Image and Vision Computing 30 (2012) 774–784

Contents lists available at SciVerse ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Regression-based intensity estimation of facial action units[☆]

Arman Savran [a,*], Bulent Sankur [a], M. Taha Bilge [b]

[a] Electrical-Electronic Engineering Department, Bogazici University, Istanbul, Turkey
[b] Department of Psychology, Bogazici University, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 6 July 2011
Received in revised form 6 November 2011
Accepted 27 November 2011

Keywords:
Action unit intensity estimation
3D facial expression recognition
Facial Action Coding System
Feature selection
AdaBoost.RT
SVM regression

ABSTRACT

Facial Action Coding System (FACS) is the de facto standard in the analysis of facial expressions. FACS describes expressions in terms of the configuration and strength of atomic units called Action Units: AUs. FACS defines 44 AUs and each AU intensity is defined on a nonlinear scale of five grades. There has been significant progress in the literature on the detection of AUs. However, the companion problem of estimating the AU strengths has not been much investigated. In this work we propose a novel AU intensity estimation scheme applied to 2D luminance and/or 3D surface geometry images. Our scheme is based on regression of selected image features. These features are either non-specific, that is, those inherited from the AU detection algorithm, or are specific in that they are selected for the sole purpose of intensity estimation. For thoroughness, various types of local 3D shape indicators have been considered, such as mean curvature, Gaussian curvature, shape index and curvedness, as well as their fusion. The feature selection from the initial plethora of Gabor moments is instrumented via a regression that optimizes the AU intensity predictions. Our AU intensity estimator is person-independent and when tested on 25 AUs that appear singly or in various combinations, it performs significantly better than the state-of-the-art method which is based on the margins of SVMs designed for AU detection. When evaluated comparatively, one can see that the 2D and 3D modalities have relative merits per upper face and lower face AUs, respectively, and that there is an overall improvement if 2D and 3D intensity estimations are used in fusion.

1. Introduction

Automated measurement of facial actions has many potential applications for intelligent human–computer interfaces (HCI) and in behavioural science. As discussed thoroughly in ref. [1], computer perception of emotional states can be useful for HCI in various ways, and these states can be inferred via the extracted facial actions. As an application example consider an automatic tutoring system: If the tutor is capable of understanding whether the user is bored or irritated, it can switch to a more affective response mode. Automatic monitoring of operator alertness and of driver fatigue are two other practical domains. Several applications may be envisioned in the entertainment area, from computer games to dolls which respond according to the player's mood. Automatic facial expression recognizers can also be useful to aid human judgement, in testing the veracity of a subject's responses, for instance to detect deception. Furthermore, facial actions can be used for generating performance-driven facial animations.

Facial Action coding System (FACS) [2] is the most common facial action measurement methodology, and it involves 44 Action Units (AUs) related to visually discernible facial muscle activations. Being composed of very extensive set of rules, FACS requires certified human coders and coding of face images is a very tedious and time consuming process. This is the main motivation for the development of automatic coders for facial behavior research and other FACS applications.

Although there is already a substantial literature on automatic expression and action unit recognition [3], it still continues to be an active area of study due to the challenging nature of the problem. However, in contrast to AU detection, there is much less work in the literature on AU intensity estimation. FACS defines AU intensities on a five-point ordinal scale, i.e., from lowest A to strongest level E intensity. The main benefit of estimating AU strengths is that the qualified AUs would yield more information about mental state and emotional involvement of a subject. Moreover, since humans can express their feelings in different ways under different situations, information conveyed by AU intensities can be exploited to adapt emotion recognizers to a particular user and context. An example adaptation framework is proposed in ref. [4] where landmark coordinates are employed instead of AUs. The measurement of intensities can also facilitate and even improve FACS coding. For instance, if the perceived expression is surprise, and if AU 5—Upper Lid Raiser is present, then, it can only be at level B [2].

In this study we focus on person-independent AU intensity estimation using 3D as well as 2D modality. Our work, as far as we know, is the first one in the area of AU intensity estimation using

[☆] This paper has been recommended for acceptance by Stefanos Zafeiriou.
[*] Corresponding author. Tel.: +90 212 359 64 14; fax: +90 212 287 24 65.
E-mail address: arman.savran@boun.edu.tr (A. Savran).

0262-8856/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2011.11.008

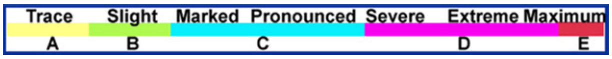

Fig. 1. Relationship between the scale of evidence and intensity scores [2].

3D data. We do a detailed investigation of AU intensity estimation using 3D modality, 2D modality and eventually consider their fusion. Notice that there are two factors that make the estimation problem challenging. First, person-independent implies that we have to combat against variabilities due to subjects. Second, without video we are deprived of the rich dynamic information. On the other hand, the advantage of subject-independent static estimation is its potential applications in a wider range, i.e., when we have only one image of a person. We pursue regression-based approach to intensity estimation, and apply regression in a person-independent data-driven expression analysis framework, i.e., without using face modelling. The data-driven approach facilitates the problem as we avoid model preparation, facial landmarking and model fitting stages, and allows fair comparison of the 3D and 2D modalities since the results are not biased by face modelling. Finally, we remark that we experiment with 25 AUs, a variety larger by a factor of three, of AU types treated so far in the literature [5].

The paper is organized as follows. In Section 2 we give the preliminaries: we describe FACS intensity scoring problem, survey previous work, and explain the expression database that we worked on. Section 3 presents how 2D and 3D features are extracted. We employ Gabor Wavelets and use different types of surface curvature data. Section 4 develops regression-based intensity estimators. We first examine prediction using AU detector decision scores, then perform non-linear regression on image features. We select different sets of features for each AU according to their intensity prediction capabilities. In Section 5 we explain our feature selection based fusion method to investigate the fusion of 3D and 2D modalities as well as the fusion of different 3D geometry features. Section 6 is devoted to the experimental results and their discussions. Finally, we draw our conclusions in Section 7. Preliminary parts of our work on intensity estimation were presented in [6].

2. Preliminaries

2.1. FACS Intensity Scoring

FACS has developed certain conventions and rules for scoring intensities of Action Units. Scoring is done on a five-point ordinal scale, A–B–C–D–E, if evidence of an AU is present. The interpretations of these levels are as follows: level A refers to a trace of the action; B, slight evidence; C, marked or pronounced evidence; D, severe or extreme action; and E, maximum evidence. The relationship between the scale of evidence and the scoring levels is a bit different for some AUs. Scoring criteria depend upon the scale of evidences, and the evidence can be assessed in terms of the degree of appearance change or in terms of the number of appearance changes. Scoring criteria are listed in the FACS manual [2] for each AU, though sometimes modified criteria are used depending on the AU combinations.

Each AU intensity level, as denoted by a letter, refers to a range of appearance changes, and not to a single strength of AU. Notice that the intensity scale is not divided into uniform intervals; for example, levels C and D cover a larger range of appearance changes. The relationship between the scale of evidence and intensity scores is depicted in Fig. 1.

FACS manual states that scoring of lower intensities, that is, levels A and B, require particularly careful examination, and level A actions can only be scored reliably by very experienced coders. While scoring of lower intensities may not be easy, distinguishing level E AUs can be difficult as well since the intense muscular contractions of the level E combine with the person's individual physical characteristics causing variability on the appearance changes across different people. Samples of low level (B) and high level (E) AU 5—Upper Lid Raiser images are shown in Fig. 2. Samples of some other AUs are shown in Figs. 7 and 8 as well.

2.2. Previous Work

There are two main paradigms employed for detection of Action Units and for recognition of expressions. In one paradigm, facial images are analyzed based on face models, called model-driven methods. This paradigm presupposes automatic and reliable extraction of a high number of facial landmarks [7] or fitting of models like Active Appearance Models (AAMs) [8]. However, landmark detection itself can be problematic, and though AAM is able to better cope with facial shape extraction (i.e., locating facial landmarks), they do not allow for accurate person-independent analysis of expressions; in fact, surmounting the limitation of these methods due to person-dependence is still a research topic. In the other paradigm, facial images are analyzed without incorporating any prior information about faces, hence they are called data-driven methods. Since no facial information is utilized, methods in this paradigm do not require landmark detection or model fitting, thus enabling practical and

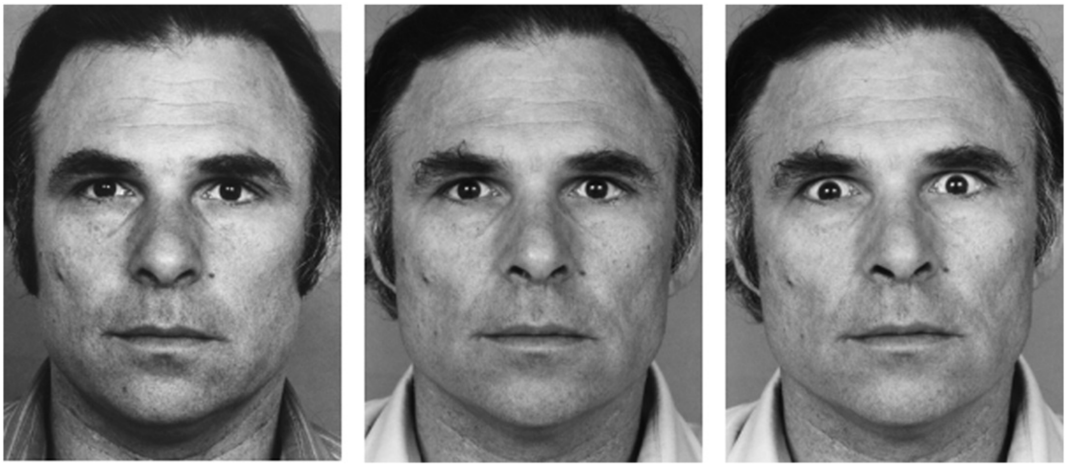

Fig. 2. AU 5—Upper Lid Raiser. Neutral (left), level B (middle) and level E (right) samples are shown [2].

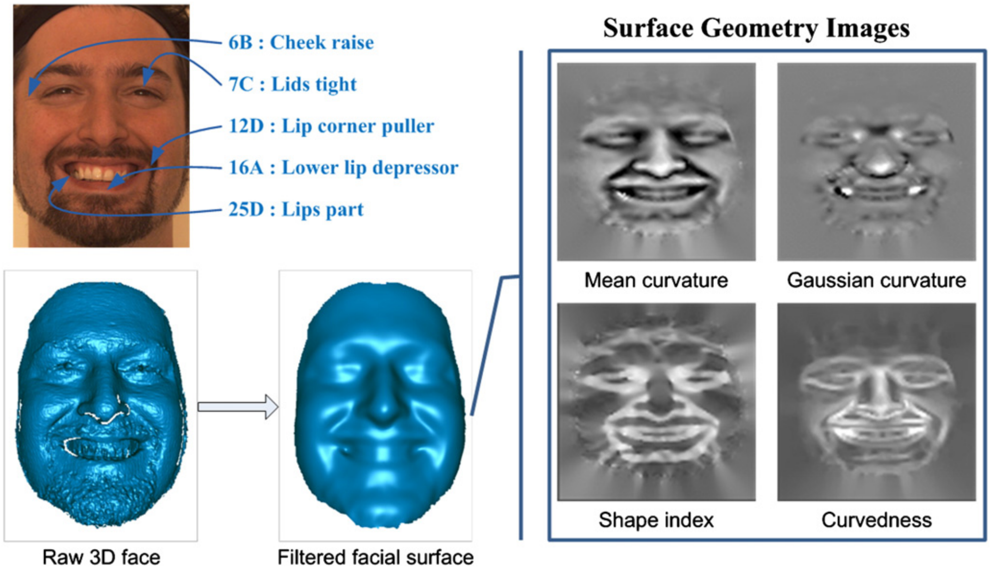

Fig. 3. Various images of a sample FACS coded face expressing happiness emotion. Several pre-processing operations are applied to obtain the filtered 3D facial surface from 3D raw data, and then different types of surface geometry images are generated.

robust recognizers. Moreover, data-driven methods are not biased or limited by the designed face models. However, without incorporation of facial information, data-driven methods may need more data to cope against confounding factors like pose, texture and lighting variations. A seminal work in this vein is [5] where Bartlett et al. apply Gabor wavelets and select AU specific subsets via AdaBoost algorithm. This method has also been employed together with DynamicBayesian Networks by Tong et al. [9] to exploit dynamic and semantic relationships between AUs for improved performance.

Although substantial amount of work on AU detection exists in the literature, AU intensity estimation studies remain quite limited. One of the early works on AU intensity estimation has been done by Pantic and Rothkrantz [10], who developed an expert system using geometric features extracted from dual-view (frontal and profile) images. However, their study is person-dependent and requires detection of a high number of landmarks. Most of the other works on expression intensity are based on the relationship between AU intensities and the likelihood scores and of AU classifiers. For instance, Bartlett et al. [5] investigated correlations between intensity levels and SVM classifier margins in their Gabor filter-based detectors. They reported moderate to high correlations between them for several AUs. One criticism of using classifier scores is that they do not necessarily incorporate all of the relevant intensity information. Yang et al. [11] have used the output scores of RankBoost based expression classifiers to better deal with intensity variations. They train RankBoost classifiers with frames that sweep from onset of an expression to its apex for the purpose of ranking image pairs according to expression intensity levels. Although they have obtained better image pair ordering performance than the linear SVM-margin approach as in [5], it is still not clear how the capability of image pair ranking in sequences monotonically increasing in intensity can be transferred to the estimation of intensities. Recently Mahoor et al. [12] studied measurement of AU 6 and AU 12 intensities over six subjects via person-specific AAMs. They approached to intensity estimation as a classification problem and applied six levels of SVM classifiers based on one-against-one technique. For feature extraction they performed AU specific dimensionality reduction by applying regularized locality preserving indexing on appearance data, and used delta features (i.e., by subtracting neutral face features).

In recent years use of 3D data for expression analysis has attracted the attention of researchers since it is purportedly more robust to pose and illumination variations. For instance, Wang et al. [13] have divided 3D faces into regions using64 manually marked points, and extracted regional histograms of surface curvatures. Other 3D methods proposed for emotion identification also have this handicap of depending on an excessive number of feature points [14–17]. There has been recently also some work on expression recognition using simultaneous 2D and 3D video [18,19]. Both of these two independent works perform model-based analysis on 2D luminance images, by either ASMs orAAMs, to track more than 80 facial points from which 2D and 3D data features are extracted. However, these studies are mostly on prototypical emotional expressions and consider AU detection in a limited way. For instance, only 11 singly occurring posed AUs are recognized in ref. [19] using a rule-based classification. A comprehensive evaluation of pure 3D AU detection vs. pure 2D detection over a set of 25 AUs was presented in our previous work [20] where data-driven analysis was employed for 2D-to-3D comparisons to avoid any bias that could arise by the adopted face modelling. It has been shown that working with 3D facial surface data indeed improves AU recognition performance, and this not only under pose and illumination perturbations but even under controlled illumination and frontal face poses.

However, to the best of our knowledge there is no previous work that tackles the intensity prediction problem using 3D faces. The merits of 3D face data for AU detection have been shown in [20], and we conjecture that 3D would also be beneficial for AU intensity estimation.

2.3. Expression database

We aim to estimate the intensity scores of AUs in a completely person-independent manner (i.e., not trained on or normalized for any one individual) using still images. We worked on the Bosphorus Database [1] [20] which has not only the intensity scores for all of its face samples and all AUs, but also a comparatively richer repertoire of AUs. It contains 4666 face scans and 105 subjects. Ground-truth FACS codes of the expressions were attributed by a certified FACS coder. The majority of the subjects are aged between 25 and 35,

[1] This database is available at http://bosphorus.ee.boun.edu.tr/.

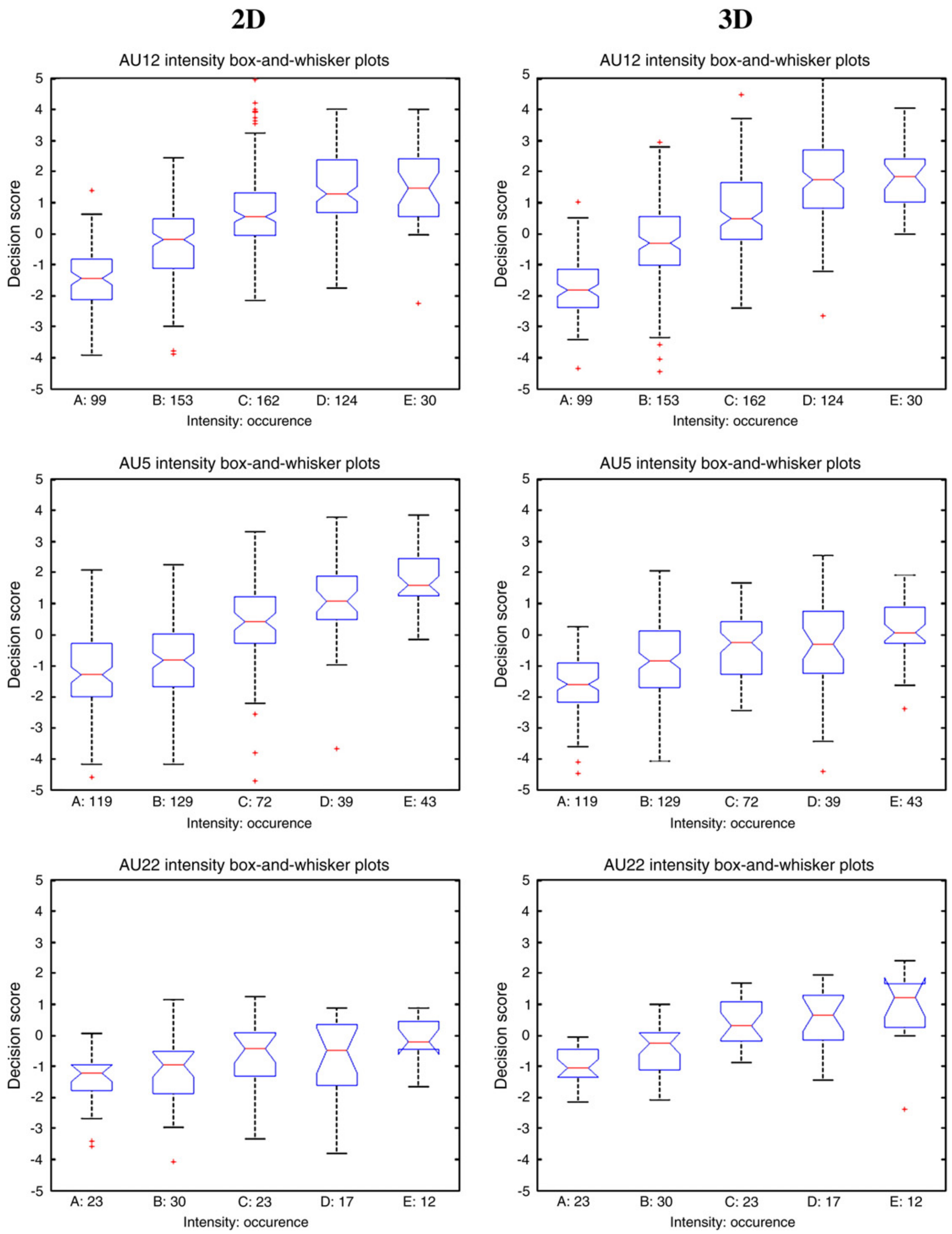

Fig. 4. Distributions of decision scores (RBF-SVM margins) of AU 12—Lip Corner Puller, AU 5—Upper Lid Raise and AU 22 Lip Funneler, for 2D and 3D data modalities shown as box-and-whisker plots (central mark: median, box: interquartile range, whiskers: extreme values, "+": outlier).

mostly Caucasian, and the cohort consists of 60 men and 45 women in total. 29 professional actors and actresses are employed for acting the expressions while the rest were recruited from students and staff. 35 men had beard/moustache (19 of them are intense and 16 of them are moderate). 71 subjects were recorded with 54 different face scans (neutral, AU, universal emotion) while a minority of 34 subjects had 31 scans having fewer number of expressions.

3D faces were acquired with a structured light system and the companion 2D face images with a normal light camera. The images were acquired under good illumination conditions, in almost frontal poses, i.e., with mild 3D rotations. The color images have 1600 × 1200$_?$ resolution and the number of points on 3D faces varies roughly between 30 K and 50 K depending on the size of the face and due to the 1/6$_?$ down-sampling of the depth maps. The decimation is automatically made by the acquisition system on each dimension.

3. Feature extraction

3.1. Gabor wavelets

As shown in a recent study [21] on smile detection, Gabor wavelets are one of the best among relevant image features such as box filters, edge orientation histograms, and local binary patterns. We extract Gabor magnitudes on 96 × 96$_?$ size images using eight wavelet

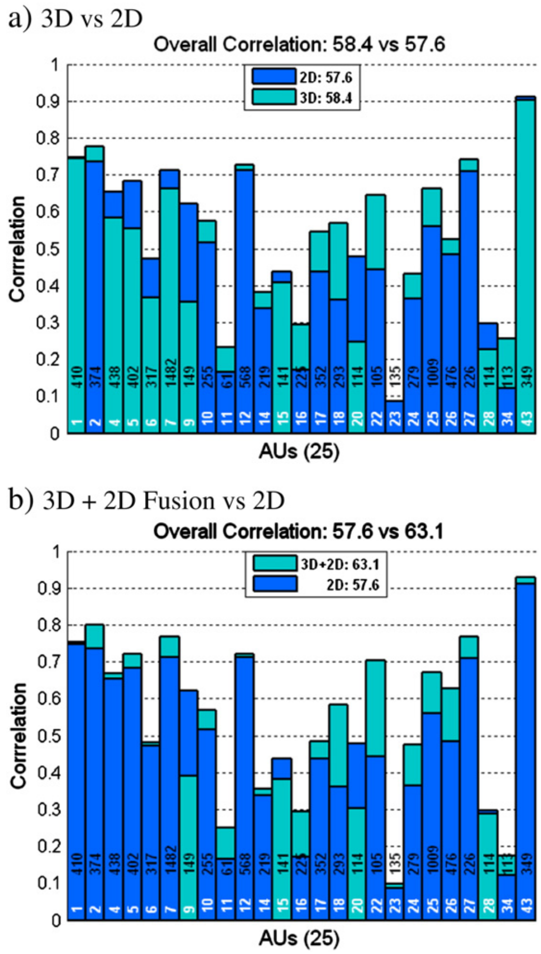

Fig. 5. Performance (correlation) comparison between 3D vs 2D vs fusion. The AU code and the total number of occurrences are inscribed in the bars.

directions and nine scales so that their Gabor wavelengths vary in the range of 2 to 32 pixels in half octave intervals. Although the resulting feature vector has $9 \times 8 \times 96 \times 96 = 663{,}552_?$ components, not all of them are informative and in fact only a very small portion is selected as described in Section 4.

Gabor wavelets are extracted separately from both 2D luminance and 3D surface geometry images. To perform Gabor analysis on 2D luminance images, first eye centers are located and then images are aligned accordingly using 2D rotation, translation and scaling, which is a typical procedure in 2D facial image registration. To perform Gabor analysis on 3D faces, we convert densely sampled surface data to 2D surface geometry images as described in Section 3.2, and then apply Gabor wavelets. Notice that, in this work we solely want to prove the potential of 3D surface data in estimating AU intensities. Thus, to preclude any misalignment effects due to automatic registration on the comparison of methods, in this work both 2D luminance images and 3D data have been normalized using manually determined landmarks. We want to emphasize two points: First, this intervention is done with the sole purpose of precluding misalignment effects from any automatic registration scheme. In other words, we desire to put into evidence the potentials of 2D and 3D for AU estimation in the absence of any other confounding factor. Second, our method still remains a data-driven method since no manually extracted information is used for any AU intensity estimation. Finally in real applications, both methods would suffer to varying degrees the consequences of automatic (non-manual) alignment.

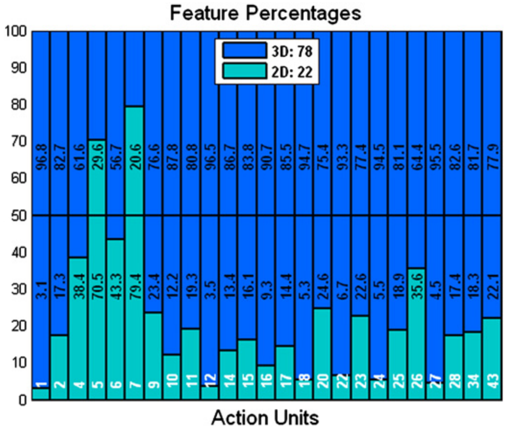

Fig. 6. Percentages of 3D and 2D features in modality fusion via AdaBoost.RT feature selection.

3.2. 3D geometry features

Since 3D acquisitions are typically noisy, several noise filtering steps are applied to remove the spikes, to smooth the data and to fill in the holes. Using this pre-processed 3D data, different surface geometry values are estimated and orthographically projected onto 2D domain. Then surface geometry values are re-sampled on the $96 \times 96_?$ pixel square image analysis domain. The transformation that maps the projected world coordinates to the image coordinates involves separate scaling of horizontal and vertical dimensions to properly fit faces to the square image region.

Re-sampling of the surface geometry data is achieved very rapidly, even for high resolution meshes, by utilizing graphics hardware. The final step is the extrapolation of the mapped values outside the 2D domain of the face surfaces. This is needed to smooth the abrupt passage from the delineated region of support of the 3D face and its background within the analysis window. A satisfactory extrapolation is achieved by an efficient image in-painting algorithm [22]. In the collection of images in Fig. 3 we see a filtered 3D face and several surface geometry images generated from it.

We consider the four basic local shape features based on surface curvatures to analyze facial surface deformations. We also tried direct use of depth data, but it performed much worse in our experiments. Curvature related data types also have the advantage of being rotation and translation invariant. Fig. 3 displays these basic data types in the form of face surface images mapped as 2D images.

1: **Mean Curvature (*H*):** Mean curvature at a surface point is the mean of the principal curvatures, i.e., the maximal k_1 and minimal k_2 curvatures, and it is an extrinsic measure of curvature. Principal curvatures are extracted by the local analysis of the differential structure of the surface and they measure the amount of surface bending in different directions. In this respect, they are good surface feature candidates for the analysis of facial deformations.

2: **Gaussian Curvature (*K*):** Gaussian curvature is the product of the principal curvatures, and it is an intrinsic measure of curvature.

3: **Shape Index (*S*):** Shape index has been developed [23] to measure the local shape by a single number in a continuous range. In contrast to mean and Gaussian curvature at a point, shape index is scale-independent and directly represents the local shape, but at the expense of curvedness information

4: **Curvedness (*C*):** Curvedness has been proposed [23] to represent the scale-dependent aspect of a 3D structure. In other words, while the difference in the amount of curvature between surfaces can be measured, this is not possible with the shape index.

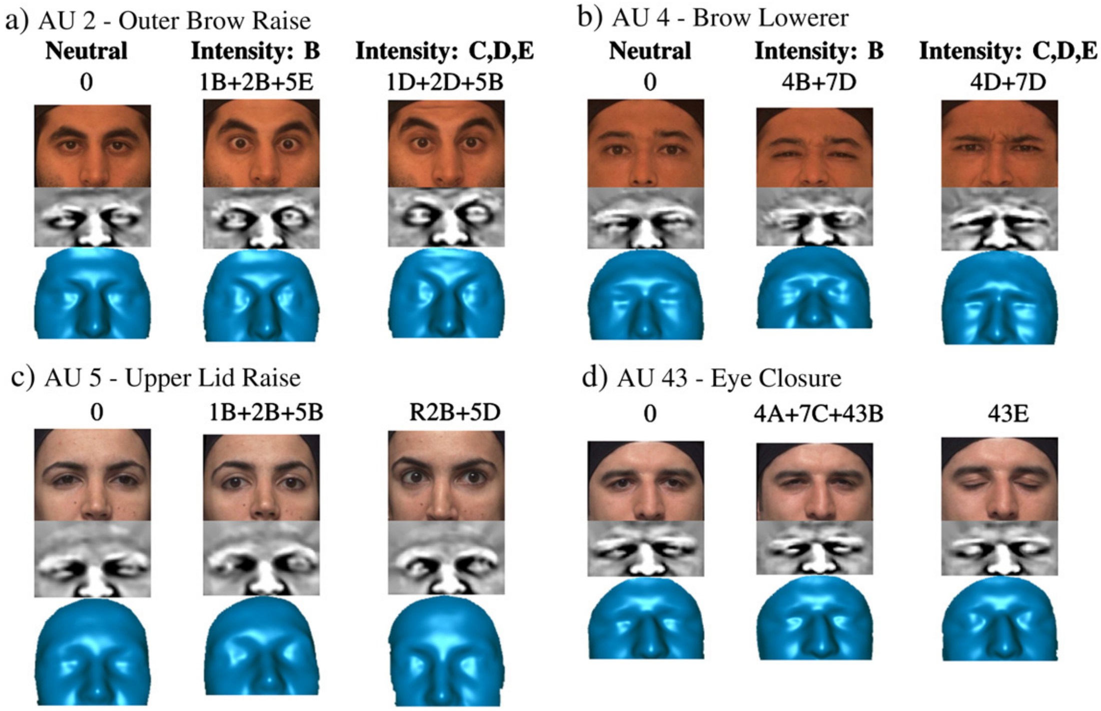

Fig. 7. Color, surface mean curvature and 3D surface images are shown for low (level B) and high (level C, D or E) intensity instances of several upper face action units together with the neutrals from the same subject. Some of these samples involve AU combinations and the corresponding FACS codes are written above each sample.

These local shape features are calculated from the maximal k_1 and minimal k_2 principal curvatures as follows:

$$H = \frac{k_1 + k_2}{2} \quad , \quad K = k_1 \times k_2$$
$$S = \frac{1}{2} - \frac{1}{\pi} \arctan \frac{k_1 + k_2}{k_1 - k_2} \quad , \quad C = \sqrt{\frac{k_1^2 + k_2^2}{2}} \tag{1}$$

4. Regression-based intensity estimation

We formulate the estimation of AU intensity levels as a regression problem. The dependent variable is the intensity in ordinal scale varying from one to five. The explanatory variables are either AU detector decision scores or selected image features. Since the output of the regressor is continuous, the outputs are quantized into five discrete intensity levels.

4.1. Regression on AU Detector Decision Scores

It was shown in ref. [5] that distances to SVM margins (separating hyperplanes) used for AU detection are correlated with intensity levels of AUs. This indicates that AU detector decision scores can also be used to estimate AU intensities. Fig. 4 shows the scatter of AU decision scores (RBF-SVM margins) for 2D and 3D modalities as box-and-whisker plots. As expected the FACS and SVM scores are correlated, in that the higher the assigned AU intensity the bigger the SVM score is. However, there is substantial overlap between adjacent intensity grades. Note that the medians of the distributions do differ significantly at the 5% significance level if their notches (interquartile ranges) do not overlap. There could be a number of explanations for these score distribution overlaps. First, whatever technique is employed some degree of overlap is perhaps unavoidable; that is, a strict separation of intensities should not be expected since FACS does not define a quantitative measure between levels. Second, in person-independent intensity estimation, one is confronted with more variability since different subjects tend to enact AUs differently and facial surface and texture vary from subject to subject. Third, a perfect ground-truth human FACS scoring by experts cannot be realistically expected due to the very detailed nature of the annotating procedure, and it is prone to judgement errors. Finally, the SVM algorithm that we use is designed to detect AUs, but not to estimate their intensity. These factors make the AU estimation problem more challenging. When we observe score distributions as a function of data modality (i.e., 2D vs 3D), we can better discern some of the difficulties of the problem. For the scatter of AU 12-Lip Corner Puller intensities (first row of Fig. 4), the trends are quite similar for 2D and 3D. In the case of AU 5-Upper Lid Raise, the score distributions of 3D overlap much more compared to those of 2D, implying a shortcoming of 3D measurements around eye regions, which is known to have poor reflectivity, and in this case as an oddity, level D scores do not follow the upward trend. On the other hand, in AU 22-Lip Funneler, an opposite pattern is observed, indicating that 2D cannot properly capture funneling distortions. Shortcomings of each modality will be partly compensated for when we resort to fusion of 2D and 3D in Section 5.

We estimate the AU intensity levels, $f(r)$, using logistic regression on SVM scores r:

$$f(r) = \frac{1}{1 + e^{-(a+br)}} \tag{2}$$

The logistic regression has a mapping range in [0 ; 1] and hence the target values in the range [0 ; 5] (level A to E) are scaled accordingly.

4.2. Regression on image features

Although we have used the distances to the SVM hyperplanes as indicative of AU intensity, this proportionality between SVM scores and intensities is not guaranteed since the support vectors are chosen for the classification task but not for intensity level estimation. We therefore consider an alternative regression in the feature space of selected Gabor wavelet magnitudes of luminance or of curvature field.

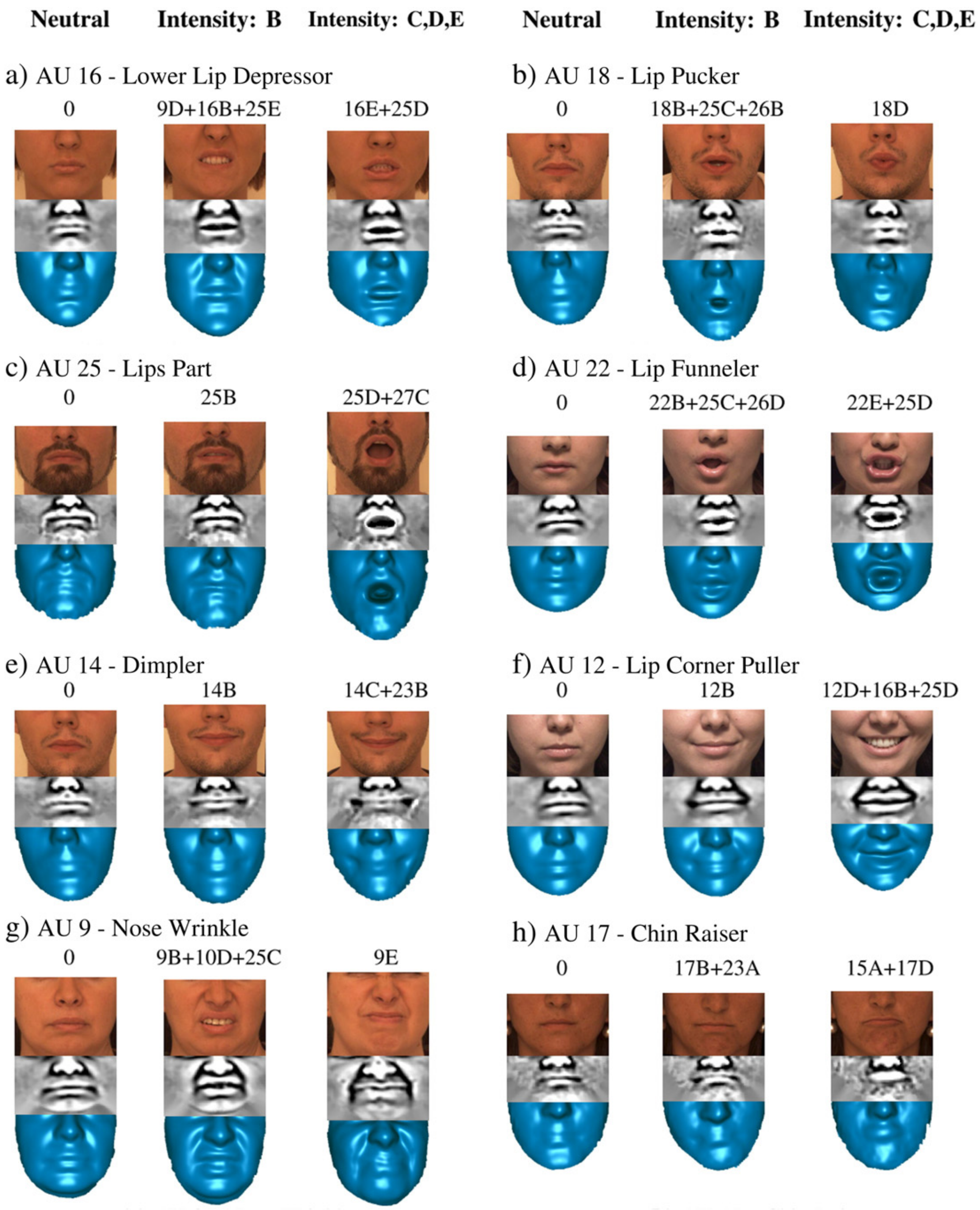

Fig. 8. Color, surface mean curvature and 3D surface images are shown for low (level B) and high (level C, D or E) intensity instances of several lower face action units together with the neutrals from the same subject. Some of these samples involve combinations and the corresponding FACS codes are written above each sample.

This regression problem is not straightforward since we have a high number of explanatory variables (features), and the dependent variable (annotator's scores) are noisy, as there is considerable overlap between intensity grades as discussed in Section 4.1. Hence, we apply SVM regression based on Vapnik's ε-insensitive loss function [24]. ε-SVM regression is appropriate because, first, high dimensionality of the input space is elegantly handled, and second, the ε-insensitive loss function is robust and generates a smooth mapping.

Another consideration is the non-linearity between the scale of evidence and intensity levels, as depicted in Fig. 1. This relationship points out to the possible benefits of non-linear modeling. Notice that there are also other sources of non-linearity, such as those due to combinations of AUs in the form of co-articulations. SVMs are also excellent tools for effectively learning various types of complex mappings by means of kernels. The SVM regression function has the form:

$$f(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{3}$$

where $\mathbf{x}$ is the feature vector, $k(\mathbf{x}_i,\mathbf{x})$ the kernel function, $f(\mathbf{x})$ is the predicted intensity level and $\mathbf{x}_i$ are the support vectors.

In our study we investigated both linear SVM and SVM with non-linear kernels of the Gaussian radial basis function (RBF) variety. The advantage of RBF is its ability in handling various types of non-linearity using only two parameters, namely spread parameter and SVM capacity. Therefore, we optimize these two hyper-parameters (capacity and spread) together with insensitivity range ($[-\varepsilon,\varepsilon]$) for each AU by performing cross-validation over the training sets.

4.3. Feature selection for regression

As described in Section 3.1 the original Gabor feature space has dimension 663,552, which is excessive to effectively train any classifier or regressor; furthermore most of the features are not informative and computations would be excessive. A state-of-the-art technique in data-driven AU detection is to apply AdaBoost feature selection scheme [20]: each weak classifier is trained with only one feature in every cycle of the algorithm, and the weights assigned to the training samples are adaptively updated.

For our regression problem, we can use the AdaBoost-generated feature sets as in the AU detection problem [20]. However, since these features are selected according to the best discrimination between negative and positive samples only, they may not be the best choices for intensity prediction. In other words we may find better feature sets if they are selected them according to their intensity prediction capabilities.

For this purpose we investigated two AdaBoost-based feature selection methods. Drucker's AdaBoost [25] regression technique incorporates sample distributions by an ad hoc adaption of the classification AdaBoost, where samples get weights proportional to their error. In our implementation of Drucker's AdaBoost for feature selection the weak learners were chosen as weighted linear least squares regressors to be trained on each feature. Thus the sample distribution weights that vary during the iterations were incorporated in the weighted least squares estimation and its simple structure permitted computationally efficient training over a high number of features. However, our preliminary experiments with Drucker's AdaBoost did not result in successful estimations and therefore we had to resort to another AdaBoost variant.

AdaBoost for thresholded-regression, namely AdaBoost.RT, was proposed to overcome some shortcomings of the previous boosting methods for regression [26]. By means of thresholding, the regression problem is reduced to a simple binary classification problem, i.e., only errors greater than a threshold in regression are treated as classification errors and play a role in training. The purpose of error thresholding is the same as ε-insensitive loss function used in SVMs (see Section 4.2, i.e., robustness). In our work we slightly modified the original AdaBoost.RT by using absolute error instead of relative absolute error to demarcate between well and poorly predicted samples. The authors in [26] divided the error by the target value, and thus put less emphasis on the error levels at high target values, which had motivations originating from their problem of hydrology engineering. However, since we do not want to bias our selected features in this way, alternatively we have normalized target intensity levels into the range [0,1]. Another difference is that, while we have used linear least squares regression as weak learners, they employed regression trees and artificial neural networks since they were not interested in feature selection. We designed AdaBoost.RT for feature selection as outlined below.

1. Inputs:
   - Set of m training samples $(\mathbf{x}_1, y_1)$, …, $(\mathbf{x_m}, y_m)$, where output $y \in R$
   - Feature vector composed of K features: $\mathbf{x_i} = [x_i^1, \ldots x_i^K]$
   - Weak regressor, $f(x)$: *Weighted Linear Least Squares Regression*
   - Threshold ϕ ($0<\phi<1$) for demarcating correct and incorrect predictions
2. Initialize:
   - Iteration $t=1$
   - Distribution weights of training samples $D_t(i)=1/m$ for all i
   - Error rate $\varepsilon_t=0$
3. Iterate while $t \leq T$ (number of selected features)
   - For each feature k (not selected)
     - Train weak regressor $f_t^k(x) \rightarrow y$ on feature x^k, providing it with distribution D_t
     - Calculate absolute error for each training sample as

       $$AE_t^k(i) = |f_t^k\left(x_i^k\right) - y_i|$$

     - Calculate the error rate of $f_t^k(x^k)$: $\varepsilon_t^k = \sum_{i:\ AE_t^k(i)>\phi} D_t(i)$
   - Choose the feature p_t and corresponding hypothesis $f_t^{p_t}$ with minimum error rate $\varepsilon_t^{p_t}$, $p_t = arg\,min_k(\varepsilon_t^k)$
   - Set $\beta_t = (\varepsilon_t^{p_t})^n$, where $n=$ power coefficient (e.g. linear, square or cubic)
   - Update distribution D_t as (includes normalization to be a probability distribution)

     $$D_{t+1} \propto D_t(i) \times \begin{cases} \beta_t, & AE_t^{p_t}(i) \leq \phi; \\ 1, & \text{otherwise.} \end{cases}$$

4. Selected feature set: $\{p_t\}_{t=1}^T$
5. The final hypothesis:

$$f_{final}(\mathbf{x}) = \frac{\sum_t \left(log\frac{1}{\beta_t}\right) f_t^{p_t}(x^{p_t})}{\sum_t \left(log\frac{1}{\beta_t}\right)}$$

There are several advantages of AdaBoost.RT. First, it is not as sensitive to noise and outliers, since sample weights are not updated proportionally to sample errors, thanks to the regression threshold. Second, unlike Drucker's AdaBoost, the weight updating parameter gives more emphasis to more difficult samples. Finally, it is able the handle weak learners even with error rates greater than 0.5, i.e., the algorithm does not have to terminate in these circumstances unlike many other boosting algorithms. In other words, we can keep on collecting features to any desired number even beyond the point where features do not necessarily satisfy the constraint of being better than 50%. This property may become especially useful in our case as our actual goal is to select features to be used for SVM-based regression algorithms rather than using the final hypothesis of the AdaBoost (fifth stage in the algorithm written above). Hence, even though some of the features with large error rates do not contribute much for the AdaBoost regression, we may benefit from them eventually in subsequent algorithms that use these selected features, i.e., SVMs as in our case.

On the other hand, a disadvantage of AdaBoost.RT is that, we have to find proper threshold values since the estimations are quite sensitive to the chosen threshold. If it is too low, majority of the samples are deemed incorrectly predicted; when it is too high, it is possible that only a few hard samples will get boosted, and because it is likely that they will be outliers, the estimator will be unstable. We have determined the threshold by cross validation. Also, there are several options for the power coefficient n. If we increase it, more weight is assigned to the harder samples for very low error rates ε_t. Since in our cross validation experiments higher values of n with different threshold values did not provide any improvement, we employed the linear model $n=1$.

5. Fusion by feature selection

We investigated the potential of fusing different types of features as they may contain complementary information. We implemented the feature fusion by AdaBoost feature selection. First, all the features to be fused, from different modalities and/or from their different representations of 3D, are pooled, and then we let AdaBoost to pick features from the pool efficiently. To make this point clear, we emphasize that we implement two-stage AdaBoost: The first stage is the feature selection from the raw 663,552-feature set, which is run separately for 2D, for 3D mean curvature, for 3D Gaussian curvature, etc. all resulting in sets of 200 features. The second stage pools various 200-feature sets and then runs AdaBoost on this modest pool. Otherwise the resulting merged feature pools, say 2D pool plus 3D pool, would become impractically large.

We have investigated two fusion tasks:

- **Fusion of 3D Geometry with 2D Luminance Data.** We expect that 3D and 2D modalities contain complementary information because factors like skin pigmentation and facial hair change the facial albedo, factors that can not be captured in3D modality. Most of the albedo variations occur on lips, eyes and eyebrows, facial features that have high leverage in recognition of expressions. Moreover, acquisition noise and the consequent smoothing operations on 3D may cause loss of some details like wrinkles. For this purpose we apply feature selection on the combined set of Gabor features of geometry and luminance data.
- **Fusion of 3D Geometry Features.** We explore fusion of mean curvature, Gaussian curvature, shape index and curvedness for possible performance improvement for 3D estimation. We not only fuse these three types of features, but also experiment their paired fusions to find out the most complementary combinations.

6. Experimental results and discussions

For testing our AU intensity predictors, we used 2902 images from the Bosphorus database. This subset of the Bosphorus database involves 25 AUs which occur at all the five intensity levels (see Section 2.1) and in AU various combinations. All of the samples of these 25 AUs available in the database are used in the experiments. Some sample cases are shown in Figs. 7 and 8 using 2D, 3D surface and mean curvature images. We train and test 25 intensity estimators (one for each AU) by 10-fold subject cross validation, and of course training subjects are not seen in test sets. To measure the performance of the intensity estimators we evaluate the Pearson linear correlation coefficient between AU intensity estimates and the discrete ground-truth AU intensity levels. The overall performance is calculated by weighted averaging proportional to the number of positive AU samples.

6.1. Best 3D geometry features for intensity estimation

We first evaluate the performance of the four basic 3D geometry features, either separately or in collaboration with each other. Table 1 compares their average correlation performances over the 25 AUs. Note that although the differences in average performances are small, there are individual AUs for which the performance difference is quite large. These results were obtained by ε-SVM-RBF regression on 200 Gabor magnitudes that are selected by AdaBoost.RT for each AU. We fused the features with AdaBoost.RT-based feature selection as explained in Section 5.

The reading of the table needs some clarification. The first column shows the performance for individual feature types (no fusion). Here it appears that the mean curvature is the best single feature type and the shape index is the worst. The bottom row of the table gives the performance results of nine fusion combinations. When we analyze the fusion results, we observe that each fusion combination performs better than its component features. The best performing combination pairs are the ones with mean curvature as it has already a markedly higher single performance. It is interesting to note that the fusion of the shape index and curvedness features yields the biggest increment in the performance. Recall from Section 3.2 that these two shape indicators are complementary to each other since they represent completely scale-dependent and scale-independent aspects of the local shape. The highest performance is obtained when mean curvature, Gaussian curvature and curvedness are fused, and no further improvement has been obtained by including the shape index, i.e., when all of four feature types are fused. Therefore, in the rest of the paper, we consider only this fusion triple when assessing the 3D modality in AU intensity estimation. The inferiority of the shape index may be related to the loss of curvedness information as explained in Section 3.2, since increasing intensities of facial actions can bend facial surfaces locally more, which increases the curvedness.

6.2. Assessment of the intensity estimation methods

In this section we evaluate the performances of seven intensity estimation methods for 3D, 2D and their fusion schemes. The correlations calculated over all AUs are listed in Table 2 where 2D represents luminance features, and 3D represents fusion of surface mean curvature, Gaussian curvature and curvedness features. The correlation results are grouped into three main categories: i) Estimators based on the use of SVM margins of the AU detectors either directly or in regression from; ii) Estimators based on the regression of features selected by AdaBoost for the sole purpose of AU detection; iii) Estimators based on regression on features selected specifically for intensity estimation via AdaBoost.RT. The scores of various feature selection methods indicate that 2D and 3D are on par, but that their fusion boosts performances. A more detailed 3D vs. 2D analysis as in Section 6.3 reveals that individual AU performances can vary considerably between 2D and 3D; a fact that the performance scores averaged over 25 AUs do not reveal.

In the first column in Table 2 we see the correlation performance of SVM-margins method which uses AU detector decision scores, and in the second column the performance of the logistic regression on these SVM-margins. Notice that in a previous study Bartlett et al. [5] have obtained a correlation performance of 0.53 with 2D luminance images over six AUs using linear SVM-margins. Using 2D data and RBF-SVM margins over the same set of six AUs (1, 2, 4, 5, 10 and 20), we gain 9 points (0.62); however, when the number of AUs rises to 25, the average performance falls back to 0.52; the application of logistic regression hardly improves it to 0.53 (Table 2).

On the other hand, regression on Gabor features instead of regressing on SVM-margins of AU detectors provides us with substantial gains in prediction performance. We provide results with respect to two types of automatic feature selection. The two middle columns show the performances of features selected by AdaBoost which is

Table 1
Average correlation (percentages) of the estimated intensities with the scores of the FACS annotator for different 3D geometry features and their fusions. ε-SVM-RBF regression is applied on 200 Gabor magnitudes that are selected by AdaBoost.RT for each AU. "✓" in a cell signifies that feature type is used in the fusion.

Single feature types		Feature type combinations								
Mean Crv.	**57.3**	✓	✓	✓				✓	✓	✓
Gaussian Crv.	53.5	✓			✓	✓		✓	✓	✓
Curvedness	53.0		✓			✓	✓		✓	✓
Shape Index	51.6			✓	✓		✓	✓		✓
		58.0	57.9	57.6	55.6	55.1	56.0	57.6	**58.4**	58.4

Table 2
Average correlation (percentages) of the estimated intensities with the scores of the FACS annotator. For all the methods 200 Gabor magnitudes are employed. While 2D represents luminance features, 3D represents fusion of surface mean curvature, Gaussian curvature and curvedness features.

	SVM Margins		AdaBoost Features		AdaBoost.RT Features		
Data	Direct	Logistic	ε-SVM-L	ε-SVM-R	Ada.RT	ε-SVM-L	ε-SVM-R
2D	52.2	52.8	53.9	**57.6**	47.8	52.2	55.1
3D	52.5	52.9	52.1	57.0	53.6	55.8	**58.4**
3D + 2D	56.0	56.4	58.9	62.1	58.2	60.7	**63.1**

actually used in the role of AU detection, and the last three columns belong to the AdaBoost.RT which picks features according to intensity regression criteria. We determined AdaBoost.RT thresholds according to cross validation performances (averaged over all AUs) as 0.05 for 3D and 0.03 for 2D (intensity levels are in the range [0 1]). It is observed that AdaBoost.RT features perform better than AdaBoost features when 3D data are used, either singly or in fusion with 2D. However, when 2D modality is employed alone AdaBoost.RT performs worse. This may be due to the difficulty of discriminating AU intensity levels based on luminance data by linear regression model since luminance not a direct measure of deformation and is subject to effects of shading, facial hair, and albedo variations.

Finally, we see that ε-SVM regressors yield higher correlations than AdaBoost.RT regression, and the best results are obtained by the non-linear RBF modeling. The best 2D, 3D and modality fusion results are 57.6, 58.4$_?$, $_?$ and 63.1$_?$ respectively.

6.3. Evaluation of 3D and 2D modalities for AU intensity estimation

In this section we analyze comparatively the performance of 2D and 3D data modalities for AU intensity estimation. Fig. 5a presents individual AU scores to reveal performance differentials between modalities since the overall performance averages eclipse these interesting aspects. The number of the available AU samples is inscribed on each AU bar. We see that, with 2D data, the upper face AUs, AU 4—Brow Lowerer, AU 5—Upper Lid Raiser, AU 6—Cheek raise, and AU 7—Lids Tight as well as AU 9—Nose Wrinkler and AU 20—Lip Stretcher achieve noticeably higher correlation than 3D data. On the other hand, 3D data seem to be more accurate for many lower face AUs, especially for AU 16—Lower Lip Depressor, AU 17—Chin Raiser, AU 18—Lip Pucker, AU 22—Lip Funneler, AU 10—Upper Lip Raiser, AU 25—Lips Part and AU 34—Puff, as well as for AU 2—Outer Brow Raise. Table 3 compares the intensity estimation performances on lower and upper face. While 2D achieves better performance (70.6) than 3D on upper face, 3D is better on lower face (52.6). Our opinion on the clear superiority of 3D on the lower part and 2D superiority on the upper part of the face is that, 3D sensing noise can be excessive in the eye region and 3D misses the eye texture information. On the other hand, larger deformations on the lower face make 3D more advantageous. Nevertheless, we see that correlations on upper face are significantly higher than the lower face for both modalities. This points out to the difficulties in intensity estimation for the lower face AUs.

In Fig. 5b we see that by means of modality fusion we are able to improve correlations beyond the best of 2D and 3D modalities. Correlation score with fusion on upper face is 74.4 whereas it is 54.3 for lower face (see Table 3). Fig. 6 shows the percentages of 3D and 2D features that are picked via AdaBoost.RT feature selection from the two modalities. The 3D percentage averaged over the 25 AUs is 78%$_?$. This percentage is obtained after averaging over 250 experimentation runs, i.e., 25 AUs×10 folds. For upper and lower face AUs, the percentage of captured 3D features were 61%$_?$ and 85%$_?$ respectively. Though 3D percentages are higher for both parts, notice the percentage of 2D features is much higher in upper face (39%$_?$) compared to their percentage in lower face (15%$_?$). Especially the two eye related AUs, AU 7—Lids Tight and AU 5—Upper Lid Raiser have about 80%$_?$ and 70%$_?$ 2D percentages respectively. This indicates again the importance of luminance on the upper face.

Table 3
Average correlation (percentages) of the estimated intensities over lower and upper AUs for 2D, 3D and their fusion.

Face part	#	2D Lum.	3D Geom.	3D + 2D Fusion
Lower AUs	4834	47.4	52.6	54.3
Upper AUs	3772	70.6	65.8	74.4

However, we also observe that for few AUs, e.g. AU 9—Nose Wrinkler and AU 20—Lip Stretcher, 2D features alone give considerably better result than 3D + 2D fusion. This may be due to inconsistencies between the weak learners employed in AdaBoost feature selection and the SVMs used as final estimators. AdaBoost selects features based on the assumptions made by the weak learners, which are linear estimators in our case. Thus, the selected features are not necessarily the optimal features for the SVM estimators. Because of this reason, 3D features might have been selected instead of many 2D features according to the linear estimation capability in the fusion process, even though those neglected 2D features might have been more beneficial when non-linear SVM estimation is performed.

7. Conclusion

In this paper we have comparatively investigated person-independent intensity estimation of 25 AUs from still images on 2D and 3D modalities. Our intensity estimator operates in a data-driven manner, thus does not require the aid of landmarks and allows unbiased (without face modelling) comparison of the modalities. There is only one other person-independent study in the literature on of AU intensity estimation, which in particular uses SVM margins and Gabor features and addresses only eight AUs [5]. Our proposed intensity estimator based on regression of appearance features proves to be superior to that based on SVM margins; on the average AU correlation performances, we obtained about 5.5%$_?$ improvement in 2D and about 11%$_?$ improvement by also incorporating 3D modality. To the best of our knowledge this is the first study to employ regression for intensity estimation, whether for subject-independent or for subject-dependent estimation.

Non-linear regression over the Gabor features by RBF-SVMs improves the results for both 2D and 3D data modalities. However, when it comes to feature selection, regression-based selection by AdaBoost.RT was more beneficial for the 3D modality, whereas the use of the selection by AdaBoost focused to AU detection yielded better results for the 2D modality. This may be because the luminance data is not a direct measurement of facial deformations in contrast to 3D capture of the facial surface. Luminance data are under the influence of several confounding factors like facial hair, illumination and pose. More training data may be needed for the 2D modality to be able to reliably select intensity discerning features.

For 3D estimation, we considered the basic four local shape indicators via Gabors as the features. It was found that the best intensity estimation performance was obtained by the mean curvature representation. However, with experiments on various feature fusion combinations, we obtained the top performance with fusion of mean curvature, Gaussian curvature and curvedness features. Shape index seems to be not as useful as the three representations above, and it did not also contribute to fusion performances.

Our experiments show that 3D is not necessarily better than 2D; in fact, while 3D show improvements on some AUs it incurs into performance drops on some other AUs. We have conjectured that these drops in performance may be because of 3D acquisition noise in eye regions, because texture is missing, and also because FACS groundtruths were scored on 2D appearance data, which could have created a bias toward 2D modality. Especially, performances and percentages of the fused features for the eye related AUs reveal the importance of eye texture. Eventually, fusion of the two modalities boosts the 3D estimation performances further.

As a future work, we will consider assessment of AU detection and intensity estimation in spontaneous expressions, which is important for development of real-life systems. This is a more challenging problem for several reasons. Spontaneous expressions are accompanied by uncontrolled head movements. They typically happen in relatively lower intensities, i.e., are more subtle than the posed ones. Although 3D spontaneous databases are currently not available and

3D acquisition devices have some drawbacks, such as light projection onto subject's face and higher cost of real-time 3D video, recent progress [27,28] points out the possibility of such databases. Therefore, our work will progress toward 3D spontaneous expressions.

Acknowledgements

This work is supported by the Bogazici University fund 09HA202D, by the TUBITAK project 107E001 grants and by the Turkish State Planning Organization (DPT) under the TAM Project, number 2007K120610. Part of the calculations reported in this paper were performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TR-Grid e-Infrastructure).

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, J. Taylor, Emotion recognition in human–computer interaction, Signal Process. Mag. IEEE (January 1, 2001) 32–80.

[2] P. Ekman, W.V. Friesen, J.C. Hager, Facial Action Coding System, The Manual on CD ROM, , 2002.

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

[4] N. Doulamis, An adaptable emotionally rich pervasive computing system, European Signal Processing Conference (EUSIPCO), Florence, Italy, 2006.

[5] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, J.R. Movellan, Automatic recognition of facial actions in spontaneous expressions, J. Multimed. 1 (6) (2006) 22–35.

[6] A. Savran, B. Sankur, M.T. Bilge, Estimation of facial action intensities on 2d and 3d data, European Signal Processing Conference (EUSIPCO), Barcelona, Spain, 2011.

[7] S. Koelstra, M. Pantic, I. Patras, A dynamic texture-based approach to recognition of facial actions and their temporal models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1940–1954.

[8] S. Lucey, A.B. Ashraf, J. Cohn, Investigating spontaneous facial action recognition through AAM representations of the face, Face Recognition Book, Pro Literatur Verlag, Mammendorf, Germany, 2007.

[9] Y. Tong, W. Liao, Q. Ji, Facial action unit recognition by exploiting their dynamic and semantic relationships, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1683–1699.

[10] M. Pantic, L.J.M. Rothkrantz, An expert system for recognition of facial actions and their intensity, AAAI/IAAI, 2000, pp. 1026–1033.

[11] P. Yang, Q. Liu, D.N. Metaxas, Ieee rankboost with *l-1* regularization for facial expression recognition and intensity estimation, International Conference of Computer Vision (ICCV), 2009.

[12] M. Mahoor, S. Cadavid, D. Messinger, J. Cohn, A framework for automated measurement of the intensity of non-posed facial action units, IEEE CVPR Workshop on Human Communicative Behaviour Analysis, 2009.

[13] J. Wang, L. Yin, X. Wei, Y. Sun, 3D facial expression recognition based on primitive surface feature distribution, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, 2006.

[14] H. Soyel, H. Demirel, Facial expression recognition using 3d facial feature distances, in: International Conference on Image Analysis and Recognition, (ICIAR), International Conference on Image Analysis and Recognition, (ICIAR), Montreal, Canada, 2007.

[15] I. Mpiperis, S. Malasiotis, V. Petridis, M.G. Strintzis, 3D facial expression recognition using swarm intelligence, IEEE International Conference on Accoustics, Speech and Signal Processing (ICASSP), Las Vegas, Nevada, USA, 2008.

[16] I. Mpiperis, S. Malassiotis, M. Strintzis, Bilinear elastically deformable models with application to 3d face and facial expression recognition, IEEE International Conference on Automatic Face and Gesture Recognition (FG), Amsterdam, Netherlands, 2008.

[17] H. Tang, T. Huang, 3D facial expression recognition based on automatically selected features, IEEE CVPR Workshop on 3D Face Processing, Anchorage, Alaska, USA, 2008.

[18] Y. Sun, M. Reale, L. Yin, Recognizing partial facial action units based on 3d dynamic range data for facial expression recognition, IEEE International Conference on Automatic Face and Gesture Recognition (FG), Amsterdam, Netherlands, 2008.

[19] F. Tsalakanidou, S. Malassiotis, Real-time 2d+3d facial action and expression recognition, Pattern Recognit. 43 (5) (2010) 1763–1775.

[20] A. Savran, B. Sankur, M.T. Bilge, Comparative evaluation of 3d vs. 2d modality for automatic detection of facial action units, Pattern Recognit. 45 (2) (2012) 767–782.

[21] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, J. Movellan, Toward practical smile detection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 2106–2111.

[22] A. Telea, An image inpainting technique based on the fast marching method, Graph. Tools 9 (1) (2004) 25–36.

[23] J.J. Koenderink, A.J. van Doorn, Surface shape and curvature scales, Image Vis. Comput. 10 (8) (1992) 557–564.

[24] V. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

[25] H. Drucker, Improving regressors using boosting techniques, International Conference on Machine Learning, San Francisco, CA, USA, 1997.

[26] D.L. Shrestha, D.P. Solomatine, Experiments with adaboost.rt, an improved boosting scheme for regression, Neural Comput. 18 (2006) 1678–1710.

[27] D. Modrow, C. Laloni, G. Doemens, G. Rigoll, A novel sensor system for 3d face scanning based on infrared coded light, SPIE, 2008.

[28] N. Karpinsky, S. Zhang, High-resolution, real-time 3d imaging with fringe analysis, J. Real-Time Image Process. (2010) 1–12.