Automated Facial Expression Recognition System

Andrew Ryan
Naval Criminal Investigative Services, NCIS
Washington, DC United States
andrew.h.ryan@navy.mil

Jeffery F. Cohn, Simon Lucey, Jason Saragih, Patrick Lucey, & Fernando De la Torre
Carnegie Mellon University
Pittsburgh, Pennsylvania United States
jeffcohn@cs.cmu.edu, slucey@cs.cmu.edu, jsaragih@andrew.cmu.edu, ftorre@cs.cmu.edu

Adam Rossi
Platinum Solutions, Inc
Reston, Virginia United States
adam.rossi@platinumsolutions.com

***Abstract*—Heightened concerns about the treatment of individuals during interviews and interrogations have stimulated efforts to develop "non-intrusive" technologies for rapidly assessing the credibility of statements by individuals in a variety of sensitive environments. Methods or processes that have the potential to precisely focus investigative resources will advance operational excellence and improve investigative capabilities. Facial expressions have the ability to communicate emotion and regulate interpersonal behavior. Over the past 30 years, scientists have developed human-observer based methods that can be used to classify and correlate facial expressions with human emotion. However, these methods have proven to be labor intensive, qualitative, and difficult to standardize. The Facial Action Coding System (FACS) developed by Paul Ekman and Wallace V. Friesen is the most widely used and validated method for measuring and describing facial behaviors. The Automated Facial Expression Recognition System (AFERS) automates the manual practice of FACS, leveraging the research and technology behind the CMU/PITT Automated Facial Image Analysis System (AFA) system developed by Dr. Jeffery Cohn and his colleagues at the Robotics Institute of Carnegie Mellon University. This portable, near real-time system will detect the seven universal expressions of emotion (*figure 1*), providing investigators with indicators of the presence of deception during the interview process. In addition, the system will include features such as full video support, snapshot generation, and case management utilities, enabling users to re-evaluate interviews in detail at a later date.**

Keywords-automated facial expression recognition system; Biometric systems; utilizing facial features; shape and appearance modeling; expression recognition; facial action processing; constrained local models; facial expression recognition; support vector machines; spontaneous facial behavior

I. INTRODUCTION

Interrogations are a critical practice in the information gathering process, but the information collected can be severely compromised if the interviewee attempts to mislead the interviewer through the use of deception. Being able to quantitatively assess an interview subject's emotional state and changes in emotional state would be a tremendous advantage in being able to guide an interview and assess the truthfulness of the interviewee.

Figure 1. Demonstrates the seven universal expressions of emotion. Each of these expressions is racially and culturally independent.

Recent advances in facial image processing technology have facilitated the introduction of advanced applications that extend beyond facial recognition techniques. This paper introduces an Automated Facial Expression Recognition System (AFERS): A near real-time, next generation interrogation tool that has the ability to automate the Facial Action Coding System (FACS) process for the purposes of expression recognition.

The AFERS system will analyze and report on a subject's facial behavior, classifying facial expressions with one of the seven universal expressions of emotion [1].

II. FACIAL ACTION CODING SYSTEM

In behavioral psychology, research into systems and processes that can recognize and classify a subject's facial expressions have allowed scientists to more accurately assess and diagnose underlying emotional state. This practice has since opened the door to new breakthroughs in areas such as pain analysis and depression treatment. In 1978, Paul Ekman and Wallace V. Friesen published the Facial Action Coding System (FACS), which, 30 years later, is still the most widely used method available. Through observational and electromyographic study of facial behavior, they determined how the contraction of each facial muscle, both singly and in unison with other muscles, changes the appearance of the face.

978-1-4244-4170-9/09/$25.00 ©2009 IEEE

Rather than using the names of the active muscles, FACS measures these changes in appearance using units called Action Units (AUs). ***Figure 2*** illustrates some of these Action Units and the appearance changes they describe. The benefits of using AUs are two-fold. First, individually and in combination they provide a way to unambiguously describe nearly all possible facial actions. Second, combinations of AUs refer to emotion-specified facial expressions. Happy, for instance, is distinguished by the combination AU 6 and AU 12. AU 6, *orbicularis oculi* contraction, raises the cheeks and causes wrinkling lateral to the eyes. AU 12, *zygomatic major* contraction, pulls the lip corners obliquely into a smile. Seven expressions appear universally in Western- and non-Western, literate, and pre-literate cultures [1].

While FACS is an efficient, objective method to describe facial expressions, it is not without its drawbacks. Coding a subject's video is a time- and labor-intensive process that must be performed frame by frame. A trained, certified FACS coder takes on average 2 hours to code 2 minutes of video. In situations where real-time feedback is desired and necessary, manual FACS coding is not a viable option.

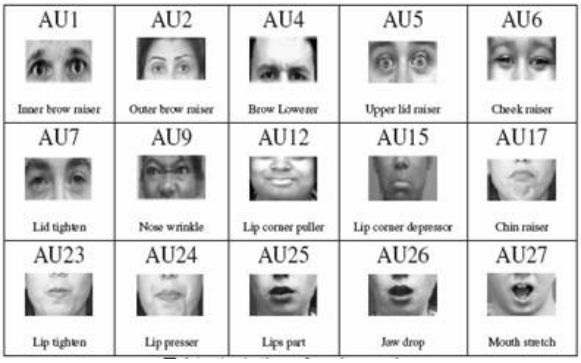

Figure 2. Sample AUs and the appearance changes they describe [2].

III. AFERS System Overview

AFERS is designed to operate in a platform independent manner allowing it to be hosted on various hardware platforms and to be compatible with most standard video cameras. AFERS employs shape and appearance modeling using constrained local models for facial registration and feature extraction and representation, and support vector machines for expression classification. AFERS provides both pre- and post-analysis capabilities and includes features such as video playback, snapshot generation, and case management. In addition to the AFERS processing algorithms, the implementation features a plug-in architecture that is capable of accommodating future algorithmic enhancements as well as additional inputs for behavior analysis.

AFERS is built upon both Java and C++ technologies. The user interface, video processing and analytics engine are built using Java and the expression recognition engine is built using C++. The two technologies are bridged via the Java Native Interface (JNI). ***Figure 3*** depicts a high level overview of each of these components, and their interaction with each other during the expression recognition process.

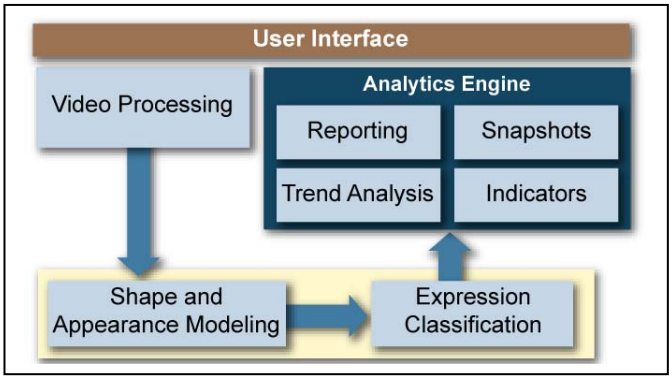

Figure 3. High level overview of the AFERS components.

IV. Video Processing

The video processing component of the AFERS application is responsible for sequencing the inputted video into individual frames at a rate of 25 frames per second. Once the video is sequenced the video processing component places the frames onto a queue for input into the shape and appearance modeling component.

V. Shape & Appearance Modeling

The successful automatic registration and tracking of non-rigidly varying geometric landmarks on the face is a key ingredient to the analysis of human spontaneous behavior. Until recently, popular approaches for accurate non-rigid facial registration and tracking have centered upon inverting a synthesis model (or in machine learning terms a generative model) of how faces can vary in terms of shape and appearance. As a result, the ability of such approaches to register an unseen face image is intrinsically linked to how well the synthesis model can reconstruct the face image.

Perhaps the most well known application of inverting a synthesis model for non-rigid face registration can be found in the active appearance model (AAM) work first proposed by [3]. Other closely related methods can be found in the morphable models work of Blanz and Vetter [4]. AAMs have become the *de facto* standard in non-rigid face alignment/tracking [5].

An example of a shape and appearance model using AAM can be seen in ***Figure 4***. Shape is represented by the *x,y* coordinates of facial features and appearance by the texture within that shape. Any face can be represented by a mean shape and appearance and their modes of variation. Because the models are generative, new faces and expressions can be generated from the model.

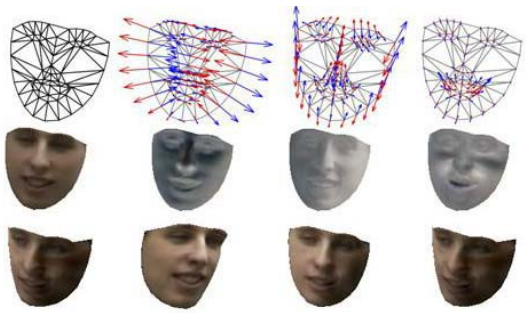

Figure 4. Row 1 shows the mean face shape (left) and 1st three shape modes and row 2 the mean appearance (left) and 1st three appearance modes for an AAM. The AAM is invertible and can be used to synthesize new face images. Four example face faces generated with an AAM are shown in row 3.

AAMs are learned from training data. That is, they are person-specific. In our experience, about 5% of face images must be hand labeled for model training. For many biometric and forensic applications, generic models that can be used with previously unknown persons are needed. AAMs have inherent problems when attempting to fit generically to face images. This problem can be directly attributed to the balance that shape and appearance models require in their representational power (i.e., the model's ability to synthesize face images). If the representational power is too constrained, the method can do a good job on a small population of faces but cannot synthesize faces outside that population. On the other hand, if the representational power is too unconstrained, the model can easily synthesize all faces but can also synthesize non-face objects. Finding a suitable balance between these two extremes in a computationally tractable manner has not been easily attained through an invertible synthesis paradigm [6]. Hence, pre-training on face images for person-specific models has been needed. Person-specific models, while capable of precise tracking, are ill-suited for use with unknown persons. AFERS is intended for just such use, person-specific shape and appearance modeling.

A. Constrained local models (CLM)

Accurate and consistent tracking of non-rigid object motion, such as facial motion and expressions, is important in many computer vision applications and has been studied intensively in the last two decades. This problem is particularly difficult when tracking subjects with previously unseen appearance variations. To address this problem, a number of registration/tracking methods have been developed based on local region descriptors and a non-rigid shape prior. We refer to this family of methods collectively as a constrained local model (CLM). Our definition of CLMs is much broader than that given by Cristinacce and Cootes [7] who employ the same name for their approach. Cristinacce and Cootes' method can be thought of as a specific subset of the CLM family of models. Probably, the best-known example of a CLM can be found in the seminal active shape model (ASM) work of Cootes and Taylor. Instantiations of CLMs differ primarily in the literature with regards to: (i) whether the local experts employ a 1D or 2D local search, (ii) how the local experts are learnt, (iii) how the source image is normalized geometrically and photometrically before the application of the local experts, and (iv) how one fits the local experts' responses to conform to the global non-rigid shape prior. Disregarding these differences, however, all instantiations of CLMs can be considered to be pursuing the same two goals: (i) perform an exhaustive local search for each landmark around their current estimate using some kind of patch-expert (i.e., feature detector), and (ii) optimize the global non-rigid shape parameters such that the local responses for all of its landmarks are minimized

A major advantage of CLMs over conventional methods for non-rigid registration, such as AAMs, lies in their ability to: (i) be discriminative and generalize well to unseen appearance variation; (ii) offer greater invariance to global illumination variation and occlusion; (iii) model the non-rigid object as an ensemble of low dimensional independent patch-experts; and (iv) not employ complicated piece-wise affine texture warp operations that might introduce unwanted noise [8].

For the AFERS project, we are extending the CLM framework in several ways. We are simplifying the optimization in a way that allows the optimization algorithm to be parallelized for faster performance. We are replacing the original non-linear patch experts that were suggested in [7] with linear support vector machines (SVM). This approach further increases performance and improves accuracy of model fitting. And we are using a composite warp in place of an additive warp that increases robustness to changes in scale. These extensions of the CLM framework will enable sufficiently fast model fitting to support the demands of real-time expression recognition.

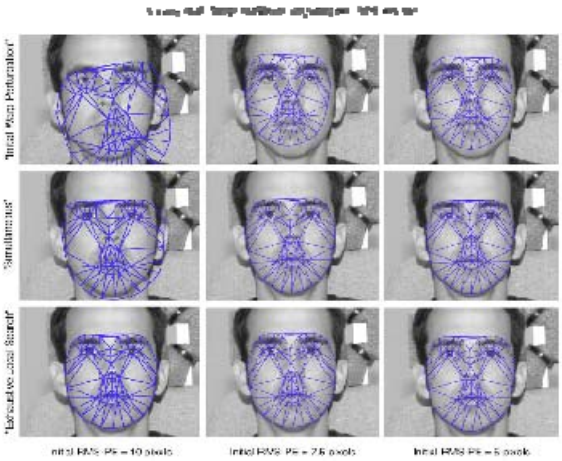

Figure 5. Examples of alignment performance on a single subject's face. Rows 1, 2 and 3 illustrate the alignment for the initial warp perturbation, simultaneous (AAM), and our constrained local model (using exhaustive local search algorithm), respectively. Columns 1, 2, and 3 illustrate the alignment for initial warp perturbation of 10, 7.5 and 5 pixels RMS-PE, respectively.

In initial work, we evaluated our approach to CLM by comparing it with one of the leading approaches to AAM [9] in face images from the Multi-PIE database [10]. Multi-PIE consists of face images of 337 participants of Asian, Caucasian, and African-American background that were recorded under multiple pose, illumination, and expression conditions on as many as four occasions over several months. The database samples some of the variability that AFERS is intended to

manage. The ELS algorithm for CLM was compared against two well-known AAM fitting approaches, namely the ‘‘simultaneous” (SIM) and ‘‘project-out” algorithms. The ELS algorithm obtained real-time fitting speeds of over 35 fps, compared to the SIM algorithm’s speed of 2–3 fps. In addition, the ELS algorithm achieved superior alignment performance to the SIM algorithm in nearly all comparisons. For an example, please see ***Figure 5***. (For further explanation and results, see [11]).

VI. Representation of Facial Features

Once the CLM has estimated the shape and appearance parameters, we can use this information to derive features from the face for expression recognition. From the initial work conducted in [12] we extract the following features:

PTS: Similarity normalized shape, s_n, refers to the vertex points for the *x*- and *y*- coordinates of the face shape, 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 s_n can be obtained by synthesizing a shape instance of s that ignores the similarity parameters *p*. An example of the normalized shape features, PTS, is given in ***Figure 6***.

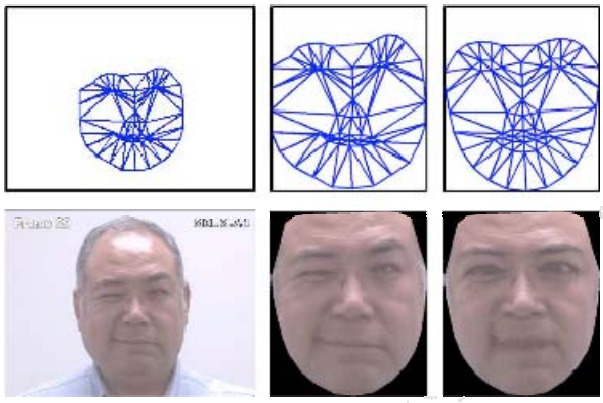

Figure 6. Example of AAM derived representations (a) Top row: input shape, Bottom row: input image, (b) Top row: Similarity Normalized Shape (sn), Bottom Row: Similarity Normalized Appearance(an), (c) Top Row: Base Shape (s0), Bottom Row: Shape Normalized Appearance(a0)

APP: Canonical normalized appearance a_0 refers to where all the non-rigid shape variation has been normalized with respect to the base shape s_0. This is accomplished by warping each triangle patch appearance in the source image so that it aligns with the base face shape. If we can remove all shape variation from an appearance, we obtain a representation referred to as shape-normalized appearance, a_0. This canonical normalized appearance a_0 differs from the similarity normalized appearance a_n in that it removes the non-rigid shape variation and not the rigid shape variation. The resulting features yield an approximately 27,000 dimensional raw feature vector. A mask is applied to each image so that the same number of pixels is used for each. To reduce the dimensionality of the features, we use a 2D discrete cosine transform (DCT). Lucey et al. [11] found that using M = 500 gave the best results. Examples of the reconstructed images with M = 500 are shown in ***Figure 7***. Note that regardless of the head pose and orientation, the appearance features are projected back onto the normalized base shape, so as to make these features more robust to such variability.

PTS+APP: combination of shape and appearance features $s_n + a_0$ refers to the shape features being concatenated to the appearance features.

Figure 7. Top row shows the first three frames of an image sequence. The followings show reconstructed images using 100, 200, and 500 DCT coefficients, repectively. Note that regardless of the head pose and orientation, the appearance features are projected back onto the normalized base shape, so as to make these features more robust to such variability.

VII. Expression Recognition

A leading approach to pattern recognition is that of support vector machines (SVM) [13]. 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 unlabelled test observation x_ by

$$\begin{cases} \text{true} & \text{if } \mathbf{w}^T\mathbf{x}^* \geq b \\ \text{false} & \text{otherwise} \end{cases}$$

where *w* is the vector normal to the separating hyperplane and *b* is the bias. Both *w* and *b* are estimated so that they minimize the structural risk of a train-set, thus avoiding the possibility of over-fitting to the training data. Typically, *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. For AFERS, we will use a linear kernel due to its ability to generalize well to unseen data in many pattern recognition tasks and its efficiency.

Figure 8 gives an example of AFERS processing of an image sequence. Input video is processed using CLM. Shape and appearance parameters are estimated for each video frame and then input to an SVM for expression recognition. AFERS will be tested in two publically available datasets, Cohn-Kanade AU-Coded Facial Expression Database [2] and MMI [15], and in GEMEP [14].

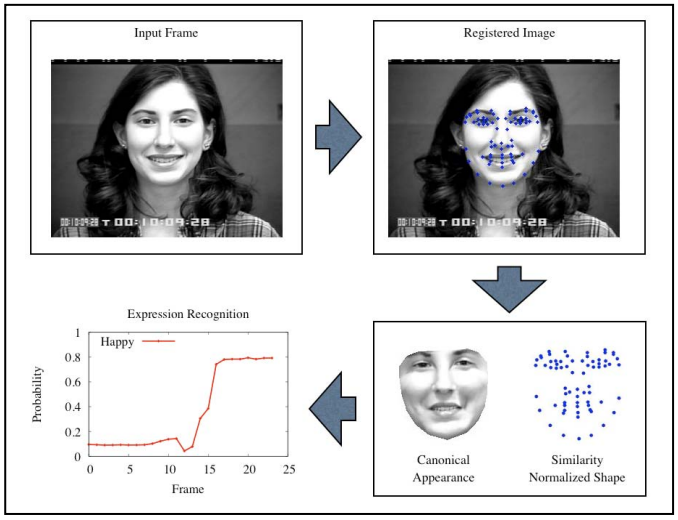

Figure 8. Automatic Facial Expression Recognition System. Similarity normalized shape and and canonical appearance are estimated for each video frame. Parameters are then inputted to SVMs to recognize emotion expression on a frame-by-frame basis.

VIII. Analytics Engine

During runtime, the AFERS application provides operators with several real-time outputs of the expression recognition process, along with snapshot generation and interrogation reporting.

A. Current FACS Emotion Response Indicator

AFERS displays the current expression response demonstrated by the subject and determined by the automated FACS coding. Each time the subject's expression changes, even if only for a fraction of a second (as with microexpressions), the results are updated within the user interface in real-time.

B. Trend Analysis

AFERS also provides a polygraph chart-like output that tracks and displays the historical changes in the subject's expression. As with the emotion response indicator, every time the subject's expression changes, a new data point is plotted. The operator can then perform quick analysis of the trends in a subject's magnitude and duration of expressions, and correlate them to the questions being asked.

C. Snapshot Generation

The AFERS snapshot generation feature allows an operator to manually generate a snapshot of the subject's current emotional response during an interrogation. Each snapshot captures the time, frontal photo of the subject along with the current facial expression, and active action units. AFERS' snapshots can also be configured to be collected automatically, based on the criteria the operator has established (such as any time the subject's expression transitions from neutral to fear, or any time a certain set of action units become active).

D. Cumulative Interrogation Report

At the conclusion of the interrogation, AFERS generates a detailed report that includes all critical information tracked during the interrogation, including vital information about the subject and the date, time, and duration of the interrogation. The complete listing of snapshots generated and expressions recorded, along with the time during the interrogation where they occurred, is reported as well.

IX. AFERS Case File

Due to its intended domain of use in law enforcement interrogation, the initial AFERS software required case management capabilities that were non-proprietary and compact enough to be stored on a laptop. In addition, since multiple installations of the AFERS hardware were envisioned, the ability to archive and transfer cases from one system to another with minimal effort and data conflicts was necessary.

The case file is built upon a simple XML schema with a one-to-many case to subject relationship, and a one-to-many subject to interrogation relationship. The choice to use XML allows for future case management applications to intake AFERS case data, regardless of whether or not the expression information is relevant to the given application's functionality. Also, tied to the case file are the recorded interrogation audio/video and all snapshots gathered throughout the course of the AFERS runtime.

X. Future Applications

While the initial AFERS system is projected to aid in structured scenarios such as interviews and as a supporting aid to the polygraph, AFERS is designed with the ability to provide analysis in unstructured scenarios. With additional research and refinement of the AFERS processing algorithms along with the introduction of additional contextual models, AFERS has the goal of being extended to directly detect the presence of deception.

XI. Conclusion

Extensive efforts have been made over the past two decades in academia, industry, and government to discover more robust methods of assessing truthfulness, deception, and credibility during human interactions. The empirical foundation for AFERS has evolved from the original work by Paul Ekman and Wallace V. Friesen into a robust research and development effort. The potential for AFERS exists not just in the criminal investigative arena but equally in its ability to bolster investigation in national security, counterintelligence, and counterterrorism missions.

ACKNOWLEDGMENT

The research and development of the AFERS application is supported by the Technical Support Working Group through funding from the Investigative Support and Forensics subgroup to Platinum Solutions, Inc. Thanks to Dr. Andrew Ryan from the Naval Criminal Investigative Service for his sponsorship of this initiative.

REFERENCES

[1] Keltner, D., and Ekman, P.: 'Facial expression of emotion', in Lewis, M., and Haviland, J.M. (Eds.): 'Handbook of emotions' (Guilford, 2000, 2nd edition), pp. 236-249.

[2] Kanade, T., Cohn, J.F., and Tian, Y.: 'Comprehensive database for facial expression analysis', Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000, FG '00, pp. 46-53

[3] Edwards, G.J., Taylor, C.J., and Cootes, T.F.: 'Interpreting face images using active appearance models'. Proc. IEEE International Conference on Automatic Face and Gesture Recognition, Zurich, Switzerland,1998.

[4] Blanz, V., and Vetter, T.: 'Face recognition based on fitting a 3D morphable model', IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25, pp. 1063 - 1074.

[5] Matthews, I., and Baker, S.: 'Active appearance models revisited', International Journal of Computer Vision, 2004, 60, (2), pp. 135-164.

[6] Gross, R., Baker, S., and Matthews, I.: 'Generic vs. person specific active appearance models', Image and Vision Computing, 2005, 23, (11), pp. 1080-1093.

[7] Cristinacce, D., and Cootes, T.F.: 'Feature detection and tracking with constrained local models'. Proc. British Machine Vision Conference 2006 .

[8] Lucey, S., Wang, Y., Cox, M., Sridharan, S., and Cohn, J.F.: 'Efficient constrained local model fitting for non-rigid face alignment', Image and Vision Computing Journal, in press, 2009.

[9] Matthews, I., Xiao, J., and Baker, S.: '2D vs. 3D deformable face models: Representational power, construction, and real-time fitting', International Journal of Computer Vision, 2007, 75, (1), pp. 93–113.

[10] Gross, R., Matthews, I., Cohn, J.F., Kanade, T., and Baker, S.: 'Multi-PIE'. Proc. Eighth IEEE International Conference on Automatic Face and Gesture Recognition, Amsterdam 2008.

[11] Lucey, P., Cohn, J.F., Lucey, S., Sridharan, S., and Prkachin, K.: 'Automatically detecting action units from faces of pain: Comparing shape and appearance features'. Proc. 2nd IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB), Miami, FL, 2009.

[12] Ashraf, A.B., Lucey, S., Cohn, J.F., Chen, T., Prkachin, K.M., and Solomon, P.: 'The painful face: Pain expression recognition using active appearance models', Image and Vision Computing, 2009.

[13] Hsu, C.W., Chang, C.C., and Lin, C.J.: 'A practical guide to support vector classification'. Department of Computer Science, National Taiwan University, 2005.

[14] Bänziger, T., and Scherer, K.: 'Using actor portrayals to systematically study multimodal emotion expression: The GEMEP Corpus '. Proc. Affective Computing and Intelligent Interaction (ACII 2007).

[15] Pantic, M., Valstar, M., Rademaker, R., and Maat, L.: 'Web-based database for facial expression analysis' (2005).