5%
<15 deg	100.0%	99.4%	99.1%	98.5%	99.1%	96.9%
<20 deg	100.0%	99.9%	99.7%	99.8%	99.8%	99.2%
>20 deg	0.0%	0.1%	0.3%	0.2%	0.2%	0.8%

quantify how much head movement occurred, we used the 3D parameters from the AAM (see Section III-C for details) to estimate the pitch, yaw and roll. The histograms of these parameters are shown in Figure 4. As you can see from this figure, there is quite a bit of variance in terms of the pitch, yaw and roll. Upon inspection of the data, it appeared that a lot of head movement occurred when a patient was in pain. To gauge this relative to the pain score, we have generated Table I to display the variation in head position as a function of pain. From this it appears when a patient was in pain (pain score ≥ 1), the variance of head position for pitch, yaw and roll was much greater than when a person was not in pain. In terms of how much variation there was, we have produced Table II which shows the proportion of frames that differed from the fully frontal view. As can be seen from this table, close to 90% were within 10 degrees of being fully frontal and over 99% were within 20 degrees from the fully frontal view.

III. Automatic Detection System

In our system, we employ an Active Appearance Model (AAM) based system which uses AAMs to track the face and extract visual features. We then use support vector machines (SVMs) to classify individual AUs and pain. An overview of our system is given in Figure 5. We describe each of these modules in the following subsections.

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 [19].

The shape $\mathbf{s}$ 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 $\mathbf{p}^T = [\mathbf{p}_s^T, \mathbf{p}_o^T]$. 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 [19] 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 [20]. Figure 6 shows the AAM in action, with the 68 point mesh being fitted to the patient's face in every frame.

B. Feature Extraction

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 [14], [21], we extracted the following features:

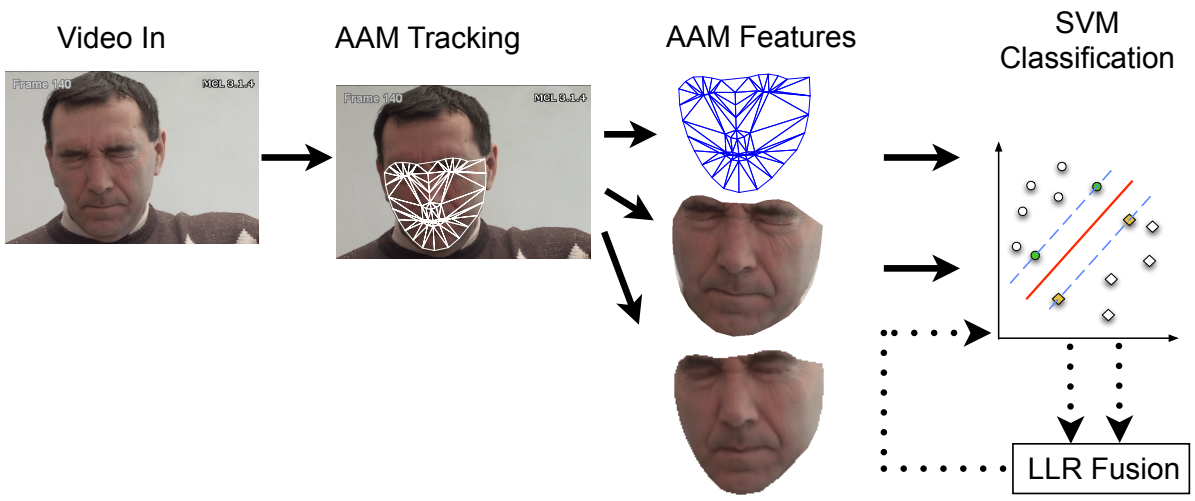


Fig. 5. Block diagram of our automatic system. The face is tracked using an AAM and from this we get both shape and appearance features. Both these features are used for classifying individual AUs using a linear SVM. The SVMs output for the AUs can be fused together using linear logistical regression (LLR). LLR calibrates the score into a log-likeihood score so that the scores are normalized into the same domain so that they can be combined easily. This calibration is a supervised process.

- **SPTS:** The similarity normalized shape, $\mathbf{s}_n$, refers to the 68 vertex points in s_n for both the x- and y- coordinates, resulting in a raw 136 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 $\mathbf{p}$. An example of the similarity normalized shape features, SPTS, is given in Figure 6(3rd row).
- **SAPP:** The similarity normalized shape, $\mathbf{a}_n$, refers to the where all the rigid geometric variation (translation, rotation and scale) has been removed. It achieves this by using $\mathbf{s}_n$ calculated above and warps the pixels in the source image with respect to the required translation, rotation and scale. An example of the similarity normalized shape features, SAPP, is given in Figure 6(4th row). This is the type of approach is employed by most researchers [16], [18], as only coarse registration is required (i.e. just face and eye locations). From viewing the examples, it can be seen that when head movement is experienced some of the face is partially occluded which can affect performance, also some non-facial information (such as the background) is included due to occlusion.
- **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. An example of these features, CAPP, is given in Figure 6(Bottom row).

C. 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 Equation 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 Equation 3 into 3D linear shape models. In this work we stack the 2D AAM shape vectors in all N images into a measurement matrix:

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

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

Fig. 6. Example of the output of the AAM tracking and the associated shape and appearance features: (Top row) the original sequence, (Second row) the AAM tracked sequence, (Third row) the similarity normalized shape features (SPTS), (Fourth row) the similarity normalized appearance features (SAPP),(Bottom row) the canonical normalized appearance features (CAPP).

vertices n). 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 couild 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:

$$\begin{aligned} \mathbf{M} &= \tilde{\mathbf{M}}\mathbf{G}, \text{and} \\ \mathbf{B} &= \mathbf{G}\tilde{\mathbf{B}} \end{aligned} \tag{6}$$

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 [22] for full details.

D. Support Vector Machine Classification

Support vector machines (SVMs) have been proven useful in a number of pattern recognition tasks including face and facial action recognition. SVMs attempt to find the hyperplane 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,

$$\mathbf{w}^T\mathbf{x}^* \quad \begin{matrix} >^{true} \\ <_{false} \end{matrix} \quad b \tag{7}$$

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 to the training data. Typically, $\mathbf{w}$ is not defined explicitly, but through a linear sum of support vectors. A linear kernel was used in our experiments due to its ability to generalize well to unseen data in many pattern recognition tasks [23]. LIBSVM was used for the training and testing of SVMs [24].

E. Fusion of Scores Using Linear Logistical Regression (LLR)

In classification, a decision is based on a score from a classifier such as a SVM. In the case of the SVM the score relates to the distance from the decision hyperplane, which works well for a single decision. However, these scores have no real meaning when comparing them from different SVMs. As such, comparing or combining these scores does not make sense and can lead to erroneous results. Calibrating the scores into a common domain is required so that comparisons and fusion can take place. Logistical linear regression is one method of doing this [25].

Given we have N AU detectors with output scores $(s_1, s_2, \ldots, s_N)$, LLR calibrates all the individual scores

through learning the weights $(a_0, a_1, \ldots, a_N)$ via logistic regression so that $F_N = a_0 + a_1 s_1 + a_2 s_2 + \ldots + a_N s_N$, where the constant a_0 improves the calibration through regularization.

To train the weights, a set of supervised training scores and an objective function needs to be set. In [25], they used a logistic regression objective that is normalized with respect to the proportion of positive examples to negative examples $(K : L)$, which are weighted to the synthetic prior $P = 0.5$ [3]. The objective is stated in terms of a cost, which must be minimized:

$$C_{wlr} = \frac{P}{K}\sum_{j=1}^{K}\log(1+e^{-f_j-\mathrm{logit}P}) + \frac{1-P}{L}\sum_{j=1}^{L}\log\left(1+e^{g_j+\mathrm{logit}P}\right) \quad (8)$$

where the fused target and non-target scores are respectively:

$$f_j = a_0 + \sum_{i=1}^{N} a_i s_{ij}, g_j = a_0 + \sum_{i=1}^{N} a_i r_{ij} \quad (9)$$

and where

$$\mathrm{logit}P = \log\frac{P}{1-P} \quad (10)$$

and s_{ij} is an N by K matrix of scores that each of the N component systems calculated for each of the K target trails, and r_{ij} is an N by L matrix of scores that each of the N component systems calculated for each of the L non-target trials.

The fused score f is then used for detection. The FoCal package was used for calibrating and fusing the various AU SVM scores together using LLR [25].

F. Performance Measurement

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 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)[4]. 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 [26], [18].

[3]The value of P has a small effect and 0.5 is a reasonable choice for the task of AU and pain detection.

TABLE III
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	N	SPTS	SAPP	CAPP
4	1074	$\mathbf{67.8} \pm 1.4$	$\mathbf{45.9} \pm 1.5$	$\mathbf{47.7} \pm 1.5$
6	5612	$\mathbf{78.9} \pm 0.6$	$\mathbf{79.4} \pm 0.5$	$\mathbf{83.8} \pm 0.5$
7	3366	$\mathbf{66.3} \pm 0.8$	$\mathbf{66.1} \pm 0.8$	$\mathbf{68.0} \pm 0.8$
9	423	$\mathbf{53.4} \pm 2.4$	$\mathbf{76.5} \pm 2.1$	$\mathbf{87.3} \pm 1.6$
10	525	$\mathbf{80.4} \pm 1.7$	$\mathbf{85.7} \pm 1.5$	$\mathbf{73.0} \pm 1.9$
12	6956	$\mathbf{78.5} \pm 0.5$	$\mathbf{79.1} \pm 0.5$	$\mathbf{82.8} \pm 0.5$
20	706	$\mathbf{69.3} \pm 1.7$	$\mathbf{56.4} \pm 1.9$	$\mathbf{58.0} \pm 1.9$
25	2433	$\mathbf{74.7} \pm 0.9$	$\mathbf{63.2} \pm 1.0$	$\mathbf{65.6} \pm 1.0$
26	2199	$\mathbf{52.7} \pm 1.1$	$\mathbf{55.8} \pm 1.1$	$\mathbf{55.4} \pm 1.1$
43	2454	$\mathbf{89.9} \pm 0.6$	$\mathbf{78.8} \pm 0.8$	$\mathbf{88.3} \pm 0.7$
AVG		$\mathbf{74.4} \pm 0.8$	$\mathbf{72.0} \pm 0.8$	$\mathbf{75.3} \pm 0.8$

IV. Spontaneous Action Unit Detection

A. AU Detection Results

We conducted detection for ten AUs (4, 6, 7, 9, 10, 12, 20, 25, 26 and 43)[5]. The results for the AU detection with respect to the similarity-normalized shape (SPTS) and appearance (SAPP) and the canonical appearance (CAPP) features are shown in Table III. In terms of the overall average accuracy of the AU detection (bottom line of the table), the SAPP (72.0) features performed worse than the SPTS (74.4) and the CAPP (75.3) features. The differences may not be large, but they are significantly significant ($p < 0.05$). This result is quite interesting because in the majority of works conducted in the field (see Section I.B) have used these features for AU and emotion detection. However, it is not surprising as the pain data used in these experiments contains quite a bit of head motion which corresponds to poor image registration as can be seen in Figure 6 (fourth row). Conversely, it is not surprising that the CAPP features achieved the best performance as they couple together both the shape and appearance representations. This synthesized view captures both the geometric and shape features of the face so that no non-face information is incorporated in the representation. It must be noted though that for the majority of the time the patient's were relatively frontal $(\pm 20^o)$, so that is why the results for the SAPP were as close as they were.

In terms of individual AU detection, it can be seen depending on the AU, the best performing feature set varies. When comparing SPTS and SAPP, the SPTS features yielded the higher detection rates for AUs 4, 20, 25 and 43. Conversely, for

[4]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

[5]These AUs had more than 20 frames coded, all other AUs with less than this were omitted due to lack of sufficient training data).

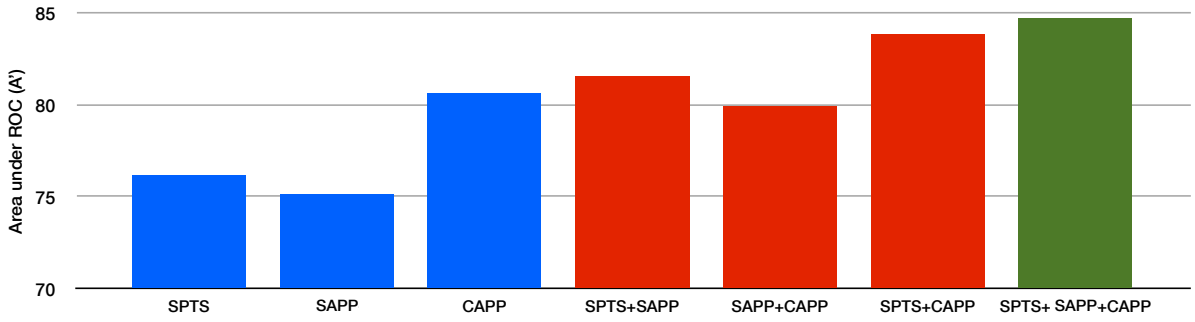


Fig. 7. The performance of the various features for the task of pain detection (blue = single features, red = fused 2 features, green = fuse all 3 features). The upper-bound error for all feature sets varied from approximately ± 0.67 to 0.80.

TABLE IV
Results showing the area underneath the ROC curve for the combination of the similarity-normalized shape (SPTS) and appearance (SAPP) and canonical appearance (CAPP) features using LLR fusion. Note the average is a weighted one, depending on the number of positive examples.

AU	SPTS+SAPP	SAPP+CAPP	SPTS+CAPP	ALL
4	**50.7** $\pm$ 1.5	**47.9** $\pm$ 1.5	**48.5** $\pm$ 1.5	**53.7** $\pm$ 1.5
6	**85.5** $\pm$ 0.5	**82.9** $\pm$ 0.5	**85.9** $\pm$ 0.5	**86.2** $\pm$ 0.5
7	**67.8** $\pm$ 0.8	**69.1** $\pm$ 0.8	**68.5** $\pm$ 0.8	**70.0** $\pm$ 0.8
9	**80.0** $\pm$ 2.0	**70.1** $\pm$ 2.2	**71.3** $\pm$ 2.2	**79.8** $\pm$ 2.0
10	**63.3** $\pm$ 2.1	**75.5** $\pm$ 1.9	**78.1** $\pm$ 1.8	**75.4** $\pm$ 1.9
12	**83.6** $\pm$ 0.5	**82.4** $\pm$ 0.5	**83.8** $\pm$ 0.4	**85.6** $\pm$ 0.4
20	**54.6** $\pm$ 1.9	**67.1** $\pm$ 1.8	**67.8** $\pm$ 1.8	**66.8** $\pm$ 1.8
25	**56.0** $\pm$ 1.0	**73.2** $\pm$ 0.9	**64.0** $\pm$ 1.0	**73.3** $\pm$ 0.9
26	**53.5** $\pm$ 1.1	**52.7** $\pm$ 1.1	**52.2** $\pm$ 1.1	**52.3** $\pm$ 1.1
43	**85.5** $\pm$ 0.7	**88.0** $\pm$ 0.7	**91.9** $\pm$ 0.6	**90.9** $\pm$ 0.6
AVG	**74.3** $\pm$ 0.8	**75.7** $\pm$ 0.8	**76.2** $\pm$ 0.7	**78.0** $\pm$ 0.7

AUs 9 and 10 where the SAPP features obtained significantly better performance. The other AUs, 6, 7 and 12 achieved comparable rates. Other than the poor registration of the SAPP features, another explanation of these results can stem from the AAM 2-D mesh. For AU4 (brow lowering), 20 (lip stretcher), 25 (lips part) and 43 (eye closing), the areas of the face in which movement pertaining to these AUs occurs lie on the 2-D mesh. So it is intuitive that the most discriminating features for these actions would relate to the shape features. For AUs 9 (nose wrinkler) and 10 (lip raiser), these correspond with a lot of textural change in terms of wrinkles and not so much in terms of contour movement, which would suggest why the SAPP features performed better than the SPTS for these even with the poor registration. Though again we see the benefit of the canonical view where the textural features are synthesized back to the base mesh where most AU obtained an improvement in performance (although there seems some degradation with AU10, which suggests that the AAM misses important information around the upper lip when transforming the appearance back to the base mesh, also with AU4 where the mask used sometimes cuts off the top of the eyebrows).

These results are backed up by the experience of human FACS coders, where the relative importance of shape and appearance varies with type of AU. Specific examples are that of brow lowering (AU 4), where FACS coders look for strong changes in shape and variable changes in appearance. The mixed contribution of appearance features results from individual differences in facial furrows and wrinkles. Some people have a smooth brow at rest, while others have permanent facial furrows of different intensity and shape. Such individual differences can complicate the use of appearance features for AU detection. Cheek raising (AU 6), on the other hand, produces changes in shape that are easily confusable with closely related actions (AU 7 especially). Thus, the information value of shape or appearance for human FACS coders varies by action unit.

From these results, it would seem that there exists complimentary information in all the AAM representations. To test this hypothesis, we fused all these features together using LLR fusion [25]. The results are given in Table IV. As can be seen from the results this seems to be the case as the fusion of all the AAM representations yields the best performance. Again the difference is not great but is significant at $p < 0.05$ when comparing them across all combinations.

The improvement is rather more pronounced when you compare the fusion of all representations (ALL) result to just the SPTS in Table III, where the difference is 6.0 (72.0 vs 78.0), which suggests that applying an AAM approach for spontaneous AU detection would yield better performance than current methods used today, which is intuitive and backs up literature suggesting as much [18].

V. Automatic Pain Detection

The results for automatically detecting pain are given in Figure 7, which shows a clearer view of the trend we observed in the AU detection results in Section IV.A. For the individual feature sets (in blue) we see the SAPP (75.1) features achieving the lowest performance rate, followed by the SPTS (76.9) and then the CAPP (80.9) features again yielding the best results.

When we combine the different feature sets (in red), we again see the benefit of fusing the various representations together showing that there exists complimentary information (although the SAPP+CAPP features is slightly lower than the CAPP features). This is highlighted by when all three representations are fused together (in green). This result is very significant when we compare the similarity-normalized features (SAPP) which most researchers use and compare them

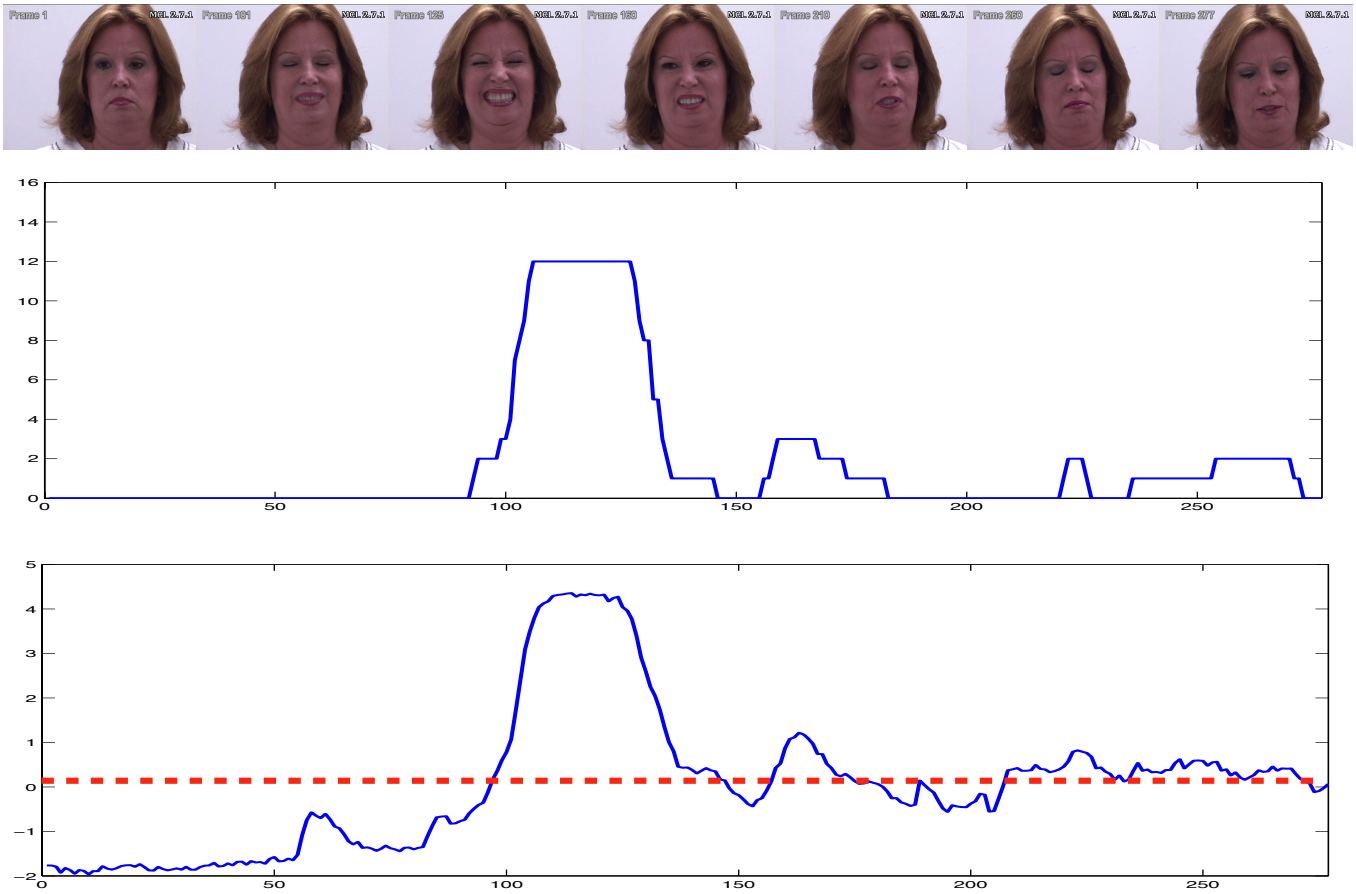


Fig. 8. (Top) the frames which coincide with actions of interest, namely: (a) frame 1, (b) frame 101, (c) frame 125, (d) frame 160, (e) frame 210, (f) frame 260 and (g) frame 277. (Middle) The frame-level FACS coded pain intensity defined by Prkachin and Solomon (as described in Section II.A.). (Bottom) The output scores from the SVM for the combined AAM feature where the horizontal red line denotes the threshold which the scores have to be above for the patient to be deemed to be in pain.

to the combined AAM representations as an improvement of nearly 10% in the area underneath the ROC curve is achieved (75.1 vs 84.7). This highlights the importance of good registration when dealing with spontaneous expressions.

In terms of the relevance to the task of pain detection, these results raise some very interesting issues. The most important one is of *context*. If this system is going to be used for a patient who is mobile and expresses a broad gamut of emotions, the current system will 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. However, if the context is very limited (such as pain/no-pain), then this proposed system would be of use. An example would be in a hospital setting (such as an ICU ward) where the patient is severely impaired, with limited ability to express emotions other than pain/no-pain. This system would then be able to automatically monitor when a patient is in distress and alert care-givers to these periods.

This scenario raises the issue of accuracy, and how much pain does a person have to be in for this to trigger an alert. An example of this is shown in Figure 8, where we see that our system can easily detect the period (frames 90-140) when the person is in major pain (i.e. pain intensity ≥ 10) but for the more subtle pain intensities the decision is still rather ambiguous. However, this may not be important though as intensities of 10 and greater may only be required. So in this context, the application of this system would be of much use, but it is very hard to estimate what would be required in a clinical setting without trailing it.

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.

VI. Conclusions and Future Work

In this paper, we have looked at automatically detecting pain at a frame-by-frame level based on facial action units. For this work we used the UNBC-McMaster database which contains patients with shoulder injuries portraying real or spontaneous pain. A major challenge associated with this was the problem of major facial deformation and head motion caused by the pain, which makes registering the face and facial features a challenging one. This is quite problematic using the

common approach of registering the face via the similarity-transform (normalize for scale, rotation and translation), as the subsequent features miss some important facial information. We have shown this for both AU and pain detection and we show that this can be somewhat overcome by using an AAM approach we can yield significant improvements in terms of area underneath the ROC curve (72.0 vs 78.0 for average AU detection and 75.1 vs 84.7 for pain detection).

The importance of having a system which can automatically detect pain is very important as it could greatly improve the efficiency and overheads associated with monitoring patient progress in a hospital setting. To this end, we have also raised the issue of context and where it would be practical to use such a system and what it would detect (only pain of intensity ≥ 10) and how it would report it (i.e. second, minutes etc.).

As we have noted on several occasions throughout the paper, head motion is a common occurrence though out the dataset. However, it is also indicative of someone in distress. In future work we plan to look at using this as a key future in detecting pain. In addition to this, we hope to look at other modes of information that can be quantified such as eye gaze and body movement (guarding and restlessness). Measuring the overall expressiveness as a combination of these modes maybe the next step in gaining a more robust and accurate objective of pain. The utilization of the system where a patient is in bed needs to be examined as well. This introduces added complexities as the face will be also partially occluded due to the angle of the patient's face to the camera. Using techniques like those described in this paper suggest a potential solution.

REFERENCES

[1] R. Cornelius, *The Science of Emotion*. Upper Saddler River: Prentice Hall, 1996.

[2] T. Hadjistavropoulos and K. Craig, "Social influences and the communication of pain," in *In Pain: Psychological perspectives*. New York, USA: Erlbaum, 2004.

[3] A. Williams, H. Davies, and Y. Chadury, "Simple pain rating scales hide complex idiosyncratic meanings," *Pain*, vol. 85, pp. 457–463, 2000.

[4] D. Wong and C. Baker, "Pain in Children: Comparision of Assessment Scales," *Pediatric Nursing*, vol. 14, no. 1, pp. 9–17, 1988.

[5] D. Turk and R. Melzack, "The measurement of pain and the assessment of people experiencing pain," in *Handbook of pain assessment*.

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Faces of pain: automated measurement of spontaneous facial expressions of genuine and posed pain," in *Proceedings of the International Conference on Multimodal Interfaces*. New York, NY, USA: ACM, 2007, pp. 15–21.

[14] 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.

[15] 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.

[16] 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.

[17] T. Kanade, J. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 46–53.

[18] 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.

[19] 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.

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

[21] 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.

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

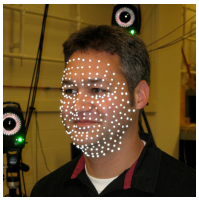
[24] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[25] N. Brummer and J. du Preez, "Application-Independent Evaluation of Speaker Detection," *Computer Speech and Language*, 2005.

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

Patrick Lucey received his B.Eng (Hons) degree in Electrical Engineering from the University of Southern Queensland, Australia, in 2003 and his PhD degree from the Queensland University of Technology, Brisbane, Australia in 2008. He is currently a Post-Doctoral Research Fellow within the Robotics Institute at Carnegie Mellon University as well as the Department of Psychology at the University of Pittsburgh. In 2006, he was a research intern within the Human Language Technology at IBM T.J. Watson Research Centre in Yorktown Heights in New York, USA. In 2007, his paper on "pose-invariant lipreading" was awarded the best student paper at the "INTERSPEECH" conference. In 2008, he was co-organizer for the International Conference on Auditory-Visual Speech Processing (AVSP). His areas of research include, affective computing applied for medical and sporting applications, human-computer-interaction and audio-visual speech recognition. Patrick is a member of the IEEE.

Jeffrey F. Cohn PhD, Professor of Psychology at the University of Pittsburgh and Adjunct Faculty at the Carnegie Mellon University Robotics Institute. He has led interdisciplinary and inter-institutional efforts to develop advanced methods of automatic analysis of facial expression and prosody; and applied those tools to research in human emotion, social development, non-verbal communication, psychopathology, and biomedicine. He co-chaired the 2008 IEEE International Conference on Automatic Face and Gesture Recognition and the 2009 International Conference on Affective Computing and Intelligent Interaction. He has co-edited two recent special issues of the Journal of Image and Vision Computing on social signal processing and automatic facial expression recognition.



Iain Matthews received a BEng degree in electronic engineering in 1994, and a PhD in computer vision in 1998, from the University of East Anglia. He was a member of Systems faculty of the Robotics Institute at Carnegie Mellon University until 2006 conducting research in face modelling and tracking. He then spent two years at Weta Digital as part of the team that developed the facial mocap system for the movie Avatar. Since 2008 he has been a Senior Research Scientist at Disney Research Pittsburgh where he leads the computer vision group. He also holds an adjunct faculty position at the Robotics Institute at CMU. Iain is a member of the IET, IEEE and the IEEE Computer Society.

Simon Lucey is a Senior Research Scientist in the CSIRO ICT Centre and a current "Futures Fellow Award" recipient (2009 - 2013) from the Australian Research Council. Previous to joining the CSIRO, Simon was an Assistant Research Professor in the Robotics Institute at Carnegie Mellon University, and was a faculty member there from 2005 to October 2009. Before that he was a Post Doctoral Fellow in the Electrical and Computer Engineering (ECE) department at Carnegie Mellon University. Dr. Lucey's research interests are in computer vision and machine learning with specific interests in their application to human behaviour (particularly with reference to faces and bodies). He received his Ph.D. in 2003 on the topic of audio-visual speaker and speech recognition from the Queensland University of Technology (QUT) and his undergraduate degree in Electrical and Electronic Engineering from the University of Southern Queensland (USQ), Australia. To his credit he has over 50 publications in international conferences, journals and book chapters.

Jessica Howlett received her B.Eng(Hons) in Computer Systems Engineering in 2007, and her MEngSc (Computer and Communications Engineering) in 2008 from the Queensland University of Technology, Brisbane, Australia. Her Master's project thesis was in the area of Pain Expression Recognition, undertaken within the SAIVT Laboratory at QUT.

Sridha Sridharan has a BSc (Electrical Engineering) degree and obtained a MSc (Communication Engineering) degree from the University of Manchester Institute of Science and Technology (UMIST), UK and a PhD degree in the area of Signal Processing from University of New South Wales, Australia. He is currently with the Queensland University of Technology (QUT) where he is a full Professor in the School of Engineering Systems. Professor Sridharan is the Deputy Director of the Information Security Institute and the Leader of the Research Program in Speech, Audio, Image and Video Technologies at QUT. In 1997, he was the recipient of the Award of Outstanding Academic of QUT in the area of Research and Scholarship. In 2006 he received the QUT Faculty Award for Outstanding Contribution to Research. Professor Sridharan is a Senior Member of the IEEE.

Kenneth M. Prkachin is a Professor and was the founding Chair of Psychology at the University of Northern British Columbia. He earned his Ph.D. in Clinical Psychology from the University of British Columbia in 1978. His research is generally focused on the role of emotion and its expression in health and illness, with specific applications to the study of pain and risk of heart disease. He has published several articles on the nature and properties of the facial expression of pain.