Learning to combine local models for Facial Action Unit detection

Shashank Jaiswal, Brais Martinez and Michel F. Valstar
School of Computer Science, The University of Nottingham
{psxsj3,pszbm1,michel.valstar}@nottingham.ac.uk

***Abstract*— Current approaches to automatic analysis of facial action units (AU) can differ in the way the face appearance is represented. Some works represent the whole face, dividing the bounding box region in a regular grid, and applying a feature descriptor to each subpatch. Alternatively, it is also common to consider local patches around the facial landmarks, and apply appearance descriptors to each of them. Almost invariably, all the features from each of these patches are combined into a single feature vector, which is the input to the learning routine and to inference. This constitutes the so-called feature-level fusion strategy. However, it has recently been suggested that decision-level fusion might provide better results. This strategy trains a different classifier per region, and then combines prediction scores linearly. In this work we extend this idea to model-level fusion, employing Artificial Neural Networks with an equivalent architecture. The resulting method has the advantage of learning the weights of the linear combination in a data-driven manner, and of jointly learning all the region-specific classifiers as well as the region-fusion weights. We show in an experiment that this architecture improves over two baselines, representing typical feature-level fusion. Furthermore, we compare our method with the previously proposed linear decision-level region-fusion method, on the challenging GEMEP-FERA database, showing superior performance.**

I. INTRODUCTION

In any method aimed at automatic learning and recognition of facial muscle activations (FACS Action Units, AUs [4])) or facial expression, the face images are normally pre-processed to eliminate spurious sources of variability, most notably misalignment. These pre-processing steps include the detection of the face bounding box [18][11], the detection of the facial landmarks [10][19] (although optional, it is an important step) and the face registration. The first two steps focus on finding some inner-facial structures, i.e., the facial landmarks in most cases, that can be put in correspondence for all faces. Registering face images to a reference face image reduces the variations in the image due to in-plane rotation, translation and scaling. This is an attempt to minimize the effect of variations in the image data caused by factors other than facial expressions.

After registration, features are typically extracted from the registered face images. Two main aspects are considered in the literature: which features to extract (e.g. LBP [20][16], HOG [1][2], etc), and where to extract them from. Regarding the latter, two approaches are common in the literature: the so-called holistic and part-based approaches. The holistic approach represents the appearance of the full face bounding box and extract features directly from it. The bounding box is usually split into a number of rectangular blocks to encode spatial information (e.g. [7]). Alternatively, the face appearance can be represented by a combination of local image patches around each facial landmark [22]. In both cases, a set of feature vectors are extracted from different image patches (be it subdivisions of the bounding box or patches around landmarks). These vectors are concatenated into a single feature vector, which is the input to the training and inference routines.

Recently, Jiang et al. [5] proposed alternative strategies for two of these aspects: 1) where to extract the features from, and 2) how to combine region-specific features. The authors proposed to extract features from non-rectangular regions defined by the locations of facial landmarks. Specifically, the facial landmarks were used to define a mesh over the whole face, and a feature vector was extracted from each of the regions enclosed by the mesh (see Fig. 2). This strategy combines the benefits of encoding the appearance of the whole face, as in holistic approaches, with the parts of the face representing the same facial area across examples, as for part-based approaches. The second improvement proposed in [5] relates to the way all the region-specific features are combined. Instead of simply concatenating them into a single vector, Jiang et al. propose to perform decision-level fusion. To this end, a separate classifier is trained per region, and then the output scores are combined into the final decision. Jiang et al. showed that both of these strategies resulted in a significant performance boosts, either when used in isolation, and when both were applied. The latter case was shown to consistently outperform any other strategy compared against.

However, the way each part was combined in the decision-level fusion strategy was somewhat heuristic. Specifically, it involved first training a classifier for each region, with the final decision being a linear combination of region-specific scores. The coefficients of the linear combination were computed during training using a validation set, taking the performance as the weights, so that better-performing classifiers were given a larger weight, while classifiers performing approximately at random would be given 0 weight. However, in this work we argue that a data-driven approach, in which the ideal weights learnt from the training set, results in better performance. To implement and compare this, we propose to use Artificial Neural Networks (ANNs), and use an otherwise equivalent architecture. The benefits of this approach are: the weights for combining region-specific classifiers are learnt automatically without resorting

The work of Valstar and Martinez is funded by European Union Horizon 2020 research and innovation programme under grant agreement No 645378. The work of Valstar and Jaiswal is also supported by MindTech Healthcare Technology Co-operative. We are also grateful for access to the University of Nottingham High Performance Computing Facility.

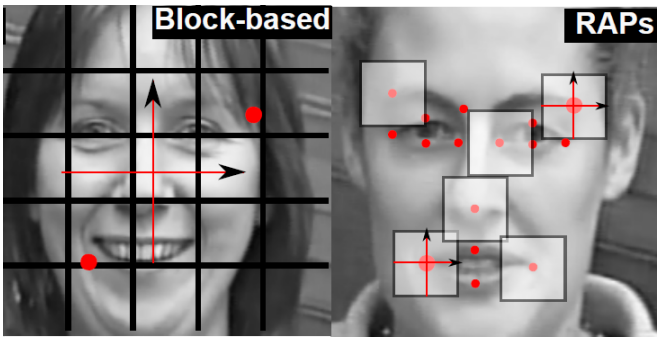

Fig. 1. Feature extraction methods: holistic/block-based (left) and part-based/RAPs (right). Image taken from [5].

to heuristics, and the contributions of all the region-specific classifiers are learnt in combination.

We conduct an experiment to demonstrate that the ANN architecture equivalent to decision-level fusion shows superior performance when compared to feature concatenation for facial AU recognition. Furthermore, we perform an experiment on the GEMEP-FERA [17] dataset, which has state-of-the-art complexity, in order to compare the performance of our method with that of Jiang et al. [5].

II. RELATED WORK

There has been a long-standing discussion regarding which features offer the best performance for automatic AU analysis. Every work has claimed the superiority of some combination of features, registration method, performance measure and dataset. However, there has not been a uniform conclusion and different works still resort to different features. Some features types are more commonly used with specific image representation strategies. Specifically, it is common to use LBP/LPQ features in a holistic manner in combination with a tiling approach [6]. I.e., the whole face bounding box is divided into a regular lattice (see Fig. 1), the feature representation is computed on each tile, and then the features are concatenated into a single vector. The use of histogram-based features is common to alleviate the effect of poor registration of holistic approaches, while tiling is aimed at maintaining some spatial information within the feature dimensionality.

Alternatively, some works use a part-based model, computing a feature representation of a small patch around each facial landmark, and then all of resulting feature vectors are concatenated into a single one. This strategy is often combined with HOG/SIFT features (e.g. [3]). There is a long tradition in computer vision for using these features in part-based approaches and thus it is a natural choice. However, whether they offer the best performance for part-based representations is not clear, and other feature representations might be also similarly adequate.

Holistic methods are capable of representing the whole face appearance and not only patches around points. Furthermore, they are sensitive to flexible shape deformations (e.g. the lip stretching which is associated with a smile). However, they offer a poor registration, in the sense that since face images are only globally aligned, each pixel will refer to slightly different parts of the face on different examples. Thus, each feature encoding appearance will refer to a slightly different part of the face, and extracting generalising fine-grained patterns from holistically-computed appearance feature vectors is challenging. Part-based models offer instead a much better registration. However, they are less sensitive to flexible movements, and they do not represent the whole face. Furthermore, works such as Lucey et al. [9], where the registration is taken to the extreme and all faces are registered with a piecewise affine transformation into neutral frontal pose (thus maximising registration and yet eliminating an important amount of the expressive information). Remarkably, this strategy offers good performance, and highlights the paramount importance of a good face registration strategy [5]. However, whether there is a better intermediate option in which less expressive information is lost, is a reasonable question.

The work by Jiang et al. [5] offered an alternative solution. It consists of using the facial landmarks to create a mesh, as in typical works on active appearance models [12]. Then, face regions are defined by merging some of these triangles. Which triangles to fuse is manually defined, but the decisions are based on domain knowledge relating different regions to the facial muscles responsible for the AU. This strategy showed superior performance in combination with different feature extraction approaches and for two state-of-the-art databases [5]. Thus, we adopt this strategy in this work.

However, the problem of how to combine information coming from different regions has received much less attention. Action Units are localised within the face by definition. Thus, restricting the regions of the face used as input is reasonable, particularly when considering the low number of examples available for training. However, AUs often co-occur, and information from other regions can be used as well for a given AU, thus making it infeasible to define a hand-crafted combination of regions for each AU. For example, AU12 (lip stretcher, typical of smiles) co-occurs often with AU6 (the squinting typical of spontaneous smiles). Thus, the face regions relevant to AU12 include that on the outer part of the eyes.

Thus, learning which parts of the face to give importance to and which not is an interesting problem that has received little attention so far. One such approach was presented for facial expressions in Zhong et al. [21], where they employed a sparse multi-task learning approach to find relevant parts capturing expressive behaviour (i.e., for any expression) and expression-specific behaviour (i.e., for specific expressions). An alternative for the case of AU was presented by Jiang et al. [5], who used a decision-level fusion approach to combine region-specific classifiers. In this work we follow the same approach, but replicating the structure for the case of ANNs and replacing the weighted combination of regional decisions with a decision layer learned from data. Our study hence is one of the first to study how to learn region-specific weights for the case of AU detection.

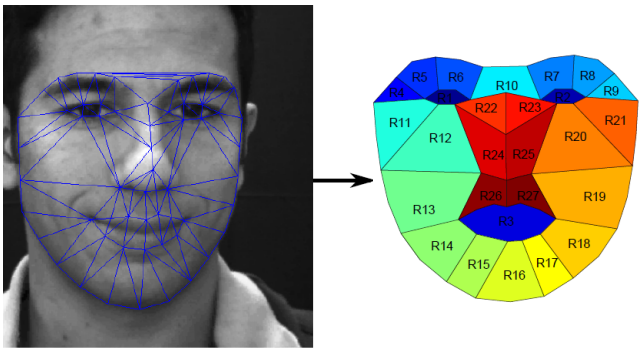

Fig. 2. Delaunay triangulation (left) using facial landmarks which is utilized in defining a set of 27 facial regions (right). Image on the left has been taken from [5].

III. METHODOLOGY

A. Facial regions

In this work, we have used the approach described by Jiang et al. [5] to divide the face into distinct non-rectangular regions using facial landmarks. The motivation for this approach is to divide the entire face into a set of homogenous regions which can capture the entire face and also offer good registration properties across faces. Defining these regions in terms of the facial landmarks guarantees that the facial regions represented are homogeneous across faces. In particular, the regions are construct using a Delaunay triangulation, dividing the face in the same manner as that of Active Appearance models (See figure 2). Since some of the resulting triangular regions are too small to encode any relevant information, they are merged together to form larger facial regions. The sets of triangles which are fused, capture semantics of facial actions as they were defined using domain knowledge regarding FACS [4] AU definitions (each AU corresponds to local appearance changes caused by activation of localised facial muscles). The resulting facial regions are polygons of variable shapes and sizes.

In this work, we follow the region definition described in Jiang et al. [5], resulting in 27 distinct facial regions, shown in Fig. 2. It is interesting to note that this representation covers the whole face except the forehead. Although some expressive information can be attained from the wrinkles of the forehead, there are no facial landmarks enclosing this region. However, partial occlusion of the forehead due to facial hair is common, which might complicate their usage for inference.

B. Decision Level Fusion in Artificial Neural Networks

In order to learn a classifier for facial AU recognition, we use an Artificial Neural Network (ANN) whose topology is designed to fuse features coming from different regions at the top layer, i.e., the decision layer. The final decision value for a test image is a weighted sum of the scores for each region of the face. Thus, the face region level classifiers and the weight for each classifier are learnt jointly in this ANN. This network topology is shown on the left-hand side of Fig. 3. It consists of n input layers corresponding to n feature vectors, computed from n facial regions. Each input layer is connected to a separate unit in the hidden layers. Hence the hidden layer has n units, each connected to a separate input layer. This architecture is the ANN equivalent of late or decision-level fusion.

The baseline against which we compare this architecture, consists of a fully connected ANN (see the right-hand side of Fig. 3), which is most typically used for classification purposes. The fully connected architecture corresponds to the feature-level fusion case.

The hidden layer, both in the proposed architecture and in the baseline architecture, is followed by an output layer of size 1 unit, so that each AU is learnt independently. Despite ANNs being able to provide multi-dimensional outputs, this would make it difficult to balance the training set in terms of positive and negative examples.

C. Feature extraction and Learning

In order to learn the classifiers for facial AU detection, appearance feature descriptors are extracted from each facial region described in section III-A. Due to the varying shape of each of these regions, we employ histogram-based features, which result in a representation of constant dimensionality. Hence, we get a set of feature vectors from each training example. These feature vectors are used as input to our proposed ANN described in section III-B. The network parameters are learnt using the back propagation algorithm.

In the above method, one feature vector from each facial region is fed into the corresponding input units. These input units are connected to a single hidden unit. Hence each group of these input and hidden units acts as a region specific sub-classifier for the target AU detection. The weights in the final layer indicates the relative importance of each sub-classifier learnt by the ANN.

IV. EVALUATION

Databases: In order to demonstrate the advantage of our proposed approach, we performed experimental evaluation of our models using the MMI [15], Cohn-Kanade (CK) [8] and the GEMEP-FERA [17] databases. Both the CK and the MMI datasets contain videos of expressions posed on command and thus are not naturalistic. The recording conditions are controlled, so that the subjects keep a frontal view to the camera at all time. While the videos within the MMI dataset contain full activation episodes, the videos within the CK contain instead only the activation of the expression until its apex. They are considered to be easy datasets, and they serve the purpose of comparing the proposed architecture with the baseline architectures. Instead, the GEMEP-FERA dataset is more challenging. While the expressions are also posed on command, the subjects are professional actors who were free to decide how to display a set of emotions. Thus, the conditions are more realistic. Different from MMI and CK, the subjects do not keep a frontal head pose with respect to the camera. Thus, this is considered to be a state-of-the-art dataset in terms of its challenging conditions. It is within

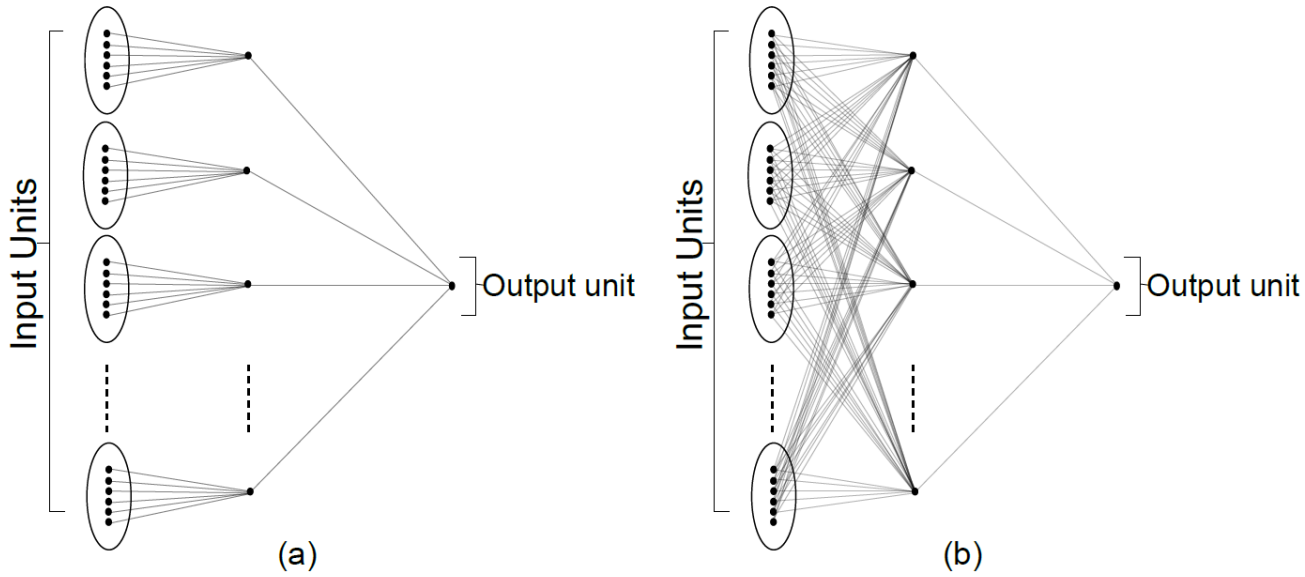

Fig. 3. Topology of our proposed ANN (left) as compared to a conventional ANN (right) normally used for classification tasks. In contrast to a conventional ANN, our ANN is not fully interconnected between the input layer and the hidden layer.

this dataset that we compare our approach to the performance attained by Jiang et al. [5].

Pre-processing: In all of our experiments, the training and testing images were pre-processed by detecting facial landmarks using the method described in Xiong & De La Torre [19]. This algorithm provides 49 inner-facial landmarks (i.e., no landmarks on the face contour are computed). Using these landmarks, a face registration step is performed. In particular, the facial landmarks that are stable under expressions (i.e., the corners of the eyes and the nose landmarks) are used to register the face to some anchor landmarks (we used the mean shape as the anchor landmarks). The registration is attained using a Procrustes transformation, which accounts for in-plane rotations, translations and uniform scaling.

Then, appearance feature descriptors are extracted from each of the 27 facial regions described in section III-A. For simplicity, we used the Local Binary Patterns (LBP) and, in particular, the uniform LBP version [14] as the feature descriptor. The uniform LBP is a 59-dimensional histogram. We restricted ourselves to histogram-based features due to the irregular shape of the regions where the features are computed.

Artificial Neural Network parameters: The ANNs that we employed in our experiments has a logistic sigmoid activation function in the hidden units and in the output units. For learning the parameters of the network, mean-squared error was used as the loss function. An L_2 regularisation is used, and no early stopping is performed. The network optimization was done using the Scaled Conjugate Gradient Backpropagation algorithm [13].

Experiments: We conducted 2 different experiments for evaluating the performance of our models for AU recognition task. In the first experiment, we compared the performance of our proposed partially connected ANN against fully connected ANNs, which are generally used for classification tasks and correspond to the feature-level fusion strategy. Our second experiment compares the performance of our approach with the approach used by Jiang et al. [5].

For the first experiment we used approximately 3000 images extracted from the video sequences from the MMI and Cohn-Kanade datasets. These datasets were used to compare the performance of our proposed architecture with respect to the baseline architectures. Our baseline model consists of a ANN which takes the concatenated features from all facial regions defined by the facial landmarks. In consequence, as opposed to our proposed ANN, the baseline ANN is fully connected, i.e. each input unit is connected to each and every hidden unit. We tried 2 different sizes of the hidden layer for the baseline models. One baseline has a hidden layer with 27 hidden units (same as our proposed ANN model) and the other baseline network has a hidden layer with 1000 hidden units. A 3-fold subject-independent cross-validation was used to evaluate the performance on these datasets. Since the learning of any ANN is sensitive to initialization, therefore, we trained and tested our model 10 times for each AU and the median performance was reported. We used the area under the Precision-Recall Curve (AUC) as the performance measure for this experiment.

Table I shows the performance comparison of the baseline models (with 27 and 1000 hidden units) and our proposed model (partially connected with 27 hidden units) for 14 different AUs. Our approach shows superior performance in 11 out of the 14 AUs considered. The average performance over all AUs is also higher for our approach. From the table,

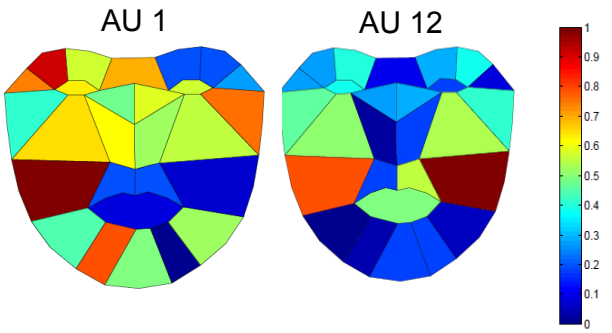

Fig. 4. Visualization of the relative importance of each facial region captured in the learnt weights between the hidden layer and the output layer of our proposed ANN. On the left is the visualization for AU1 and on the right is the visualization for AU12. For illustration purpose, the weights here have been normalized to lie between 0 and 1.

it is clear that the decision level fusion approach in ANNs performs better than the standard feature concatenation approach.

In the second experiment, we computed the performance of our approach on the GEMEP-FERA dataset and compared it with the approach used in Jiang et al. [5]. For this experiment, we used only the training partition of this dataset. The experiment was done for 12 different AUs and a leave-one-subject-out cross validation was used to evaluate the performance for each AU. Since Jiang et al. [5] reported the results attained using the 2AFC scores as the performance measure, we also adopt this criteria when reporting our results. Table II shows the performance comparison of the two approaches. Our approach shows higher performance in 8 out of 12 AUs. The average performance over all the 12 AUs is also significantly higher, which shows that the joint learning of the region-specific classifiers and the classifier weights using our proposed ANN performs better than learning the classifier and their corresponding weights separately using the method proposed in Jiang et al. [5].

Similar to Jiang et al. [5], we are also capable of visualising which parts of the face are important for detecting any given AU. It is important to note that these regions do not need to be restricted to the regions where the corresponding facial muscle produces an appearance change. Instead, the relevant regions often correspond to AUs that frequently co-occur with the target AU. Examples of this visualisation process are shown in Fig. 4, for AU1 and AU12. We found however that often the weights that were high on one side of the face were low on the same region on the other side of the face. This might reflect that the information between both sides is more commonly redundant, and thus the regularisation term primes solutions where all the information is extracted from only one region rather than having high weights for both of the regions. This effect reflects the fact that each of the region classifiers are indeed learnt in conjunction to those of the other regions. However, it is clear that such learned region-fusion weights would not allow detection of uni-lateral AUs on the side of the face with low

AU	Fully Connected ANN (27 hidden units)	Fully Connected ANN (1000 hidden units)	Partially Connected ANN (Our method)
1	0.73	0.74	**0.76**
2	0.57	0.60	0.60
4	0.57	0.57	**0.58**
5	0.43	**0.47**	0.43
6	0.53	0.52	**0.56**
7	0.16	**0.18**	0.16
9	0.66	0.67	**0.70**
10	0.18	0.16	0.18
12	0.69	0.72	**0.73**
15	0.25	0.35	**0.39**
17	0.53	0.53	**0.57**
18	0.07	0.08	**0.12**
20	0.51	0.51	**0.53**
25	0.83	0.83	**0.85**
Mean	0.48	0.49	**0.51**

TABLE I

PERFORMANCE (AUC) COMPARISON OF A FULLY CONNECTED ANN WITH THE PARTIALLY CONNECTED ONE (OUR METHOD), ON THE MMI+COHN-KANADE DATABASE.

AU	Jiang et al. [5]	Our method
1	0.64	**0.81**
2	**0.72**	0.71
4	0.53	**0.65**
6	0.72	**0.82**
7	**0.68**	0.64
10	**0.66**	0.60
12	0.74	**0.79**
15	0.53	**0.67**
17	0.70	**0.73**
18	**0.75**	0.72
25	0.57	**0.60**
26	0.53	**0.59**
Mean	0.65	**0.69**

TABLE II

PERFORMANCE (2AFC) COMPARISON OF OUR METHOD WITH JIANG ET AL. [5], ON THE GEMEP-FERA DATABASE.

weights. For e.g., a face image in which AU1 is activated only on the left-hand side, will not get detected. The outcome of these weights also reflects the fact that relatively little data of asymmetric facial expressions is available.

V. CONCLUSIONS AND FUTURE WORKS

This paper provides further evidence of the good performance attained by decision-level fusion strategies, attaining superior performance compared to feature-level fusion. It is also shown that an Artificial Neural Networks approach can learn all the part-based classifiers and the weights corresponding to the decision-level layer jointly. Further experiments are required however, including varying features, further combinations (e.g. part-based approach and decision-level fusion) and experiments on more datasets. Further experimentations with different ANN architectures and parametrization might also result in better performance. It also remains to be explored whether a decision-level fusion

strategy could be successfully applied to other AU-related problems such as AU intensity estimation. Finally, it is possible to extend the architecture to learn a combination of region-specific parts, but where the parts are computed using more than one appearance representation approach (i.e., include the holistic tile-based and the local representations). Furthermore, geometric features could be also included in the fusion.

REFERENCES

[1] M. Dahmane and J. Meunier. Emotion recognition using dynamic grid-based hog features. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 884–888, March 2011.

[2] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon. Emotion recognition using phog and lpq features. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 878–883, March 2011.

[3] X. Ding, W.-S. Chu, F. De la Torre, J. F. Cohn, and Q. Wang. Facial action unit event detection by cascade of tasks. In *International Conference on Computer Vision*, 2013.

[4] P. Ekman, W.V. Friesen, and J.C. Hager. *Facial Action Coding System (FACS): Manual*. A Human Face, Salt Lake City (USA), 2002.

[5] B. Jiang, B. Martinez, M. F. Valstar, and M. Pantic. Decision level fusion of domain specific regions for facial action recognition. In *International Conference on Pattern Recognition*, 2014.

[6] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic. Dynamic appearance descriptor approach to facial actions temporal modelling. *Trans. Systems, Man and Cybernetics, Part B*, 44(2):161–174, 2014.

[7] Bihan Jiang, M.F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 314–321, March 2011.

[8] Takeo Kanade, Yingli Tian, and Jeffrey F. Cohn. Comprehensive database for facial expression analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–, 2000.

[9] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, and K. M. Prkachin. Automatically detecting pain in video through facial action units. *Trans. Sys., Man and Cybernetics, Part B*, 41:664–674, 2011.

[10] Brais Martinez, Michel F. Valstar, Xavier Binefa, and Maja Pantic. Local evidence aggregation for regression based facial point detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2013.

[11] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool. Face detection without bells and whistles. In *ECCV*, 2014.

[12] Iain Matthews and Simon Baker. Active appearance models revisited. *Int'l Journal of Computer Vision*, 60(2):135–164, 2004.

[13] Martin Fodslette Møller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, 6(4):525–533, 1993.

[14] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[15] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Multimedia and Expo, IEEE International Conference on*, pages 5 pp.–, July 2005.

[16] Caifeng Shan, Shaogang Gong, and Peter W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing*, 27(6):803 – 816, 2009.

[17] M.F. Valstar, Bihan Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 921–926, March 2011.

[18] Paul Viola and Michael J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.

[19] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 532–539, 2013.

[20] Guoying Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, June 2007.

[21] Lin Zhong, Qingshan Liu, Peng Yang, Bo Liu, Junzhou Huang, and D.N. Metaxas. Learning active facial patches for expression analysis. In *Computer Vision and Pattern Recognition, IEEE Conference on*, pages 2562–2569, 2012.

[22] Yunfeng Zhu, F. De la Torre, J.F. Cohn, and Yu-Jin Zhang. Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *Affective Computing, IEEE Transactions on*, 2(2):79–91, April 2011.