The First Facial Expression Recognition and Analysis Challenge

Valstar, Michel F.; Jiang, Bihan; Mehu, Marc; Pantic, Maja; Scherer, Klaus R.

The First Facial Expression Recognition and Analysis Challenge

Michel F. Valstar, Bihan Jiang, Marc Mehu, Maja Pantic, and Klaus Scherer

Abstract— **Automatic Facial Expression Recognition and Analysis, in particular FACS Action Unit (AU) detection and discrete emotion detection, has been an active topic in computer science for over two decades. Standardisation and comparability has come some way; for instance, there exist a number of commonly used facial expression databases. However, lack of a common evaluation protocol and lack of sufficient details to reproduce the reported individual results make it difficult to compare systems to each other. This in turn hinders the progress of the field. A periodical challenge in Facial Expression Recognition and Analysis 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. In this paper we present the first challenge in automatic recognition of facial expressions to be held during the IEEE conference on Face and Gesture Recognition 2011, in Santa Barbara, California. Two sub-challenges are defined: one on AU detection and another on discrete emotion detection. It outlines the evaluation protocol, the data used, and the results of a baseline method for the two sub-challenges.**

I. INTRODUCTION

Computers and powerful electronic gadgets surround us in ever increasing numbers, and the computing aspect is increasingly hidden behind 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. An important functionality of these interfaces will be the capacity to perceive and understand intentions and emotions as communicated by facial expressions.

Facial Expression Recognition and Analysis (FERA), in particular FACS AU detection [4] and discrete emotion detection, has been an active topic in computer science for some time now, and many promising approaches have been reported [11], [19]. Arguably the first manuscript on automatic facial expression recognition being published in 1974 [12]. The first survey of the field was published in 1992 [13] and has been folluwed up by several others since [19], [11]. The question is, do the approaches proposed to date actually deliver what they promise? To help answer that question, we are of the opinion that it is time to take stock of how far the field has progressed in an objective manner.

Valstar, Jiang, and Pantic are with the Department of Computing, Imperial College London, UK michel.valstar@imperial.ac.uk, m.pantic@imperial.ac.uk

Mehu and Scherer are with the Swiss Center for Affective Sciences, University of Geneva, Switzerland marc.mehu@unige.ch, klaus.scherer@unige.ch

Maja Pantic is also with the Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands

Researchers often do report on the accuracy of their proposed approaches using a number of popular facial expression databases (e.g. The Cohn-Kanade database [6], the MMI-Facial Expression Database [18], and the JAFFE database [8]). 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 and Analysis 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.

This paper describes the first such challenge, organised under the name of FERA2001, which will be held in conjunction with the 9th IEEE International Conference on Automatic Face and Gesture Recognition. The challenge will allow a fair comparison between systems vying for the title of 'state of the art'. To do so, it uses a partition of the GEMEP corpus [1], developed by the Geneva Emotion Research Group (GERG) at the University of Geneva led by Klaus Scherer.

The challenge is divided in two sub-challenges that reflect two popular approaches to facial expression recognition: an AU detection sub-challenge and an emotion detection sub-challenge. The AU detection sub-challenge calls for researchers to attain the highest possible F1-measure for 12 frequently occurring AU (see Table I). The emotion detection sub-challenge calls for systems to attain the highest possible classification rate for the detection of five discrete emotions: anger, fear, joy, relief, and sadness. The set of emotions is not a subset of the basic emotions postulated by Ekman [3], though it contains some. Table II lists a definition of each emotion.

The majority of existing automatic facial expression recognition literature can be divided based on the types of features they use in three ways: those that use appearance-based features (e.g. [5], [2]), those that use geometric feature-based approaches (e.g. [17], [7]), and those that use both (e.g. [15]). Both appearance- and geometric feature-based approaches have their own advantages and disadvantages, and we expect that systems that use both will ultimately result in the highest accuracy.

Another way existing systems can be classified is in the way they make use of time. Some systems only use the temporal dynamics information encoded directly in their features (e.g. [5], [20]), others only employ machine learning techniques to model time (e.g. [16], [14]), while again

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

others employ both (e.g. [17]). Currently it is unknown what approach is most successful.

A full survey of the field is out of scope for this paper, and we encourage the interested reader to review a number of excellent surveys for a more detailed description of the current state of the art [19], [11].

The remainder of this paper is structured as follows: in section II we will list the requirements for a suitable challenge data set, and describe the data set chosen for the first FERA challenge. In section III we describe the challenge protocol for both the AU detection and emotion detection sub-challenges. Section IV then describes the baseline method and the baseline results for the two sub-challenges. We conclude the paper with a summary of this paper in section V.

II. THE GEMEP-FERA2011 DATASET

For the challenge, we will use part of the GEMEP database [1]. The GEMEP corpus [1] consists of over 7000 audiovisual emotion portrayals, representing 18 emotions portrayed by 10 actors who were trained by a professional director. As the basis of their expression, the actors were instructed to utter 2 pseudo-linguistic phoneme sequences or a sustained vowel 'aaa'. Of the total number of recordings, 1260 portrayals were selected and included in a rating study to evaluate inter-judge reliability and recognition accuracy. Baenziger and Scherer [1] showed that portrayed expressions of the GEMEP are recognized 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. The data has not been made publicly available yet, and is thus ideal for a fair challenge. A detailed description of the GEMEP corpus can be found in [1].

A. Partitioning

A subset of the GEMEP corpus was annotated in terms of facial expression using the FACS and that subset was used in the AU detection sub-challenge. To be able to objectively measure the performance of the participants' entries, we split the dataset into a training set and a test set. A total of 158 portrayals (87 for training and 71 for testing) was selected for the AU sub-challenge. All portrayals represented actors speaking one of the 2 pseudo-linguistic phoneme sequences so 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.

For the emotion sub-challenge, a total of 289 portrayals were selected (155 for training and 134 for testing). Approximately 17% of the portrayals in the emotion sub-challenge represented the actors uttering the sustained vowel 'aaa' while the remaining portrayals represented the actors speaking one of the 2 pseudo-linguistic phoneme sequences. The training set included 7 (3 men) actors 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 between 3 and 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 with seen and unseen subjects of the test set. Videos displaying subjects that are also present in the training set belong to the seen test set, the others in the unseen test set.

B. Availability

The training set was made available through a website[1] employing user-level access control to all participants directly after the challenge's call for participation was made. 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 private institutions. When a signed EULA is received by the FERA2011 organisers, the account of that particular participant was activated. The participant could then download two zip files: one containing all training data for the AU detection sub-challenge and the other containing all 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. Again, one zip file contained all test videos for the AU detection sub-challenge and the other the videos for the emotion detection sub-challenge.

To continue providing a facial expression recognition benchmark after the challenge is over, the GEMEP-FERA2011 dataset will remain available through its website. The procedure for obtaining benchmark scores will be identical to that for the challenge, as described in section III.

[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 only difference will be that the test partition is always available (but still without labels, of course).

III. THE FERA2011 CHALLENGE

The challenge consists of 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 register for the challenge and sign the EULA to gain access to the training data. Secondly they train their systems. In the third stage participants download the test partition and generate the results of their systems. Fourthly they send their results to the FERA2011 organisers who calculate their scores, and finally the participants submit a paper describing their approach and reporting their scores.

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, or 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 we make the annotation of AD50 available together with the other AU labels, participants are able to exclude sections of speech from their training sets for these two AUs. Likewise, for the computation of the scores, any detections of AU25 and AU26 during speech will be discarded. For emotions, the label files contain a single word indicating what emotion was displayed in the corresponding video.

In training their systems, participants are encouraged to use other databases of FACS AU coding to train their AU detection systems. Examples of this are the MMI Facial Expression database [18], as well as the Cohn-Kanade database [6]. Because of the nature of the emotion categories in this challenge, it is not possible to use other training data for the emotion recognition sub-challenge. To assess how well their systems perform before the test partition is available, participants are encouraged to perform a cross-validation evaluation on the training data.

The test partition is made available one week before the challenge's paper submission deadline. Again, it consists of one zip file for the AU detection sub-challenge and one zip file for the emotion recognition sub-challenge, each containing a similar directory structure. Only this time, there are no labels associated with the test videos. Participants create predictions with their trained systems, which should be formatted in exactly the same way as the training labels, and should be send to the FERA2011 organisers by email, who then respond with the computed scores. To allow participants to identify major faults in their programmes, they are allowed two submissions of their results.

The scores are computed in terms of F1-measure for AU detection and classification rate for emotion detection. To obtain the overall score 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 categories 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 prediction has to be specified for every frame, for every AU, as being either present or absent). The classification rate for emotions is computed based on a per-video prediction (event-based detection). It is calculated per emotion as the fraction of the number of videos correctly classified as that emotion divided by the total number of videos of that emotion in the test set.

IV. BASELINE EVALUATION

This is the first time that the GEMEP data is used for automatic facial expression recognition, which means that there are no other works that participants can compare their methods with, and no means to check whether the results obtained are reasonable. Therefore, in this work we provide baseline recognition results using appearance based features, which will allow participants to make this comparison. Standard Viola & Jones face detection, followed by similar Haar-cascade eye detection is applied to each face. The eye locations are used to register for scale and in-plane head rotation. The features used are Local Binary Patterns appearance descriptors (LBP, [9]). As classifier we employ standard Support Vector Machines (SVMs) with a radial

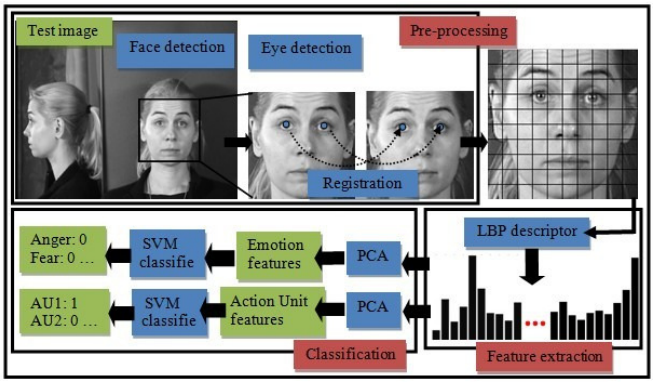

Fig. 1. Overview of the FERA2011 baseline system for detection of 12 Action Units and 5 emotions.

basis function kernel. We reduce the dimensionality of our facial expression representation using Principal Component Analysis (PCA). Fig 1 gives an overview of the baseline system's approach.

A. Feature extraction

Local Binary Patterns (LBP) were first introduced by Ojala et al. in [9], and proved to be a powerful means of texture description. By thresholding a 3×3 neighbourhood of each pixel with the central value, the operator labels the pixels. Considering the 8-bit result as a binary number, a 256-bin histogram of the LBP labels computed over a region is used as a texture descriptor.

Ojala et al. [10] 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 of the operator, by introducing the notion of a uniform Local Binary Pattern. 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 [10]. The 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 Local Binary Patterns greatly reduces the length of feature vector. The number of possible patterns for a neighbourhood of P pixels is 2^P for the basic LBP while only $P + 2$ for LBP^u. An early stage experiment is conducted to find the optimal parameters for this application, resulting in $P = 8$, and $R = 1$. Hence, we adopt $LBP^u_{8,1}$ descriptor in this paper.

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 labelled image $f(x, y)$ can be defined as:

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

where n is the possible labels produced by 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 occurrences of the patterns without any indication about their locations. To also consider shape information of faces, face images were divided into small regions to extract LBP histograms (as shown in figure 1). The LBP features extracted from each sub-region are concatenated into a single, spatially enhanced feature histogram. The final histogram is used as a feature vector to represent face image. A grid size of 10×10 is used in the experiments.

B. Training AU detectors

A separate binary Support Vector Machine (SVM) classifier was trained for each AU independently. We divided the set A of AUs into two groups G: upper-face AUs $G_u = \{AU1, AU2, AU4, AU6, AU7\}$, and lower-face AUs $G_l = \{AU10, AU12, AU15, AU17, AU18, AU25, AU26\}$. The training set for a certain AU consisted of selected frames that included this AU (positive examples), selected frames in which any of the other AUs from the same group was active, plus selected frames displaying a neutral expression.

To select which frames could be used to train each classifier we adopt the method used in [5], that selects from every video in the training set only frames with distinct AU combinations. Because this method relies on the availability of labelled AU temporal phases, which are not included in the GEMEP-FERA2011 dataset, we had to modify this method slightly: First we segment each video into temporal blocks with distinct AU combinations. These blocks usually last multiple frames. We then pick the middle frame of each block with a distinct AU combination. If a video has multiple blocks with the same AU combination, we take 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 upper-face AUs and lower-face AUs. To wit, for each AU $a \in G_u$ we concatenate the histograms of the top-five rows of LBP blocks, while for each AU $a \in G_l$ we concatenate the histograms of the bottom five rows. To reduce the dimensionality of the descriptors we apply PCA, retaining 95% of the energy. Features were then normalised to lie in the range $[-1, 1]$.

We employed an RBF kernel, which means we need to set two parameters: the RBF scale parameter σ, and the SVM slack variable ζ. Parameter optimisation is achieved using a 5-fold cross-validation on the training set. During parameter optimisation we optimise for the F1-score, not the classification rate, as it is the F1 score that will be used as the challenge score. We also make sure that we split the folds along subject divides, i.e. we make sure that 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	Random
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

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

from the same subject never appears in both the training and evaluation sets. As reported in [5], for AU detection this can lead to a performance increase of up to 9% F1-measure, compared to randomly splitting the data.

C. AU Detection Results

Table III shows the results of the AU detection measured in F1-measure. The table shows results for three different partitions of the test data: the first is the partition of the test data of 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 videos of subjects that are also part of the training set. Participants would thus be able to train subject specific detectors for this partition, and thus obtain a higher score. The third partition is simply the entire (overall) test set. It is the performance on the overall partition that will be used to rank participants.

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 overall partition, and are shown in the last column of Table III. It shows that the baseline method does not outperform a naive approach for all AUs. This may be due to the fact that while we choose parameters for optimal F1 measure, the training of an SVM inherently uses classification rate as the value to optimise.

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 end of the video, we defined that every frame of a video shares the same label. The appearance of the facial expressions however do change within the video, and we cannot pinpoint emblematic frames. We therefore use every frame of a video as train and test data.

For the emotion classifiers all 10 rows are used. To reduce the dimensionality of the two feature sets PCA was applied. The number of principal components retained was chosen to encode 90% of the variance in the original data.

The emotion detection sub-challenge is a 5-class forced choice problem. We train a single one-versus-all SVM classifier for each emotion. The five resulting classifiers each give a prediction $y_{e,j}$ about the presence of emotion e for frame j in a test video. To decide the label Y of a video of n frames, we find the emotion with the largest number of frames classified:

$$Y = \arg\max_{e} \sum_{j=1}^{n} y_{e,j} \tag{3}$$

E. Emotion Detection Results

Classification rates by the baseline method for the emotion detection sub-challenge are shown in Table V. In addition, we provide confusion matrices for the person independent (Table VI), person specific (Table VII), and overall partitions (Table VIII). Rows are predicted results, columns the ground truth. As with the AU detection sub-challenge the performance on the overall partition is used to rank participants in the emotion detection sub-challenge.

Again, to assess the quality of the baseline method, we have compared our results to a naive emotion detector, which in this case assigns a uniform random label to each video in the test set. The results show that this time the baseline approach well exceeds the random method.

V. CONCLUSIONS

This paper describes the first challenge on Facial Expression Recognition and Analysis, held in conjunction with the 9th IEEE International Conference on Face and Gesture Recognition, March 2011, Santa Barbara, California. The

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	Random
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

TABLE VI

CONFUSION MATRIX FOR PERSON INDEPENDENT EMOTION RECOGNITION.

pred truth	Anger	Fear	Joy	Relief	Sadness
Anger	12	11	5	0	8
Fear	0	1	0	0	0
Joy	0	3	14	8	1
Relief	1	0	0	5	2
Sadness	1	0	1	3	4

challenge consists of a FACS Action Unit detection sub-challenge and an emotion recognition sub-challenge. This work outlines the data used for the challenge as well as the challenge protocol. In addition, we've provided a description of a baseline system that uses Local Binary Pattern features, Principal Component Analysis, and Support Vector Machines to either detect the activation of AUs per frame, or recognise emotions in an entire video. The results of the baseline indicate that the data has the right level of difficulty: it is by no means impossible to detect the desired events, but the task is challenging.

VI. ACKNOWLEDGMENTS

This work has been funded in part by the European Community's 7th Framework Programme [FP7/20072013] under the grant agreement no. 231287 (SSPNet). In addition, the work of Maja Pantic and Michel Valstar has been funded in part by EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching.

TABLE VII

CONFUSION MATRIX FOR PERSON SPECIFIC EMOTION RECOGNITION.

pred truth	Anger	Fear	Joy	Relief	Sadness
Anger	12	5	1	1	0
Fear	0	4	0	0	0
Joy	0	1	8	1	1
Relief	0	0	0	7	0
Sadness	1	0	2	1	9

TABLE VIII

CONFUSION MATRIX FOR EMOTION RECOGNITION ON THE OVERALL TEST SET.

pred truth	Anger	Fear	Joy	Relief	Sadness
Anger	24	16	6	1	8
Fear	0	5	0	0	0
Joy	0	4	22	9	2
Relief	1	0	0	12	2
Sadness	2	0	3	4	13

REFERENCES

[1] 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.

[2] 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.

[3] P. Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3/4):169–200, 1992.

[4] P. Ekman, W. V. Friesen, and J. C. Hager. *FACS Manual*. A Human Face, May 2002.

[5] 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*, 2011. In print.

[6] 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.

[7] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F. de la Torre, and J. Cohn. Aam derived face representations for robust facial action recognition. In *Automatic Face and Gesture Recognition, 2006. FGR 2006. 7th International Conference on*, pages 155 –160, 2006.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] F. I. Parke. *A parametric model for human faces*. PhD thesis, The University of Utah, 1974.

[13] 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.

[14] 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.

[15] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2), 2001.

[16] 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.

[17] 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.

[18] 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*, page 65, 2010.

[19] 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.

[20] 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.