Facial Action Units Intensity Estimation by the Fusion of Features with Multi-kernel Support Vector Machine

Zuheng Ming, Aurélie Bugeau, Jean-Luc Rouas, Takaaki Shochi

Facial Action Units Intensity Estimation by the Fusion of Features with Multi-kernel Support Vector Machine

Zuheng Ming[1], Aurélie Bugeau[1], Jean-Luc Rouas[2], Takaaki Shochi[1]
[1] LaBRI, University of Bordeaux, UMR 5800, F-33400 Talence, France
[2] LaBRI, CNRS, UMR 5800, F-33400 Talence, France

***Abstract*— Automatic facial expression recognition has emerged over two decades. The recognition of the posed facial expressions and the detection of Action Units (AUs) of facial expression have already made great progress. More recently, the automatic estimation of the variation of facial expression, either in terms of the intensities of AUs or in terms of the values of dimensional emotions, has emerged in the field of the facial expression analysis. However, discriminating different intensities of AUs is a far more challenging task than AUs detection due to several intractable problems. Aiming to continuing standardized evaluation procedures and surpass the limits of the current research, the second Facial Expression Recognition and Analysis challenge (FERA2015) is presented. In this context, we propose a method using the fusion of the different appearance and geometry features based on a multi-kernel Support Vector Machine (SVM) for the automatic estimation of the intensities of the AUs. The result of our approach benefiting from taking advantages of the different features adapting to a multi-kernel SVM is shown to outperform the conventional methods based on the mono-type feature with single kernel SVM.**

I. INTRODUCTION

Facial expressions are some of the most direct, naturally preeminent means for human beings to regulate interactions with each other [1]. They hold almost 55% information in the emotion communication in life, while the speech and the nonverbal content occupied about 38% and 7% separately [2]. Motivated by the development of automatic recognition of the facial expression, many valuable applications are emerging in the domains of human-computer interaction, social robots, mobile devices, cars, consumer photography and also in the domains of medicine and psychology such as automatic pain detection, medical assistance as well as developing auxiliary tools for neuroscience and behavior research ([3],[4],[5]). Therefore researchers in computer vision and machine learning have been increasingly interested in the topic of facial expression recognition and analysis in recent years, with the aim of creating machines with interfaces that are better aligned to human communication [6].

Before Ekman et al. [7] proposed the Facial Action Coding System (FACS), the recognition of facial expressions was really difficult as the facial expressions include complex motions and the range of the facial behavior is extremely wide. Ekman et al ([7], [8]) define the FACS as a comprehensive set of atomic non-overlapping facial muscle action named Action Units (AUs). Each facial expression can be decried as the varied combinations and strength of AUs. Besides the 64 AUs defined for the upper, lower face and the head position, the FACS also represents the intensity variation of AUs by assigning the letters A (trace) to E (maximum), in practice, the letters are often replaced by the number 1-5 to describe the corresponding intensity variation.

Thus, by using the FACS, human expert coders can manually code nearly any facial expression and determine their temporal changes (onset, peak and offset). However, even for the experienced expert, the annotation is a laborious and error-prone work when the amount of video to be coded is huge. Hence automatic recognition and measurement is essential to the real use.

Limited by the datasets, the early researches of facial expression recognition focused on the basic, posed expressions, specifically happy, sadness, fear, anger, disgust and surprise ([9], [10]), which can be easier distinguished in configuration, intensity and timing from those that occur spontaneously. However, many facial actions are common in the natural situation but occur rarely in the posed, prototypical expressions. Thus, the researchers emphasized the importance of the analysis of the spontaneous facial behavior after the progress of the recognition of basic posed expressions [11]. Since FACS AUs anatomically based on the actions of one or a few facial muscles can describe nearly all possible facial actions, it is widely used in the detection of the natural emotion state such as pain detection [12], and psychological studies[13]. By detecting one or more of the AUs, the facial expression in question can be produced by the combination of the obtained AUs ([14], [15], [6]).

While a full expression cannot be described only by the presence or absence of the AUs, the meaning and function of spontaneous facial expression depends largely on intensities of AUs. For example, most of the smile genuineness impression is created by the intensity and the facial motion of the smile, not just the activation of AU6 [16]. Thus, measuring the full intensities of the AUs allows to determine the more complex expression and emotional states. This is also necessary to produce a comprehensive recognition system of expression in the real life. However, the estimation of the intensities of AUs is a far more challenging problem than the AUs detection [1], as the high dependencies of the facial morphology and expressiveness of the subject, the affect of the co-occurrences of AUs and the lighting condition, head position etc.. In this context, the second Facial Expression Recognition and Analysis challenge (FERA 2015) [17] added the sub-challenge for estimating the intensities of AUs for the

This work was supported by The Investments for the Future Programme IdEx Bordeaux - CPU (ANR-10-IDEX-03-02)

natural scenario, aiming to help the field to progress beyond its current limitations.

Many systems have been proposed to estimate the intensities of the AUs ([18], [19], [1]). The representation of the face image and the classification / regression model after registering facial images are the most essential parts of the systems. Many features exist to represent the face. A good feature of the natural facial image can not only differ well the objective from others but also with better efficiency. Unfortunately, the different features always have their own advantages and drawbacks. The appearance-based features such as Local Gabor Binary Patterns (LGBP) [20], Local Gabor Binary Patterns from Three Orthogonal Planes (LGBPTOP) [6] and Histogram of Oriented Gradients (HOG) [21] are robust to the illumination variation and the misalignment, meanwhile the use of histogram to present these features lose the sensibility of the subtle movement of facial action. In contrary, this drawback of appearance-based features can be well complemented by the shape-based features such as Active Appearance Model (AAM) geometrical features [22] which use the landmarks to describe the deformation of the facial region. The geometrical features are also hardly satisfying as the lack of the texture information. A good representation of the facial image can be obtained when combining the advantages of different features. Nevertheless, in the conventional systems either based on Support Vector Machine (SVM) or Support Vector Regression (SVR) for estimating the intensities of AUs, the different kinds of features can only be concatenated to adapt to the single kernel [23]. This is not reasonable since the different features suit the different kernels (e.g. the intersection kernel fits the histogram-based feature [24]). The recent applications of [25] and developments based on SVM have shown that using multiple kernels as a convex linear combination of other basis kernels instead of a single one can enhance interpretability of the decision function and improve classifier performance [26]. Therefore, we propose a method using the fusion of the features with the multi-kernel SVM for estimating the intensities of the AUs in the context of the second FERA2015 challenge.

This paper is organized as follows: section II provides an overview of the related work; section III describes the features used in this work; section IV presents the estimation model based on multi-kernel SVM; section V describes the experiment setup and reports the experimental results; finally, draw the conclusion of this work.

II. RELATED WORK

Due to the development of the datasets provided publicly for the research, the automatic estimation of the intensities of AUs has emerged recently in the field of automatic analysis of the facial expressions. The general framework of the automatic estimation systems of the intensities of AUs has three stages: 1) the facial registration; 2) feature extraction as the facial image representation; 3) models based on machine learning for estimating the final intensities. In terms of the second stage, the estimation methods can be divided into geometric-based methods ([14], [18]), appearance-based methods ([6]) and methods based on the fusing the two features [27]. In terms of the third stage, the methods can be divided into the classification-based methods ([28], [18]) and the regression-based methods ([29], [30]). With the recent rise of Deep Learning neural models, a new perspective of facial expression recognition has appeared [31].

The geometric-based methods use the landmarks locating on the face to extract the shape features of the facial components. For instance, Pantic et al.[14] proposed a fully automatic schema for tracking 20 landmarks locating in the region of eyes, nose and mouth. The points were automatically detected in the first frame and then a particle filtering schema using factorized likelihoods and combining a rigid and a morphological model was applied to track the facial points. The AUs displayed in the input video were the recognized by the SVM trained on the features selected by the AdaBoost.

The appearance-based methods focus on extracting the texture feature of the facial image. Having a good ability to detect the wave-like structures and also robust to the to the misalignment and the variation of the illumination, Gabor wavelet features are widely used to represent the facial texture. The LBP [10] features were originally proposed for texture analysis, while due to their tolerance to illumination changes and the computational simplicity, they have become very popular for face analysis recently. Zhang et al.[32] proposed to use appearance features extracted in a multi-layer architecture by applying the LBP operator to the Gabor magnitude response images to form the LGBP features. This has been shown to be very robust to illumination variations and misalignment. Almaev et al [6] later extended the LGBP features to spatio-temporal volumes on Three Orthogonal Planes (i.e. LGBPTOP), which have also obtained a good performance in AUs detection.

The classification-based methods treat the estimation of the intensities levels of AUs as a multi-class classification problem. These methods use a classier such as SVM, to nominal data. Mavadati et al.[18] employed the SVM for intensities classification of 13 AUs. They also proposed a new dataset DISFA of spontaneous facial expression for training the model. The input features were the concatenations of facial shape and facial appearance by using AAM. Due to the excessive number of features, the manifold learning method was selected to reduce the dimension of features.

The regression-based methods use the regression models such as SVR, logistic-regression-based model etc. to estimate the intensities of AUs on a continuous scale. Jeni et al. [30] proposed a model based on SVR for AUs intensities estimation. The authors extracted the features from the image patches around the facial landmarks for 14 AUs.

III. FEATURES

Aiming to take advantages of the appearance-based features and the geometric-based features, three different features such as LGBP, geometrical features and HOG are used in this work.

A. LGBP

The LGBP feature is an appearance-based feature, in which the LBP operates on the magnitude response of Gabor filtering of an image instead of the natural image [32]. In the basic LBP, it codes each pixel of an image by thresholding its surrounding eight neighbors by its value. Thus, the LBP has eight digit binary numbers, which correspond to the 256 possibilities in the histogram. In [10] notes that only 59 patterns called uniform LBP account for about 90% of the LBP response. Therefore, instead of 256 patterns, only 59 are selected. Since each area of the face contains different valuable information, the face is divided into 4x4 blocks, and a histogram is built for each block. In addition, three spatial frequencies and six orientations were defined to produce a total of 18 Gabor filters, and the magnitude responses result in the 18 Gabor images. Finally, a feature vector of 59x16x18 dimensions is formed for each facial image.

B. Geometrical Features

The Geometrical Features used in this work are based on 49 landmarks detected with the Cascaded Regression facial point detector proposed by Xiong and De La Torre [33] as shown in Fig. 1. Before extracting the geometrical features,

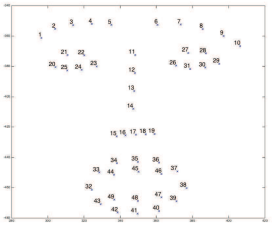


Fig. 1. 49 facial landmarks used to compute geometric features

facial landmarks of every video frame are aligned with a mean shape using a set of stable points. Stable points are defined as those not affected by AU activations, such as the points with notation of 20, 23, 26, 29 (eye corners region) and 11 - 19 (nose region). The mean facial shape is computed prior to geometric features extraction. The alignment is performed by computing a non-reflective affine transformation, which minimizes the difference between stable point coordinates of the two shapes. All mean shape landmark coordinates are then subtracted from the corresponding aligned shape points resulting in a set of aligned facial points which form the first 49x2 = 98 geometric features.

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

For the next set of features the facial landmarks have been split into three groups representing left eye (points 20 - 25) with left eyebrow (points 1 - 5), right eye (points 26 - 31) with right eyebrow (points 6 - 10), and the mouth region (points 32 - 49). For each of these groups a set of features representing Euclidean distances between points in the group are computed resulting in 37 features. Furthermore, the angles in radians for each consecutive triplet points within the groups are also extracted (32 features).

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

In total there are 316 geometric features extracted on each frame.

C. HOG

The HOG descriptor was introduced initially for the human detection [21]. It has attracted the attention in the field of facial expression analysis as it represents both the appearance and shape information with a simplicity of computation. The HOG feature counts the occurrences of gradient orientations in a localized portion of image. Before applying the HOG operators, the images are divided into small cells, and the histogram of the gradients is computed for each cell.

In this work, 49 cells with the size 40*40 pixels around the landmarks described above are defined. Next, the horizontal gradient filter [-1 0 1] with 59 orientation bins are applied to calculate the histograms. Finally the obtained histogram of each cell are concatenated to form the HOG feature vector of size 49*59.

IV. Model based on multi-kernel SVM

SVM has been recognized as an effective algorithm in numerous pattern recognition and facial expression recognition applications. Benefiting from the kernel function, the SVM can be applied to both linear and non linear classification. SVM is originally designed for binary classification. Nevertheless, it can be also used as multi-class SVM via some classification strategies: one-against-rest strategy and one-against-one strategy. Due to the computation complexity and the unbalanced problem, we use the one-against-one strategy in this work. In this strategy, $C(C-1)/2$ binary classifiers are employed to classify a sample between every possible pair of classes and the sample is finally classified to the class with the most votes [34].

Here, the classes in the SVM correspond to the five levels of intensity of a specific AU plus a class corresponding to

the level zero of intensity indicating the absence of AU in an image. Thus, there are six classes in total on which a full automatic estimation of intensities of AUs can be realized. The decision function is [26]:

$$f(x) = \sum_{i=1}^{m} \alpha_i k(x_i, x) + b \tag{1}$$

where $f(x)$ is the decision function of the classification of the input sample, x is a sample to be classified, α_i is the dual representation of the hyperplane's normal vector, k is the kernel function resulting from the dot product in a transformed high-dimensional feature space.

It has been proved that using multiple kernels instead of a single one can enhance interpretability of the decision function and improve classifier performance. A common approach is to consider that the kernel $k(x_i, x)$ is actually a convex linear combination of other basis kernels (such as linear kernel, polynomial kernel and RBF kernel etc.) [26]:

$$k(x_i, x) = \sum_{k=1}^{M} \beta_k K_k(x_i, x), \quad \textit{with } \beta_k \geq 0, \ \sum \beta_k = 1 \tag{2}$$

where M is the total number of kernels. Each basis kernel K_k may either use the full set of variables describing x or only subset of variables. Alternatively, kernels K_k can simply be classical kernels (such as Gaussian kernel) with different parameters or may rely on different data sources associated with the same learning problem parameters[35]. In this work, two different kernels, which are Gaussian kernel (i.e. RBF kernel) and interaction kernel are integrated in the multiple kernel framework, and each kernel is associated with different source of data. Specifically, the LGBP features are associated to the intersection kernel while the geometrical feature and the HOG features are integrated into Gaussian kernel. The optimization of parameters α_i and β_k is known as the multiple kernel learning (MLK) problem. Sonnennburg et al. [36] reformulated the MLK problem as a semi-infinite linear program (SILP). It optimizes the parameters by iteratively solving the classical single kernel SVM optimization problem. We use the efficient toolbox [34] in this work.

V. Experiments

A. Dataset and Evaluation Procedure

In the context of the FERA2015, the evaluation and training of the model are performed on the dataset BP4D-Spontaneous [37] provided by the organizer. BP4D-Spontaneous dataset consists of video data of young adults responding to emotion-elicitation tasks. This dataset includes 328 digital videos of 41 participants (56.1% female, 49.1% white and ages 18-29) from the different departments of Binghamton University. The AUs in the videos were coded by the highly trained coders. 5 AUs (i.e. AU06, AU10, AU12,AU14, AU17) were selected for the intensity Estimation Sub-Challenge. In our work, an equal-interval down-sampling of the dataset is applied in order to save time when training the model.

Intra-class Correlation Coefficient (ICC) [38], specifically, ICC (3,1) is selected as the evaluation criterion in this work. Assuming there were k judges (here 2: the ground truth and predicted values), and n targets (i.e. samples), and defining the within-target sum squares (WSS), between-raters sum squares (RSS), between-target sum squares (BSS), and residual sum of squares (ESS = WSS-RSS), the ICC (3,1) is :

$$ICC = \frac{BMS - EMS}{BMS + (k-1)EMS} \tag{3}$$

where $BMS = \frac{BSS}{n-1}$ is between-class mean square and $EMS = \frac{ESS}{(k-1)(n-1)}$ is residual mean squares. This score ranges from 0 to 1, but sometimes negatives can occur [38]. Besides, the 5-fold cross-validation is also used in the evaluation procedure.

B. Experimental Results

We first compare the estimation results of intensities of AUs based on the single kernel SVM with different features and the multi-kernel SVM with different fusions of features, as well as the estimation result of the concatenation of features associated with the single kernel SVM. Secondly, we present the effects of the kernels weights in the multi-kernel SVM.

Comparison of estimation results. The estimation results based on the different SVM-based models associated with the different features or the fusions of features are demonstrated in Table I. In order to have an idea of the benchmark of the system, the baseline result of the FERA2015 challenge is also listed in Table 1 although the test samples are probably not exactly the same. In Table I, the linear kernel, the intersection kernel and the RBF kernel are denoted as 'LINEAR', 'INT' and 'RBF' and the geometrical features are denoted as 'GEO'. The features in the same brackets mean the concatenation of the two features as an input of a kernel. The first two groups of rows correspond to the results of the single kernel SVM, and the third group presents the results of the multi-kernel SVM. In this experiment, the multi-kernel consists of the intersection and RBF kernels. The first feature (or concatenation of features) listing in the third group of the table is for the intersection kernel while the second corresponds to the RBF kernel. The best

TABLE I
AUs INTENSITIES ESTIMATION RESULTS

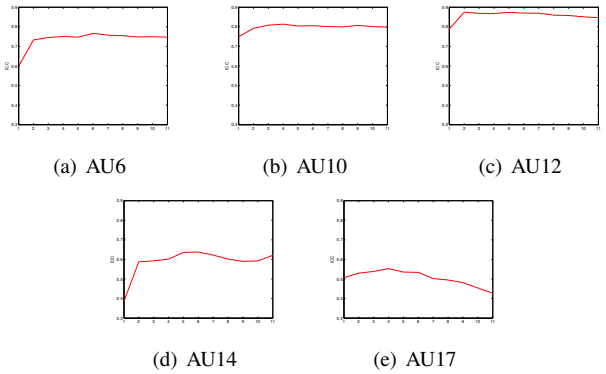
ICC	AU6	AU10	AU12	AU14	AU17
LGBP—LINEAR(*)	0.6940	0.6410	0.6700	0.3250	0.1850
GEO—LINEAR(*)	0.6900	0.6960	0.6530	0.4530	0.2780
LGBP—INT	0.7622	0.7898	0.8424	0.6295	0.4274
GEO—RBF	0.6365	0.7830	0.8122	0.4696	0.5067
HOG—INT	0.7360	0.7754	0.8297	0.5424	0.5097
HOG—RBF	0.7619	0.7530	0.8401	0.5617	0.5127
(LGBP+GEO)—RBF	**0.7909**	0.7925	0.8791	0.6682	0.4805
LGBP+GEO	0.7665	0.8137	0.8752	0.6358	0.5337
(weight of INT β_1)	(0.60)	(0.30)	(0.10)	(0.50)	(0.30)
LGBP+(GEO+HOG)	0.7884	0.8143	**0.8818**	**0.6748**	0.5296
(weight of INT β_1)	(0.10)	(0.10)	(0.20)	(0.20)	(0.02)
(LGBP+HOG)+GEO	0.7890	**0.8278**	0.8487	0.6658	**0.5437**
(weight of INT β_1)	(0.20)	(0.02)	(0.02)	(0.10)	(0.30)

Note: '*' are the baseline results provided by the FERA2015 challenge.

estimation results (denoted in bold) of all AUs except AU6 belong to the multi-kernel SVM associated to the fusion of different features. The best estimation performance of AU14 obtained by the multi-kernel SVM improves of about 21% compared to the single RBF kernel SVM combined with the geometrical feature. For AU17, the estimation result of multi-kernel SVM improves of about 12% in compared with the intersection kernel SVM with LGBP feature. Even for AU12, which is relatively easy to recognize for all methods, the multi-kernel based method continues to improve about 4% compared to the best result obtained by the single kernel SVM with one feature. Furthermore, in the second group of rows, the concatenation of the LGBP and geometrical feature with a single RBF kernel SVM naturally improves the estimation performance compared to the single feature based methods shown in the first group. This method even gains a slightly better result than multi-kernel based method in the case of AU6. It suggests that the two different appearance-based and geometric-based features can well complement to each other to play their advantages and alleviate their drawbacks. This point can be also presented in the case of multi-kernel SVM. The multi-kernel SVM using the HOG as an appearance feature to supplement the geometrical feature shows better results compared to the multi-kernel without HOG.

Table I also suggests that the AU12 and AU10 are the best estimated AUs by whichever method, and AU17 is the worst one overall. For AU17, it can be partly explained: as the number of samples of AU17 is obviously less than others, the model was not trained as well as others. In terms of AU12, which is an action corresponding to the 'lip corner puller' as a smile expression with the deformation around the mouth region which can be relative easy to be tracked by the geometrical features, the reult is better than others. This result also coincides with the result presented in [18].

Effects of the weights of kernels. Here we demonstrate the relationship between the weight of the intersection kernel and the performance (measured by ICC) of the estimation of the intensities of AUs as shown in Fig. 2. The intersection kernel weight β_1 of the multi-kernel varies from 0 to 1. When $\beta_1 = 0$ the multi-kernel equals to the RBF kernel, while when $\beta_1 = 1$ the multi-kernel is in fact an intersection kernel. From the Fig. 2, we can see that along with the β_1 changing gradually from 0 to 1, the value of ICC rises gradually until the point of the maximum value. The ICC then begins to decrease until β_1 reaches 1, except for AU14 for which ICC rises slightly again in the end. In this transition process, we observe that the best estimations of the AUs all appeared when the SVM fused the two different kernels. This well proves the effectiveness of the multi-kernel based SVM for the problem of the estimation the intensities of AUs.



(a) AU6 (b) AU10 (c) AU12

(d) AU14 (e) AU17

Fig. 2. The performance (i.e. ICC) of the estimation of the intensities of AUs in function of the weight β_1 of the intersection kernel in the multi-kernel SVM associated to the LGBP and geometrical features.

VI. CONCLUSION AND FUTURE WORK

In this work, a method based on multi-kernel SVM associated with the fusion of different features was proposed for the estimation of the intensities of AUs. To the best of our knowledge, this method has not yet been applied for the problem of the prediction of the intensities of AUs. From the estimation result, the effectiveness of this model has been appropriately proved in the domain of the facial expression estimation. In particular, from the performance of the estimation of AUs intensities along with the transition of the kernels weights, we can clearly see that the multi-kernel takes the

advantages of the classical basis kernels. Indeed, the multi-kernel SVM is capable to outperform the single kernel SVM. We have also highlighted that even with only one Gaussian kernel (i.e. RBF kernel), the concatenation of the appearance-based feature LGBP and geometric-based feature gained a better performance compared to using one feature only. It proves that the different features with intrinsic advantages and drawbacks can complement mutually. Seeking a solution to exploit their superiority and supplement each other was our initial motivation to use the multi-kernel based SVM. In this work, we only adopted two classical kernels in the framework of multi-kernel SVM. In future works, we will study later how to use a more sophisticated model which can be adapted to more features to predict the spontaneous AUs or facial expression with higher accuracy and efficiency.

VII. ACKNOWLEDGMENTS

This study has been carried out in the frame of The Investments for the Future Programme IdEx Bordeaux - CPU (ANR-10-IDEX-03-02). The authors would like to thank all subjects who participated in this study.

REFERENCES

[1] O. Rudovic, V. Pavlovic, and M. Pantic. Context-sensitive Dynamic Ordinal Regression for Intensity Estimation of Facial Action Units. *IEEE TPAMI*, PP(99):1–1, 2014.

[2] A. Mehrabian and M. Wiener. Decoding of inconsistent communications. *Journal of personality and social psychology*, 6(1):109, 1967.

[3] M. Pantic. Machine analysis of facial behaviour: Naturalistic and dynamic behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3505–3513, 2009.

[4] Z. Hammal and J. F. Cohn. Automatic detection of pain intensity. In *Proc. ICMI*, NY, USA, 2012.

[5] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan. Toward practical smile detection. *IEEE TPAMI*, 31(11):2106–2111, Nov 2009.

[6] T. R. Almaev and M. Valstar. Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Proc. IEEE ACII*. IEEE, 2013.

[7] P. Ekman and W. V. Friesen. *Manual for the facial action coding system*. Consulting Psychologists Press, 1978.

[8] J. C. Hager, P. Ekman, and W. V Friesen. Facial action coding system. *Salt Lake City, UT: A Human Face*, 2002.

[9] Z. Zhang, M. Lyons, M. Schuster, and S. Akamatsu. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron. In *Proc. IEEE ICFG*, 1998.

[10] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE TPAMI*, 29(6):915–928, 2007.

[11] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE TPAMI*, 31(1):39–58, 2009.

[12] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews. Painful data: The unbc-mcmaster shoulder pain expression archive database. In *Proc. IEEE ICFG*, 2011.

[13] L. F. Barrett. Was darwin wrong about emotional expressions? *Current Directions in Psychological Science*, 20(6), 2011.

[14] M. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. In *Proc. IEEE CVPR'W*, 2006.

[15] B. Jiang, M. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Proc. IEEE ICFG*, 2011.

[16] S.D. Gunnery, J.A. Hall, and M.A. Ruben. The deliberate duchenne smile: Individual differences in expressive control. *Journal of Nonverbal Behavior*, 37(1):29–41, 2013.

[17] 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 *Proc. IEEE ICFG*, 2015.

[18] S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE TAC*, 4(2):151–160, 2013.

[19] S. Kaltwang, O. Rudovic, and M. Pantic. Continuous pain intensity estimation from facial expressions. In *Advances in Visual Computing*, pages 368–377. Springer, 2012.

[20] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multikernel learning. *IEEE TSMC-B*, 42(4):993–1005, 2012.

[21] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE CVPR*, volume 1, 2005.

[22] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE TPAMI*, 23(6):681–685, 2001.

[23] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *Proc. IEEE ICFG*, 2011.

[24] S. Maji, A. C Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *Proc. IEEE CVPR*, 2008.

[25] F. R. Bach, G. R. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proc. ICML*, 2004.

[26] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, et al. Simplemkl. *JMLR*, 9:2491–2521, 2008.

[27] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE TPAMI*, 27(5):699–714, 2005.

[28] M. H. Mahoor, S. Cadavid, D. S. Messinger, and J. F. Cohn. A framework for automated measurement of the intensity of non-posed facial action units. In *CVPR'W*, 2009.

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

[30] 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 *Proc. IEEE ICFG*, 2013.

[31] P. Liu, S. Han, Z. Meng, and Y. Tong. Facial expression recognition via a boosted deep belief network. In *Proc. IEEE CVPR*, 2014.

[32] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang. Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In *Proc. IEEE ICCV*, volume 1, 2005.

[33] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *Proc. IEEE CVPR*, 2013.

[34] C. Chang and C. Lin. Libsvm: a library for support vector machines. *TIST*, 2(3):27, 2011.

[35] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *JMLR*, 5:27–72, 2004.

[36] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *JMLR*, 7:1531–1565, 2006.

[37] 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. *Image and Vision Computing*, 32(10):692–706, 2014.

[38] P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.