# ESTIMATION OF FACIAL ACTION INTENSITIES ON 2D AND 3D DATA

*Arman Savran*[1], *Bülent Sankur*[2], *and M. Taha Bilge*[3]

Electrical and Electronics Engineering[1,2] Department of Psychology[3]
Boğaziçi University, Istanbul, Turkey
arman.savran@boun.edu.tr[1] bulent.sankur@boun.edu.tr[2] taha.bilge@boun.edu.tr[3]

## ABSTRACT

The paradigm of Facial Action Coding System (FACS) offers a comprehensive solution for facial expression measurements to researchers. FACS defines atomic expression components called Action Units (AUs) and describes their strength on a five-point scale. Despite considerable progress in AU detection, the AU intensity estimation has not been much investigated. We propose SVM-based regression on AU feature space, and investigated person-independent estimation of 25 AUs that appear singly or in various combinations. Our method is novel in that we use regression for estimating intensities and comparatively evaluate the performances of 2D and 3D modalities. The proposed technique shows improvements over the state-of-the-art person-independent estimation, and that especially the 3D modality offers significant advantages for intensity coding. We have also found that fusion of 2D and 3D can boost the estimation performance, especially when modalities compensate for each other's shortcomings.

## 1. INTRODUCTION

Automated measurement of facial actions has many potential applications for intelligent human-computer interfaces and in behavioral science. Extracted facial actions can be used, for instance, to infer the emotional state of a person, to detect whether the driver appears too tired, or even they can be used for generating performance-driven facial animations. Facial Action coding System (FACS) [2] is the most common facial action measurement methodology which involves 44 Action Units (AUs) related to facial muscle activations that can be visually discerned. Being composed of very extensive set of rules, FACS requires certified human coders and coding is a very time consuming process. These issues also motivates development of automatic coders to be used in facial behavior research.

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 measurement of intensities can be useful in behavior research and for improved FACS coding. For instance, if the expression is surprise, and if AU 5 - Upper Lid Raiser is available, then, it should only be at B level. Second, estimating strength of the AUs we yield more information about mental state and emotional involvement of a subject. Furthermore, AU intensity outputs can be a basis for studying AU dynamics.

Most of the the works on expression intensity have investigated relationships between classification decision scores and intensities. For instance, Bartlett *et al.* [1] investigated correlations between intensity levels and SVM classifier margins of their Gabor filter-based detectors. They reported moderate to high correlations for several AUs. One criticism of using classifier scores is that they do not incorporate intensity knowledge. Yang *et al.* [9] have used the output scores of RankBoost based expression classifiers to better deal with intensity variations. They train RankBoost classifiers with onset to apex ranked image sequences, in order to rank the image pairs according to their emotion intensities. They obtained better image pair ordering performance than the linear SVM-margin approach in *et al.* [1]. However, though related, correctly identifying ranking of image pairs in a sequence whose intensity increases monotonically is quite a different problem than estimating intensities directly from single images. Besides, some additional techniques should be figured out to convert ordering of image pairs into intensities, also in a way that we have consistent intensity measurements between sequences. Recently Mahoor *et al.* [5] studied measurement of AU 6 and AU 12 intensities over six subjects via person-specific AAMs. They approach to intensity estimation as a classification problem and apply six level SVM classifiers by one-against-one technique. For feature extraction they perform AU specific dimension reduction by applying regularized locality preserving indexing on appearance data, and use delta features (*i.e.*, by neutral face feature subtraction).

The main novelty of our work is in the application of a nonlinear regression scheme for AU intensity estimation. A second contribution is the detailed investigation of 3D modality for improved AU intensity estimation and the comparison and eventual fusion of 3D with 2D modality data. Finally, we experiment with 25 AUs, a much bigger variety, by a factor of three, of AU types treated in previous works in the literature.

## 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 interpretation of these levels are as follows: the A level refers to a trace of the action; B, slight evidence; C, marked or pronounced; D, severe or extreme; and E, maximum evidence. 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. The relationship between the scale of evidence and the scoring levels is a bit different for some AUs. Scoring criteria are listed in the FACS manual [2] for each

AU, though sometimes modified criteria are used depending on the AU combinations.

By definition, each level denoted by a letter refers to a range of appearance changes, that is, they do not correspond to a single strength of AU. Notice that the intensity scale is not divided into uniform intervals; the C and D levels cover a larger range of appearance changes. The relationship between the scale of evidence and intensity scores is depicted in Figure 1.

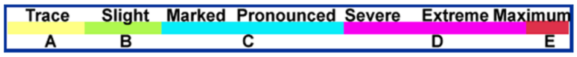


Figure 1: The scale of evidence and intensity scores [2].

FACS manual states that scoring of lower intensities, the A and B levels, requires particularly careful examination, and A level actions can be scored reliably only by very experienced coders. While scoring of lower intensities may not be easy, distinguishing the E level AUs can be difficult as well since the intense muscular contractions of the E level combine with the person's individual physical characteristics causing variability on the appearance changes across different people. Examples of some low level and high level AU samples are shown in Figure 5.

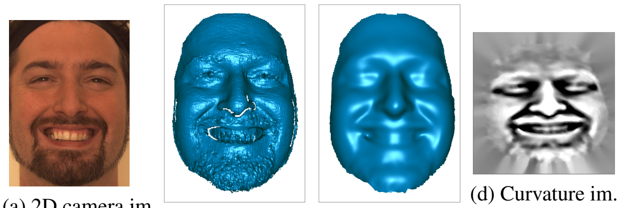
### 2.2 Expression Database

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. There are two reasons that makes this type of estimation problem more challenging. First, person-independent estimation means existence of additional variability due to different subjects. Second, without video we are deprived of the rich dynamic information. We worked on the Bosphorus Database [7] which not only has the intensity scores for all of its face samples and all AUs, but also it has a rich repertoire of AUs.

Bosphorus database contains 105 subjects with various expressions for which ground-truth FACS codes were attributed by a certified FACS coder. The images were acquired under good illumination conditions, in almost frontal poses, *i.e.*, with mild 3D rotations. In this database, faces have also their 3D scans, captured by a structured light system. The number of points on the resulting 3D faces varies between 30K and 50K. This gives us the opportunity to estimate the AU intensities based on 3D geometry data.

## 3. FEATURE EXTRACTION

We apply Gabor wavelets and extract AU specific feature sets. The procedure is as follows. First, the 2D face images are registered and normalized to $96 \times 96$ size images by means of translation, rotation and scaling using the eye centers. Then, Gabor features are extracted from normalized faces in eight directions and nine scales. In these scales, the wavelengths vary in half octave intervals from 2 to 32 pixels. This yields a redundant set of $9 \times 8 \times 96 \times 96 = 663,552$ features, out of which we select only 200 Gabor magnitudes using AdaBoost feature selection scheme where the weak classifiers are nearest mean classifier trained for each feature to decide the presence of the target AU. This process is repeated for each AU to select AU specific features.



(a) 2D camera im. (b) Raw 3D face (c) Filtered face (d) Curvature im.

Figure 2: Illustration of pre-processing steps of 3D surface data to produce surface curvature images. Input data (b) is filtered (c) and its surface curvature is computed. The curvature information is projected onto 2D domain and extrapolation is performed (d). The FACS code for this particular expression is 6B+7C+12D+16A+25D.

Since we are interested in measuring AU intensities on 3D faces as well, we replicate the same set of operations. 3D faces are first smoothed, registered and normalized in 3D. However, the Gabor feature extraction is not done directly on 3D surfaces, but on 2D maps of the mean curvature maps computed from 3D data [7]. Figure 2c illustrates the raw 3D data, its smoothed version and the curvature field in 2D. In addition, we evaluate the fusion of 2D and 3D modalities. For this purpose, we apply AdaBoost to select 200 features from the combined feature set of Gabor magnitudes that are extracted from luminance and curvature data.

## 4. REGRESSION BASED INTENSITY ESTIMATION

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

### 4.1 Regression on SVM Margins

It was shown in [1] 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. Figure 3 shows the scatter of AU 12 decision scores (RBF-SVM margins) for 2D and 3D modalities as box-and-whisker plots. As expected higher AU intensities correspond to bigger SVM scores. Also, there are substantial overlaps between some adjacent intensity grades. Note that the medians of the distributions do differ significantly at the 5% significance level if their notches (first quartile-to-median and median-to-third quartile ranges) do not overlap. One explanation for these overlaps is that the SVM algorithm was designed to detect an AU, but not necessarily to estimate its intensity. Furthermore, whatever technique is employed substantial overlaps is perhaps unavoidable due to the fact that a strict separation between intensities is difficult to achieve since FACS does not define a quantitative measure between levels. Finally, in person-independent intensity estimation, one is confronted with more of variability since different subjects can enact AUs differently and facial surface and texture vary from subject to subject. These factors make the estimation problem more challenging. Though not shown here, we have also ob-

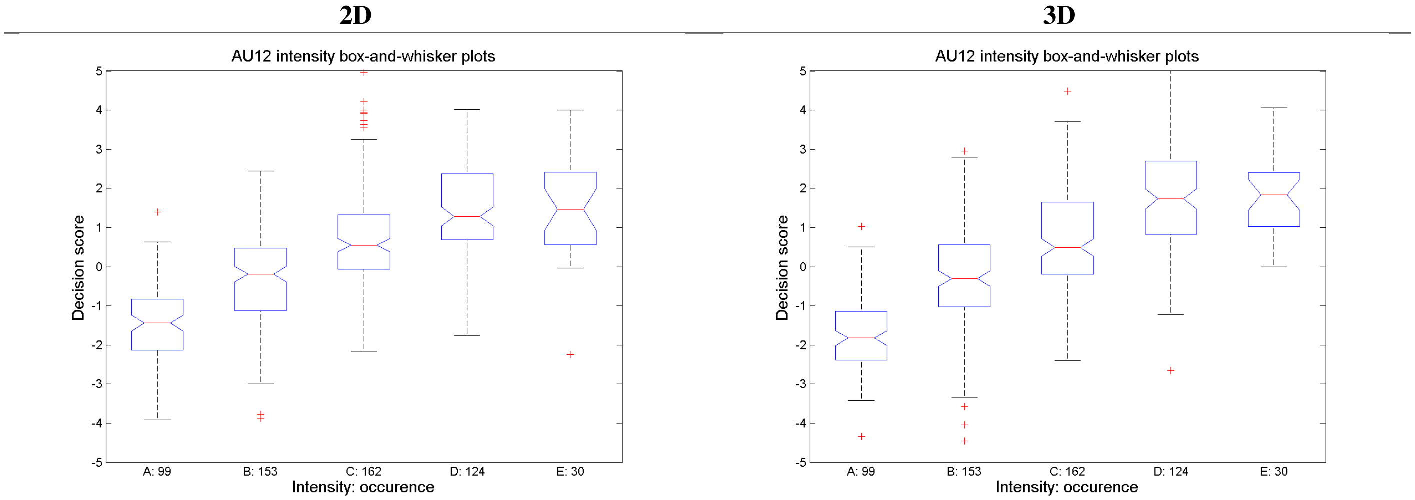


Figure 3: Distributions of decision scores (RBF-SVM margins) of AU 12 - Lip Corner Puller for 2D and 3D data modalities shown as box-and-whisker plots (central mark: median, box: interquartile range, whiskers: extreme values, '+': outlier)

served that overlap of the scores considerably changes for some AUs depending on the 3D and 2D modalities. 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 $r$:

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

### 4.2 Regression on Image Features

Although the scores, *i.e.*, the distance to the hyperplanes in SVMs designed for AUs imply stronger evidence for these AUs, proportionality of SVM scores to intensities is not guaranteed since the support vectors were 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 mean curvature field.

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 $\varepsilon$-insensitive loss function [8]. $\varepsilon$-SVM regression is appropriate because, first, high dimensionality of the input space is not an issue for SVMs, and second, the $\varepsilon$-insensitive loss function is robust and generates a smooth mapping.

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

$$f(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) + b \tag{2}$$

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. Recall that $\mathbf{x}$ represents the vector of 200 Gabor features that were also used for AU detection.

In our study we investigated both linear SVM and SVM with nonlinear kernels of the Gaussian radial basis function (RBF) variety. Advantage of RBF is its ability in handling various types of non-linearities despite having single spread parameter. Depending on this parameter and SVM capacity many non-linearities can be captured. Therefore, we optimize these two hyper-parameters also together with insensitivity range ([-$\varepsilon$,$\varepsilon$]) for each AU by performing cross-validation over the training sets.

## 5. EXPERIMENTAL RESULTS AND DISCUSSIONS

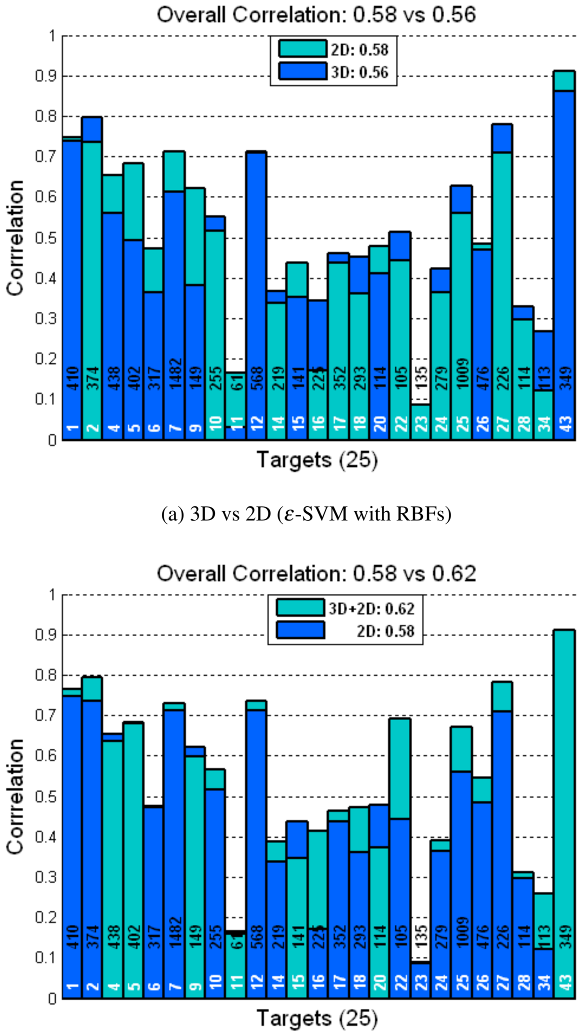
For testing our AU intensity predictors, we used 2902 images from the Bosphorus database. This subset of the Bosphorus dataset involves 25 AUs which occur in various intensities and combinations. Some sample images are shown in Figure 5. We train and test 25 AU detectors and intensity estimators by 10-fold subject cross validation such that training subjects are not seen in the test sets. The performance of 2D and 3D AU detectors are measured by the Area under the Curve measure, denoted as AuC, where the curve is that of Receiver Operating Characteristics (ROC), that is hit versus false alarm rates as a function of the threshold. This area measure is equivalent to theoretical maximum achievable correct rate of a detector. To measure the performance of the intensity estimators we evaluate correlation coefficient between AU intensity estimates and the discrete ground-truth AU intensity levels. The overall performances are calculated by weighted average according to the number of positive AU samples.

The correlations calculated over all AUs are listed in Table 1. In the first column we see the performance of SVM-margins method, and in the second column the performance of the logistic regression on SVM-margins. Notice that in a previous study Bartlett *et al.* [1] have obtained a correlation performance 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 AUs (1, 2, 4, 5, 10 and 20) we have obtained 0.62, however, over 25 AUs the average performance is 0.51. When we apply logistic regression,

|  | *SVM Margins* |  | *Image Features* |  |
|---|---|---|---|---|
| **Data** | **Direct** | **Logistic** | **$\varepsilon$-SVM-Lin** | **$\varepsilon$-SVM-RBF** |
| **2D** | 0.51 | 0.53 | 0.54 | 0.58 |
| **3D** | 0.50 | 0.51 | 0.53 | 0.56 |
| **3D+2D** | 0.53 | 0.55 | 0.59 | **0.62** |

Table 1: Correlation of the estimated intensities with the scores of the FACS annotator.

we obtain 0.53 (Table 1). On the other hand, we see that the improvements with $\varepsilon$-SVM regressor with image features yield higher correlations, and the non-linear RBF modeling achieves 0.58 correlation. The best overall result, 0.62, is obtained by fusion using $\varepsilon$-SVM with RBFs. We also experimented with different number of features and found that using more than 200 features does not improve: with 400 features the results are 0.59, 0.57 and 0.62 for 2D, 3D and fusion respectively.



(a) 3D vs 2D ($\varepsilon$-SVM with RBFs)

(b) Fusion vs 2D ($\varepsilon$-SVM with RBFs)

Figure 4: Performance (correlation) comparison between 3D vs 2D vs fusion. The AU code and the total number of occurrences are inscribed in the bars.

When we compare the performance of data modalities we see that 3D brings some improvement on averaged results only when used in conjunction with 2D data. Overall averaging may hide some important information, hence we compare intensity estimation for each AU over 2D and 3D data modalities in Figure 4a. The number of the available AU samples are inscribed on each AU bar. We see that, with 2D data, most of the upper face AUs, AU 4 - Brow Lowerer, AU 5 - Upper Lid Raiser, AU 6 - Cheek Raise, AU 7 - Lids Tight and AU 43 - Eye Closure, as well as AU 9 - Nose Wrinkler achieve noticeably higher correlation than 3D data. On the other hand, 3D data seems to be more convenient for many lower face AUs, especially for AU 16 - Lower Lip Depressor, AU 18 - Lip Pucker, AU 22 - Lip Funneler, AU 25 - Lips Part, AU 27 - Mouth Stretch and AU 34 - Puff, as well as for AU 2 - Outer Brow Raise.

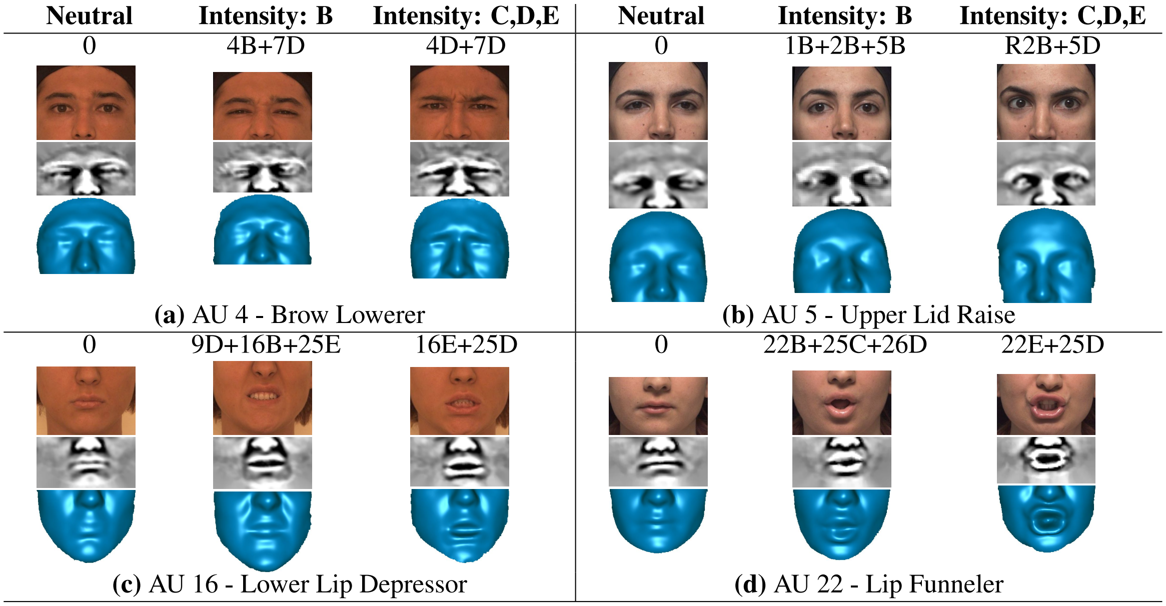
In contrast to intensity estimation, the improvements in overall AU detection performances by 3D data are much more substantial. The AuC detection results averaged over 25 AUs are 93.5%, 95.5% and 96.6% for 2D (luminance) data, 3D data (mean curvature) and for their fusion, respectively. Nevertheless, from Figure 4a it is understood that the this overall performance contrast between detection and estimation is actually not due to the inferiority of 3D modality. In fact, the advantages and disadvantages of the 3D modality for intensity estimation conforms to the results of detection for most of the AUs, however higher performance drops on certain AUs that have much more samples, such as AU 7, inverts the overall performances. We present the AU detection problem with 3D and 2D modalities in detail in [7]. One expects normally that 3D data would be more informative for he intensity estimation problem. Explanation for why this promise is not fulfilled are as follows. 3D sensing noise is excessive in the eye region and 3D misses the eye texture information. Moreover, the ground-truth data in manual FACS scoring is generated based on the observation of 2D appearances, which may generate a bias in favor of 2D. It is imaginable that the FACS annotator could have defined the intensity labels slightly differently using 3D data.

From Figure 4b we see that by means of modality fusion we are able to preserve the highest correlations of 2D and 3D modalities in general. However, interestingly, even though the correlation values of AU 22 are around 0.5 for 2D and 3D, it is boosted to 0.7 with fusion. These results shows the importance of fusion with 3D.

## 6. CONCLUSION AND FUTURE WORK

In this paper we investigated person-independent intensity estimation of 25 AUs from still images comparatively on 2D and 3D modalities. Our intensity estimator operate in a data-driven manner, thus do not require the aid of landmarks. The only other person-independent study in the literature on estimation of AU intensities apply SVM margins and Gabor features and address eight AUs [1]. Our proposed intensity estimator based on regression of appearance features proves to be superior to that based on SVM margins, both for 2D and 3D data modalities. To the best of our knowledge we are the first one to employ regression for intensity estimation, whether for subject-independent or for subject-dependent estimation.

Our 3D experiments show improvements on some AUs but also performance drops on some other AUs, both in the detection and intensity estimation problems. However, when



**(a)** AU 4 - Brow Lowerer

**(b)** AU 5 - Upper Lid Raise

**(c)** AU 16 - Lower Lip Depressor

**(d)** AU 22 - Lip Funneler

Figure 5: 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 and lower face action units together with the neutrals from the same subject.

3D is fused with 2D luminance images, the overall performance increases significantly. We have observed that whenever a modality is better for detection of an AU, its intensity estimation is also superior in the same modality. However, the performance drop in intensity estimation for certain AUs with 3D data is more pronounced as compared to the performance differential for detection. As discussed in Section 5, we have conjectured that this may be because of 3D acquisition noise in eye regions, since texture is missing, and also because FACS ground-truths were scored on 2D appearance data, which could have created a bias toward 2D modality.

There are several directions of future work of this problem. First of all, while we have used features optimally selected for AU detection, it is possible to redesign features specifically for intensity estimation.

Person-independent AU intensity estimation must deal with the confounding factor of subject variability. To reduce the portion of variability due to attributes of individuals, one can subtract the neutral face of the subject, whenever available. A future study can reveal if intensity estimation benefits from the subtraction of the neutral.

Assessment of AU detection and intensity estimation in spontaneous expressions 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. However, though 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 [6, 4] points out the possibility of such databases. Therefore, our work will progress toward 3D spontaneous expressions.

## 7. ACKNOWLEDGMENT

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.

## REFERENCES

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

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

[3] B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36(1):259–275, 2003.

[4] N. Karpinsky and S. Zhang. High-resolution, real-time 3d imaging with fringe analysis. *Journal of Real-Time Image Processing*, pages 1–12, 2010.

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

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

[7] A. Savran, B. Sankur, and M. T. Bilge. Facial action unit detection: 3d versus 2d modality. In *CVPR Workshop on Human Communicative Behavior Analysis*, San Francisco, California, USA, 2010.

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

[9] P. Yang, Q. Liu, and D. N. Metaxas. Rankboost with $l_1$ regularization for facial expression recognition and intensity estimation. In *ICCV*, September 2009.