The Painful Face – Pain Expression Recognition Using Active Appearance Models

Ahmed Bilal Ashraf [*]
bilal@cmu.edu

Simon Lucey [*]
slucey@cs.cmu.edu

Jeffrey F. Cohn [*] [†]
jeffcohn@cs.cmu.edu

Tsuhan Chen [*]
tsuhan@cmu.edu

Zara Ambadar [†]
ambadar@pitt.edu

Ken Prkachin [‡]
kmprk@unbc.ca

Patty Solomon [§]
solomon@mcmaster.ca

Barry-John Theobald [¶]
b.theobald@uea.ac.uk

ABSTRACT

Pain is typically assessed by patient self-report. Self-reported pain, however, is difficult to interpret and may be impaired or not even possible, as in young children or the severely ill. Behavioral scientists have identified reliable and valid facial indicators of pain. Until now they required manual measurement by highly skilled observers. We developed an approach that automatically recognizes acute pain. Adult patients with rotator cuff injury were video-recorded while a physiotherapist manipulated their affected and unaffected shoulder. Skilled observers rated pain expression from the video on a 5-point Likert-type scale. From these ratings, sequences were categorized as no-pain (rating of 0), pain (rating of 3, 4, or 5), and indeterminate (rating of 1 or 2). We explored machine learning approaches for pain-no pain classification. Active Appearance Models (AAM) were used to decouple shape and appearance parameters from the digitized face images. Support vector machines (SVM) were used with several representations from the AAM. Using a leave-one-out procedure, we achieved an equal error rate of 19% (hit rate = 81%) using canonical appearance and shape features. These findings suggest the feasibility of automatic pain detection from video.

[*]Carnegie Mellon University (CMU), USA

[†]University of Pittsburgh, USA

[‡]University of Northern British Columbia, Canada

[§]McMaster University, Canada

[¶]University of East Anglia, Norwich, UK

ICMI'07, November 12-15, 2007, Nagoya, Aichi, Japan.

Categories and Subject Descriptors

I.2.1.0 [**Vision and Scene Understanding**]: [video analysis, motion, modeling and recovery of physical attributes]; J.4 [**Computer Applications**]: Social and Behavioral Sciences—*psychology*

General Terms

Machine Learning, Pattern Recognition, Computer Vision, Expression Recognition

Keywords

Active Appearance Models, Support Vector Machines, Pain, Facial expression, Automatic facial image analysis

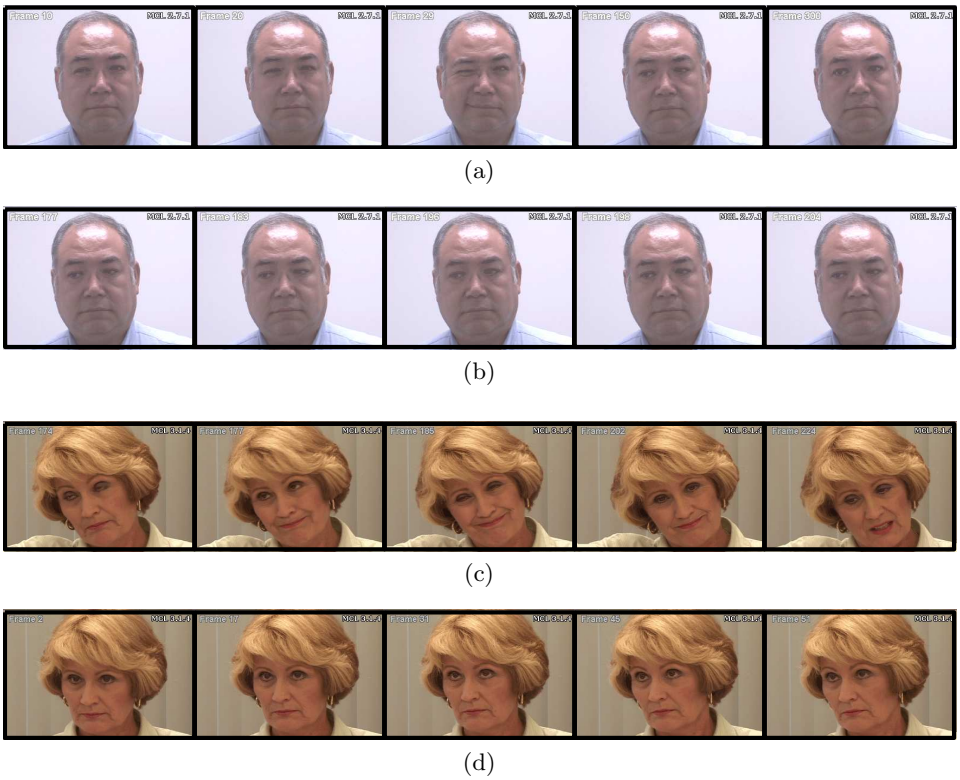
1. INTRODUCTION

Pain is difficult to assess and manage. Pain is fundamentally subjective and is typically measured by patient self-report, either through clinical interview or visual analog scale (VAS). Using the VAS, a patient indicates the intensity of pain by marking a line on a horizontal scale, anchored at each end with words such as "no pain" and "the worst pain imaginable". This and similar techniques are popular because they are convenient, simple, satisfy a need to attach a number to the experience of pain, and often yield data that confirm expectations.

Self-report measures, however, have several limitations [7, 12]. These include idiosyncratic use, 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 [9]. Moreover, self-report measures cannot be used with young children, with many patients in postoperative care or transient states of consciousness, and with severe disorders requiring assisted breathing, among other conditions.

Significant efforts have been made to identify reliable and valid facial indicators of pain [8]. These methods require manual labeling of facial action units or other observational measurements by highly trained observers [4, 11]. Most must be performed offline, which makes them ill-suited for real-time applications in clinical settings.

In the past several years, significant progress has been made in machine learning to automatically recognize facial



(a)

(b)

(c)

(d)

Figure 1: Examples of temporally subsampled sequences. (a) and (c) : Pain ; (b) and (d) : No Pain

expressions related to emotion [16, 17]. While much of this effort has used simulated emotion with little or no head motion, several systems have reported success in facial action recognition in real-world facial behavior, such as people lying or telling the truth, watching movie clips intended to elicit emotion, or engaging in social interaction [3, 5, 18]. In real-world applications and especially in patients experiencing acute pain, out-of-plane head motion and rapid changes in head motion and expression are particularly challenging. Extending the approach of [14], we applied machine learning to the task of automatic pain detection in a real-world clinical setting involving patients undergoing treatment for pain.

In machine learning, the choice of representation is known to influence recognition performance [10]. Both appearance and shape representations have been investigated. Examples of appearance based representations are raw pixels and Gabor filters [1, 2]. A drawback of appearance based approaches is that they lack shape registration, and thus cannot locate vital expression indicators, such as the eyes, brows, eyelids, and mouth.

Shape-based representations, which include Active Shape Models [6] and Active Appearance Models(AAM) [6, 15] address this concern. They decouple shape and appearance and perform well in the task of expression recognition, especially in the context of non-rigid head motion [19]. In our previous work [14], we explored various representations derived from AAMs and concluded that this decoupling between shape and appearance was indeed beneficial for action unit recognition. In this paper we extend our work based on AAM representations to pain expression recognition.

2. IMAGE DATA

Image data were from the UNBC-McMaster Shoulder Pain Expression Archive, which includes video clips for 129 subjects (63 male, 66 female). Each subject was seen in "active" and "passive" conditions. In the active condition, subjects initiated shoulder rotation on their own; in passive, a physiotherapist was responsible for the movement. Camera angle for active tests was approximately frontal to start; camera angle for passive tests was approximately 70 degrees to start.

Following each movement, subjects completed a 10 cm Visual Analog Scale to indicate their level of subjective pain. The scale was presented on paper, with anchors of "no pain" and "worst pain imaginable". Video of each trial was then rated by an observer with considerable training in the identification of pain expression, using a 0(no pain) – 5(strong pain) Likert-type scale. Reliability of coding was established with a second coder, proficient in Facial Action Coding System (FACS) [11]. The Pearson correlation between the observers' ratings on 210 randomly selected tests was 0.80. Correlation between observer rating and subject self-reported pain on VAS was 0.74, for the tests used for recognition of pain from no pain. This correlation suggests moderate to high concurrent validity for pain intensity. For the current study, we included subjects in the active condition who had one or more pain ratings of both pain (0) and no-pain (3, 4, or 5). Fifty eight subjects met this initial criterion. Of these 58, 31 were excluded for reasons including face out of frame, glasses, facial hair, bad image quality, occlusion, and hair in face. The final sample consisted of 21 subjects.

Videos were captured at a resolution of 320x240, out of

which the face area spanned an average of approximately 140x200 (28000) pixels. Sample pain sequences are shown in Figure 1.

3. AAMs

In this section we briefly describe Active Appearance Models (AAMs). The key dividend of AAMs is that they provide a decoupling between shape and appearance of a face image. Given a pre-defined linear shape model with linear appearance variation, AAMs have been demonstrated to produce excellent alignment of the shape model to an unseen 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 [6]. Keyframes within each video sequence were manually labeled, while the remaining frames were automatically aligned using a gradient-descent AAM fit described in [15].

3.1 AAM Derived Representations

The *shape* $\mathbf{s}$ of an AAM [6] is described by a 2D triangulated mesh. In particular, the coordinates of the mesh vertices define the shape $\mathbf{s}$ (see row 1, column (a), of Figure 2 for examples of this mesh). These vertex locations correspond to a source appearance image, from which the shape was aligned (see row 2, column (a), of Figure 2). Since AAMs allow linear shape variation, the shape $\mathbf{s}$ can be expressed as a base shape $\mathbf{s}_0$ plus a linear combination of m shape vectors $\mathbf{s}_i$:

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

where the coefficients $\mathbf{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$ are the shape parameters. These shape parameters can typically be divided into similarity parameters $\mathbf{p}_s$ and object-specific parameters $\mathbf{p}_o$, such that $\mathbf{p}^T = [\mathbf{p}_s^T, \mathbf{p}_o^T]$. Similarity parameters are associated with the geometric similarity transform (i.e., translation, rotation and scale). The object-specific parameters, are the residual parameters representing geometric variations associated with the actual object shape (e.g., mouth opening, eyes shutting, etc.). Procrustes alignment [6] is employed to estimate the base shape $\mathbf{s}_0$.

Once we have estimated the base shape and shape parameters, we can normalize for various variables to achieve different representations as outlined in the following subsections.

3.1.1 Similarity Normalized Shape, $\mathbf{s}_n$

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

3.1.2 Similarity Normalized Appearance, $\mathbf{a}_n$

This representation contains appearance from which similarity variation has been removed. Once we have similarity normalized shape $\mathbf{s}_n$, as computed in section 3.1.1, a similarity normalized appearance $\mathbf{a}_n$ can then be synthesized by employing a piece-wise affine warp on each triangle patch appearance in the source image (see row 2, column (b), of Figure 2) so the appearance contained within $\mathbf{s}$ now aligns with the similarity normalized shape $\mathbf{s}_n$.

3.1.3 Shape Normalized Appearance, $\mathbf{a}_0$

If we can remove all shape variation from an appearance, we'll get a representation that can be called as shape normalized appearance, $\mathbf{a}_0$. $\mathbf{a}_0$ can be synthesized in a similar fashion as $\mathbf{a}_n$ was computed in section 3.1.2, but instead ensuring the appearance contained within $\mathbf{s}$ now aligns with the base shape $\mathbf{s}_0$. We shall refer to this as the face's *canonical appearance* (see row 2, column (c), of Figure 2 for an example of this canonical appearance image) $\mathbf{a}_0$.

3.2 Features

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

S-PTS: *similarity normalized shape* $\mathbf{s}_n$ representation (see Equation 1) of the face and its facial features. There are 68 vertex points in $\mathbf{s}_n$ for both $x-$ and $y-$ coordinates, resulting in a raw 136 dimensional feature vector.

S-APP: *similarity normalized appearance* $\mathbf{a}_n$ representation. Due to the number of pixels in $\mathbf{a}_n$ varying from image to image, we apply a mask based on $\mathbf{s}_0$ so that the same number of pixels (approximately $27,000$) are in $\mathbf{a}_n$ for each image.

C-APP: *canonical appearance* $\mathbf{a}_0$ representation where all shape variation has been removed from the source appearance except the base shape $\mathbf{s}_0$. This results in an approximately $27,000$ dimensional raw feature vector based on the pixel values within $\mathbf{s}_0$.

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

4. SVM CLASSIFIERS

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

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

where $\mathbf{w}$ is the vector normal to the separating hyperplane and b is the bias. Both $\mathbf{w}$ and b are estimated so that they minimize the structural risk of a train-set, thus avoiding the possibility of overfitting to the training data. Typically, $\mathbf{w}$ is not defined explicitly, but through a linear sum of support vectors. As a result SVMs offer additional appeal as they allow for the employment of non-linear combination functions through the use of kernel functions, such as the *radial basis function* (RBF), *polynomial* and *sigmoid* kernels. A linear kernel was used in our experiments due to its ability to generalize well to unseen data in many pattern recognition tasks [13]. Please refer to [13] for additional information on SVM estimation and kernel selection.

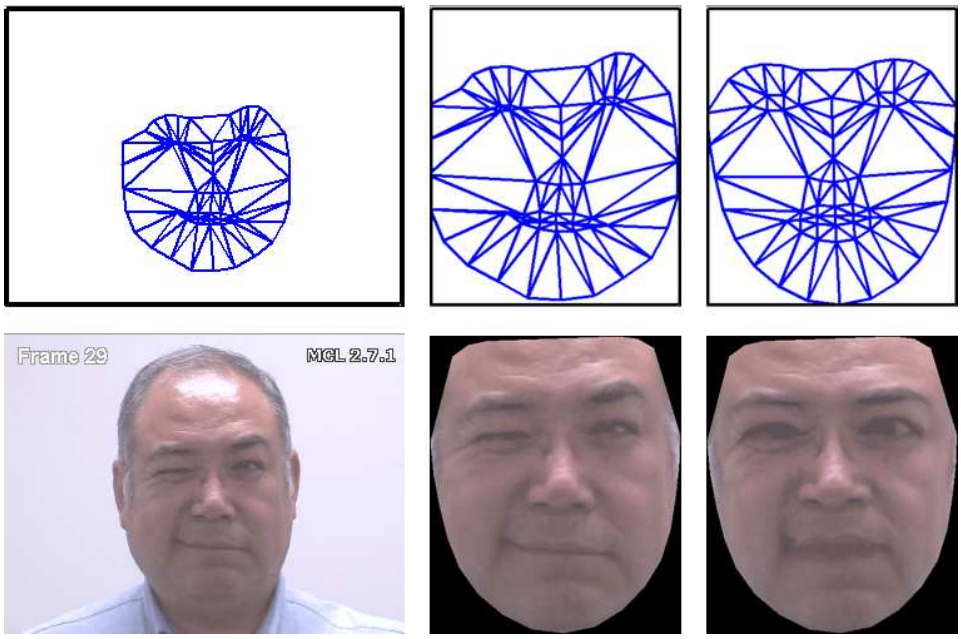
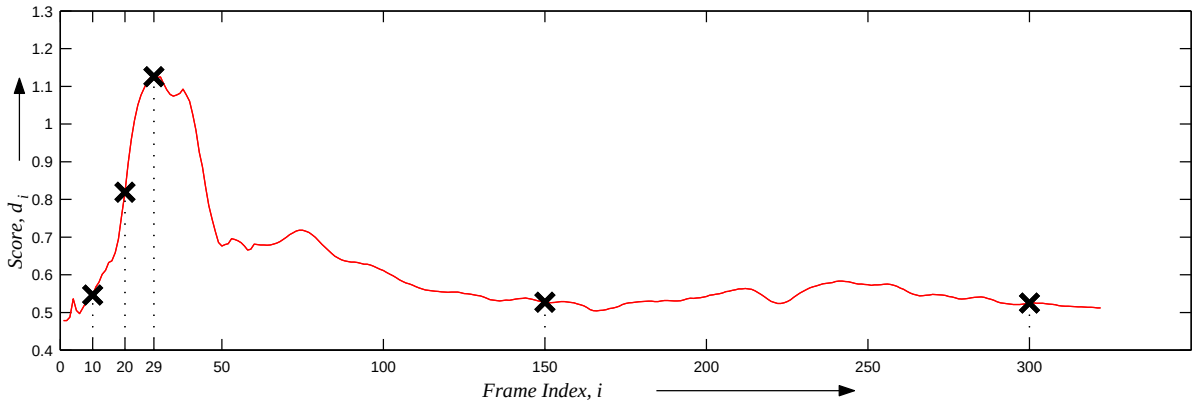


Figure 2: Example of AAM Derived Representations (a) Top row :input shape(s), Bottom row : input image, (b) Top row: Similarity Normalized Shape(s_n), Bottom Row: Similarity Normalized Appearance(a_n), (c) Top Row: Base Shape(s_0), Bottom Row: Shape Normalized Appearance(a_0)

Frame # 10 *Frame # 20* *Frame # 29* *Frame # 150* *Frame # 300*

(a)



(b)

Figure 3: Example of Video Sequence Prediction (a) Sample frames from a pain-video sequence with their frame indices, (b) Scores for individual frames. Points corresponding to the frames shown in (a) are highlighted. For the example sequence shown, a cumulative score $D_{sequence} = \sum_{i=1}^{T} d_i = 195$ was computed and compared to the derived equal error threshold of -3 (as explained in Section 5.2) to yield an output decision of 'Pain'

5. EXPERIMENTS

5.1 Pain Model Learning

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

S-PTS: *similarity normalized shape* $\mathbf{s}_n$

S-APP: *similarity normalized appearance* $\mathbf{a}_n$

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

To check for subject generalization, a leave-one-subject-out strategy was employed for cross validation. Thus, there was no overlap of subjects between the training and testing set. The number of training frames from all the video sequences was prohibitively large to train an SVM, as the training time complexity for SVM is $O(m^3)$, where m is the number of training examples. In order to make the step of model learning practical, while making the best use of training data, each video sequence was first clustered into a preset number of clusters. Standard K-Means clustering algorithm was used, with K set to a value that reduces the training set to a size manageable by SVM.

Linear SVM training models were learned by by iteratively leaving one subject out, which gives rise to N number of models, where N is the number of subjects. Since the ground truth was available at the level of an entire video sequence, all the clusters belonging to a pain containing sequence were considered as positive (pain) examples, while clusters belonging to 'no pain' sequence were counted as negative (no pain) training examples.

5.2 Video Sequence Prediction

While learning was done on clustered video frames, testing was carried out on individual frames. The output for every frame was a score proportional to the distance of test-point from the separating hyperplane.

In order to predict a video as a pain sequence or not, the output scores of individual frames were summed together to give a cumulative score for the entire sequence:

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

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

Having computed the sequence level cumulative score in Equation 3, we seek a decision rule of the form:

$$D_{sequence} \underset{\text{no pain}}{\overset{\text{pain}}{\gtrless}} Threshold \tag{4}$$

Using this decision rule, a threshold was set such that false accept rate equaled false reject rate (i.e., equal error rate point was computed). Output scores for a sample sequence along with the implementation of decision rule is shown in Figure 3. The score values track the pain expression, with a peak response corresponding to frame 29 shown in Figure 3(a).

Features	EER	Hit Rate
S-PTS	0.2899	0.7102
S-APP	0.5345	0.4655
C-APP + S-PTS	0.1879	0.8121

Table 1: Results of experiments performed in section 5. Column 1 indicates the features used for training, Columns 2 and 3 represent the corresponding Equal-Error and Hit Rates respectively

6. RESULTS

Table 1 shows equal-error rates (EER) using each of the representations. It is intuitively satisfying that the results highlight the importance of shape features for pain expression. The best results (EER = 0.1879 , Hit Rate : 0.8121) are for canonical appearance combined with similarity normalized shape (C-APP + S-PTS). This result is consistent with our previous work [14], in which we used AAMs for facial action unit recognition.

It is surprising, however, that similarity normalized appearance features (S-APP) performed at close-to-chance levels despite the fact that this representation can be fully derived from canonical appearance and similarity normalized shape. S-PTS combined with C-APP may add additional dimensions that aid in finding an optimal separating hyperplane between two classes.

7. DISCUSSION

In this paper, we have explored various face representations derived from AAMs for recognizing facial expression of pain. We have demonstrated the utility of AAM representations for the task of pain-no pain classification. We have also presented a method to handle large training data when learning SVM models. Some conclusions of our experiments are:

- Use of concatenated similarity normalized shape and shape normalized appearance (S-PTS + C-APP) is superior to either similarity normalized shape (S-PTS) or similarity normalized appearance (S-APP) alone.
- Shape features have an important role to play in expression recognition generally, and pain detection particularly.
- Automatic pain detection through video appears to be a feasible task. This finding suggests an opportunity to build interactive systems to detect pain in diverse populations, including infants and speech impaired adults. Additional applications for system development include pain-triggered automatic drug dispensation and discriminating between feigned and real pain.
- Two limitations may be noted. One was the exclusion of subjects wearing glasses or having facial hair, which limits generalizability. A second limitation was the availability of ground-truth at the level of video recorded tests instead of the level of individual frames. Not all frames in a pain sequence actually correspond to pain. Providing a frame-level ground-truth could assist the model-learning step and thus improve recognition performance. This limitation can be overcome

either by expert labeling of frames or by automatic refinement of ground-truth by exploiting the information embedded in no-pain video sequences. Incorporating temporal information is also expected to positively influence the performance. Based on our initial findings, further improvements on above lines can pave way towards developing automatic pain detection tools for clinical use.

8. ACKNOWLEDGMENTS

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

9. REFERENCES

[1] M. Bartlett, G. Littlewort, C. Lainscesk, I. Fasel, and J. Movellan. Machine learning methods for fully automatic recognition of facial expressions and facial actions. *IEEE International Conference on Systems, Man and Cybernetics*, pages 592–597, October 2004.

[2] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscesk, I. Fasel, and J. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2:568–573, June 2005.

[3] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pages 223–228, 2006.

[4] J. F. Cohn, Z. Ambadar, and P. Ekman. Observer-based measurement of facial expression with the facial action coding system. In J. A. Coan and J. B. Allen, editors, *The handbook of emotion elicitation and assessment. Oxford University Press Series in Affective Science*, pages 203–221. Oxford University Press, New York, NY, 2007.

[5] J. F. Cohn and K. L. Schmidt. The timing of facial motion in posed and spontaneous smiles. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:1–12, 2004.

[6] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *PAMI*, 23(6):681–685, 2001.

[7] R. R. Cornelius. *The science of emotion.* Prentice Hall, Upper Saddler River, New Jersey, 1996.

[8] K. D. Craig, K. M. Prkachin, and R. V. E. Grunau. The facial expression of pain. In D. C. Turk and R. M., editors, *Handbook of pain assessment.* Guilford, New York, 2nd edition, 2001.

[9] A. C. d. C. Williams, H. T. O. Davies, and Y. Chadury. Simple pain rating scales hide complex idiosyncratic meanings. *Pain*, 85:457–463.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification.* John Wiley and Sons, Inc., New York , NY , USA, 2nd edition, 2001.

[11] P. Ekman, W. V. Friesen, and J. C. hager. Facial action coding system: Research Nexus. Network Research Information, Salt Lake City , UT, 2002.

[12] T. Hadjistavropoulos and K. D. Craig. Social influences and the communication of pain. In *Pain: Psychological perspectives*, pages 87–112. Erlbaum, New York, 2004.

[13] C. W. Hsu, C. C. Chang, and C. J. Lin. A practical guide to support vector classification. *Technical Report*, 2005.

[14] S. Lucey, A. B. Ashraf, and J. Cohn. Investigating spontaneous facial action recognition through AAM representations of the face. In K. Kurihara, editor, *Face Recognition Book.* Pro Literatur Verlag, Mammendorf, Germany, April 2007.

[15] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004.

[16] M. Pantic, A. Pentland, A. Niholt, and T. S. Huang. Human computing and machine understanding of human behavior: A survey. *Proceedings of the ACM International Conference on Multimodal Interfaces*, 2006.

[17] Y. Tian, J. F. Cohn, and T. Kanade. Facial expression analysis. In S. Z. Li and A. K. Jain, editors, *Handbook of face recognition*, pages 247–276. Springer, New York , NY, 2005.

[18] M. F. Valstar, M. Pantic, Z. Ambadar, and J. F. Cohn. Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. *Proceedings of the ACM International Conference on Multimodal Interfaces*, pages 162–170, November 2006.

[19] J. Xiao, S. Baker, I. Matthews, and T. Kanade. (In press). 2d vs. 3d deformable face models: Representational power, construction, and real-time fitting. *International Journal of Computer Vision, Springer Science.*