Meta-Analysis of the First Facial Expression Recognition Challenge

Michel Valstar, *Member, IEEE,* Marc Mehu, Bihan Jiang, Maja Pantic, *Fellow, IEEE*, Klaus Scherer

***Abstract*—Automatic Facial Expression Recognition has been an active topic in computer science for over two decades, in particular Facial Action Coding System (FACS) Action Unit (AU) detection and classification of a number of discrete emotion states from facial expressive imagery. Standardisation and comparability has received some attention; for instance, there exist a number of commonly used facial expression databases. But lack of a commonly accepted evaluation protocol and typically lack of sufficient details needed to reproduce the reported individual results make it difficult to compare systems. This in turn hinders the progress of the field. A periodical challenge in Facial Expression Recognition would allow such a comparison on a level playing field. It would provide an insight on how far the field has come, and would allow researchers to identify new goals, challenges and targets. This paper presents a meta-analysis of the first such challenge in automatic recognition of facial expressions, held during the IEEE conference on Face and Gesture Recognition 2011. It details the challenge data, evaluation protocol, and the results attained in two sub-challenges: AU detection and classification of facial expression imagery in terms of a number of discrete emotion categories. We also summarise the lessons learned and reflect on the future of the field of facial expression recognition in general and on possible future challenges in particular.**

***Index Terms*—Facial expression analysis, challenges, FACS analysis, discrete emotion recognition**

I. Introduction

COMPUTERS and other powerful electronic devices surround us in ever increasing numbers, with their ease of use continuously being improved by user friendly interfaces. Yet to completely remove all interaction barriers, the next-generation computing (a.k.a. pervasive computing, ambient intelligence, and human computing) will need to develop human-centred user interfaces that respond readily to naturally occurring, multimodal, human communication [40]. An important functionality of these interfaces will be the capacity to perceive and understand the user's cognitive appraisals, action tendencies, and social intentions that are usually associated with emotional experience. Because facial behaviour is believed to be an important source of such emotional and interpersonal information [2], automatic analysis of facial expressions is crucial to human-computer interaction.

Manuscript received 1st of June, 2011. M.F. Valstar is with the Mixed Reality Lab, University of Nottingham, UK. Bihan Jiang and Maja Pantic are with the Department of Computing, Imperial College London, UK. On submission of manuscript M.F. Valstar was with Imperial College. E-mail: Michel.Valstar@imperial.ac.uk. M. Pantic is also with the Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, The Netherlands. E-mail: m.pantic@imperial.ac.uk. Marc Mehu and Klaus Scherer are with the Swiss Center for Affective Sciences, University of Geneva, Switzerland

Facial Expression Recognition, in particular FACS AU detection [18] and classification of facial expression imagery in a number of discrete emotion categories, has been an active topic in computer science for some time now, with arguably the first work on automatic facial expression recognition being published in 1973 [25]. Many promising approaches have been reported since then [41], [68]. The first survey of the field was published in 1992 [45] and has been followed up by several others [20], [41], [68]. However, the question remains as to whether the approaches proposed to date actually deliver what they promise. To help answer that question, we felt it was time to take stock, in an objective manner, of how far the field has progressed.

Researchers often do report on the accuracy of the proposed approaches using a number of popular, publicly available facial expression databases (e.g. The Cohn-Kanade database [26], the MMI-Facial Expression Database [43], [60], or the JAFFE database [33]). However, only too often publications fail to clarify exactly what parts of the databases were used, what the training and testing protocols were, and hardly any cross-database evaluations are reported. All these issues make it difficult to compare different systems to each other, which in turn hinders the progress of the field. A periodical challenge in Facial Expression Recognition would allow this comparison in a fair manner. It would clarify how far the field has come, and would allow us to identify new goals, challenges, and targets.

Two main streams in the current research on automatic analysis of facial expressions consider facial affect (emotion) inference from facial expressions and facial muscle action detection [39], [42], [57], [68]. These streams stem directly from the two major approaches to facial expression measurement in psychological research [10]: message and sign judgment. The aim of the former is to infer what underlies a displayed facial expression, such as affect or personality, while the aim of the latter is to describe the outward "surface" of the shown behaviour, such as facial movement or facial component shape. Thus, a frown can be judged as possibly caused by anger in a message-judgment approach and as a facial movement that lowers and pulls the eyebrows closer together in a sign-judgment approach. While message judgment is all about interpretation, with the ground truth being a hidden state that is often impossible to measure, sign judgment is agnostic, independent from any interpretation attempt, leaving the inference about the conveyed message to higher order decision making. Most facial expression analysis systems developed so far adhere to the message judgment approach. They attempt to recognise a small set of prototypic emotional facial expressions said to relate directly to a small number of discrete affective

states such as the six basic emotions proposed by Ekman [15], [42], [57], [68]. Even though automatic classification of face imagery in terms of the six basic emotion categories is considered largely solved, reports on novel approaches are published even to date (e.g., [28], [35], [50], [54]). While in truth such systems recognise prototypical facial expressions but not actually recognise emotions, for brevity we will refer to this process as 'emotion recognition'.

In sign judgment approaches [9], a widely used method for manual labelling of facial actions is the Facial Action Coding System (FACS) [18]. FACS associates facial expression changes with actions of the muscles that produce them. It defines 9 different action units (AUs) in the upper face, 18 in the lower face, and 5 AUs that cannot be classified as belonging to either the upper or the lower face. Additionally, it defines so-called action descriptors, 11 for head position, 9 for eye position, and 14 additional descriptors for miscellaneous actions. AUs are considered to be the smallest visually discernible facial movements. AU intensity scoring is defined on a 5-level ordinal scale by FACS. It also defines the make up of AUs temporal segments (onset, apex and offset), but goes short of defining rules how to code them in a face video, or what rules govern the transitions between the temporal segments. Using FACS, human coders can manually code nearly any anatomically possible facial expression, decomposing it into the specific AUs and their temporal segments that produced the expression.

As AUs are independent of any interpretation, they can be used as the basis for any higher order decision making process including recognition of basic emotions [18], cognitive states like (dis)agreement and puzzlement [11], psychological states like pain [13], and socio-cultural signals like emblems (i.e., culture-specific interactive signals like wink, coded as left or right AU46), regulators (i.e., conversational mediators like exchange of a look, coded by AUs for eye position), and illustrators (i.e. cues accompanying speech like raised eyebrows, coded as AU1+AU2) [17]. Hence, AUs are extremely suitable to be used as mid-level parameters in an automatic facial behaviour analysis system as they reduce the dimensionality of the problem [62] (thousands of anatomically possible facial expressions [17] can be represented as combinations of 32 AUs).

In terms of feature representation, the majority of the automatic facial expression recognition literature can be divided in three ways: those that use appearance-based features (e.g. [7], [24], [35]), those that use geometric feature-based approaches (e.g. [28], [59]), and those that use both (e.g. [3], [56]). Both appearance- and geometric feature-based approaches have their own advantages and disadvantages, and we expect that systems that use both for this challenge will result in the highest accuracy.

Another way existing systems can be classified is in the way they make use of temporal information. Some systems only use the temporal dynamics information encoded directly in the utilised features (e.g. [24], [69]), others only employ machine learning techniques to model time (e.g. [52], [58]), while others employ both (e.g. [59]). Currently it is unknown what approach could guarantee the best performance.

Fig. 1. An example of the GEMEP-FERA dataset: one of the actors displaying an expression associated with the emotion 'anger'.

This paper describes the first facial expression recognition challenge, organised under the name of FERA 2011, which was held in conjunction with the 9th IEEE International Conference on Automatic Face and Gesture Recognition. The challenge provided a fair comparison between systems vying for the title of 'state of the art'. To do so, it used a partition of the GEMEP corpus [6], developed by the Geneva Emotion Research Group (GERG).

This data is described in in section III of this paper. An overview of recent literature in the field is provided in section II. In section IV we describe the challenge protocol for both the AU detection and emotion recognition sub-challenges. The baseline method against which FERA 2011 participants could compare their results is described in section V. We provide a summary description of the participants' systems in section VI. A detailed analysis of the results attained in this challenge is given in section VII. We conclude the paper with a discussion of the challenge and its results in section VIII.

II. Overview of existing works

Below we present a short overview of the main streams of automatic recognition of prototypical facial expressions associated with discrete emotional states, and of automatic detection of FACS Action Units. For detailed surveys, we refer the reader to [42], [68].

A. Emotion recognition

Emotion recognition approaches can be divided in two groups based on the type of features used, either appearance based features or geometry based features. Appearance features describe the texture of the face caused by expression, such as wrinkles and furrows. Geometric features describe the shape of the face and its components such as the mouth or the eyebrows.

Within the appearance based techniques, the theory of non-negative matrix factorisation (NMF) has recently led to

a number of promising works. A technique called graph-preserving sparse nonnegative matrix factorisation (GNSMF) was introduced by Zhi et al. [71], and applied to the problem of six basic emotions recognition. The GSNMF is an occlusion-robust dimensionality reduction technique that can be employed either in a supervised or unsupervised manner. It transforms high-dimensional facial expression images into a locality-preserving subspace with sparse representation. On the Cohn-Kanade database, it attains a 94.3% recognition rate. On occluded images it scored between 91.4% and 94%, depending on the area of the face that was occluded.

Another recent NMF technique is non-linear non-negative component analysis, a novel method for data representation and classification proposed by Zafeiriou and Petrou [67]. Based on NMF and kernel theory, the method allows any positive definite kernel to be used, and assures stable convergence of the optimisation problem. On the Cohn-Kanade database, they attained an average 83.5% recognition rate over the six basic emotions.

Other appearance features that have been successfully employed for emotion recognition are the Local Binary Pattern (LBP) operator [50], [70], Local Gabor Binary Patterns [35], Local Phase Quantisation (LPQ) and Histogram of oriented Gradients [14], and Haar filters [31].

Most geometric feature based approaches use Active Appearance Models (AAMs) or derivatives of this technique to track a dense set of facial points (typically 50-60). The locations of these points are then used to infer the shape of facial features such as the mouth or the eyebrows and thus to classify the facial expression. A recent example of an AAM based technique is that of Asthana et al., who compare different AAM fitting algorithms and evaluate their performance on the Cohn-Kanade database, reporting a 93% classification accuracy [4].

Another example of a system that uses geometric features to detect emotions is that by Sebe et al. [48]. Piece-wise Bézier volume deformation tracking was used after manually locating a number of facial points. They experimented with a large number of machine learning techniques. Surprisingly, the best result was attained with a simple k-Nearest Neighbour technique that attained a 93% classification rate on the Cohn-Kanade database.

Sung and Kim used AAMs to track facial points in 3D videos [54]. They introduce Stereo Active Appearance Models (STAAM), which improves the fitting and tracking of standard AAMs by using multiple cameras to model the 3D shape and rigid motion parameters. A layered generalised discriminant analysis classifier, which is based on linear discriminant analysis, is then used to combine the 3D shape and registered 2D appearance. Unfortunately, although the approach appears to be promising, it was evaluated for only three expressions, and no results on a benchmark database (such as the Cohn-Kanade or MMI Facial Expression databases) were presented.

Current challenges in automatic discrete emotion recognition that remain to be addressed are dealing with out-of-plane head rotation, spontaneous expressions, and recognising mixtures of emotions. Out of plane rotation and mixtures of emotions are two problems that are likely to coincide when moving to spontaneous, real-world data. While some progress has been made in dealing with occlusions and tracking facial points in imagery of unseen subjects (e.g. [46], [71]) , these two elements remain a challenge as well.

B. Action Unit Detection

Action Unit detection approaches can be divided in a number of ways. Just as for emotion recognition it is possible to divide them into systems that employ appearance based features, geometric features, or both. Another way of dividing them is how they deal with the temporal dynamics of facial expressions: frames in a video can either be treated as being independent of each other (this includes methods that target static images), or a sequence of frames can be treated by a model that explicitly encompasses the expression's temporal dynamics.

A recently proposed class of appearance based features that have been used extensively for face analysis are dense local appearance descriptors. First a particular appearance descriptor is computed for every pixel in the face. To reduce the dimensionality of the problem and the sensitivity to alignment of the face the descriptor responses are then summarised by histograms in pre-defined sub-regions of the face. For AUs, this approach was followed by Jiang et al., using LBP and LPQ [24].

Another successful appearance descriptor is the Gabor Wavelet filter. Littlewort et al. [31] select the best set of Gabor filters using GentleBoost, and train SVMs to classify AU activation. Some measure of AU intensity is provided by evaluating for a test instance the distance to the separating hyperplane provided by the trained SVM. Haar-like features were used in an AdaBoost classifier by Whitehill and Omlin [64].

An example of an appearance-based approach that explicitly models a facial expression's temporal dynamics is that of Koelstra et al. [27]. In their work, they propose a method that detects AUs and their temporal phases onset, apex and offset using Free-Form Deformations and Motion History Images as appearance descriptors and Hidden Markov Models as machine learning technique.

In the geometric feature category, Valstar and Pantic [61] automatically detect 20 facial points and use a facial point tracker based on particle filtering with factorised likelihoods to track this sparse set of facial points. From the tracked points both static and dynamic features are computed, such as the distances between pairs of points or the velocity of a facial point. With this approach they are able to detect both AU activation and the temporal phases onset, apex, and offset.

Simon et al. use both geometric and appearance based features, and include modelling of some of the temporal dynamics of AUs in a proposed method using segment-based SVMs [51]. Facial features are first tracked using a person-specific AAM so that the face can be registered before extracting SIFT features. PCA is applied to reduce the dimensionality of this descriptor. The proposed segment-based SVMs method combines the output of static SVMs for multiple frames and uses structured-output learning to learn the beginning and end

time of each AU. The system was evaluated for 8 AUs on the M3 database (previously called RU-FACS), attaining an average of 83.75% area under the ROC curve.

When facing real-world data, researchers have to face problems such as very large data sizes or low AU frequencies of occurrence. In their work, Zhu et al. focus on the automatic selection of an optimal training set using bi-directional bootstrapping from a dataset with exactly such properties [72]. The features used are identical to those described used by Simon et al. [51]. The proposed dynamic cascades with bi-directional bootstrapping applies gentleBoost to perform feature selection and training instance selection in a unified framework. On the M3 database the system attained an average 79.5% area under the ROC curve for 13 AUs. For an overview of more recent work by researchers at CMU see [29].

Current challenges in AU detection include handling of out-of-plane head rotations and occlusion, two conditions that occur frequently in real-world data of spontaneous expressions. Because AUs are more localised in the face than expressions of discrete emotions, the problem of occlusion is much bigger for AUs than for emotions. Likewise, out-of-plane head rotations can cause self-occlusions of parts of the face that display some AUs, making the problems caused by out-of-plane head-poses harder than it is for emotions. Another issue of moving to data of spontaneous expressions is that the co-occurrences between AUs becomes much harder to model, compared to the limited number of co-occurrence patterns in databases of posed expressions such as the Cohn-Kanade database.

Besides AU detection, the detection of an AU's temporal phase transitions (onset, apex, and offset), as well as its intensity are partially unsolved problems. Being able to predict these variables would allow researchers to detect more complex, higher level behaviour such as deception, cognitive states like (dis)agreement and puzzlement, or psychological states like pain [11], [13]

III. THE GEMEP-FERA DATASET

To be suitable to base a challenge on, a dataset needs to satisfy two criteria. Firstly, it must have the correct labelling, which in our case means frame-by-frame AU labels and event-coding of discrete emotions. Secondly, the database cannot be publicly available at the time of the challenge. The GEMEP database [6] is one of the few databases that meets both conditions, and was therefore chosen for this challenge.

The GEMEP corpus consists of over 7000 audiovisual emotion portrayals, representing 18 emotions portrayed by 10 actors who were trained by a professional director. The actors were instructed to utter 2 pseudo-linguistic phoneme sequences or a sustained vowel 'aaa'. Figure 1 shows an example of one of the male actors displaying an expression associated with the emotion *anger*. A study based on 1260 portrayals showed that portrayed expressions of the GEMEP are recognised by lay judges with an accuracy level that, for all emotions, largely exceeds chance level, and that inter-rater reliability for category judgements and perceived believability and intensity of the portrayal is very satisfactory [6]. At the time of organising the challenge, the data had not been made publicly available yet, making it a suitable dataset to base a fair challenge on. A detailed description of the GEMEP corpus can be found in [6].

TABLE I
ACTION UNITS INCLUDED IN THE AU DETECTION SUB-CHALLENGE. TEST SET S DENOTES SEEN SUBJECTS, WHILE TEST SET U DENOTES UNSEEN SUBJECTS. NUMBER OF VIDEOS: N_{total} = 158; $N_{training}$ = 87; N_{test} = 71

AU	Description	Train	Test S	Test U	Total
1	Inner brow raiser	48	9	28	85
2	Outer brow raiser	48	12	21	81
4	Brow lowerer	34	10	26	70
6	Cheek raiser	37	8	27	72
7	Lid tightener	43	14	30	87
10	Upper lip raiser	48	13	21	82
12	Lip corner puller	56	16	33	105
15	Lip corner depressor	30	6	11	47
17	Chin raiser	49	14	31	94
18	Lip pucker	28	12	20	60
25	Lips part	67	22	37	126
26	Jaw drop	46	12	23	81

The GEMEP-FERA dataset is a fraction of the GEMEP corpus that has been put together to meet the criteria for a challenge on facial Action Units and emotion recognition. By no means does the GEMEP-FERA dataset constitute the entire GEMEP corpus. In selecting videos from the GEMEP corpus to include in the GEMEP-FERA dataset, the main criterium was the availability of a sufficient number of examples per unit of detection for training and testing. It was important that the examples selected for the training set were different than the examples selected for the test set.

A. Partitioning

For the AU detection sub-challenge, we used a subset of the GEMEP corpus annotated with the Facial Action Coding System [18]. The twelve most commonly observed AUs in the GEMEP corpus were selected (see Table I). To be able to objectively measure the performance of the competing facial expression recognition systems, we split the dataset into a training set and a test set. A total of 158 portrayals (87 for training and 71 for testing) were selected for the AU sub-challenge. All portrayals are recordings of actors speaking one of the 2 pseudo-linguistic phoneme sequences. Consequently, AU detection is to be performed during speech. The training set included 7 actors (3 men) and the test set included 6 actors (3 men), half of which were not present in the training set. Even though some actors were present in both training and test set, the actual portrayals made by these actors were different in both sets.

For the emotion sub-challenge, portrayals of five emotional states were retained: anger, fear, joy, sadness, and relief. Four of these five categories are part of what Ekman called basic emotions [16] as they are believed to be expressed universally by specific patterns of facial expression. The fifth emotion, relief, was added to provide a balance between positive and negative emotions but also to add an emotion that is not typically included in previous studies on automatic emotion recognition. Emotion recognition systems are usually modelled on the basic emotions, hence adding "relief" made the task more challenging.

A total of 289 portrayals were selected for the emotion sub-challenge (155 for training and 134 for testing). Approximately 17% of these were recordings of actors uttering the sustained vowel 'aaa' while the remaining portrayals were recordings of actors speaking one of the 2 pseudo-linguistic phoneme sequences. The training set included 7 actors (3 men) with 3 to 5 instances of each emotion per actor. The test set for the emotion sub-challenge included 6 actors (3 men), half of which were not present in the training set. Each actor contributed 3 to 10 instances per emotion in the test set.

The actors who were not present in the training sets were the same for both sub-challenges. Details about the training and test sets can be found in table I (AU sub-challenge) and table II (Emotion sub-challenge). The tables distinguish between videos depicting seen and unseen subjects of the test set. Videos of subjects that are also present in the training set belong to the seen test set, the others to the unseen test set.

B. Availability

The training set was made available through a website[1] employing user-level access control. Upon registering for the challenge, participants were requested to sign an End User License Agreement (EULA), which states, among other things, that the data can only be used for the challenge, and that it cannot be used by commercial parties. When a signed EULA was received by the FERA 2011 organisers, the account of that particular participant was activated. The participant could then download two zip files: one containing the training data for the AU detection sub-challenge and the other containing the training data for the emotion detection sub-challenge.

The test data was distributed through the same website. However, it was only made available 7 working days before the submission deadline. This was done to ensure that the results submitted are fair, by not allowing the participants enough time to manually reconstruct the labels of the test data.

To continue to provide a facial expression recognition benchmark, the GEMEP-FERA 2011 dataset will remain available online. The procedure for obtaining benchmark scores will be identical to that for the challenge, as described in section IV. The only difference will be that the test partition is made available as well (but still without labels, of course).

IV. Challenge Protocol

The challenge is divided into two sub-challenges. The goal of the AU detection sub-challenge is to identify in every frame of a video whether an AU was present or not (i.e. it is a multiple-label binary classification problem at frame level). The goal of the emotion recognition sub-challenge is to recognise which emotion was depicted in that video, out of five possible choices (i.e. it is a single label multi-class problem at event level).

The challenge protocol is divided into five stages. First, interested parties registered for the challenge and signed the EULA to gain access to the training data. Then they trained their systems. In the third stage, the participants downloaded the test partition and generated the predictions for the sub-challenges they were interested in. They then sent their results to the FERA 2011 organisers who calculate their scores. In the case of the FERA 2011 challenge, the participants then submitted a paper describing their approach and reporting their scores to the FERA 2011 workshop. Researchers who intend to follow this benchmark protocol after the FERA 2011 challenge are assumed to submit a paper to another relevant outlet.

Because of concerns regarding the ease with which the emotion labels can be guessed from the video data, the organisers introduced a secondary test for the emotion sub-challenge held the day before the FERA 2011 workshop. The secondary test set contained 50 previously unreleased GEMEP videos displaying one of the five discrete emotions used in the challenge. Participants had the choice to either send their end-to-end programs to the organisers, who then run the secondary test for them, or they could choose to perform the test on their own hardware on-site the day before the workshop. The scores for this secondary test set were not to influence the participant ranking in the emotion detection sub-challenge, but they were announced during the FERA 2011 workshop, on the FERA 2011 web-site, and in this manuscript. All participants but one performed this secondary test.

The training data is organised as two zip files, one for each sub-challenge. When unpacked, the zip-files contain a directory structure in which every folder contains a single video and a single text-file with the corresponding labels. For AUs, the label file is n_f rows by 50 columns, where n_f indicates the number of frames in that video. Each column corresponds to the label for the AU with the same number, e.g. the second column contains the labels for AU2. Zeros indicate the absence of an AU, and a one indicates the presence (activation), of an AU for the corresponding frame. Columns corresponding to non-existing AUs (e.g. AU3) are all zero. During speech (coded as AD50), there is NO coding for AU25 or AU26. Because the annotation of AD50 is made available together with the other AU labels, participants are able to exclude sections of the videos containing speech from their training sets for these two AUs. Likewise, for the computation of the scores, any detections of AU25 and AU26 during speech is discarded. For emotions, the label files contain a single word indicating what emotion was displayed in the corresponding video.

Participants were encouraged to use other facial expression databases annotated in terms of FACS AUs to train their proposed AU detection systems. Examples of such databases, which are publicly available, are the MMI Facial Expression database [60] and the Cohn-Kanade database [26]. Because of the nature of the emotion categories used in this challenge (i.e. the categories are not limited to standard six-basic-emotions categories and the displays are not short-lived posed prototypical facial expressions of emotions but professionally acted audiovisual displays of emotions), the participants were not encouraged to use other training data for the emotion recognition sub-challenge. To assess how well systems perform before the test partition was made available, participants were encouraged to perform a cross-validation evaluation on the training data.

[1] http://gemep-db.sspnet.eu

TABLE II
EMOTIONS INCLUDED IN THE EMOTION DETECTION SUB-CHALLENGE. TEST SET S DENOTES SEEN SUBJECTS, WHILE TEST SET U DENOTES UNSEEN SUBJECTS. NUMBER OF VIDEOS: $N_{total} = 289$; $N_{training} = 155$; $N_{test} = 134$

Emotion	Definition	Train	Test S	Test U	Total
Anger	Extreme displeasure caused by someone's stupid or hostile action	32	14	13	59
Fear	Being faced with an imminent danger that threatens our survival or physical well-being	31	10	15	56
Joy	Feeling transported by a fabulous thing that occurred unexpectedly	30	20	11	61
Relief	Feeling reassured at the end or resolution of an uncomfortable, unpleasant, or even dangerous situation	31	18	8	57
Sadness	Feeling discouraged by the irrevocable loss of a person, place, or thing	31	18	7	56

The test partition was made available one week before the FERA 2011 paper submission deadline. In the test data, there were no labels associated with the test videos. Participants predicted the labels by means of their trained systems and send them to the FERA 2011 organisers by email, who then computed the correctness of the predictions (the scores). To allow the participants to identify and correct major faults in the programmes, they were allowed two submissions of predictions.

The scores are computed in terms of F1-measure for AU detection and in terms of classification rate for emotion detection. For the AU-detection sub-challenge, we first obtain the F1-score for each AU independently, and then compute the average over all 12 AUs. Similarly, for the emotion-recognition sub-challenge the classification rate is first obtained per emotion, and then the average over all 5 emotions is computed. The F1-measure for AUs is computed based on a per-frame detection (i.e. an AU activation prediction has to be specified for every frame, for every AU). The classification rate for emotion categories is computed based on a per-video prediction (event-based detection).

V. BASELINE SYSTEM

The FERA 2011 challenge was the first event where the GEMEP data was used for automatic facial expression recognition, which means that there was no existing work that participants could compare their methods to, and thus there was no means available to participants to check whether their obtained results were reasonable. To overcome this problem, the FERA 2011 organisers provided results of a baseline system for both sub-challenges. The baseline approach used static local-appearance-based features and statistical machine learning techniques. The baseline system was designed as to make it easy to reproduce the baseline results.

The publicly available OpenCV[2] implementation of the Viola & Jones face detector [63] was used to determine the rough location of the face. The height and width of the face-box output by the Viola & Jones face detector is rather unstable, varying by approx. 5% std. even for videos in which the face hardly moves. Also, the face detector does not provide any information about the head pose. To facilitate the appearance descriptor to correlate better with the shown expression instead of with variability in head pose and face detector output, we first perform face registration based on the location of the eyes. To detect the eyes, we use the OpenCV implementation of a Haar-cascade object detector, trained for either a left- or a right eye. After the left eye location p_l and right eye location p_r are determined, the image is rotated so that the angle α, defined as the angle between the line connecting the eyes and the horizontal axis of the image, is 0 degrees. The image is then scaled to make the distance between p_r and p_l 100 pixels, and the face box is cropped to be 200 by 200 pixels, with p_r at position $\{p_r^x, p_r^y\} = \{80, 60\}$. The local appearance descriptors are subsequently extracted from such registered images.

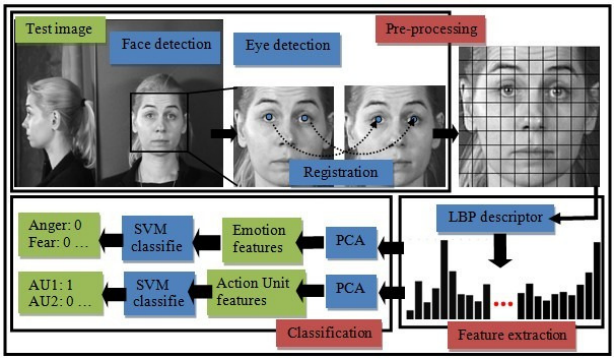

Fig. 2. Overview of the FERA 2011 baseline system for detection of 12 Action Units and 5 emotions.

As dense local appearance descriptors we chose to use uniform LBPs [37]. They have been used extensively for face analysis in recent years, e.g. for face recognition [1], emotion detection [50], or detection of facial muscle actions (FACS Action Units) [24]. As classifier we employ standard Support Vector Machines (SVMs) with a radial basis function kernel. We reduced the dimensionality of our facial expression representation using Principal Component Analysis (PCA). Fig 2 gives an overview of the baseline system.

A. Feature extraction

LBPs were first introduced by Ojala et al. in [36], and proved to be a powerful means of texture description. For every pixel the LBP operator creates a label by thresholding a 3×3 neighbourhood of that pixel with the value of the pixel itself. Considering the 8-bit result as a binary number, a 256-bin histogram is generated over the LBP response in a region of interest. This histogram is used as the texture descriptor.

Ojala et al. [38] later extended the basic LBP to allow a variable number of neighbours to be chosen at any radius from the central pixel. They also greatly reduced the dimensionality

[2]http://opencv.willowgarage.com/wiki/, DOA 02-06-2011

of the LBP operator, by introducing the notion of a uniform LBP. A local binary pattern is called uniform if it contains at most two bitwise transitions from 0 to 1 or vice versa when the binary string is considered circular [38]. The LBP operator for the general case based on a circularly symmetric neighbour set of P members on a circle of radius R, is denoted by $LBP^u_{P,R}$. Superscript u reflects the use of uniform patterns. Parameter P controls the quantisation of the angular space and R determines the spatial resolution of the operator. Bilinear interpolation is used to allow any radius and number of pixels in the neighbourhoods.

Using only rotation invariant uniform LBPs greatly reduces the length of the feature vector. The number of possible patterns for a neighbourhood of P pixels is 2^P for the basic LBP while being only $P + 2$ for LBP^u. An early stage experiment was conducted to find the optimal parameters for this application, resulting in $P = 8$, and $R = 1$. Hence, we adopted $LBP^u_{8,1}$ descriptor in our baseline system.

The occurrence of the rotation invariant uniform patterns over a region is recorded by a histogram. After applying the LBP operator to an image, a histogram of the LBP-labelled region of interest in the image can be defined as:

$$H_i = \sum_{x,y} I(f(x,y) = i), i = 0, ..., n-1. \quad (1)$$

where n is the number of possible labels produced by the LBP operator and

$$I(A) = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

An LBP histogram computed over the whole face image represents only the frequency of the patterns without any indication about their locations. To take the spatial information into account as well, we divide the registered face region into smaller sub-regions and extract separate LBP histograms from each of them (as shown in figure 2). The LBP features extracted from each sub-region are subsequently concatenated into a single, spatially-enhanced feature histogram. This was used as a feature vector representing the shown facial expression. A grid size of 10×10 was used in the experiments, as this was empirically found to be the best division of the face region into sub-regions for AU detection. Figure 3 shows the results of this test for three upper face AUs. The data used for this study was taken from the MMI Facial Expression Database [60].

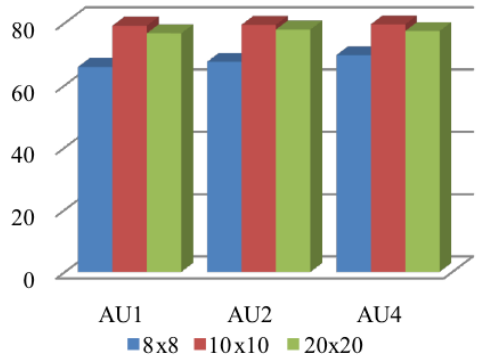

Fig. 3. Results for AU detection using different grid sizes used to extract LBP features.

B. Training AU detectors

Binary Support Vector Machine (SVM) classifiers were trained for each AU independently. Because of the independence assumption, we only need to look at the appearance changes caused by a single AU at a time. This meant we could divide the set A of AUs into two groups: upper-face AUs $G_u = \{AU1, AU2, AU4, AU6, AU7\}$, which only cause appearance changes in the upper half of the face, and lower-face AUs $G_l = \{AU10, AU12, AU15, AU17, AU18, AU25, AU26\}$ that only affect the lower face. The training set for an AU consisted of frames where that particular AU was present (positive examples), frames in which any of the other AUs from the same group was active (either negative or positive examples, depending on whether the target AU was present as well), plus frames displaying an expressionless face (negative examples).

To select the frames to be used to train a classifier, we adopted the method proposed in [24], which selects from every video in the training set only frames displaying distinct AU combinations. Because this method relies on the availability of labelled AU temporal phases (onset, apex, and offset of AUs), which are not available for the GEMEP-FERA 2011 dataset, we modified this method slightly. First we segmented each video into periods with distinct AU combinations. These segments usually have a duration of multiple frames. We then pick the middle frame of each block. If a video has multiple blocks with the same AU combination, we took the training frame from the first occurrence of this combination. Note that when we select frames for $A_i \in G_j$ with $j \in \{u, l\}$, we only look at AU combinations of G_j.

A different set of features was used for the upper-face AUs and the lower-face AUs. To wit, for each AU $a \in G_u$ we concatenate the histograms of the top-five rows of the 10 x 10 LBP grid, while for each AU $a \in G_l$ we concatenate the histograms of the bottom five rows. To reduce the dimensionality of the feature set we apply PCA, selecting m_a eigenvectors for the subsequent analysis such that their cumulative energy is 95%. Features are then normalised to lie in the range $[-1, 1]$.

For the 1-vs-all frame-based AU classification, we employ SVMs with a RBF kernel. Two parameters: the RBF scale parameter σ, and the SVM slack variable ζ are determined by means of a 5-fold cross-validation on the training set. During parameter optimisation, we optimised for the F1-score, not the classification rate, as it is the F1 score that was used as the evaluation measure in the challenge. We also make sure that we split the folds along subject divides, i.e. we make sure that data from the same subject never appears in both the training and evaluation sets. As reported in [24], for AU detection this can lead to a performance increase of up to 9% F1-measure, compared to randomly splitting the data.

TABLE III
F1-MEASURE FOR ACTION UNIT DETECTION RESULTS ON THE TEST SET FOR THE BASELINE METHOD. PERFORMANCE IS SHOWN FOR THE PERSON INDEPENDENT (PI), PERSON SPECIFIC (PS), AND OVERALL PARTITIONS. THE LAST COLUMN SHOWS RESULTS OF A NAIVE CLASSIFIER ON THE OVERALL TEST SET.

AU	PI	PS	Overall	Naive
1	0.634	0.362	0.567	0.506
2	0.675	0.400	0.589	0.477
4	0.133	0.298	0.192	0.567
6	0.536	0.255	0.463	0.626
7	0.493	0.481	0.489	0.619
10	0.445	0.526	0.479	0.495
12	0.769	0.688	0.742	0.739
15	0.082	0.199	0.133	0.182
17	0.378	0.349	0.369	0.388
18	0.126	0.240	0.176	0.223
25	0.796	0.809	0.802	0.825
26	0.371	0.474	0.415	0.495
Avg.	0.453	0.423	0.451	0.512

C. AU Detection Results

Table III shows the results of the AU detection baseline system measured in terms of F1-measure. The table shows results for three different partitions of the test data. The first is the partition of the test data for which the test subjects are not present in the training data (Person Independent partition). This partition shows the ability of AU detection systems to generalise to unseen subjects. The second partition of the test data consists of recordings of subjects that appear both in the training and in the test set. Participants would thus be able to train subject specific detectors for this partition. The third column shows the results for the entire, unpartitioned test set, which we call the overall performance. It is this performance on the whole test set that is used to rank participants in the AU-detection sub-challenge.

To assess the quality of the baseline method, we have also computed the results for a naive AU detector. The best strategy for a naive classifier in the situation of sparse positive examples (i.e. sparse AU activation), is to score all frames as active. The results are computed over the full (overall) test set, and are shown in the last column of Table III. As can be seen, the baseline method does not outperform a naive approach in all cases. One reason for this may be the fact that while we choose parameters for optimal F1 measure, the training algorithm of SVMs inherently uses the classification rate as the value for which it optimises.

D. Training Emotion detectors

The emotion detection sub-challenge calls for the detection of five discrete emotion classes. Each video has a single emotion label $e \in E$, where $E = \{Anger, Fear, Joy, Relief, Sadness\}$. Since the videos do not display any apparent neutral frames at the beginning or the end of the video, we defined that every frame of a video shares the same label. The appearance of the facial expression however does change over the course of a video. We therefore use every frame of a video to train and test our algorithm on.

For the emotion classifiers all 10×10 sub-regions of the LBP grid described in section V-A is used. To reduce the dimensionality of the feature set we apply PCA, selecting m_e eigenvectors for the subsequent analysis such that their cumulative energy is 90%.

The emotion recognition problem is a 5-class forced choice problem. We trained a single one-versus-all SVM classifier with an RBF kernel for each emotion. Two parameters: the RBF scale parameter σ, and the SVM slack variable ζ are determined by means of a 5-fold cross-validation on the training set. Each of the five resulting classifiers gives a prediction $y_j^e \in \{-1, 1\}$ for the presence of emotion e for frame j in a test video. The label Y of a video of n frames is the emotion class e into which the largest number of frames have been classified:

$$Y = \arg\max_e \sum_{j=1}^{n} y_j^e \tag{3}$$

TABLE IV
2AFC SCORE FOR ACTION UNIT DETECTION ON THE TEST SET FOR THE BASELINE METHOD. PERFORMANCE IS SHOWN FOR THE PERSON INDEPENDENT (PI), PERSON SPECIFIC (PS), AND OVERALL PARTITIONS.

AU	PI	PS	Overall
1	0.845	0.613	0.790
2	0.818	0.640	0.767
4	0.481	0.607	0.526
6	0.690	0.568	0.657
7	0.572	0.530	0.556
10	0.577	0.627	0.597
12	0.738	0.700	0.724
15	0.555	0.567	0.563
17	0.679	0.661	0.646
18	0.620	0.599	0.610
25	0.544	0.669	0.593
26	0.457	0.555	0.500
Avg.	0.631	0.611	0.628

TABLE V
CLASSIFICATION RATES FOR EMOTION RECOGNITION ON THE TEST SET FOR THE BASELINE METHOD. PERFORMANCE IS SHOWN FOR THE PERSON INDEPENDENT (PI), PERSON SPECIFIC (PS), AND OVERALL PARTITIONS. LAST COLUMN SHOWS OVERALL RANDOM RESULTS.

Action Unit	PI	PS	Overall	Naive
Anger	0.857	0.923	0.889	0.222
Fear	0.067	0.400	0.200	0.160
Joy	0.700	0.727	0.710	0.161
Relief	0.313	0.700	0.462	0.115
Sadness	0.267	0.900	0.520	0.200
Average	0.441	0.730	0.556	0.172

E. Emotion Detection Results

Classification rates attained by the baseline method for the emotion recognition sub-challenge are shown in Table V. Again, to assess the quality of the baseline method, we have compared the baseline method results to a naive emotion detector, which in this case assigns a uniform random label to each video in the test set. As can be seen from Table V, the baseline approach well exceeds the naive emotion detector.

VI. PARTICIPANT SYSTEMS

In total 12 participants contributed to the challenge. We will now describe the systems of participants that were entered in the emotion recognition and/or the AU detection sub-challenge.

A. Action Unit Detection Systems

Senechal et al. [49] propose a system that combines shape and appearance information using multi-kernel SVMs. The shape information is obtained using AAMs and the appearance using Local Gabor Binary Pattern (LGBP) histograms. For the AAM coefficients a Radial Basis Frequency kernel is used, and for the LGBP features a histogram intersection kernel. The SVM output is temporally filtered by taking for every frame the average over a short time window.

Wu et al. [65] evaluated the performance of their Computer Expression Recognition Toolbox (CERT, [31]) for the AU detection problem. CERT uses a Viola&Jones style face and facial feature detection system, which is used to register the face. A bank of Gabor filters is applied to the registered face and AUs are detected using SVMs.

B. Action Unit Detection and Emotion Detection Systems

Baltrusaitis et al. [5] presented a system based on the mind-reading work of El Kaliouby & Robinson [19]. Their real-time system operates on three increasingly longer temporal levels: first AUs, head actions and shoulder actions are detected on a timescale of 5 frames. To detect face, head, and shoulder gestures the action information is fed into Hidden Markov Models that operate on a 15-frame level. Finally, Dynamic Bayesian Networks are used to infer the five discrete emotions, again at the 15-frame temporal level.

Chew et al. [8] argue that, given sufficiently accurate registration, the pixel intensity information of the face is all that is needed to recognise facial actions and applying linear filters such as LBPs to the face image is not necessary. They attain highly accurate registration using Saragih et al.'s version of Constrained Local Models [46]. SVMs are trained on the pixel information after canonical normalising the face area, which removes all non-rigid shape variation with respect to a reference shape.

Gehrig & Ekenel's proposed system uses Discrete Cosine Transform histograms in a manner similar to the baseline system's LBP approach. The histograms derived from 10×10 non-overlapping blocks in a registered face are normalised on a per-block basis and used as input to Kernel Partial Least Squares regression.

C. Emotion Recognition Systems

Dhall et al. [14] use Pyramid Histogram of Oriented Gradients and LPQ appearance features to detect emotions. To avoid using frames with similar appearance, facial feature points are first tracked using a Constrained Local Model tracker. The resulting face shapes are clustered and used to select key frames from which appearance features are extracted. However, face registration is achieved using a face detector rather than using the tracked facial point locations. Finally, emotions are detected using SVMs and Largest Margin Nearest Neighbour classifiers.

Dahmane & Meunier [12] recognise emotions in a system similar to the baseline approach described above. Instead of LBPs, their system uses Histogram of Oriented Gradient features.

Littlewort et al. [30] present a system that is based on CERT. From the CERT outputs of AU and head orientation predictions, dynamic features called Extremes of Displacement, Velocity and Accelaration (EDVA) are computed. The EDVA features are then used as input to a Multinomial Logistic Regression classifier to detect the emotions. The authors also experiment with detecting the emotions anger, fear, joy, and sadness directly using existing CERT models.

Meng et al. [34] start from the dynamic appearance descriptor Motion History Histogram and the static appearance descriptor LBP. The former is extended to also encode local texture, whilst the latter is extended to also encode dynamic appearance. The two new spatio-temporal appearance descriptors are combined using multi-kernel SVMs to distinguish between the five emotions.

Srivastava et al. [53] use Accumulated Motion Images (AMIs, essentially Motion Energy Images) and geometric features extracted from tracked facial points. Two separate one-vs-all multi-class SVMs are trained for the AMI and the geometric features. During testing, a confidence is calculated for both multi-class SVMs by subtracting for each the highest output of a component of a multi-class SVM from the second highest output. The multi-class SVM with the highest confidence is used to decide what emotion was displayed.

Tariq et al. [55] use an ensemble of features consisting of Hierarchic Gaussianisation, SIFT, and Optical Flow to recognise emotions. For the subject-dependent data partition, they learned a specific model for each subject and a face recognition system. This approach proved to be highly successful.

Yang & Bhanu [66] present an approach that uses so-called Emotion Avatar Images. All frames of the input video (face images) are first registered using the SIFT flow algorithm [32], which performs global alignment of the face, yet retains the facial motion caused by facial expression. Such registered frames within one video are then mapped onto a person-independent face model, which is built based on the entire training set. The final result is the Emotion Avatar Image, a single image that represents all the expression-related facial motion present in the input video. From this image LBPs and LPQ features are derived, and used as input to Support Vector Machines which are trained to distinguish between the 5 target emotion classes.

VII. Competition Results

The number of parties who showed interest in participating in the FERA 2011 challenge indicates that the facial expression analysis field is of a moderate size. The challenge data was downloaded by 20 teams, of which 15 participated in the challenge and submitted a paper to the FERA 2011 workshop. Of the 15 papers, 11 papers were accepted for publication, based on a double-blind peer review process. In total, 10 teams participated in the emotion recognition sub-challenge, and five teams took part in the AU detection sub-challenge (three teams participated in both sub-challenges).

Demographic statistics are as follows: Teams were from many countries and often spanned multiple institutes. The participating institutes were dispersed over 9 countries (USA,

TABLE VI
AVERAGE CLASSIFICATION RATES OVER ALL EMOTIONS FOR THE EMOTION RECOGNITION SUB-CHALLENGE AND AVERAGE F1-MEASURE OVER ALL AUS FOR THE AU DETECTION SUB-CHALLENGE. HIGH SCORES ARE PRINTED IN BOLD.

Participant	AU detection			Emotion detection			
	Person-independent	Person-specific	Overall	Person-independent	Person-specific	Overall	Secondary
ANU [14]	N.A.	N.A.	N.A.	0.649	0.838	0.734	0.700
ISIR [49]	**0.633**	**0.576**	**0.620**	N.A.	N.A.	N.A.	N.A.
KIT [21]	0.543	0.473	0.523	0.658	0.944	0.773	0.760
MIT-Cambridge [5]	0.470	0.422	0.461	0.448	0.433	0.440	0.480
Montreal [12]	N.A.	N.A.	N.A.	0.579	0.870	0.700	**0.96**
NUS [53]	N.A.	N.A.	N.A.	0.636	0.730	0.672	0.640
Riverside [66]	N.A.	N.A.	N.A.	**0.752**	0.962	**0.838**	0.860
QUT [8]	0.530	0.460	0.510	0.624	0.554	0.600	0.00
UCLIC [34]	N.A.	N.A.	N.A.	0.609	0.837	0.700	0.740
UCSD1 [30]	N.A.	N.A.	N.A.	0.714	0.837	0.761	0.640
UCSD2 [65]	0.604	0.539	0.583	N.A.	N.A.	N.A.	N.A.
UIUC-UMC [55]	N.A.	N.A.	N.A.	0.655	**1.00**	0.798	0.780
Baseline	0.453	0.423	0.451	0.440	0.730	0.560	N.A.

Australia, Canada, Germany, Singapore, Sweden, UK, Belgium, and France). In total 53 researchers participated in the challenge, with a median of 6 researchers per paper. Five entries were multi-institute endeavours. This indicates that the research community is not entrenched in local enclaves, instead there appears to be a large amount of cooperation and communication between researchers of automatic facial behaviour understanding. With four authors being psychologists, the field can claim a certain degree of interdisciplinary collaboration as well.

A. Emotion Recognition Sub-Challenge

Table VI shows the scores attained in the emotion recognition sub-challenge. As can be seen, 9 out of 10 participating systems outperform the baseline approach on the full test set. The winning team, Yang & Bhanu of the University of California Riverside, attained an overall 83.8% classification result [66].

It is interesting to note the person-specific results obtained by the multi-institute team of the University of Illinois Urbana Champaign/University of Missouri Columbia [55]. The proposed method, which included an automatic face recognition module, attained a perfect emotion recognition score on the subject-dependent test set.

As expected, most participating systems scored higher on the person specific test-set than the person-independent test-set. In general, the performance on the person-specific partition was very high, with 7 out of 10 teams scoring over 80%, and 3 out of 10 teams scoring over 90%. When, in addition, we take into consideration that these results were obtained using a relatively small training set, this may well lead us to conclude that inferring discrete affective states from face videos of known users for whom *a priori* training data is available can be considered to be a solved problem.

The secondary on-site emotion recognition test was introduced to perform a sanity check regarding the reported results. That is, it was used to ensure nobody had either grossly inflated their performance results by guessing the emotion labels of the original test set, or had in fact relied on some form of manual processing of the data. Participants were allowed to apply bug-fixes to their original entry, which in at least one case led to a significant improvement in results [12]. For reasons unknown to the challenge organisers, the team of QUT chose not to perform the secondary test.

TABLE VII
F1 MEASURES PER AU, FOR EVERY PARTICIPANT IN THE AU-DETECTION SUB-CHALLENGE. LAST COLUMN SHOWS AVERAGE OVER ALL PARTICIPANTS, AND HIGH SCORES ARE PRINTED IN BOLD.

AU	ISIR	KIT	MIT-Camb.	QUT	UCSD	**Avg**
1	**0.809**	0.606	0.681	0.780	0.634	0.702
2	**0.731**	0.520	0.635	0.723	0.636	0.649
4	0.582	0.529	0.446	0.433	**0.602**	0.518
6	**0.833**	0.822	0.739	0.658	0.759	0.762
7	**0.702**	0.554	0.323	0.553	0.604	0.547
10	0.475	0.467	0.328	0.468	**0.565**	0.460
12	0.803	0.798	0.658	0.778	**0.832**	0.774
15	**0.245**	0.065	0.114	0.156	0.193	0.155
17	**0.557**	0.518	0.300	0.471	0.499	0.469
18	0.431	0.329	0.127	**0.448**	0.345	0.336
25	**0.850**	0.800	0.815	0.311	0.815	0.718
26	**0.576**	0.515	0.475	0.537	0.515	0.524

B. Action Unit Detection

The results for the AU detection sub-challenge are shown per partition in Table VI, and overall results per AU for each team are shown in table VII. The winner of the AU detection sub-challenge was the team of Senechal et al., from the Institut des Systemes Intelligents et de Robotique, Paris [49]. Their method attained an F1 measure of 63.3%, averaged over all 12 AUs. This is well above the baseline's 45.3%, but still very far off from a perfect AU recognition.

Looking at individual AUs, we can see that AU1, AU2, AU6, and AU12 are consistently detected well by all participants, while AU4, AU5, AU10, AU17, AU18, and AU26 were consistently detected with low accuracy. AU25, parting of the lips, is detected with high accuracy by all participants except QUT [8]. The authors noted in [8] that this may have been due to an inability to deal with speech effectively. AU7, narrowing of the eye aperture caused by contraction of the orbicularis occuli muscle (pars palpebralis), was only detected with high accuracy by Senechal et al.

Contrary to what would normally be expected, Table VI shows that performance on the person specific partition was consistently worse than on the person independent part. Unfor-

tunately, given the relatively small size of the test partition, this is probably simply because the videos selected for the person specific part may have been that much more challenging than those included in the person independent part.

VIII. Discussion

The first facial expression recognition and analysis challenge has been a great community effort and a resounding success, both in terms of the attained results as well as the level of participation. We hope to have established a new benchmark for facial expression analysis, which should allow researchers in the field to objectively gauge their performance. To keep this benchmark available in the future, the FERA 2011 organisers are keeping the GEMEP-FERA database available through their online repository, and they will continue to provide the service of calculating researchers' test scores.

One of the opportunities that arise from hosting a challenge like FERA 2011, is that one can learn what are the current trends in a field. For instance, five teams participated in the AU detection sub-challenge, compared to 10 teams for the emotion recognition sub-challenge. This indicates that despite the criticism on the practical use of discrete emotion classification and the theory behind it, it remains the most popular approach for computer scientists.

With respect to machine learning techniques we noticed a strong trend to use Support Vector Machines (SVMs). Out of 12 teams, 10 teams used SVMs. Perhaps more significantly, three teams used multiple kernel SVMs [44], including the AU detection winner [49]. Surprisingly, only 1 team modelled time [5], and, although such techniques have proven very popular in recent literature, it is also the only team that has used probabilistic graphical models.

In terms of features, the following was observed: Four teams encoded appearance dynamics, and there were four teams that combined appearance and geometric features, including the AU detection winners. Although modelling of depth would improve the ability to deal with out of plane head rotations, only a single team infers 3D from 2D images. This appears to be successful, as the team that employed it also won the AU detection challenge. Unfortunately, from their work [49] it is not possible to assess exactly how big the influence of this 3D inference was. Geometric features on their own were neither popular nor successful: there was only a single team that relied solely on Geometric features, and they were ranked very low in the emotion recognition sub-challenge.

Considering the short interval between the call for participation and the submission deadline (less than three months, including Christmas), participation levels were high. The organisers also noted a high enthusiasm among the teams, with an attitude that would be best described as collaborative competitive: Researchers were both interested in winning as well as in learning what really works for this problem. We therefore conclude that a follow-up of this challenge would find broad interest in the automatic human behaviour analysis community.

A follow-up challenge using a larger dataset should be organised in order to address the two following issues: First, the scores for AU detection on the person-specific partition were worse than on the person-independent partition. It shows that the two partitions can not be said to represent the same underlying distribution (i.e. all possible ways of expressing the five emotions by all subjects in the dataset). Essentially, the two partitions differ too much in their level of difficulty, and this is caused by not having enough data to sample the two partitions from. Secondly, it is important for a fair challenge to minimise the possibility for participants to cheat and this can be implemented by using a large dataset that is difficult to manually annotate in the time provided for training the algorithms.

Another issue that arose during the challenge is the choice of performance measure. It is well known that in a heavily unbalanced data, such as that of the AU detection sub-challenge, the classification rate is not a suitable measure. A naive classifier based on the prior probability of the classes will give an over-optimistic representation of the problem and is very likely to outperform systems that try to detect both classes with equal priority. In the literature, people therefore often use two measures, the F1-measure and the area under the receiver-operator characteristic curve (AUC). The F1-measure is a single scalar value that represents the harmonic mean of precision and recall (i.e. it equally favours precision and recall), and can be computed with binary predictions. The AUC can only be computed if real ordinal predictions are provided by the classifier.

To avoid restricting participants' choice of classifiers to those that provide a real valued output, we thus opted to use the F1-measure. Unfortunately, the baseline results showed that even this measure may be misleading. The naive approach of attaining the highest F1-measure in the case of the AU-detection test set would be to assign the positive (i.e. AU is active) label to all test instances. As Table III shows, this actually results in a higher F1-score than that attained by the baseline system. This is because the F1-measure is wholly determined by the number of true positives, false positives, and false negatives. The number of true negative examples thus plays no role, while they are credited in the AUC. The AUC may thus be a better performance measure to be used in a future challenge, at the (probably minimal) cost of restricting participants' choice of classifiers.

As pointed out in section VII-A, inferring discrete emotions from video of known subjects may well be considered solved. Any progress in this area will probably be only marginal and perhaps best left to industry. Recently, there has been more interest in the automatic recognition of dimensional affect [22], [23], [47]. A future challenge may well focus on this.

The detection of AUs, however, is still far from solved, and this should definitely remain a focus in future events. One thing to address in the future is the number of AUs included in the test set. For FERA 2011, there were only 12 AUs that needed to be detected. In the future, it would be desirable to have a dataset that will allow a competition on detection of all 31 AUs, plus possibly a number of FACS Action Descriptors [18]. Besides addressing the detection of the activation of AUs, it would be a good thing to move towards the detection of the intensities and temporal segments of AUs, as it is these

characteristics that prove to be crucial in many higher-level behaviour understanding problems [11], [13], [17].

ACKNOWLEDGMENTS

This work has been supported by the European Communitys 7th Framework Programme [FP7/20072013] under grant agreement no. 231287 (SSPNet). The work of Valstar was also supported by EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching. The work by Maja Pantic is also funded in part by the European Research Council under the ERC Starting Grant agreement no. ERC-2007- StG-203143 (MAHNOB) . Work on the GEMEP-FERA dataset was supported by the Swiss National Science Foundation (FNRS 101411-100367), by the National Center of Competence in Research (NCCR) Affective Sciences financed by the Swiss National Science Foundation (51NF40-104897) and hosted by the University of Geneva, and by EU Networks of Excellence HUMAINE (IST 507422).

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikäinene. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2037–2041, 2006.
[2] N. Ambady and R. Rosenthal. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta–analysis. *Psychological Bulletin*, 11(2):256–274, 1992.
[3] A. B. Ashraf, S. Lucey, J. F. Cohn, T. Chen, Z. Ambadar, K. M. Prkachin, and P. E. Solomon. The painful face - pain expression recognition using active appearance models. *Image and Vision Computing*, 27(12):1788 – 1796, 2009.
[4] A. Asthana, J. Saragih, M. Wagner, and R. Goecke. Evaulating aam fitting methods for facial expression recognition. In *Proc. Int'l. Conf. Affective Computing and Intelligent Interaction*, 2009.
[5] T. Baltrusaitis, D. McDuff, N. Banda, M. Mahmoud, R. El Kaliouby, P. Robinson, and R. Picard. Real-time inference of mental states from facial expressions and upper body gestures. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, 2011.
[6] T. Bänziger and K. R. Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. In K. R. Scherer, T. Bänziger, and E. B. Roesch, editors, *Blueprint for Affective Computing: A Sourcebook*, Series in affective science, chapter 6.1, pages 271–294. Oxford University Press, Oxford, 2010.
[7] M. Bartlett, G. Littlewort-Ford, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behaviour. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 223–230, 2006.
[8] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, 2011.
[9] J. Cohn and P. Ekman. Measuring facial action by manual coding, facial emg, and automatic facial image analysis. In R. R. . K. S. J. A. Harrigan, editor, *Handbook of nonverbal behavior research methods in the affective sciences*, pages 9–64. Oxford University Press, 2005. New York.
[10] J. F. Cohn. Foundations of human computing: Facial expression and emotion. *Proc. ACM Int'l Conf. Multimodal Interfaces*, 1:610–616, 2006.
[11] D. Cunningham, M. Kleiner, C. Wallraven, and H. Bülthoff. The components of conversational facial expressions. *Proc. ACM Int'l Symposium on Applied Perception in Graphics and Visualization*, pages 143–149, 2004.
[12] M. Dahmane and J. Meunier. Emotion recognition using dynamic grid-based hog features. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, pages 884–888, 2011.
[13] A. C. de C. Williams. Facial expression of pain: An evolutionary account. *Behavioral and Brain Sciences*, 25(4):439–488, 2002.
[14] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using phog and lpq features. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, pages 878–883, 2011.
[15] P. Ekman. *Face of man: Universal expression in a new guinea village*. Garland, 1982. New York.
[16] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3/4):169–200, 1992.
[17] P. Ekman and W. Friesen. The repertoire of nonverbal behavioral categories – origins, usage and coding. *Semiotica*, 1:49–98, 1969.
[18] P. Ekman, W. V. Friesen, and J. C. Hager. *FACS Manual*. A Human Face, May 2002.
[19] R. El Kaliouby and P. Robinson. Generalization of a vision-based computational model of mind-reading. *Affective Computing and Intelligent Interaction*, pages 582–589, 2005.
[20] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003.
[21] T. Gehrig and H. Ekenel. Facial action unit detection using kernel partial least squares. *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2092–2099, 2011.
[22] H. Gunes and M. Pantic. Automatic, dimensional and continuous emotion recognition. *International Journal of Synthetic Emotions*, 1(1):68–99, 2010.
[23] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. In *Proc. of IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 827–834, 2011.
[24] B. Jiang, M. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, pages 314–321, 2011.
[25] T. Kanade. *Picture Processing System by Computer Complex and Recognition of Human Faces*. PhD thesis, Tokyo University, 1973.
[26] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. *Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000. Proceedings*, pages 46–53, 2000.
[27] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture based approach to recognition of facial actions and their temporal models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32:1940–1954, 2010.
[28] I. Kotsia and I. Pitas. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16(1):172–187, 2007.
[29] F. D. la Torre and J. F. Cohn. Facial expression analysis. In V. K. T.B. Moeslund, A. Hilton and L. Sigal, editors, *Guide to Visual Analysis of Humans: Looking at People*. Springer, 2011.
[30] G. Littlewort, J. Whitehill, T. Wu, N. Butko, P. Ruvolo, J. Movellan, and M. Bartlett. The motion in emotion-a cert based approach to the fera emotion challenge. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, pages 897–902, 2011.
[31] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett. The computer expression recognition toolbox (cert). In *Proc. IEEE Int'l. Conf. Automatic Face and Gesture Recognition*, pages 298–305, 2011.
[32] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(5):978–994, 2010.
[33] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 200–205, 1998.
[34] H. Meng, B. Romera-Paredes, and N. Berthouze. Emotion recognition by two view svm_2k classifier on dynamic facial expression features. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, pages 854–859, 2011.
[35] S. Moore and R. Bowden. Local binary patterns for multi-view facial expression recognition. *Computer Vision and Image Understanding*, 115(4):541 – 558, 2011.
[36] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on featured distribution. *Pattern Recognition*, 29(1):51–59, 1996.
[37] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.
[38] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution grey-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.
[39] M. Pantic and M. Bartlett. Machine analysis of facial expressions. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 377–416. I-Tech Education and Publishing, 2007. Vienna, Austria.
[40] M. Pantic, A. Nijholt, A. Pentland, and T. Huang. Human-centred intelligent human-computer interaction (hci2): How far are we from attaining it? *Int'l Journal of Autonomous and Adaptive Communications*

Systems, 1(2):168–187, 2008.
[41] M. Pantic and L. Rothkrantz. Automatic analysis of facial expressions: the state of the art. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1424 – 1445, 2000.
[42] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proc. IEEE*, 91(9):1370–1390, 2003.
[43] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. Int'l Conf. Multimedia & Expo*, pages 317–321, 2005.
[44] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 2008.
[45] A. Samal and P. A. Iyengar. Automatic recognition and analysis of human faces and facial expressions: a survey. *Pattern Recogn.*, 25:65–77, January 1992.
[46] J. Saragih, S. Lucey, and J. Cohn. Face alignment through subspace constrained mean-shifts. In *Proc. Int'l. Conf. Computer Vision*, pages 1034–1041, 2009.
[47] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. Avec 2011 the first international audio/visual emotion challenge. *Springer Lecture Notes on Computer Sciences*, pages 415–424, 2011.
[48] N. Sebe, M. Lew, I. Cohen, T. Gevers, and T. Huang. Authentic facial expression analysis. *Image and Vision Computing*, 25(12):1856–1863, 2007.
[49] T. Senechal, V. Rapp, L. Prevost, H. Salam, R. Seguier, and K. Bailly. Combining lgbp histograms with aam coefficients in the multi-kernel svm framework to detect facial action units. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, pages 860–865, 2011.
[50] C. Shan, S. Gong, and P. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing Journal*, 27(6):803–816, 2009.
[51] T. Simon, M. H. Nguyen, F. De La Torre, and J. Cohn. Action unit detection with segment-based svms. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2737 –2744, june 2010.
[52] T. Simon, M. H. Nguyen, F. D. L. Torre, and J. F. Cohn. Action unit detection with segment-based svms. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:2737–2744, 2010.
[53] R. Srivastava, S. Roy, S. Yan, and T. Sim. Accumulated motion images for facial expression recognition in videos. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, pages 903–908, 2011.
[54] J. Sung and D. Kim. Real-time facial expression recognition using staam and layered gda classifier. *Image and Vision Computing Journal*, 27(9):1313–1325, 2009.
[55] U. Tariq, K.-H. Lin, Z. Li, X. Zhou, Z. Wang, V. Le, T. Huang, X. Lv, and T. Han. Emotion recognition from an ensemble of features. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, pages 872–877, 2011.
[56] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2), 2001.
[57] Y. L. Tian, T. Kanade, and J. F. Cohn. *Handbook of Face Recognition*. Springer, 2005. New York.
[58] Y. Tong, J. Chen, and Q. Ji. A unified probabilistic framework for spontaneous facial action modeling and understanding. *Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, Dec 2010.
[59] M. Valstar and M. Pantic. Combined support vector machines and hidden markov models for modeling facial action temporal dynamics. In *ICCV-HCI'07*, pages 118–127, 2007.
[60] M. Valstar and M. Pantic. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, pages 65–70, 2010.
[61] M. Valstar and M. Pantic. Fully automatic recognition of the temporal phases of facial actions. *Systems, Man, and Cybernetics, Part B*, 42(1):28–43, 2012.
[62] M. F. Valstar and M. Pantic. Biologically vs. logic inspired encoding of facial actions and emotions in video. *IEEE Int'l. Conf. on Multimedia and Expo*, pages 325–328, 2006.
[63] P. Viola and M. Jones. Robust real-time object detection. *Int J Comput Vis*, 57(2):137–154, 2002.
[64] J. Whitehill and C. Omlin. Haar features for facs au recognition. In *Proc. Int'l. Conf. Automatic Face and Gesture Recognition*, 2006.
[65] T. Wu, N. Butko, P. Ruvolo, J. Whitehill, M. Bartlett, and J. Movellan. Action unit recognition transfer across datasets. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, pages 889–896, 2011.
[66] S. Yang and B. Bhanu. Facial expression recognition using emotion avatar image. In *Proc. IEEE Int'l Conf. Automatic Face and Gesture Analysis*, pages 866–871, 2011.
[67] S. Zafeiriou and M. Petrou. Nonlinear non-negative component analysis algorithms. *Image Processing, IEEE Transactions on*, 19(4):1050 –1066, april 2010.
[68] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.
[69] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary pattern with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2(6), 2007.
[70] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary pattern with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2(6):915–928, 2007.
[71] R. Zhi, M. Flierl, Q. Ruan, and W. Kleijn. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 41(1):38 –52, feb. 2011.
[72] Y. Zhu, F. De la Torre, J. Cohn, and Y. Zhang. Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *Affective Computing, IEEE Transactions on*, PP(99):1, 2011.

Michel F. Valstar (M'09) is a Lecturer in the Mixed Reality Lab at the University of Nottingham. He received his masters degree in Electrical Engineering at Delft University of Technology in 2005 and his PhD in computer science with the intelligent Behaviour Understanding Group (iBUG) at Imperial College London in 2008. Currently he is working in the fields of computer vision and pattern recognition, where his main interest is in automatic recognition of human behaviour. In 2011 he was the main organiser of the first facial expression recognition challenge, FERA 2011. In 2007 he won the BCS British Machine Intelligence Prize for part of his PhD work. He has published technical papers at authoritative conferences including CVPR, ICCV and SMC-B and his work has received popular press coverage in New Scientist and on BBC Radio. He is also a reviewer for many journals in the field, including Transactions on Pattern Analysis and Machine Intelligence, Transactions on Affective Computing, Systems, Man and Cybernetics-B and the Image and Vision Computing journal.

Marc Mehu (2007, Ph.D. in psychology, University of Liverpool) is postdoctoral researcher at the Swiss Centre for Affective Sciences, University of Geneva. Marc Mehus research interests lie in the evolution of social behaviour, including social signals, and its role in the regulation of social relationships. More precisely, he is interested in the adaptive value of facial and vocal expressions. His Ph.D. research looked at the social function of smiling and laughter from an evolutionary perspective. His research applies a variety of methods including ethological observations in natural settings as well as experiments on person perception and social interactions. Present and future research of his involve the integration of cognitive, emotional, and social factors in communication research using recent developments of evolutionary theory, with a focus on the integration of dominance and altruism in social bonding.

Maja Pantic (M'98–SM'06–F'12) is Professor in Affective and Behavioural Computing at Imperial College London, Department of Computing, UK, and at the University of Twente, Department of Computer Science, the Netherlands. She received various awards for her work on automatic analysis of human behaviour including the European Research Council Starting Grant Fellowship 2008 and the Roger Needham Award 2011. She currently serves as the Editor in Chief of Image and Vision Computing Journal and as an Associate Editor for both the IEEE Transactions on Systems, Man, and Cybernetics Part B and the IEEE Transactions on Pattern Analysis and Machine Intelligence. She is an IEEE Fellow.

Klaus Scherer (1970, Ph.D. in psychology, Harvard University) is Director of the Swiss Center for Affective Sciences and founder of the GERG at Department of Psychology at the University of Geneva. He has conducted many research programs, financed by granting agencies and private foundations in the USA, Germany, and Switzerland, directed at the study of cognitive evaluations of emotion-eliciting events and on facial and vocal emotion expression. He reported his work in numerous publications in the form of monographs, contributed chapters, and papers in international journals. He edited several collected volumes and handbooks and coedits the Affective Science Series for Oxford University Press. He was the founding co-editor of the journal Emotion. He is a member of several international scientific societies and a fellow of the American Psychological Association and the Acoustical Society of America. He has been elected member of the Academia Europea and honorary foreign member of the American Academy of Arts and Sciences. Klaus Scherer also pursues activities directed at the practical application of scientific research findings in industry, business, and public administration. He directs several long-term applied research programs in the area of organisational behaviour and human resources, on psychological assessment, and on speech technology.