Decision Level Fusion of Domain Specific Regions for Facial Action Recognition

Bihan Jiang*, Brais Martinez*, Michel F. Valstar† and Maja Pantic*‡
*Department of Computing, Imperial College London, UK
†Mixed Reality Lab, School of Computer Science, University of Nottingham, UK
‡Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Netherlands

***Abstract*—In this paper we propose a new method for the detection of action units that relies on a novel region-based face representation and a mid-level decision layer that combines region-specific information. Different from other approaches, we do not represent the face as a regular grid based on the face location alone (holistic representation), nor by using small patches centred at fiducial facial point locations (local representation). Instead, we propose to use domain knowledge regarding AU-specific facial muscle contractions to define a set of face regions covering the whole face. Therefore, as opposed to local appearance models, our face representation makes use of the full facial appearance, while the use of facial point locations to define the regions means that we obtain better-registered descriptors compared to holistic representations. Finally, we propose an AU-specific weighted sum model is used as a decision-level fusion layer in charge of combining region-specific probabilistic information. This configuration allows each classifier to learning the typical appearance changes for a specific face part and reduces the dimensionality of the problem thus proving to be more robust. Our approach is evaluated on the DISFA and GEMEP-FERA datasets using two histogram-based appearance features, Local Binary Pattern and Local Phase Quantisation. We show superior performance for both the domain-specific region definition and the decision-level fusion respect to the standard approaches when it comes to automatic facial action unit detection.**

I. INTRODUCTION

The Facial Action Coding System (FACS) is a taxonomy of human facial expressions designed to facilitate human annotation of facial behaviour. It specifies a list of 32 atomic facial muscle actions, named Action Units (AUs), and 14 additional descriptors that account for miscellaneous actions. Automating the AU annotation process is widely regarded as an important step towards their deployment in a wide range of problems, including medical applications [12], social behaviour modelling [3] or security applications [7].

Constructing an effective face representation from images is a crucial step for successful automatic facial action analysis. One of the most widely used cues are appearance-based features, which aim to capture differences in appearance caused by muscle actions both in terms of changes of permanent facial features (e.g. the triangular shape of the mouth corner when smiling) as well as transient features such as wrinkles, bulges and furrows that only appear when an action is performed. There is a wide range of literature on automatic AU analysis focusing on the problem of finding the ideal feature descriptors. For example, the performance of Local Binary Patterns (LBP) [14], Histograms of Oriented Gradients (HOG) [5], Local Phase Quantization (LPQ) [13], and Local Gabor Binary Patterns (LGBP) [17] have been studied for this particular problem, as well as a number of dynamic appearance descriptors such as LBP-TOP [14], [23], LPQ-TOP [14], and LGBP-TOP [2].

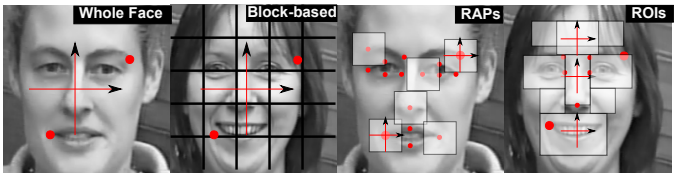

Fig. 1. Different ways to apply appearance descriptors (holistic vs. local features): whole face, block-based holistic, local appearance based on point locations, and local appearance based on regions of interest. Pixels indexed by the same local coordinates have the same semantic meaning (last two), but pixels indexed by the same global coordinates have different semantic meanings due to the face shape variation (first two)

However, which parts of the face these descriptors should be applied to is an important question that has received very little attention. The only two commonly used approaches are the holistic and local strategies for feature extraction. In a holistic approach the whole region defined by the face bounding box is encoded in the feature representation. This is typically used in combination with a block-based representation by which the face is divided into a regular grid, a feature vector is extracted from each of the grid's blocks, and the resulting vectors are concatenated to form the final feature representation (e.g. [14]). Alternatively, local approaches apply an appearance descriptor to image patches centred at a number of facial landmarks. Each landmark-localised patch produces a feature vector which is then concatenated to create the face representation [24] (see Fig. 1).

In this work, we discuss the pros and cons of these two approaches, and propose a novel feature extraction strategy based on a new definition of the regions from which local appearance features are computed, which is our first major contribution. The new definition results from the application of domain knowledge (expert knowledge of how FACS is defined), and aims at capturing AU-related changes within a region.

The second contribution of this paper addresses the way information from multiple patches is combined. Existing methods extract features from regions (where the definition of *region* can vary), and then proceed to concatenate the extracted features into a single vector. This is called feature-level fusion,

as the information coming from different sources is combined into a single feature vector. However, the activation of an AU only causes appearance changes in a subset of the face patches. This suggests that a decision-level fusion approach can be beneficial. Decision-level fusion relies on the application of a separate classifier for each region considered. This is followed by the combination of the resulting predictions into a single prediction by some function, often again a classifier or a weighted product/sum of the region-based classifiers. Our second contribution is to adopt this strategy, experimentally comparing it to existing feature concatenation schemes. In particular, we consider a linear weighted sum of the region-based decisions, for which the weights are derived from a cross-validated performance of each region's classifier.

To summarise, our main contributions are a novel way to extract facial appearance features, described in full in section III, and the proposal of a decision-level fusion strategy for combining the region-level classifiers is described in section IV. In section V we experimentally show that both of the proposed contributions result in a performance improvement. The improvement is consistent using two different appearance descriptors on two separate datasets. Section VI provides our concluding remarks and future direction. However, prior to that, we review some related work in section II.

II. Related Work

Depending on which features are used, automatic AU analysis works can be divided into appearance-based, geometry-based, motion-based, or hybrid approaches (those that combine at least two of the previous approaches). In this work we focus on appearance-based methods. In turn, existing appearance-based methods can be divided into holistic methods and local (or part-based) methods. Holistic methods try to model the appearance of the whole face by applying an appearance descriptor over the full face region. Alternatively, local approaches apply the appearance descriptor at local patches centred at the facial landmarks.

Some features can be naturally applied in a holistic manner, as Gabor magnitude features (e.g. [4]), although the resulting feature dimensionality is very large. Some recently successful features, like LBP or LPQ, are histogram-based. In these cases, representing the full face appearance by means of only one histogram becomes suboptimal. It has been shown that, then, the use of a block-based representation greatly improves the results [16], [14], [13]. In particular, in a block-based representation, the face patch is divided in a grid-like manner into blocks, and an appearance descriptor is applied to each of the blocks. Some works use overlapping blocks to improve the robustness to face registration errors [10]. Furthermore, block-based representations have also been successfully applied to non-histogram features like DCT [8]. Notably, this strategy was used in combination with the use of LGBP features by the winner of the FERA 2011 AU detection sub-challenge [17].

This approach, however, also lead to a larger feature dimensionality.

Local appearance-based approaches are typically constructed by computing an appearance representation from regions around the landmark point (a strategy noted as RAP here). Most features are suited for this approach. Furthermore, some as SIFT or HOG work best locally, since otherwise large edges due to the face structure dominate over small edges related to facial expression information. For example, the work in [24] studied the performance of Gabor, SIFT and DAISY features when applied around facial landmarks for AU detection. They showed that SIFT features worked comparably to DAISY and slightly outperformed Gabor features.

In terms of classification, the standard approach is to concatenate the features extracted from different regions around points/blocks to form a single vector. Then learning is performed over the concatenated features (e.g. [11], [22], [13]). Feature selection techniques such as GentleBoost can be applied to select the most discriminative features (e.g. [18]). Fusion techniques other than feature-level fusion have been studied in order to combine different types of features, such as geometry-based and appearance-based features [17]. However, to the best of our knowledge, it does not exist in the AU literature any work that explicitly studies fusion techniques for combining information extracted from different face regions.

III. Region Definition and Feature Extraction

Holistic and local feature extraction approaches convey somewhat different information. Holistic approaches extract information according to a coordinate system defined by whole face. On the other hand, local methods use a coordinate system defined in terms of inner-facial features, for example facial landmarks (see Fig. 1). Therefore, the level of registration attained by local methods is superior to that of holistic methods. More specifically, since holistic methods only use Procrustes analysis to perform the data registration, the physical parts of the face from which a feature is extracted can change considerably between different examples. These differences become particularly evident when dealing with non-frontal head poses. In contrast, local methods present the advantage of always encoding the appearance of the same part of the face, as long as the facial points they depend on are detected correctly. In contrast, holistic methods present the advantage of encoding the whole face appearance, while local methods fail to encode information from some portions of the face. For example, the cheeks can be a useful cue despite not containing facial landmarks.

Our proposed strategy aims to make full use of the face appearance, yet maintaining the benefit of the strong registration afforded by the facial landmark points. Therefore, our proposed descriptor is less sensitive to shape differences caused by identity and non-frontal head poses than holistic approaches. To this end, we consider a set of regions defined by facial landmarks as show in Fig. 2. It is possible to see that in this case, the whole appearance of the face is considered, including regions such as the cheeks. Furthermore, the regions considered do not have a uniform local support. For example,

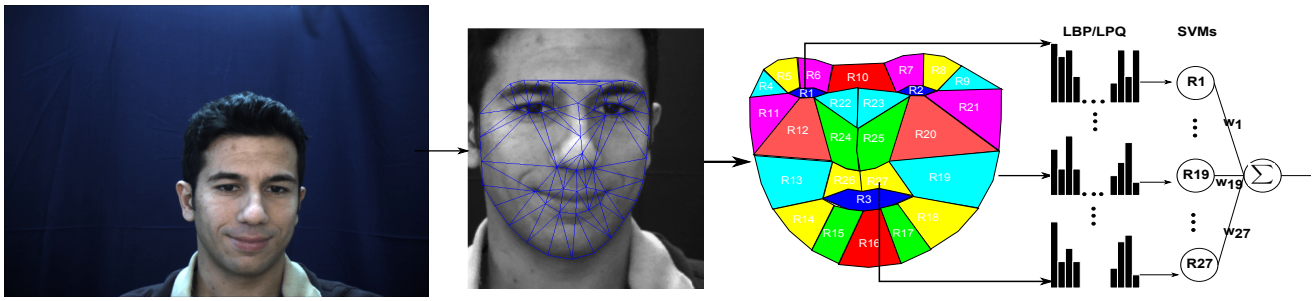

Fig. 2. Representation of the AU detection process. Each of the regions produces a separate histogram. Each of these histograms is analysed by means of a region-specific classifier, and the outputs are combined into a final score as a weighted mean.

the a large portion of the image corresponds to the lips, while the eyes cover a much smaller region.

The regions were constructed by first considering the Delaunay triangulation used for dividing the face in Active Appearance Models (AAM) [1]. Each triangle in the mesh is designed to encode information as homogeneous as possible, and represents a natural way for dividing the face region. However, for the task of AU detection some of the triangles encode very little information. Therefore, we used domain knowledge regarding FACS AU definitions for joining some of these triangles into larger regions. For example, R4 and R9 enclose the region of the face in which the contraction of the corrugator produces furrowing (AU 4). Similarly, R22 and R23 clearly show wrinkling when the depressor supercilli muscle contracts, which is again related to AU4. Similarly, AU25 is captured by R3. A full mapping can easily be derived from the FACS [6]. An overview of mappings between facial muscles and AUs can be found in Table 2.1 of [19].

Our hypothesis is that in this way relevant changes due to AUs will be encoded more homogeneously within the appearance descriptor. While for holistic approaches what is encoded within a block varies from image to image, local features encode the aspect of patches such as the corner of the mouth or the lip contour, but do not necessarily encode the aspect of the interior of the mouth. Furthermore, even when parts of the inner mouth are encoded, the appearance is combined with that of the outer parts of the mouth, producing less characteristic patterns.

Since the selected regions are now of varying size and shapes, we are restricted to histogram-based feature representations, such as LBP and LPQ features[1]. They have however been consistently among the best performing features for AU detection (e.g. [20]). In our case, we present experimental results for both LBP and LPQ features.

IV. Region-based classification

In the classical feature-level fusion all the appearance descriptors from the different regions are concatenated into a single feature vector. We propose instead a decision-level fusion where a different classifier is trained for each of the regions defined in section III. To obtain a final prediction of an AU being active or not in a given image, the outputs of the region-based classifiers are fused using a weighted sum of the individual scores.

This approach is expected to have a number of advantages. Training a separate classifier for each region allows each classifier to learn over more specific and uniform parts of the face, resulting in less class overlap in the features. By means of the weighted combination of region-based scores, parts of the face that do not display any visible changes when a target AU is active will have a reduced impact on the final decision (in essence the classifier for such a region would be trained on random noise). Finally, it reduces the dimensionality of the problem, again making the learning task easier. An example of region weighting for AU12 when applied to a holistic block-based approach is shown in figure 3. As expected, the regions around the mouth and cheeks have the highest impact, some regions around the eyes also contribute to a lesser degree (mostly due to the correlation with AU6), while the rest of the regions produce a low impact on the final decision.

In particular, we use an SVM as the binary classifier for each region. Since both LBP and LPQ are histogram-based features, we adopt a histogram intersection kernel. Therefore, the only parameter to be optimised is the soft margin parameter. This is achieved through a grid search strategy, and a subject-independent cross-validation has been carried out. A probabilistic score is computed by using the logistic function. That is to say, given the output of the SVM trained to detect AU i from the face region j, we compute:

$$p(c_i = 1|x_j) = \frac{1}{1 + e^{s_{i,j}(x_j)}} \quad (1)$$

where c_i is a binary indicator of the action of AU i, and x_j is the appearance feature representation of region j.

The per-region scores are joined together as:

$$p(c_i = 1) = \sum_{j=1}^{n} w_j p(c_i = 1|\mathbf{x}_j) \quad (2)$$

In order to find the weights w_j, we conducted a subject-independent cross-validation experiment within the training set for each region. Through this process, we computed a

[1]HOG features typically use a block-based internal representation of the represented image patch, rendering it inadequate for our purpose

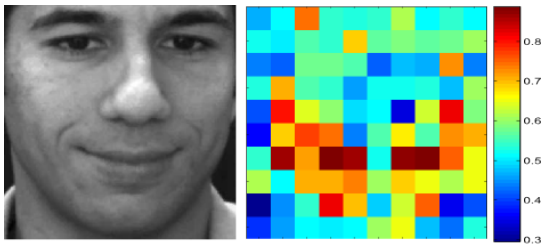

Fig. 3. The performance of each sub-classifier from 10×10 facial regions for the detection of AU12 in using block-based (BLK) regions.

performance score for each region. In our case, we used the 2-alternative forced choice task (2AFC) score. The percentage of correctly classified examples in a 2AFC evaluation framework is equivalent to the area under the ROC curve (AUC) [9], and can be computed more efficiently than the AUC itself. The weights are obtained by averaging the 2AFC score over all folds. This follows the rationale that the performance of a region based classifier on a predefined evaluation set gives a good measure of how relevant the information within the region is towards the detection of the target AU.

It is common to further apply an output-smoothing step in order to enforce temporal consistency on the predicted outputs, for example using a first order Markov chain. In our case, we avoid this step as our aim is to compare the performance of the proposed steps respect to equivalent standard procedures. However, when designing an expression recognition system one would expect to make use of the temporal consistency.

V. Experimental results

In order to show the improvement attained by each of the proposed contributions, we conducted experiments on two datasets commonly used for automatic AU analysis. In particular, we use the Denver Intensity of Spontaneous Facial Actions (DISFA) database [15], and the GEMEP-FERA challenge dataset [21]. We also show experimental results when using both LPQ and LBP features, as they are some of the most common features for automatic AU analysis. In order to show that each of the proposed steps result in a performance improvement, we conduct two experiments. The first experiment compares the performance of holistic and local feature extraction approaches against our region definition, while maintaining the standard feature-level fusion strategy. The second experiment shows that the performance is further increased by using the decision-level fusion, independently on the feature extraction strategy used and, in particular, is optimal when combined with the proposed region-based strategy. In the following we describe the experimental setting in more detail.

A. Experimental setting

The *DISFA dataset* contains videos of 27 participants in a controlled lab environment. Expressions were elicited by showing videos to the subjects. Since the subjects are looking to a screen placed in front of them, the head poses are typically frontal or near-frontal. Frame-based AU activation and intensity labels manually annotated by two FACS experts for 12 AUs are provided. The dataset also includes precise per-frame facial landmark locations for 66 points, obtained through subject-specific AAM. The accuracy of the landmarks might be unrealistic under more general conditions.

The *GEMEP-FERA challenge dataset* is a subset of the GEMEP dataset recently used for a challenge on AU detection [20]. All the expressions are acted. However, the participants are trained actors. In consequence, the expressions displayed are of similar characteristics to spontaneous ones. Non-frontal head poses are more common in this dataset, as the subjects have an unconstrained range of head motions. The dataset is split into a training set (including 7 actors), and a test set. We only use the training partition and perform a leave-one-subject-out cross validation to evaluate the performance. Since no tracked points are provided in this dataset, we automatically detected the facial landmark points using the AAM tracker proposed in [1]. We do not use subject-specific models for the tracking, making the results more generalizable. We put no effort in correcting the tracked points, except for eliminating glaring errors.

The *performance measure* used in this work is the 2-alternative forced choice task (2AFC). The percentage of correctly classified examples in a 2AFC evaluation framework is equivalent to the area under the Receiver Operator Characteristic curve (AUC), and can be computed more efficiently than the AUC.

B. Evaluation results

The first experiment shows how using our definition of regions affects performance. In this experiment we follow a standard feature-level fusion strategy instead of the proposed decision-level fusion approach. In this way, it is possible to judge the relative merits of each contribution on their own. As can be seen from Fig. 4, the proposed method results in higher average performance than both holistic and local approaches for the 4 combinations of datasets and features considered. It is interesting to see that the relative performance of the holistic approach is lower in the GEMEP-FERA dataset than on the DISFA dataset. The reason for this is likely to be the poorer face registration that Procrustes analysis can attain when the face is in a non-frontal position. GEMEP-FERA contains a lot of non-frontal head pose so a good registration is crucial for that dataset. Thus, the benefit of using inner facial structures to define the face regions is relatively higher. The average performances are 0.78, 0.77 and 0.80 2AFC for holistic, local and our approach for the combination of LBP and DISFA, and 0.78, 0.78 and 0.81 when LPQ features are used instead. For the GEMEP-FERA these numbers are 0.57, 0.60, and 0.62 for LBP, and 0.57, 0.59, 0.65 for LPQ.

Our second experiment shows the performance increase obtained by using the proposed decision-level fusion strategy. Tables I and II show the performance obtained when using our proposed definition of regions, and either a feature-level or decision-level fusion strategy. The reported experiments again

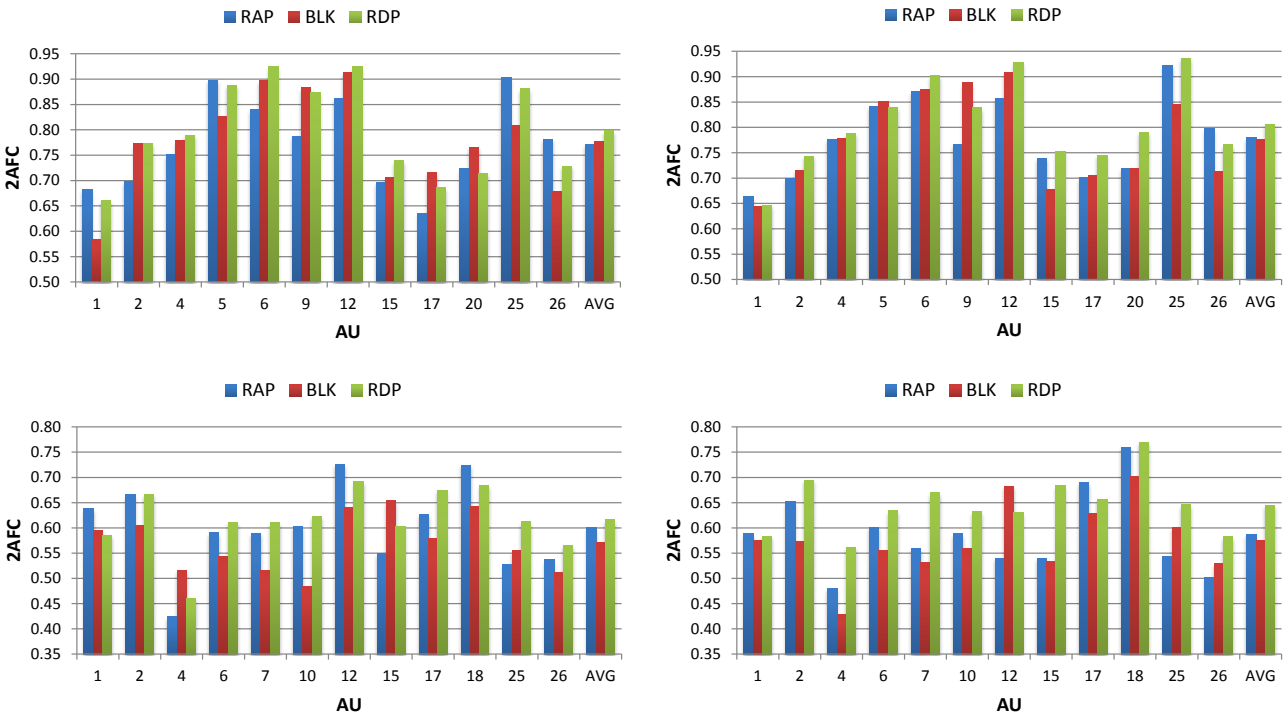

Fig. 4. Results in terms of 2AFC when using LBP (left column) and LPQ (right column) on the DISFA dataset (top row) and the GEMEP-FERA training set (bottom row). (RAP: region around points, BLK: block-based, RDP - region defined by points)

TABLE I
RESULTS (2AFC) FOR TESTING THE SYSTEM ON THE DISFA DATABASE USING LBP AND LPQ FOR FEATURE-LEVEL FUSION AND THE PROPOSED DECISION-LEVEL FUSION

	LBP		LPQ	
AU	**feat-lev**	**dec-lev**	**feat-lev**	**dec-lev**
1	0.67	**0.70**	0.59	**0.69**
2	**0.82**	0.79	0.73	**0.76**
4	0.74	**0.78**	0.68	**0.78**
5	**0.86**	0.84	0.77	**0.88**
6	0.90	**0.92**	0.88	**0.93**
9	0.80	**0.85**	0.77	**0.87**
12	0.91	**0.93**	0.90	**0.93**
15	0.63	**0.73**	0.69	**0.74**
17	**0.70**	0.68	0.73	**0.74**
20	0.72	**0.74**	0.77	**0.78**
25	**0.87**	0.85	**0.88**	0.85
26	0.76	**0.77**	0.75	**0.76**
AVG	0.78	**0.80**	0.76	**0.81**

TABLE II
RESULTS (2AFC) FOR TESTING THE SYSTEM ON THE GEMEP-FERA TRAINING SET USING LBP AND LPQ FOR FEATURE-LEVEL FUSION AND THE PROPOSED DECISION-LEVEL FUSION

	LBP		LPQ	
AU	**feat-lev**	**dec-lev**	**feat-lev**	**dec-lev**
1	0.63	**0.64**	0.61	**0.69**
2	0.68	**0.72**	0.68	**0.67**
4	0.47	**0.53**	**0.60**	0.51
6	0.64	**0.72**	0.63	**0.69**
7	0.59	**0.68**	0.71	0.71
10	0.57	**0.66**	0.60	0.66
12	0.66	**0.74**	0.62	**0.76**
15	**0.63**	0.53	0.66	0.53
17	**0.71**	0.70	0.68	**0.72**
18	0.72	**0.75**	0.72	**0.82**
25	**0.61**	0.57	0.58	**0.59**
26	**0.55**	0.53	**0.57**	0.55
AVG	0.62	**0.65**	0.64	**0.66**

include all 4 combinations of datasets and features, and for each of them the proposed decision-level fusion performs best.

We provide a summary of the performance increases of the different feature configuration approaches in figure 5. To this end, we have averaged the performance across all AUs, and all features (LBP and LPQ) and datasets (DISFA and GEMEP-FERA) used in this work. It is possible to see how both of the proposed novelties yield a performance increase. In particular, the best performing configuration that does not use any of the proposed improvements attain an averaged performance of 0.67 2AFC, while the proposed configuration yields 0.74 2AFC of average performance. This is a very large jump in relative performance of 11.1%. To put this into perspective, [13] and [2] recently proposed the use of appearance descriptors of spatio-temporal volumes as an alternative to frame-based appearance descriptors, attaining a 7% and 4% relative performance increase, respectively. Similarly, the performance boost between the baseline results on the FERA challenge and the winners of the challenge is of 13.7% [21].

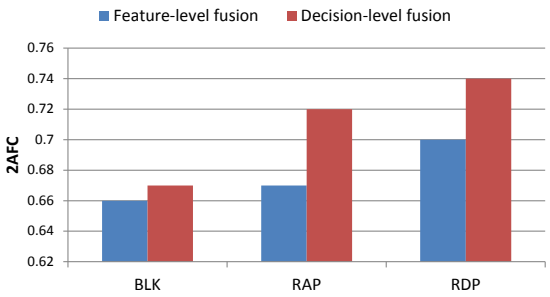

Fig. 5. Performance for different deffinition of regions, and for different fusion methods. Results are averaged across AU, and features and datasets used. (RAP: region around points, BLK: block-based, RDP - region defined by points)

VI. Conclusions

In this paper, we have proposed a novel definition of face regions from which to extract appearance features for automatic AU analysis. To this end, we proposed to use the triangulation constructed for AAM, and to merge the triangles of the resulting mesh based on domain knowledge. Furthermore, we have proposed a decision-level fusion strategy that is shown to improve over feature-level fusion for all of the feature extraction settings considered. These two contributions account to an average 11% performance increase over the baseline method, and we show that each of the proposed improvements increase the performance in every single experiment performed. In particular, the superiority of our region definition respect to holistic and local methods was shown in 4 different settings, while the superiority of the decision-level fusion was shown in 12 cases (i.e., for all combinations of feature, dataset, and region definition). The region-based approach currently does not make use of any information contained in the geometric parameters of the facial regions, such as the width or height of the mouth region. In future work we will include this information as part of our decision level fusion framework.

VII. Acknowledgments

This work has been funded by the European Community 7th Framework Programme [FP7/2007-2013] under grant agreement no. 611153 (TERESA). The work of Bihan Jiang is also funded in part by the EPSRC project EP/J017787/1 (4DFAB). The work of Brais Martinez was also funded in part by the EPSRC grant EP/H016988/1: Pain rehabilitation: E/Motion-based automated coaching. The work of Michel Valstar is supported by Horizon Digital Economy Research, RCUK grant EP/G065802/1.

References

[1] J. Alabort, G. Tzimiropoulos, S. Zafeiriu, and M. Pantic. Generic active appearance models revisited. In *Proc. IEEE Asian Conf. Computer Vision*, pages 650–663, 2012.

[2] T. Almaev and M. Valstar. Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In *Int'l Conf. on Affective Computing and Intelligent Interaction*, 2013.

[3] Z. Ambadar, J. F. Cohn, and L. I. Reed. All smiles are not created equal: Morphology and timing of smiles perceived as amused, polite, and embarrassed/nervous. *J. Nonverbal Behavior*, 33:17–34, 2009.

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

[5] S. W. Chew, P. Lucey, S. Saragih, J. F. Cohn, and S. Sridharan. In the pursuit of effective affective computing: The relationship between features and registration. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 42(4):1006–1016, 2012.

[6] P. Ekman, W. Friesen, and J. C. Hager. *Facial action coding system*. A Human Face, 2002.

[7] M. G. Frank and P. Ekman. The ability to detect deceit generalizes across different types of high-stakes lies. *Journal of Personality and Social Psychology*, 72(6):1429–1439, 1997.

[8] T. Gehrig and H. K. Ekenel. Facial action unit detection using kernel partial least squares. In *Proc. IEEE Int. Conf. Computer Vision Workshop*, 2011.

[9] D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. New York: Wiley, 1966.

[10] T. Gritti, C. Shan, V. Jeanne, and R. Braspenning. Local features based facial expression recognition with face registration errors. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 1–8, 2008.

[11] J. Hamm, C. G. Kohler, R. C. Gur, and R. Verma. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. *Journal of Neuroscience Methods*, 200(2):237–56, 2011.

[12] M. Heller and V. Haynal. The faces of suicidal depression. *Kahiers Psychiatriques Genevois*, 16:107–117, 1994.

[13] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic. Dynamic appearance descriptor approach to facial actions temporal modelling. *IEEE Transactions of Systems, Man and Cybernetics – Part B*, 2013. In press.

[14] B. Jiang, M. F. Valstar, and M. Pantic. Action unit detection using sparse appearance descriptors in space-time video volumes. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, pages 314–321, 2011.

[15] S. M. Mavadati, M. H. Mahoor, K. Bartlett, and P. Trinh. Automatic detection of non-posed facial action units. In *International Conference on Image Processing*, pages 1817–1820, 2012.

[16] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Combining AAM coefficients with LGBP histograms in the multi-kernel SVM framework to detect facial action units. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition Workshop*, pages 860–865, 2011.

[17] T. Senechal, V. Rapp, H. Salam, R. Seguier, K. Bailly, and L. Prevost. Facial action recognition combining heterogeneous features via multi-kernel learning. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 42(4):993–1005, 2012.

[18] C. Shan and T. Gritti. Learning discriminative lbp-histogram bins for facial expression recognition. In *British Machine Vision Conference*, pages 1–8, 2008.

[19] M. F. Valstar. *Timing is everything: A spatio-temporal approach to the analysis of facial actions*. PhD thesis, Imperial College London, 2008.

[20] M. F. Valstar, B. Jiang, M. Mehu, M. Pantic, and K. Scherer. The first facial expression recognition and analysis challenge. In *IEEE Int'l Conf. on Automatic Face and Gesture Recognition Workshop*, 2011.

[21] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer. Meta-analyis of the first facial expression recognition challenge. *IEEE Trans. Systems, Man and Cybernetics, Part B*, 42(4):966 –979, 2012.

[22] J. Wang and L. Yin. Static topographic modeling for facial expression recognition and analysis. *Computer Vision and Image Understanding*, 108(1):19–34, 2007.

[23] G. Y. Zhao and M. Pietikainen. Dynamic texture recognition using local binary pattern with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2(6):915–928, 2007.

[24] Y. Zhu, F. De la Torre, J. F. Cohn, and Y. Zhang. Dynamic cascades with bidirectional bootstrapping for action unit detection in spontaneous facial behavior. *IEEE Trans. Affective Computing*, pages 79–91, 2011.