Cross-dataset learning and person-specific normalisation for automatic Action Unit detection

Tadas Baltrušaitis and Marwa Mahmoud and Peter Robinson
Computer Laboratory, University of Cambridge, United Kingdom

***Abstract*— Automatic detection of Facial Action Units (AUs) is crucial for facial analysis systems. Due to the large individual differences, performance of AU classifiers depends largely on training data and the ability to estimate facial expressions of a neutral face. In this paper, we present a real-time Facial Action Unit intensity estimation and occurrence detection system based on appearance (Histograms of Oriented Gradients) and geometry features (shape parameters and landmark locations). Our experiments show the benefits of using additional labelled data from different datasets, which demonstrates the generalisability of our approach. This holds both when training for a specific dataset or when a generic model is needed. We also demonstrate the benefits of using a simple and efficient median based feature normalisation technique that accounts for person-specific neutral expressions. Finally, we show that our results outperform the FERA 2015 baselines in all three challenge tasks - AU occurrence detection, fully automatic AU intensity and pre-segmented AU intensity estimation.**

I. INTRODUCTION

Over the past few years, there has been an increased interest in machine understanding and recognition of affective and cognitive mental states, especially based on facial expression analysis [20]. As the face is considered the main channel of nonverbal communication, facial expression analysis is used in different applications to facilitate human computer interaction [3], [17]. Moreover, automatic analysis of facial expressions can be used as a tool in studying some medical conditions, such as depression [10].

Automatic detection and analysis of facial Action Units [7] (AUs) is one of the main building blocks in automatic facial expressions analysis. This includes detecting AUs as they occur on the face and estimating their intensities, which would in turn allow us to analyse their occurrence, co-occurrence and dynamics.

Yet, there are a lot of challenges in automatic detection of AUs, namely: unbalanced training datasets, individual differences (difficulty in finding a universal reference of the neutral expression), pose variation and occlusion.

In this paper, we present an automatic real-time approach for AU detection based on appearance and geometric features. We address the individual difference challenge by presenting and comparing detection results when a person-specific normalisation approach is employed. We show that our approach generalises well when trained on one dataset and tested on a different one. We argue that - when a generalisable approach is desired - cross-dataset learning can improve the performance of AU detection, especially with the difficulty of having a completely balanced dataset of natural expressions.

A high level overview of our AU detection and intensity estimation system can be seen in Figure 1.

The main contributions of our work are as follows: demonstrating how occurrence detection can be significantly improved for certain AUs by using person-specific neutral expression normalisation; demonstrating the benefits of using multiple datasets for generic model training; presenting a full AU detection pipeline that is capable of running real-time (20-30 fps). Finally, all of our training and testing code is available to the research community[1].

II. PREVIOUS WORK

AU recognition has received a lot of attention over the past years. We refer the reader to recent surveys and challenges [6], [22], [23], [25]. In the following paragraphs, we review the most relative work to ours.

Wu *et al.*[26] demonstrated that it is possible to build AU recognition systems that generalise across datasets. They, however, found that performance was hugely affected by the training dataset (and in some cases the generalisation is very poor) and that retraining approaches on target data were helpful. They also found that the datasets need to be carefully balanced in order to achieve a good performance. Our work confirms some of their findings, demonstrating that generalisation is indeed possible.

Li *et al.*[14] proposed a technique for building generic models for AU recognition by using generic domain knowledge that governs AU behaviours and showed that their results improve the model's ability to generalise across datasets. In our work, we improve the generalisability by learning the dimensionality reduction technique from a broad set of datasets.

Chu *et al.*[4] demonstrated how to personalise a generic AU classifier in an unsupervised manner. They introduced a Selective Transfer Machine (STM), to personalise a generic classifier by attenuating person-specific biases. STM achieved this effect by simultaneously learning a classifier and re-weighting the training samples that are most relevant to the test subject. In contrast to their work, our work attempts to perform person normalisation in a much simpler manner and is able to adapt to a person online.

Similar to our work, Jeni *et al.*[12] used person-normalised appearance features for AU intensity estimation, but they did not report on the effects of such normalisation.

[1] https://github.com/TadasBaltrusaitis/FERA-2015

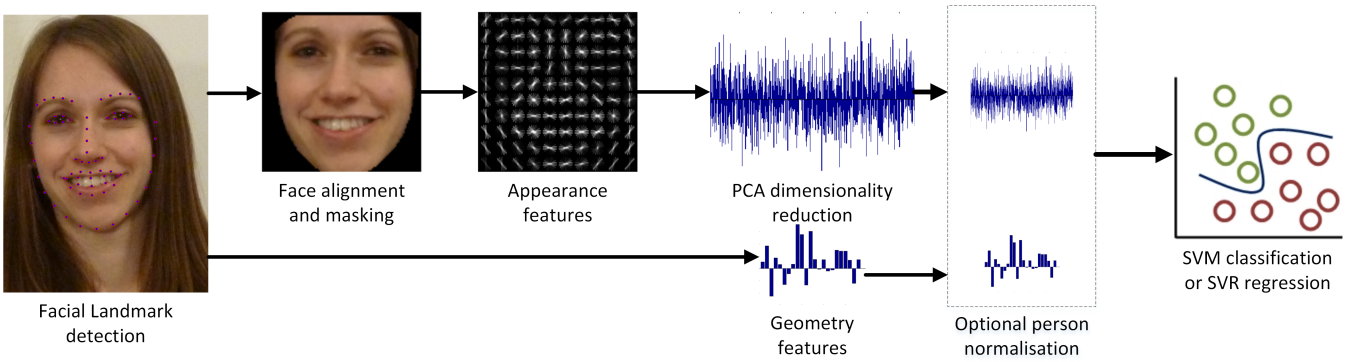

Fig. 1: Overview of the AU detection or intensity estimation pipeline.

III. Data

In our experiments, we mainly used three datasets: DISFA [16], BP4D-Spontaneous [27] and SEMAINE [18]. All three of the datasets consist of video data of people responding to emotion-elicitation tasks.

The BP4D dataset includes videos of 41 participants (21 used for training models and 20 for validating them). It includes annotations for 11 AUs for occurrence and 5 AUs for intensities. In total, the dataset consists of $150k$ high resolution AU labelled images. The annotated SEMAINE subset [24] - used in our work - contains recordings of 31 participants (15 for training and 16 for validation). It consists of one minute long recordings at $50Hz$, leading to $93k$ frames labelled for 5 AU occurrences. More information on the challenge data can be found in Valstar *et al.*[24].

DISFA contains videos of 27 participants (14 used for training and 13 for validation). It includes 4 minute-long videos of spontaneous facial expression, annotated for 12 AUs. DISFA contains over $130k$ frames. For every video frame, the intensity of 12 AUs was manually annotated on a six-point ordinal scale.

All of the datasets share three AUs in common (2, 12, and 17). SEMAINE and DISFA share AUs 2, 12, 17, 25. BP4D and DISFA share AUs 1, 2, 4, 6, 12, 15, 17. This allows us to experiment with cross-dataset training generalisation.

IV. Feature extraction

In our work, we use two main types of features: appearance and geometry ones. In order to extract them, we rely on tracking certain landmarks on a face, followed by face alignment. The following sections describe our features in more details.

A. Face tracking

We use Constrained Local Neural Field (CLNF) [2] facial landmark detector and tracker for face tracking and to extract geometry based features (explained in Section IV-D). We use the open source CLNF implementation [2]. It uses a Structural SVM face detector [13], followed by CLNF.

CLNF is an instance of a Constrained Local Model (CLM) [5], that uses more advanced patch experts and optimisation function. The model we used was trained on Multi-PIE [11] and *in-the-wild* [21] facial datasets.

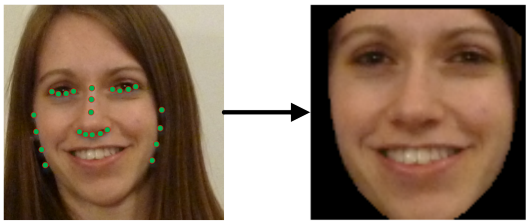

Fig. 2: Example of stable points used for alignment of the face to a common reference frame, followed by masking.

The CLM model we use can be described by parameters $\mathbf{p} = [s, \mathbf{R}, \mathbf{p}, \mathbf{t}]$ that can be varied to acquire various instances of the model: the scale factor s; object rotation $\mathbf{R}$ (first two rows of a 3D rotation matrix); 2D translation $\mathbf{t}$; a vector describing non-rigid variation of shape $\mathbf{p}$. The point distribution model (PDM) is:

$$\mathbf{x}_i = s \cdot \mathbf{R}(\bar{\mathbf{x}}_\mathbf{i} + \mathbf{\Phi}_i \mathbf{p}) + \mathbf{t}. \quad (1)$$

Here $\mathbf{x}_i = (x, y)$ denotes the 2D location of the i^{th} feature point in an image, $\bar{\mathbf{x}}_i = (X, Y, Z)$ is the mean value of the i^{th} element of the PDM in the 3D reference frame, and the vector $\mathbf{\Phi}_i$ is the i^{th} eigenvector obtained from the training set that describes the linear variations of non-rigid shape of this feature point, and the vector $\mathbf{\Psi}_i$ is the i^{th} eigenvector that describes the linear variations of non-rigid shape.

In CLM (and CLNF), we estimate the maximum *a posteriori* probability (MAP) of the face model parameters $\mathbf{p}$ given an initial location of the parameters determined by a face detection step.

B. Alignment and masking

In order to better analyse the texture of the face, we need to map it to a common reference frame and to remove changes due to scaling and in plane rotation. To do this, we used a similarity transform from the currently detected landmarks to a representation of frontal landmarks from a neutral expression (a projection of mean shape from a 3D PDM). The resulting is a 112×112 pixel image of the face

with 45 pixel interpupilary distance. To compute the similarity transformation, we used Procrustes superimposition that minimised the mean square error between aligned pixels.

In order to reduce the effect of large facial expressions (such as mouth opening and brow raises) on the similarity transform, we only use the most stable facial landmark points for the similarity transform. To determine such stable points we observed the most stable CLNF detected landmark points under facial expression on the on CK+ dataset [15] . CK+ contains videos of people showing a number of facial expressions whilst their head pose stays still and most of the expressions start from a neutral frame. The most stable points from CLNF tracker can be seen in Figure 2.

In order to remove non-facial information from the image, we also perform masking of the image (see Figure 2). This is done using a convex hull surrounding the aligned feature points. We translate the eyebrow landmarks slightly to still capture the wrinkling of the forehead. Our normalisation technique is similar to the one presented by Mavadati *et al.*[16].

C. Appearance features

Once the face is aligned to a 112×112 image, we can extract appearance features from it. In this step, Histograms of Oriented Gradients (HOGs) are extracted as proposed by Felzenswalb *et al.*[9]. We use blocks of 2×2 cells, of 8×8 pixels. This leads to 12×12 blocks of 31 dimensional histograms, leading to a 4464 dimensional vector describing the face. We use the dlib [13] implementation of HOGs.

In order to reduce the dimensionality of the HOG feature vector, we use Principal Component Analysis (PCA). Since we wanted our dimensionality reduction to be as general as possible to the problem of facial expression analysis, we did not restrict the training data to FERA 2015 dataset [24]. For constructing the data for the PCA, we used CK+ [15], DISFA [16], AVEC 2011 [23], FERA 2011 [25], and FERA 2015 datasets. Applying PCA to images (subsampling from peak and neutral expressions) and keeping 95% of explained variability lead to a reduced basis of 1379 dimensions. This allowed for more generic model training and not needing to recompute the PCA basis for specific datasets.

D. Geometry

As geometry based features, we used the non-rigid shape parameters and landmark locations in object space inferred during CLNF model tracking ($\mathbf{p}$ and $\mathbf{\Phi}_i\mathbf{p}$ from Equation 1). This led to a $23 + 204 = 227$ dimensional vector describing geometry.

Together both of the descriptors led to a 1606 dimensional vector, that describes appearance of the face. This feature vector is used in all of the following experiments.

E. Neutral expression extraction

Some facial expressions are very difficult to determine if a neutral facial expression of a person is not known. Some people appear more *smiley* or more *frowny* even if their faces are at rest [19]. We believe it is important to *calibrate* for this, by correcting for person-specific neutral expression.

We propose a simple method for estimating a neutral expression descriptor by computing a median value of face descriptors in a video sequence of a person. This relies on the assumption that neutral expression is contained in the majority of frames. This assumption holds for certain datasets (such as SEMAINE and DISFA), but not others (such as BP4D). This assumption also holds for real-life situations, where expression monitoring would take place – most of the time people are not displaying facial expressions and their interactions are dominated by neutral faces [1].

The extracted median face is then subtracted from the feature descriptor leading to a normalised feature. In this paper, we refer to this model as *dynamic* and the non-normalised one as *static*. This is because the normalised feature vector describes dynamic change from neutral rather than absolute expression.

We keep a histogram for each element in our feature vector to efficiently compute the median (as the feature value ranges are known in advance). This also allows us to use such normalisation online.

V. Classification and regression

For AU occurence detection, we used Support Vector Machines (SVM), and for AU intensity estimation, we used Support Vector Regression (SVR). In both cases, we used linear kernels as complex kernels did not improve performance and significantly slowed down training. Furthermore, we are especially interested in approaches that allow for real-time applications. In both cases we used the liblinear library [8].

Due to the unbalanced nature of AU occurrence, it was very important to re-balance the training data. This was done by under-sampling the negative AU samples from training data, leading to an equal number of positive and negative samples. This had a very large impact on the SEMAINE dataset (up to 50% increase in F1 performance on development set), especially for less frequently occurring AUs, such as 17 and 28.

The use of linear kernels also allows for very quick classification and regression. We combine the dimensionality reduction step of PCA with the support vector weights, which results in a single dot product between our feature vector and the resulting weights.

For both SVM and SVR models and in all of the experiments, we validated the C parameter during model training.

VI. Experiments

For experimental evaluation, we compare the results of using our dynamic model (using person-specific neutral expression normalisation) with static model (using no normalisation techniques). We also compare generic model training to targeted model training. And finally, we present how our system performs on FERA 2015 [24] test sets and compare our results to the two proposed baselines.

A. Dynamic model

The first task in our experiments was to test if our proposed dynamic model (person-specific neutral expression

TABLE I: Importance of per-person normalisation, showing validation results (F1). We are only displaying detection results where using a dynamic model had a significant positive effect, for the effect on other AUs was smaller or negative. Notice how some AUs are recognised much better when a person-specific appearance is taken into account, especially in DISFA and SEMAINE datasets, where a median feature is much more likely to represent a neutral expression.

	SEMAINE					DISFA							BP4D		
	AU2	AU12	AU17	AU25	AU45	AU2	AU5	AU6	AU9	AU15	AU17	AU20	AU4	AU6	Mean
Static	0.34	0.58	0.15	0.46	0.38	0.13	0.14	0.44	0.23	0.28	0.23	0.10	0.44	0.76	0.33
Dynamic	**0.59**	**0.61**	**0.44**	**0.52**	**0.42**	**0.26**	**0.17**	**0.58**	**0.46**	**0.48**	**0.43**	**0.22**	**0.53**	**0.79**	**0.46**

TABLE II: Cross-dataset generalisation, showing validation results (F1) across all datasets. Here, the model is validated on all joint-datasets, with the intention of building a single model that generalises well for different datasets. Note how training on joint datasets leads to better generalisation. Also note how BP4D and DISFA are better at generalising than SEMAINE, possibly due to the more diverse training samples and more reliable ground truth labels.

	SEMAINE			DISFA			BP4D			All
	Static models									
Training on	AU2	AU12	AU17	AU2	AU12	AU17	AU2	AU12	AU17	Mean
SEMAINE	0.31	**0.55**	0.09	0.26	0.58	0.12	0.33	0.77	0.55	0.40
BP4D	0.29	0.48	0.13	0.29	0.70	0.21	0.34	**0.87**	**0.62**	0.44
DISFA	0.30	0.52	0.15	0.13	**0.80**	0.21	0.27	0.78	0.54	0.41
Combined	**0.40**	0.54	**0.35**	**0.29**	0.76	**0.22**	**0.35**	**0.87**	0.60	**0.49**
	Dynamic models									
SEMAINE	**0.53**	**0.59**	**0.35**	0.24	0.70	0.10	0.27	0.67	0.10	0.39
BP4D	0.40	0.50	0.17	**0.35**	0.59	0.26	0.30	**0.86**	**0.58**	0.44
DISFA	0.45	0.47	0.21	0.23	**0.77**	**0.40**	0.34	0.59	0.47	0.44
Combined	0.42	0.58	0.22	0.33	0.74	0.34	**0.35**	0.83	**0.58**	**0.49**

normalisation) helps with AU detection. To test this we trained a static and dynamic feature based AU occurrence classifiers on SEMAINE, DISFA and BP4D datasets.

The results of this can be seen in Table I. It can be clearly seen that for SEMAINE and DISFA datasets the dynamic model has a large positive effect on detection scores for certain AUs. This is not the case for BP4D, where the assumption of neutral expression occurring in the majority of frames does not hold and sometimes the dynamic model leads to decreased performance. This leads us to conclude that it has to be applied selectively to certain AUs and that it can only be applied to datasets where neutral expressions are common. Hence the choice between static or dynamic features will depend on both the AUs and datasets considered. It is interesting to note that the accurate recognition of certain AUs (2, 4, 9, 15, 17, 20) seems to require knowing the neutral facial expression whilst others can be recognised quite reliably without accounting for it (6, 7, 10, 12).

B. Generic model training

In our next experiment, we wanted to see the viability of using one of the datasets to train AU occurrence detector for another dataset, and if extra training data from other datasets helps. This is especially important if we want our AU detectors to function in real world scenarios.

For this experiment, we chose the three AUs (2, 12, and 17) that occur in all three datasets of interest - SEMAINE, BP4D, and DISFA. We wanted to asses how well training on each dataset generalises to the AU prediction in general (validated on the joint dataset - SEMAINE + BP4D + DISFA).

Note that for DISFA, which has intensity labels rather than occurence ones, we binarised the labels by assuming presence at A intensity level, whereas BP4D was labelled as occurring at B-level intensities and SEMAINE at A-level.

The results of this training task can be seen in Table II. From the results we can see that some BP4D and DISFA generalise better than SEMAINE, possibly due to more balanced data distribution and more reliable ground truth labels (higher ICC scores between AU coders [24], [16]). Furthermore, we can see that using all of the datasets together leads to the most generic model. This is not surprising as we are testing on the same three datasets, but as the three datasets are quite different, these results are encouraging. Interestingly, jointly-trained static models produce the best classifier for certain AUs more often than dynamic models. This is possibly because the dynamic features extracted from BP4D are less reliable, as fewer frames contain neutral expressions.

C. Targeted model training

In the next set of experiments, we wanted to see if we can use additional labelled data from other datasets to train models that perform well on a target dataset. This is different from the previous section where the goal was to build a generic rather than a targeted model. The difference in methodology was that we validated on the target dataset rather than on joint datasets, in cases where only two datasets contain the AU labels (*e.g.* AU1) we used their combination instead of using all three datasets.

The results of targeted training are shown in Table III. We can see that the use of additional training data has a positive benefit on BP4D dataset for a number of AUs tested, and only in some instances has a negative effect. Furthermore, the difference between AU accuracies when trained on different datasets are small, meaning that our framework generalises

TABLE III: Cross-dataset training, showing testing results (F1) on SEMAINE and BP4D. We report the best performing model (static or dynamic). Note how the results of training on one dataset and testing on a different one does not lead to hugely different results (Exepct for AUs 15 and 17, which very rarely occur in DISFA); This indicates the generalisability of our approach.

	SEMAINE				BP4D						
Training on	AU2	AU12	AU17	AU25	AU1	AU2	AU4	AU6	AU12	AU15	AU17
SEMAINE	**0.54**	**0.61**	**0.44**	**0.52**	-	0.34	-	-	0.80	-	0.58
DISFA	0.45	0.54	0.23	0.46	0.44	0.34	0.50	0.76	0.83	0.26	0.54
BP4D	0.46	0.51	0.18	-	0.43	0.35	0.53	**0.79**	**0.87**	**0.44**	**0.62**
Combined	0.47	0.58	0.38	0.47	**0.45**	**0.36**	**0.55**	0.78	**0.87**	**0.44**	0.61

TABLE IV: Cross-dataset generalisation (Pearson correlation coefficients) on BP4D intensity, using the static model.

	Fully automatic			Pre-segmented		
Trained on	AU6	AU12	AU17	AU6	AU12	AU17
BP4D	0.75	**0.86**	0.50	**0.67**	0.85	0.35
DISFA	0.74	0.83	0.38	0.50	0.76	0.32
Combined	**0.76**	**0.86**	**0.52**	0.63	**0.86**	**0.38**

well. The exceptions to this are AUs 15 and 17, they occur quite rarely in the DISFA and SEMAINE datasets, possibly explaining the discrepancy.

The same experiment was performed for AU intensity estimation. The results can be seen in Table IV. Results showed that training on DISFA and testing on BP4D has a small degradation in performance, however the drop is small considering that the datasets differ from each other significantly. Finally, using both of the datasets at once is often beneficial.

D. Final results on FERA 2015 test sets

To see how well our results generalise on the test sets of the FERA 2015 challenge, we picked the best performing models on validation subsets for evaluation. In case of similar performance between models, we chose the model trained on combined datasets with the idea that it would generalise better, having been trained on more diverse data.

For SEMAINE occurrence, we used only dynamic models trained on SEMAINE. Results can be found in Table V.

For BP4D occurrence, the models were as follows: static model trained on combined datasets - AUs 4. 6, 12 and 15; static model trained on BP4D for the rest of AUs. Results can be found in Table V.

For fully continuous AU intensity, the models used were as follows: static model on combined datasets for AUs 6, 12, 17, and the rest were trained on BP4D. For segmented challenge, AUs 12 and 17 were trained on combined datasets and the rest on BP4D. Results can be found in Table VI.

As the test labels were not available for us, this represents a good evaluation of the model generalisation. Our approach outperforms both of the proposed baselines on all of the tasks, with the highest gain in performance in the pre-segmented AU intensity task. The improvement is especially big for AUs that the baseline fails to detect reliably - 15, 17, 25 and 28.

TABLE VI: Final intensity results (intra-class correlation coefficient) on FERA 2015 test dataset comparing against their proposed appearance and geometry based baselines[24].

	AU6	AU10	AU12	AU14	AU17	Mean
	Fully automatic					
Baseline geom.	0.67	**0.73**	0.78	**0.59**	0.14	0.58
Baseline app.	0.62	0.66	0.77	0.39	0.17	0.52
Ours	**0.69**	**0.73**	**0.83**	0.50	**0.37**	**0.62**
	Pre-segmented					
Baseline geom.	0.48	**0.51**	0.69	**0.59**	0.05	0.46
Baseline app.	0.33	0.48	0.60	0.50	0.11	0.40
Ours	**0.58**	0.49	**0.70**	0.52	**0.41**	**0.54**

E. Processing speed

The system is capable of running real-time (20-30 fps) on the SEMAINE and DISFA datasets on commodity hardware - dual core 3GHz Intel i3 processor and without any GPU support. The performance on the BP4D dataset is slower - 5-10 fps, due to the majority of processing being taken up by very large image loading from disk.

VII. Conclusions

In this paper, we presented a real-time AU detection and intensity estimation system. Our experiments show the benefits of a simple person-specific normalisation for certain AU detection especially on DISFA and SEMAINE datasets. We also demonstrate that the use of combined training datasets leads to better AU detectors. This is the case both when training for a specific dataset or when a generic model is needed. This is especially important if we want our methods to work in the real world. Finally, the overall results of our cross-dataset experiments and FERA 2015 test sets revealed that our methodology generalises and outperforms the two baselines on all three tasks demonstrating the effectiveness and generalisability of our system.

For future work, we would like to explore alternative ways of neutral expression estimation that do not rely on neutral expression occurring often. We would also like to explore the balancing of training data more, both in terms of positive and negative samples, but also in terms of datasets, as some of them seem to be able to generalise better than others.

ACKNOWLEDGMENTS

We acknowledge funding support from from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement 289021 (ASC-Inclusion).

TABLE V: Final occurence results on FERA 2015 test dataset comparing against their proposed appearance (BA) and geometry (BG) baselines [24].

	BP4D											SEMAINE						Mean
AU	1	2	4	6	7	10	12	14	15	17	23	2	12	17	25	28	45	
BG	0.19	0.19	0.20	0.65	**0.80**	0.80	0.80	**0.72**	0.24	0.31	0.32	0.57	**0.60**	0.09	0.45	0.25	0.40	0.44
BA	0.18	0.16	0.23	0.67	0.75	0.80	0.79	0.67	0.14	0.25	0.24	**0.76**	0.52	0.07	0.40	0.01	0.21	0.40
Ours	**0.26**	**0.25**	**0.25**	**0.73**	**0.80**	**0.84**	**0.82**	**0.72**	**0.34**	**0.33**	**0.34**	0.41	0.57	**0.20**	**0.69**	**0.26**	**0.42**	**0.48**

REFERENCES

[1] S. Afzal and P. Robinson. Natural Affect Data - Collection & Annotation in a Learning Context. *Design*, 2009.

[2] T. Baltrušaitis, L-P. Morency, and P. Robinson. Constrained local neural fields for robust facial landmark detection in the wild. In *IEEE International Conference on Computer Vision Workshops*, 2013.

[3] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. In *CVPR Workshops*, 2003.

[4] W. S. Chu, F. De la Torre, and J. F. Cohn. Selective transfer machine for personalized facial action unit detection. *CVPR*, 2013.

[5] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

[6] F. De la Torre and J. F Cohn. Facial Expression Analysis. In *Guide to Visual Analysis of Humans: Looking at People*. 2011.

[7] P. Ekman and W. V. Friesen. *Manual for the Facial Action Coding System*. Palo Alto: Consulting Psychologists Press, 1977.

[8] R. Fan, C. Kai-Wei, H. Cho-Jui, X. Wang, and C. Lin. Liblinear : A library for large linear classification. *JMLR*, 2008.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminative Trained Part Based Models. *IEEE TPAMI*, 32, 2010.

[10] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald. Social risk and depression: Evidence from manual and automatic facial expression analysis. In *FG*, 2013.

[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *IVC*, 2010.

[12] L. A. Jeni, J. M. Girard, J. F. Cohn, and F. De la Torre. Continuous au intensity estimation using localized, sparse facial feature space. In *FG*, 2013.

[13] D. E. King. Dlib-ml: A machine learning toolkit. *JMLR*, 2009.

[14] Y. Li, J. Chen, Y. Zhao, and Q. Ji. Data-free prior model for facial action unit recognition. *IEEE TAFFC*, 2013.

[15] P. Lucey, J. F. Cohn, T. Kanade, J. M. Saragih, Z. Ambadar, and I. Matthews. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, 2010.

[16] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F Cohn. Disfa : A spontaneous facial action intensity database. *IEEE T-AFFC*, 2013.

[17] D. McDuff, R. el Kaliouby, D. Demirdjian, and R. Picard. Predicting online media effectiveness based on smile responses gathered over the internet. In *FG*, 2013.

[18] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic. The SEMAINE corpus of emotionally coloured character interactions. In *IEEE International Conference on Multimedia and Expo*, 2010.

[19] D. Neth and A. M. Martinez. Emotion perception in emotionless face images suggests a norm-based representation. *Journal of vision*, 2009.

[20] P. Robinson and R. el Kaliouby. Computation of emotions in man and machines. *Phil. Trans. of the Royal Soc. B*, 2009.

[21] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV*, 2013.

[22] E. Sariyanidi, H. Gunes, and A. Cavallaro. Automatic analysis of facial affect: A survey of registration, representation and recognition. *IEEE TPAMI*, 2014.

[23] B. Schuller, M. F. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic. AVEC 2011 The First International Audio / Visual Emotion Challenge. In *ACII*, 2011.

[24] M. Valstar, J. Girard, T. Almaev, G. McKeown, M. Mehu, L. Yin, M. Pantic, and J. Cohn. FERA 2015 - Second Facial Expression Recognition and Analysis Challenge. In *IEEE FG*, 2015.

[25] M.l F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. R. Scherer. The First Facial Expression Recognition and Analysis Challenge. In *IEEE FG*, 2011.

[26] T. Wu, N. J. Butko, P. Ruvolo, J. Whitehill, M. S. Bartlett, and J. R. Movellan. Action unit recognition transfer across datasets. *IEEE FG*, 2011.

[27] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard. BP4D-Spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. *IVC*, 2014.