FERA 2015 - Second Facial Expression Recognition and Analysis Challenge

Michel F. Valstar[1], Timur Almaev[1], Jeffrey M. Girard[2], Gary McKeown[3], Marc Mehu[4], Lijun Yin[5], Maja Pantic[6,7] and Jeffrey F. Cohn[2,8]
[1] School of Computer Science, University of Nottingham, UK
[2] Department of Psychology, University of Pittsburgh, Pittsburgh, USA
[3] School of Psychology, Queen's University Belfast, Belfast, UK
[4] Department of Psychology, Webster University Private Vienna, Vienna, AT
[5] Department of Computer Science, Binghamton University, Binghamton, USA
[6] Department of Computing, Imperial College London, London, UK
[7] Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands
[8] Robotics Institute, Carnegie Mellon University, Pittsburgh, USA

***Abstract*— Despite efforts towards evaluation standards in facial expression analysis (e.g. FERA 2011), there is a need for up-to-date standardised evaluation procedures, focusing in particular on current challenges in the field. One of the challenges that is actively being addressed is the automatic estimation of expression intensities. To continue to provide a standardisation platform and to help the field progress beyond its current limitations, the FG 2015 Facial Expression Recognition and Analysis challenge (FERA 2015) will challenge participants to estimate FACS Action Unit (AU) intensity as well as AU occurrence on a common benchmark dataset with reliable manual annotations. Evaluation will be done using a clear and well-defined protocol. In this paper we present the second such challenge in automatic recognition of facial expressions, to be held in conjunction with the 11 IEEE conference on Face and Gesture Recognition, May 2015, in Ljubljana, Slovenia. Three sub-challenges are defined: the detection of AU occurrence, the estimation of AU intensity for pre-segmented data, and fully automatic AU intensity estimation. In this work we outline the evaluation protocol, the data used, and the results of a baseline method for the three sub-challenges.**

I. INTRODUCTION

Facial expression analysis is a rapidly growing field of research, due to the constantly increasing interest in applications for automatic human behaviour analysis and novel technologies for human-machine communication and multimedia retrieval. Most Facial Expression Recognition and Analysis systems proposed in the literature focus on detecting the occurrence of expressions, often either basic emotions or the Facial Action Coding System (FACS Action Units, AUs, [5]). In reality, expressions can vary greatly in intensity, and intensity is often a strong cue for the interpretation of the meaning of expressions.

Indeed McKeown et al. [12] have argued that level of intensity is the key dimension in facial expressions that distinguishes whether they are delivered for socio-communicative functionsat low levels of intensityor that they become hard-to-fake signals indicating that the expression is associated with a genuine felt emotionat high levels of intensity. If this is the case then intensity may be one of the most important features in assessing a user's psychological state from facial expressions. However, very little annotated data is available for the evaluation of AU intensity estimation approaches. In addition, despite efforts towards evaluations standards (e.g. FERA 2011 [18]), there is still a need for improved standardised evaluation datasets and procedures for the more challenging application scenarios.

In particular, there is a need for evaluation standards on large sets of data, recorded in realistic scenarios, with detailed annotations including intensity and frame-by-frame occurrence. Without such improved benchmarking procedures, facial expression recognition will continue to suffer from low comparability between proposed approaches. This is in stark contrast with more established problems in human analysis from video such as face detection and face recognition.

In these respects, the FG 2015 Facial Expression Recognition and Analysis challenge (FERA2015) shall help raise the bar for expression recognition by challenging participants to estimate AU intensity, and it will continue to bridge the gap between excellent research on facial expression recognition and low comparability of results. We do this by means of three selected tasks: the detection of FACS Action Unit occurrence (Occurrence Detection Sub-Challenge), the estimation of AU intensity when AU occurrence is known *a priori* (Pre-Segmented Intensity Estimation Sub-Challenge), and fully automatic AU intensity estimation for the most realistic scenario in which AU occurrence is not known beforehand (Fully Automatic Intensity Estimation Sub-Challenge).

II. RELATED WORK

Facial expression recognition in general and Action Unit Detection in particular has been studied extensively in the past decade. As a result, it is impossible to provide a comprehensive review of the field here. Instead we provide

This material is based upon the work supported by the National Science Foundation under grants CNS-1205664, CNS-1205195, IIS-1051103, and IIS-1051169. The work by Michel Valstar and Timur Almaev has been partially funded by NIHR MindTech Healthcare Technology Co-operative. In addition the work of Michel Valstar is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 645378.

978-1-4799-6026-2/15/$31.00 ©2015 IEEE

an overview of the relevant works only, focussing on methods that target AU occurrence detection and intensity estimation. For a general overview of the field of expression recognition we refer the reader to two excellent recent surveys [6], [22].

A. AU occurrence detection

Common binary classifiers applied to this problem include Artificial Neural Networks (ANN), Boosting techniques, and Support Vector Machines (SVM). ANNs were the most popular method in earlier works (e.g. [17], [2]). Boosting algorithms, such as AdaBoost and GentleBoost, have been a common choice for AU recognition (e.g. [7], [21]). Boosting algorithms are simple and quick to train. They have fewer parameters than SVM or ANN, and can be less prone to overfitting. They implicitly perform feature selection, which is desirable for handling high-dimensional data and speeding up inference, and can handle multiclass classification. SVMs are currently the most popular choice (e.g. [3], [19], [11]). SVMs provide good performance, can be non-linear, parameter optimisation is relatively easy, as efficient implementations are readily available, and a choice of kernel functions provides extreme flexibility of design.

B. AU intensity estimation

The goal in AU intensity estimation is to assign a per-frame label with possible integer value from 0 to 5 for each AU. This problem can be approached using either a classification or a regression learning method.

Classification-based methods: Some approaches use the confidence of a (binary) frame-based AU activation classifier to estimate AU intensity. The rationale is that the lower the intensity is, the harder the classification will be. For example, Bartlett et al. used the distance of the test sample to the SVM separating hyperplane [1], while Hamm et al. used the confidence of the decision given by AdaBoost [7].

It is however more natural to treat the problem as 6-class classification. For example, Mahoor et al. employed six one-vs.-all binary SVM classifiers [11]. Alternatively, a single multi-class classifier (e.g. ANN or a Boosting variant) could be used. The extremely large class overlap means however that such approaches are unlikely to be optimal.

Regression-based methods: AU intensity estimation is nowadays often posed as a regression problem. Regression methods penalise incorrect labelling proportionally to the difference between ground truth and prediction. Such ordinal consideration of the labels is absent in classification methods. The large overlap between classes also implies an underlying continuous nature of intensity that regression techniques are better equipped to model. Examples include Support Vector Regression ([8] and [15]). Kaltwang et al. instead used Relevance Vector Regression to obtain a probabilistic prediction [9].

III. DATA

The training, development and test data for the FERA 2015 challenge are drawn from two databases: the BP4D-Spontaneous database [24] and the SEMAINE database [13]. The training and development sets are drawn from the SEMAINE database and the original BP4D-Spontaenous database, while the test set is drawn from part of the SEMAINE database and an extended version of BP4D; both of these database subsets have not previously been publicly released. FACS is a system for human observer coding of facial expressions, decomposing expressions into atomic units of anatomically-based Action Units that correspond to specific facial muscles or muscle groups. In this challenge, FACS was used to code the occurrence and intensity of participants facial expressions.

TABLE I
OVERVIEW OF AUS INCLUDED IN THE THREE SUB-CHALLENGES

	Occurrence detection	**Intensity Estimation**
BP4D	AU1, AU2, AU4, AU6, AU7, AU10, AU12, AU14, AU15, AU17, AU23	AU6, AU10, AU12, AU14, AU17
SEMAINE	AU2, AU12, AU17, AU25, AU28, AU45	-

The challenge will focus on 14 AUs that occurred frequently in the BP4D and SEMAINE datasets. The Occurrence Detection sub-challenge requires participants to detect 11 AUs from the BP4D database and 6 from the SEMAINE database (see Table I). AUs were selected based on their frequency of occurrence and sufficiently high inter-rater reliability scores. AU intensity estimation, both for the Pre-Segmented Intensity Estimation Sub-Challenge and the Fully Automatic Intensity Estimation Sub-Challenge will be done on the BP4D data only, on a subset of 5 AUs (see Table I).

Data is split into train and test partitions. The train partition is publicly available for researchers to train and develop their AU analysis systems, and to allow participants to uniformly report performance (i.e. using cross-validation). The test partition is held back by the organisers. Participants submit their trained systems and the FERA 2015 organisers apply their systems on this held-back data to create a fair comparison.

A. BP4D Database

Both the train and test partitions of the BP4D database consist of video data of young adults responding to emotion-elicitation tasks. The datasets are described in detail below. Here we note differences between them that are most relevant to the challenge. The training data was collected first and is publicly available. The testing data is newer, part of an ongoing data collection that includes thermal imaging and peripheral physiology, and is not publicly available. The number of participants in the two partitions is 41 and 20, respectively. Some differences exist in the threshold for coding AU occurrence and intensity, and changes occurred in the mix of AU coders of the two partitions. Coders were highly trained for both, and reliability was tested throughout coding to ensure consistency.

The train partition of BP4D is selected from BP4D-Original, and the test partition from BP4D-Expanded. Below we will refer to these as BP4D-Train and BP4D-Test.

BP4D-Train The BP4D-Train dataset includes digital video of 41 participants (56.1% female, 49.1% white, ages 18-29). These individuals were recruited from the departments of psychology and computer science and from the school of engineering at Binghamton University. All participants gave informed consent to the procedures and permissible uses of their data. Participants sat approximately 51 inches in front of a Di3D dynamic face capturing system during a series of eight emotion elicitation tasks.

To elicit target emotional expressions and conversational behaviour, we used approaches adapted from other investigators plus techniques that proved promising in pilot testing. Each task was administered by an experimenter who was a professional actor/director of performing arts. The procedures were designed to elicit a range of emotions and facial expressions that include happiness/amusement, sadness, surprise/startle, embarrassment, fear/nervous, physical pain, anger/upset, and disgust.

BP4D-Test The BP4D-Test dataset includes digital video of 20 participants with similar demographics as BP4D-original. These individuals underwent similar recruitment, emotion-elicitation, and video recording procedures as those in the BP4D-Train dataset. The main difference between these datasets is that the extended dataset also collected physiological data and captured thermal images of participants. However, thermal and physiological data are not included in the FERA Challenge.

For BP4D there are 21 subjects in the training, 20 subjects in the development and 20 in the test partition. For each subject there are 8 sessions: 168 sessions in the training, 160 sessions in the development and 159 sessions in the test partition (for one subject only 7 sessions are available). There are 75,586 images in total in the training partition, 71,261 images in development and 75,726 in testing (222,573 in total).

B. SEMAINE database

The challenge uses the SEMAINE corpus [13] as the second source of data. This database was recorded to study natural social signals that occur in conversations between people and virtual humans, and to collect data for the training of the next generation of such agents. It is is freely available for scientific research purposes from http://semaine-db.eu. The scenario used in the recordings is called the Sensitive Artificial Listener (SAL) technique [4]. It involves a user interacting with emotionally stereotyped "characters" whose responses are stock phrases keyed to the user's emotional state rather than the content of what (s)he says.

For the recordings, the participants are asked to talk in turn to four emotionally stereotyped characters. These characters are Prudence, who is even-tempered and sensible; Poppy, who is happy and outgoing; Spike, who is angry and confrontational; and Obadiah, who is sad and depressive.

Video was recorded at 49.979 frames per second at a spatial resolution of 780 x 580 pixels and 8 bits per sample, while audio was recorded at 48 kHz with 24 bits per sample. To accommodate research in audio-visual fusion, the audio and video signals were synchronised with an accuracy of 25 μs using the system developed by Lichtenauer et al. [10].

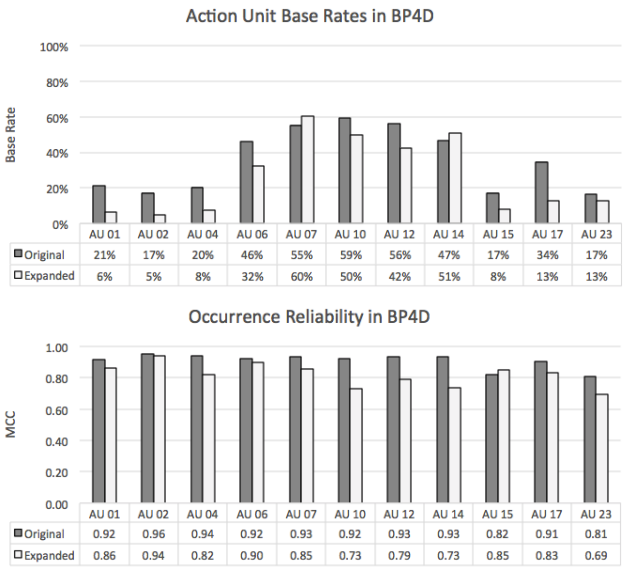

Fig. 1. AU base rates (top) and inter-coder reliability (bottom) for the BP4D data (both original and extended).

In this challenge the 24 recordings of the publicly available Solid-SAL part of the database are used as train partition. The test partition is derived from parts of the SEMAINE database that haven't been made public to date. SEMAINE sessions contain one subject each. The training partition consists of 16 sessions, the development partition has 15 sessions, and the test partition has 12 sessions. There are a total of 48,000 images in total in the training partition, 45,000 the development and 37,695 in testing (130,695 frames in total).

C. Action Unit Annotation

Action Units were annotated by a team of experts. Both databases were annotated frame-by-frame for the occurrence (i.e. activation) of AUs. In addition, BP4D was annotated frame-by-frame for the intensity of a subset of AUs.

Occurrence Annotation For BP4D-Train, coders annotated onsets when AUs reached the A-level of intensity and offsets when they dropped below it. Segments of the most facially-expressive 20 seconds of each task were selected for coding. Across all participants, AU base occurrence rates, defined as the fraction of coded frames in which an AU occurred, averaged 35.4%, and ranged from 17% for to 59%. To assess inter-coder reliability, approximately 11% of the data was independently coded by two highly trained and certified coders. Inter-coder reliability, as quantified by the Matthews Correlation Coefficient (MCC; [14]), averaged 0.91. MCC for individual AU ranged from 0.81 for AU 23 to 0.96 for AU 2. These results suggest very strong inter-coder reliability for occurrence. Fig. 1 depicts the frequency and inter-coder reliability of AU occurrence annotation on BP4D.

For BP4D-Extended, coders annotated onsets when AUs

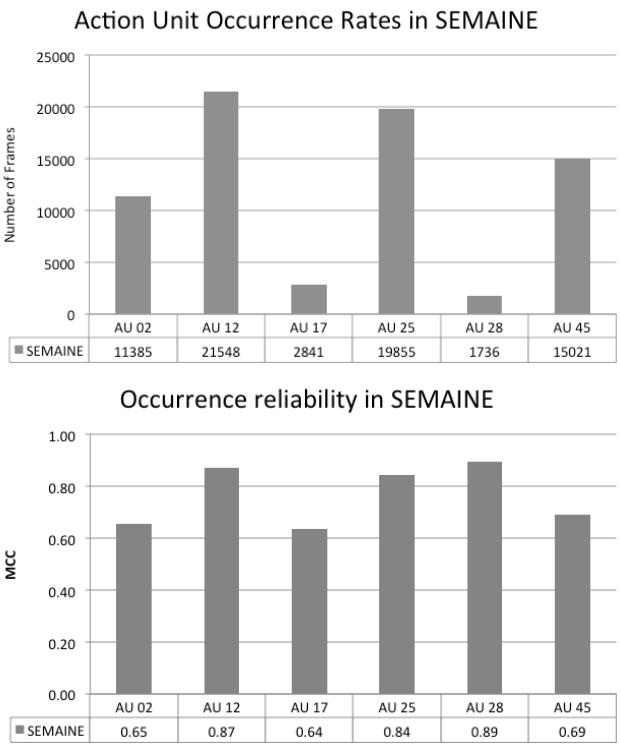

Fig. 2. AU occurrence frequency (top) and inter-coder reliability (bottom) for SEMAINE .

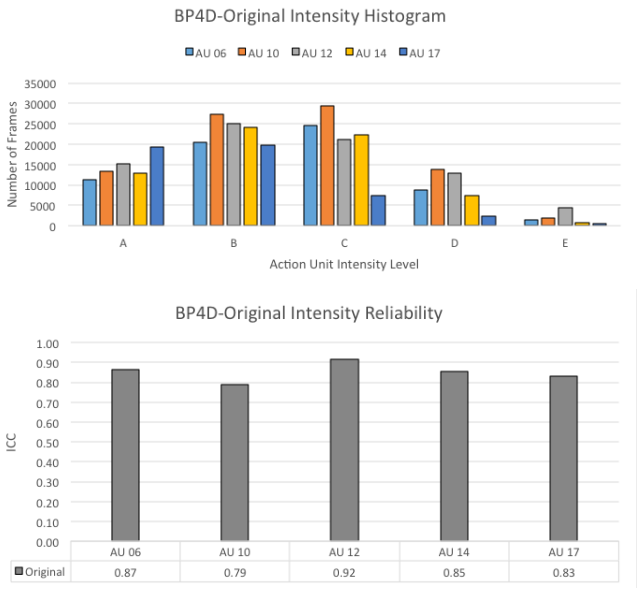

Fig. 3. AU intensity base rates (top) and inter-coder reliability (bottom) for the BP4D-Original data only.

reached the B-level of intensity and offsets when they dropped below it. Segments of the most facially-expressive 20 seconds of each task were selected for coding. Across all participants, AU base rates averaged 26.2%, ranging from 5% to 60%. To assess inter-coder reliability for occurrence, approximately 15% of the data were independently comparison coded as above. Inter-coder reliability, as quantified by MCC, averaged 0.79, ranging from 0.69 to 0.91. These results indicate strong to very strong inter-rater reliability. Across all AUs except for AU 15, inter-coder reliability for occurrence was lower in the expanded dataset than in the original dataset. These differences may be due in part to differences in threshold for determining occurrence (B-level versus A-level) and the addition of two coders in BP4D-Expanded.

For SEMAINE, one-minute segments of the most facially-expressive part of each selected interactions were coded. The same method for inter-coder reliability was employed on 10% of the data. Only AUs with good inter-coder reliability or better (i.e. MCC>0.6) were selected to be used in the challenge (see Fig. 2).

Intensity Annotation For BP4D-Original, five AUs were intensity coded in the BP4D-Original dataset: AU6, AU10, AU12, AU14, and AU17. The distribution of intensity levels was similar across the AUs. The B- and C-levels of intensity were most common for all except AU 17, which showed more A- than C-level intensity. To assess inter-coder reliability for intensity, approximately 6% of the data was independently coded by two highly trained and certified coders. Inter-coder reliability, as quantified by the intra-class correlation coefficient (ICC; [16]), averaged 0.85. ICC for individual AU ranged from 0.79 to 0.92. These results indicate strong to very strong inter-coder reliability for intensity.

IV. EVALUATION PROCEDURE

To perform a fair evaluation of participants' performance, participants are asked to submit their working programs to the challenge organisers, who will run these programs on the held-back test sets of the same two databases (BP4D and SEMAINE).

The performance measure for AU occurrence is the F1-measure, which is the harmonic mean of recall and precision. For an AU with precision P and recall R, it is calculated as:

$$F_1 = \frac{2PR}{P+R} \tag{1}$$

The performance measure for AU intensity is the Intraclass Correlation Coefficient (ICC, [16]). Given ground truth labels $\mathbf{y}$, $y_t \in \{0, 1, ...5\}$ and predictions $\hat{\mathbf{y}}$, $\hat{y}_t \in \mathcal{N}$, the ICC I is calculated as follows:

$$I = \frac{W - S}{W + (k-1)W} \tag{2}$$

where k is the number of coding sources compared; in our case $k = 2$. W and S are the Within-target Mean Squares and Residual Sum of Squares, respectively, and are computed as follows:

$$W = \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(y_{ij} - \bar{y_i})^2}{n(k-1)} = \sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(y_{ij} - \bar{y_i})^2}{n} \tag{3}$$

where $\bar{y_i} = \sum_{j=1}^{k} y_{ij}/k$ and the third term of Eq. (3) follows from $k = 2$. S is defined as:

$$S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{4}$$

To come to a single score s for the Occurrence Detection, Pre-Segmented Intensity Estimation and Fully-Automatic Intensity Estimation Sub-Challenges, labels of all test sequences will be concatenated into a sequence to calculate F1/ICC measures per AU. The average value will be used as the performance of a participant's submission:

$$s = \frac{1}{N} \sum_{a=1}^{N} f_a(\mathbf{y}, \hat{\mathbf{y}}) \tag{5}$$

where f_a is either the F1 or ICC measure for a given AU a, depending on the sub-challenge, and N is either the 14 AUs for Occurrence Detection, or the 5 AUs for intensity estimation. For AUs that do not occur in either BP4D or SEMAINE (see table I), both predictions and ground truth will be set to 0. This results in all values contributing to True Negative predictions, which do not impact the F1 measure.

The code used by the organisers to calculate participants' performance will be made available on the FERA 2015 challenge's website.

V. BASELINE SYSTEM

While baseline results exist for BP4D, this is the first time that the SEMAINE data is used for AU 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 on both datasets, for easy comparison of participants' systems. In addition, the baseline features described in this work are made available to participants, which is useful for those who wish to replicate the baseline system results and may be particularly valuable for participants who want to focus on the machine learning aspects of expression recognition.

A. Baseline Features

For the challenge baseline two types of features have been extracted: two-layer appearance features (Local Binary Gabor Patterns, LGBP [23]) and geometric features derived from tracked facial point locations. The geometric features are based on 49 landmarks detected and subsequently tracked with the Cascaded Regression facial point detector/tracker proposed by Xiong and De la Torre [20]. We only use the inner facial points as they are the main indicators of the AUs included in this challenge. For some frames facial point detection failed. When this happens the corresponding feature array contains all zeros, to indicate point detection failure to participants.

Facial landmarks $P = [p_1 \dots p_{49}]$ of every video frame are aligned with a mean shape using a set of stable points. Stable points are defined as those not affected by AU activations. In the 49 landmarks notation points 20, 23, 26, 29 (eye corners region) and 11 - 19 (nose region) are considered to be stable. The mean facial landmarks shape has been computed prior to geometric features extraction individually for each database by taking mean of 10% randomly selected video frames from every session. The alignment is performed by computing a non-reflective affine transformation, which minimises the difference between stable point coordinates of the two shapes. All mean shape landmark coordinates are then subtracted from the corresponding aligned shape points resulting in a set of aligned facial points $\tilde{P} = [\tilde{p}_1 \dots \tilde{p}_{49}]$, which form the first $49 * 2 = 98$ geometric features.

The next 98 features are composed by subtracting the aligned facial point locations of the previous frame from that of the current one. This applies to all frames except the very first one of every session, for which these features are the same as the first 98.

For the next set of features the facial landmarks have been split intro three groups representing the left eye (points 20 - 25) and left eyebrow (points 1 - 5), the right eye (points 26 - 31) and right eyebrow (points 6 - 10), and the mouth region (points 32 - 49). For each of these groups a set of features representing Euclidean distances as well as angles in radians between points within the groups is extracted.

Distances between points within a group are computed by taking the squared L2-norm between consecutive points:

$$F(i) = \|\tilde{p}_i - \tilde{p}_{i+1}\|_2^2,$$

$$i = \{1..N_p - 1\}$$

where N_p is the total number of points within the region, $\tilde{p}_i$ is the point coordinates vector and F is the feature array of the region. Hence, for each group the number of features constructed in this manner N_f is equal to:

$$N_f = N_p - 1$$

The same approach is used to calculate the angles between two lines defined by two pairs of points at a time within a group, where the two pairs share one common point. For each consecutive triplet of points Euclidean distances between them are computed first, which are then used to calculate angle between the points:

$$F(i) = \arccos\left(\frac{\tilde{p}_{12}^2 + \tilde{p}_{13}^2 - \tilde{p}_{23}^2}{2 * \tilde{p}_{12} * \tilde{p}_{13}}\right)$$

where $\tilde{p}_{ij}$ is an Euclidean distance between points i and j. The number of features extracted this way is equal to the total number of consecutive angles within the groups of points, which is equal to:

$$N_f = N_p - 2$$

There are thus 71 features in total extracted from the above face regions.

Finally, for the last 49 features we first compute median of stable points of the aligned shape. We then go through all of the aligned shape points and compute Euclidean distance between them and the median. In total there are

TABLE II

BASELINE RESULTS FOR THE OCCURRENCE SUB-CHALLENGE ON THE DEVELOPMENT PARTITION MEASURED IN F1 SCORE, 2AFC AND ACCURACY.

Action Unit	Database	Geometric			Appearance		
		F1	2AFC	Accuracy	F1	2AFC	Accuracy
AU1	BP4D	0.397	0.319	0.699	0.349	0.382	0.593
AU2	BP4D	0.317	0.364	0.699	0.265	0.465	0.571
AU4	BP4D	0.453	0.231	0.743	0.432	0.256	0.696
AU6	BP4D	0.763	0.144	0.780	0.775	0.141	0.771
AU7	BP4D	0.763	0.202	0.729	0.762	0.239	0.693
AU10	BP4D	0.831	0.851	0.796	0.804	0.815	0.744
AU12	BP4D	0.861	0.916	0.844	0.839	0.900	0.810
AU14	BP4D	0.616	0.323	0.569	0.613	0.345	0.548
AU15	BP4D	0.395	0.258	0.623	0.272	0.466	0.468
AU17	BP4D	0.617	0.267	0.652	0.538	0.383	0.518
AU23	BP4D	0.369	0.302	0.700	0.279	0.463	0.627
AU2	SEMAINE	0.235	0.243	0.653	0.343	0.237	0.863
AU12	SEMAINE	0.435	0.293	0.671	0.345	0.405	0.648
AU17	SEMAINE	0.317	0.173	0.932	0.114	0.454	0.936
AU25	SEMAINE	0.331	0.387	0.593	0.345	0.433	0.327
AU28	SEMAINE	0.457	0.108	0.975	0.308	0.238	0.975
AU45	SEMAINE	0.329	0.287	0.591	0.333	0.353	0.846
Weighted Mean	—	0.491	0.329	0.721	0.445	0.407	0.689

TABLE III

BASELINE RESULTS FOR THE OCCURRENCE SUB-CHALLENGE ON THE TEST PARTITION MEASURED IN F1 SCORE, 2AFC AND ACCURACY.

Action Unit	Database	Geometric			Appearance		
		F1	2AFC	Accuracy	F1	2AFC	Accuracy
AU1	BP4D	0.188	0.254	0.625	0.180	0.183	0.497
AU2	BP4D	0.185	0.214	0.696	0.159	0.234	0.611
AU4	BP4D	0.197	0.269	0.541	0.225	0.311	0.645
AU6	BP4D	0.645	0.177	0.753	0.671	0.136	0.720
AU7	BP4D	0.799	0.164	0.759	0.751	0.264	0.679
AU10	BP4D	0.801	0.882	0.774	0.799	0.881	0.774
AU12	BP4D	0.801	0.930	0.803	0.792	0.922	0.790
AU14	BP4D	0.720	0.293	0.673	0.666	0.321	0.612
AU15	BP4D	0.238	0.251	0.622	0.139	0.495	0.530
AU17	BP4D	0.311	0.289	0.564	0.245	0.369	0.391
AU23	BP4D	0.320	0.286	0.723	0.239	0.433	0.606
AU2	SEMAINE	0.569	0.102	0.832	0.755	0.106	0.938
AU12	SEMAINE	0.595	0.194	0.755	0.517	0.236	0.726
AU17	SEMAINE	0.091	0.168	0.926	0.066	0.261	0.927
AU25	SEMAINE	0.445	0.299	0.680	0.400	0.313	0.357
AU28	SEMAINE	0.250	0.106	0.971	0.009	0.289	0.982
AU45	SEMAINE	0.396	0.286	0.695	0.209	0.370	0.760
Weighted Mean	—	0.444	0.302	0.730	0.400	0.359	0.681

316 geometric features extracted from every video frame in both databases.

To extract appearance features the local LGBP descriptor has been adopted. LGBP takes a video frame which is first convolved with a number of Gabor filters to obtain a set of Gabor magnitude response images. This is followed by LBP feature extraction over the set of Gabor magnitude response images. The resulting binary patterns are histogrammed and concatenated into a single feature histogram. The final feature array is then composed for every frame by taking the mean of its histogram with that of up to 5 preceding frames. Prior to feature extraction, each image is split into a 4x4 regular grid to maintain some local information encoded in the features.

Preprocessing of video frames includes face localisation and segmentation by means of the Viola & Jones face detector prior to LGBP feature extraction. Fast and easy to use, it sometimes struggles to correctly detect a face on noisy data such as that used in this challenge. To keep the dimensionality of all feature vectors constant and the number of instances per video consistent with the number of frames, in this paper frames where the face detector failed to locate a face are marked with a feature vector of all zeros.

B. Baseline Results

The baseline system is kept simple on purpose since it should be easy to interpret and simple to replicate. All of the results have been obtained using linear SVM for the AU occurrence sub-challenge and linear SVR for the intensity sub-challenges using geometric and appearance features described above. In every experiment parameter search has been applied to find the best values of the SVM cost parameter C as well as SVR parameter epsilon.

Due to the high number of training samples, and in order to reduce the time required to train the models, training instances were subsampled to approach a balanced number

TABLE IV

BASELINE RESULTS FOR THE PRE-SEGMENTED INTENSITY ESTIMATION SUB-CHALLENGE ON THE DEVELOPMENT PARTITION MEASURED IN MSE AND PCC.

Action Unit	Database	Chance level	Geometric			Appearance		
		MSE	MSE	PCC	ICC	MSE	PCC	ICC
AU6	BP4D	0.879	0.529	0.696	0.684	0.828	0.473	0.372
AU10	BP4D	1.100	0.678	0.679	0.601	0.689	0.666	0.651
AU12	BP4D	1.465	0.483	0.823	0.797	0.790	0.746	0.744
AU14	BP4D	0.835	0.629	0.665	0.642	0.751	0.637	0.637
AU17	BP4D	0.695	0.694	0.151	0.042	0.668	0.210	0.108
Mean	—	0.995	0.603	0.603	0.553	0.745	0.546	0.502

TABLE V

BASELINE RESULTS FOR THE PRE-SEGMENTED INTENSITY ESTIMATION SUB-CHALLENGE ON THE TEST PARTITION MEASURED IN MSE AND PCC.

Action Unit	Database	Chance level	Geometric			Appearance		
		MSE	MSE	PCC	ICC	MSE	PCC	ICC
AU6	BP4D	0.880	0.901	0.479	0.475	0.774	0.409	0.330
AU10	BP4D	0.825	0.771	0.549	0.510	0.900	0.489	0.483
AU12	BP4D	0.790	0.450	0.691	0.685	0.945	0.613	0.604
AU14	BP4D	0.722	0.572	0.611	0.592	0.960	0.498	0.497
AU17	BP4D	0.628	0.574	0.153	0.050	0.596	0.187	0.107
Mean	—	0.769	0.654	0.497	0.462	0.835	0.439	0.404

TABLE VI

BASELINE RESULTS FOR THE FULLY AUTOMATIC INTENSITY ESTIMATION SUB-CHALLENGE ON THE DEVELOPMENT PARTITION MEASURED IN MSE AND PCC.

Action Unit	Database	Chance level	Geometric			Appearance		
		MSE	MSE	PCC	ICC	MSE	PCC	ICC
AU6	BP4D	1.914	1.103	0.699	0.690	1.020	0.720	0.694
AU10	BP4D	2.246	1.255	0.715	0.696	1.289	0.683	0.641
AU12	BP4D	2.439	2.137	0.706	0.653	2.166	0.695	0.670
AU14	BP4D	1.756	1.548	0.472	0.453	1.614	0.396	0.325
AU17	BP4D	1.019	0.960	0.365	0.278	0.991	0.303	0.185
Mean	—	1.875	1.401	0.592	0.554	1.416	0.559	0.503

TABLE VII

BASELINE RESULTS FOR THE FULLY AUTOMATIC INTENSITY ESTIMATION SUB-CHALLENGE ON THE TEST PARTITION MEASURED IN MSE AND PCC.

Action Unit	Database	Chance level	Geometric			Appearance		
		MSE	MSE	PCC	ICC	MSE	PCC	ICC
AU6	BP4D	1.992	1.004	0.698	0.670	1.366	0.644	0.622
AU10	BP4D	2.135	0.897	0.757	0.732	1.209	0.686	0.656
AU12	BP4D	2.205	0.738	0.816	0.780	1.092	0.768	0.767
AU14	BP4D	2.020	1.227	0.650	0.586	1.526	0.521	0.389
AU17	BP4D	0.656	0.806	0.184	0.144	0.819	0.225	0.168
Mean	—	1.802	0.934	0.621	0.582	1.202	0.569	0.520

of positive and negative examples while at the same time reducing the total number of training examples. Because of a high dimensionality of the appearance features, PCA has been employed for the purpose of dimensionality reduction keeping 98% of the energy. A very simple correlation-based form of feature selection has also been applied to the geometric features despite their low dimensionality. The features have been selected based on Pearson's correlation coefficient (PCC) computed for every geometric feature, measuring the correlation between the feature values and the true labels. The exact number of features selected varies depending on the AU.

Baseline results for occurrence/activation detection are shown in Tables II and III for the development and test partitions of both databases. Detection performance is measured by F1 as well as accuracy and 2AFC scores. A number of different performance measures are shown since each has their own merits, and combined they provide a deeper analysis of the results. However the challenge participants will only be judged based on F1 scores.

Whereas F1 and accuracy are well-known performance measurement, 2AFC is less well-known. The 2AFC score is a good approximation of the area under the receiver operator characteristic curve (AUC). In contrast to F1, 2AFC does take True Negative preditions into account. In this study the 2AFC has been calculated based on the SVM decision

function output values as follows:

$$2AFC(\hat{Y}) = \sum_{i=0}^{n}\sum_{j=0}^{p}\sigma(P_j, N_i)\frac{1}{n \times p}, \quad (6)$$

$$\sigma(X,Y) = \begin{cases} 1, & \text{if } X > Y \\ 0.5, & \text{if } X == Y \\ 0, & \text{if } X < Y \end{cases}$$

where $\hat{Y}$ is a vector of decision function output values, n is the total number of true negative and p the total number of true positive instances in $\hat{Y}$, and P and N are subsets of $\hat{Y}$ corresponding to all positive and negative instances, respectively.

Final scores are computed based the results of the two databases as a weighted mean based on the total number of samples in each database. Results for intensity estimation on the publicly available training/development partition are shown in Tables IV and VI corresponding to the pre-segmented intensity estimation sub-challenge as well as fully automatic one respectively. Baseline intensity estimation performance is measured by means of PCC and Mean Squared Error (MSE).

Both occurrence detection and intensity estimation baselines perform well over chance levels, but are clearly a long way of from being accurate enough to be used in real world-applications. An exception to this is the occurrence detection of AU12, which with an F1-score of 0.86 can be said to be reliable. A high 2AFC score for this AU (0.92) indicates that the classifier for AU12 also has a low number of false positives. The same can be said for AU10 detection performance. Intensity baselines overall demonstrate similar behaviour, with the best results obtained for AU10 and AU12.

VI. CONCLUSION

In this paper we have presented the Second Facial Expression Recognition and Analysis Challenge (FERA 2015) dedicated to FACS Action Units detection and intensity estimation on the highly challenging set of data. The dataset for this challenge has been composed using two facial expression databases BP4D and SEMAINE. This is the first time these datasets have been applied to FACS AU analysis except the training and development partitions of the former. The challenge addresses such significant problems of the field as expression intensity estimation as well as robust detection under non-frontal head poses, partial occlusions and environmental factors. Baseline results obtained using geometric and appearance features demonstrate a huge room for potential improvements to be brought by the challenge participants.

REFERENCES

[1] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Automatic recognition of facial actions in spontaneous expressions. *Journal of Multimedia*, 1(6):22–35, 2006.

[2] J. Bazzo and M. Lamar. Recognizing facial actions using Gabor wavelets with neutral face average difference. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 505–510, 2004.

[3] S. W. Chew, P. Lucey, S. Saragih, J. F. Cohn, and S. Sridharan. In the pursuit of effective affective computing: The relationship between features and registration. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 42(4):1006–1016, 2012.

[4] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amier, and D. Heylen. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *LREC Workshop on Corpora for Research on Emotion and Affect*, pages 1–4, Paris, France, 2008. ELRA.

[5] P. Ekman, W. Friesen, and J. C. Hager. *Facial action coding system*. A Human Face, 2002.

[6] B. Fasel and J. Luettin. Automatic facial expression analysis: a survey. *Pattern Recognition*, 36(1):259–275, 2003.

[7] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods*, 200(2):237–56, 2011.

[8] L. A. Jeni, J. M. Girard, J. Cohn, and F. D. L. Torres. Continuous au intensity estimation using localized, sparse facial feature space. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition Workshop*, 2013.

[9] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *Proceedings of the International Symposium on Visual Computing*, pages 368–377, 2012.

[10] J. Lichtenauer, J. Shen, M. Valstar, and M. Pantic. Cost-effective solution to synchronised audio-visual data capture using multiple sensors. *Image and Vision Computing*, 29(10):666–680, 2011.

[11] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 74–80, 2009.

[12] G. McKeown, I. Sneddon, and W. Curran. The underdetermined nature of laughter: the blurring of boundaries in social signals. *unknown*, 2015. under review.

[13] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3:5–17, 2012.

[14] D. M. W. Powers. Evaluation: From precision, recall and f?measure to roc, informedness, markedness, and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[15] A. Savran, B. Sankur, and M. T. Bilge. Regression-based intensity estimation of facial action units. *Image and Vision Computing*, 30(10):774–784, 2012.

[16] P. Shrout and J. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

[17] Y. Tian, T. Kanade, and J. F. Cohn. Evaluation of Gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 229–234, 2002.

[18] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition Workshop*, 2011.

[19] T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S.Bartlett, and J. R. Movellan. Multi-layer architectures of facial action unit recognition. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 2012. In print.

[20] Xuehan-Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[21] P. Yang, Q. Liua, and D. N. Metaxasa. Boosting encoded dynamic features for facial expression recognition. *Pattern Recognition Letters*, 30(2):132–139, 2009.

[22] 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.

[23] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face representation and recognition. *Computer Vision, 10th IEEE International Conference on*, 1:786–791, 2005.

[24] X. Zhang, L. Yin, J. F. Cohn, C. S., M. Reale, A. Horowitz, and J. M. Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.