Pattern Anal Applic (2019) 22:477–489
https://doi.org/10.1007/s10044-017-0645-4

THEORETICAL ADVANCES

Region-based facial representation for real-time Action Units intensity detection across datasets

Isabelle Hupont[1] · **Mohamed Chetouani**[1]

Received: 20 July 2016 / Accepted: 2 August 2017 / Published online: 7 August 2017

Abstract Most research on facial expressions recognition has focused on binary Action Units (AUs) detection, while graded changes in their intensity have rarely been considered. This paper proposes a method for the real-time detection of AUs intensity in terms of the Facial Action Coding System scale. It is grounded on a novel and robust anatomically based facial representation strategy, for which features are registered from a different region of interest depending on the AU considered. Real-time processing is achieved by combining Histogram of Gradients descriptors with linear kernel Support Vector Machines. Following this method, AU intensity detection models are built and validated through the DISFA database, outperforming previous approaches without real-time capabilities. An in-depth evaluation through three different databases (DISFA, BP4D and UNBC Shoulder-Pain) further demonstrates that the proposed method generalizes well across datasets. This study also brings insights about existing public corpora and their impact on AU intensity prediction.

Keywords Facial expressions recognition · FACS · Action Units intensity detection · Cross-dataset validation · Real-time processing

✉ Isabelle Hupont
hupont@isir.upmc.fr

Mohamed Chetouani
mohamed.chetouani@upmc.fr

[1] Institut des Systèmes Intelligents et de Robotique, Université Pierre et Marie Curie (Sorbonne Universités), CC 173 - 4 Place Jussieu, 75005 Paris, France

1 Introduction

Facial expressions are the most powerful, natural and direct way used by humans to communicate and understand each other affective state, intention and opinion. Driven by high commercial values and social interests, nowadays automatic facial expression detection is widely recognized as a key research component in different fields, ranging from Human–Machine Interaction [10] to drowsy driver detection [30], psychopathology [13] or physical pain studies [25].

There are two mainstreams for facial expression measurement: message-based and sign-based [8]. Message-based approaches seek to identify a high-level meaning for each facial expression, which often implies to label it into one or more basic emotional categories, such as "happiness", "disgust" or "sadness" [17, 18, 38]. However, basic expressions occur relatively infrequently in their prototypical forms in real-life and message-based approaches usually hide more subtle social messages. To overcome this drawback, sign-based approaches objectively describe the changes in the configuration of the face during an expression, rather than attempting to interpret its meaning.

The most widespread used sign-based approach is the Facial Action Coding System (FACS) [16], that decomposes facial expressions into small anatomically based components called Action Units (AUs). For example, AU12 codes contractions of the zygomatic major muscle (that occur e.g. during smiles) and AU4 implies contractions of the corrugator supercilii and depressor supercilii muscles (frequent in frowning episodes). For a subset of AUs, it is possible to code graded changes in intensity by using a 5-level scale ranging from "A" (trace) through "E" (maximum). Figure 1 shows examples of intensity variation in AU12 and AU4. It is important to notice that the relationship between facial appearance changes and AU intensities is not linear as, for

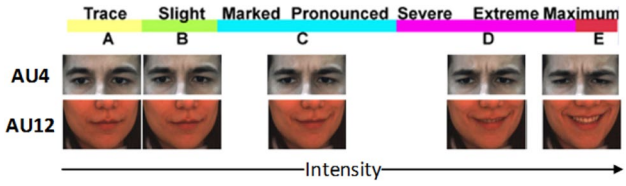

Fig. 1 Relationship between facial appearance changes and FACS intensity levels, according to the FACS manual [16]

instance, "C" and "D" levels cover a larger range of appearance changes than the other levels.

The suitability of FACS for automatic facial expression recognition has been widely acknowledged in the literature as most existing methods focus on detecting AUs and, optionally, map them to targeted emotions as a second step. However, almost all research has been directed at binary AU detection, i.e. whether facial actions are present or absent, instead of their full range intensity estimation [7, 42, 43]. Graded changes in intensity have rarely been considered, thus loosing part of the descriptive power of FACS and valuable information about the meaning and dynamics of facial expressions.

There are several reasons that make discerning intensities of AUs a far more challenging task than binary AU detection. Firstly, really few databases annotated in terms of AU intensities are available to train and validate algorithms. The reason is that the AU intensity annotation process is tedious and error-prone, and can only be reliably performed by experienced certified FACS coders. Secondly, AUs are seldom performed in isolation: the human face can show up more than 7000 AUs combinations, and this number largely increases if their intensities are also considered. Taking into account all these combinations is computationally impossible and models tend to learn the subset of most common AUs co-occurrences for the particular context of the dataset used for training [43]. Finally, typical automatic facial expression challenges, such as individual differences related to facial morphology and expressiveness across subjects, illumination variations or frequently occurring facial landmarks registration errors [6], impact even more severely the subtle AU intensity detection task. This subtlety implies fine-grained facial changes to be accurately registered and therefore demands more robust and sophisticated facial representation strategies compared to other higher-level facial expressions recognition tasks.

In this work, we present an automatic system able to perform intensity detection of the eight most commonly occurring AUs in social interaction [40]: AU1 "inner brow raiser", AU2 "outer brow raiser", AU4 "brow lowerer", AU12 "lip corner puller", AU15 "lip corner depressor", AU20 "lip stretched", AU25 "lips part" and AU26 "jaw drop" (Fig. 2). The main contributions of the paper can be summarized in the following way: we propose a method for AU intensity detection that relies on a region-based facial representation strategy conceived to mitigate the AUs co-occurrence problem and minimize the constraints on the accuracy of facial landmarks localization; the system is able to work in real-time while maintaining or even outperforming state-of-the-art approaches; and finally, we present an in-depth cross-dataset and cross-context validation study related to AU intensity detection. This research pursues exploration of the following novel research question: whether reducing AUs co-occurrences learning by separating muscular facial regions makes AU intensity detection more adaptable to new datasets and contexts.

The paper is further organized as follows. Section 2 reviews prior work. Section 3 provides a complete technical

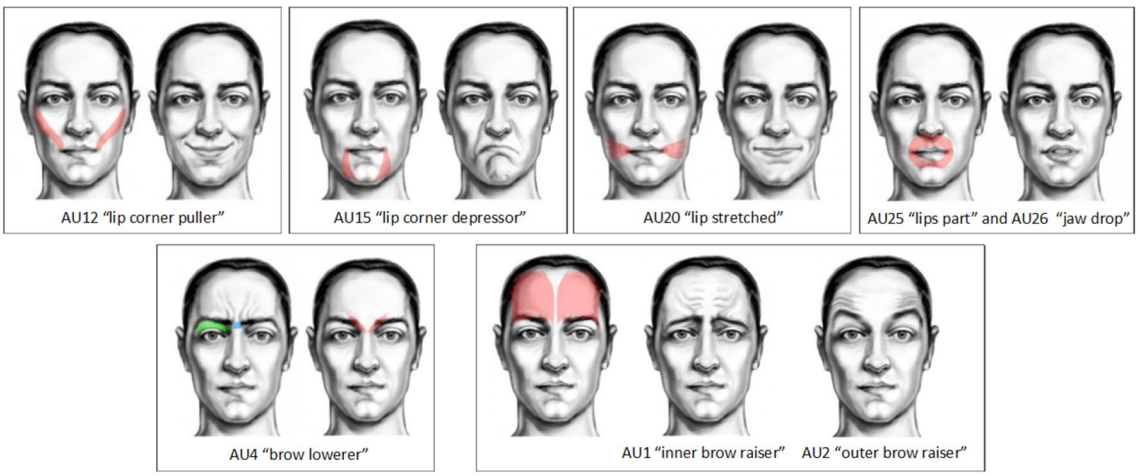

Fig. 2 AUs detected in this work. *Colored areas* represent facial muscles and regions activated during each AU. Drawings are courtesy of ARTNATOMY [11] (color figure online)

description of the method proposed for real-time AU intensity detection. Section 4 presents the performance achieved by the system, while Sect. 5 details the conducted cross-dataset experiments. Finally, Sect. 6 concludes the paper.

2 Previous work on AU intensity detection

2.1 Public databases

Nowadays, there is a large amount of visual datasets annotated in terms of presence/absence of AUs or message-based descriptors. However, there are really few databases that include AU intensity labels according to the FACS A–B–C–D–E intensity scale. The most relevant, regarding their public availability, the number of coded frames, their spontaneous (vs. posed) nature, and the quality and quantity of AUs considered, are summarized in Table 1: Denver Intensity of Spontaneous Facial Action (DISFA) database [28], Binghamton-Pittsburgh 4D Spontaneous Expression Database (BP4D) [48] and UNBC-McMaster Shoulder Pain Expression Archive (SP) [26].

As it can be seen, they differ in several factors, such as the type of AUs provided, the quality of the samples (SP resolution is much lower), the demographic makeup of participants (age, gender and ethnicity) and the number of FACS-certified annotators. The amount of annotated samples per AU and intensity, and the inter-observer reliability of manual FACS coding are also different, as depicted in Table 2. SP database has the highest inter-observer agreement (95%), but a much lower number of annotated frames. For instance, for AU12 it contains 6887 samples, while DISFA and BP4D contain 30794 and 78362 samples, respectively. Imbalance is even stronger for AU4, AU20, AU25 and AU26. More importantly, databases differ in terms of the context in which they were taken, while DISFA and BP4D were mostly recorded in typical social interaction situations (e.g. video watching, interviews), SP participants were performing painful shoulder physical tests. This variety in the databases is a valuable tool for cross-dataset validation of AU intensity detection models, but has not been fully exploited yet in the literature.

2.2 The AU intensity prediction problem

As any other facial expression recognition problem, detecting AUs intensity involves two main steps: (i) extracting the most appropriate and representative set of features from the face and (ii) use these features as inputs for a prediction model.

Regarding step (i), appearance-based approaches have been preferred to geometry-based approaches (e.g. facial angles or distances) in the literature. The reason is that the latter are especially vulnerable to even small facial landmarks extraction errors, as subtle intensity changes have to be registered. Local Binary Pattern Histograms (LBPH) [28, 45], Gabor features [12, 36, 41], Scale-Invariant Feature Transform (SIFT) [14] or Histogram of Gradients (HOG)

Table 1 Description of public databases containing AU intensity annotations

Name	AUs	Contents	Annotations	Participants	Elicitation
DISFA [28]	AU1–AU12 AU2–AU15 AU4–AU17 AU5–AU20 AU6–AU25 AU9–AU26	1 video per user (>130000 frames) Frontal faces without occlusions 1024 × 768 resolution at 20 fps under uniform illumination	Frame-level ground truth Annotations by 2 FACS coders	27 adults: 12♀ 15♂ Ages: 18–50 Ethnicities: 3 Asian, 1 African-American, 21 European, 2 Hispanic	4 minutes of emotional video clip watching
BP4D [48]	AU6–AU10 AU12–AU14 AU17	8 videos per user (>140000 frames) Frontal faces without occlusions 1040 × 1392 resolution at 25 fps under uniform illumination	Frame-level ground truth Annotations by 2 FACS coders	41 adults: 23♀ 18♂ Ages: 18-29 Ethnicities: 20 Caucasian, 4 Hispanic, 11 Asian, 6 African-American	Several tasks driven by an experimenter, such as: interviews, video viewing, insult or smell
SP [26]	AU4–AU10 AU6–AU12 AU7–AU20 AU9–AU25 AU26–AU43	2 to 16 videos per user (>48000 frames) Frontal faces, with minor rotations and without occlusions 352 × 240 resolution under uniform illumination	Frame-level ground truth Annotations by one of 3 FACS coders	25 adults: 12♀ 13♂ Ages: 4 young, 14 middle-aged, 7 elderly Ethnicities: 23 Caucasian, 2 Hispanic	Facial videos recorded while performing painful shoulder physical tests

Table 2 Number of frames per studied AU and intensity contained in the DISFA, BP4D and SP databases

Dataset	AU	"A"	"B"	"C"	"D"	"E"	Total	Reliability
DISFA [28]	AU1	2272	1749	2809	1393	555	8778	0.83 (ICC)
	AU2	7120	934	3505	836	369	12764	0.86 (ICC)
	AU4	4661	7636	6586	4328	1383	24594	0.94 (ICC)
	AU12	13943	6869	7233	2577	172	30794	0.90 (ICC)
	AU15	5180	1618	1017	47	0	7862	0.82 (ICC)
	AU20	1591	1608	1305	28	0	4532	0.83 (ICC)
	AU25	9805	13935	15693	5580	1039	46052	0.92 (ICC)
	AU26	13443	7473	3529	314	217	24976	0.93 (ICC)
SP [26]	AU4	202	509	225	74	64	1074	Inter-observer agreement of 95%
	AU12	2145	1799	2158	736	49	6887	
	AU20	286	282	118	0	20	706	
	AU25	767	803	611	138	88	2407	
	AU26	431	918	265	478	1	2093	
BP4D [48]	AU12	15157	25045	21039	12798	4323	78362	0.92 (ICC)

Annotations reliability is also provided (ICC is *Intraclass Correlation Coefficient*)

[23, 28, 31] have been therefore considered in previous AU intensity prediction works, achieving good performances.

Little attention has been paid, however, to the choice of the facial area where these descriptors should be applied. Most existing systems consider the whole face area for features computation, either by dividing it into a regular grid [15, 20] or by applying the chosen appearance descriptor to small local patches around relevant facial landmarks [14, 31, 45]. Interestingly, recent works by Mohammadi et al. [29] and Kaltwang et al. [21] assigned different weights to different facial patches by learning their most frequent co-activations for each AU intensity. Nevertheless, the whole face-based AU modeling may result in good detection, as it uses evidences of facial muscle movements over the entire face, but has the limitation of learning unwanted AUs correlations. Only a couple of pioneering works have tried to reduce the problem of AUs co-occurrences learning by separating the face area into different anatomically based regions of interest (ROIs) [19, 43]. The descriptors used for each AU detection are then computed over their corresponding ROI (e.g. mouth ROI for AU12, cheek ROI for AU6, frown ROI for AU4, etc.). ROI-based approaches also have the benefit of allowing to obtain better registered descriptors: as each region depends on its own facial landmarks, registration errors in other parts of the face are not propagated and partial facial occlusions do not affect every AU. However, these promising works have been limited to binary AU detection so far.

To tackle step (ii), the few existing AU intensity prediction works are divided into: classification-based methods and regression-based methods. Classification-based approaches consider intensities as a discrete set of classes and classify input features into one of six categories: absence of AU ("neutral" or "N"), "A", "B", "C", "D" or "E". Preferred classifiers include multiclass Support Vector Machines (SVM), used with both non-linear [27, 28] and linear [14] kernels, Markov models [15] and Dynamic Bayesian Networks [22, 23]. Some authors argue that by modeling intensities in a discrete way each level is considered independently, thus, ignoring their ordering [34]. On the contrary, in regression-based methods intensity scores of A through E? are assigned a numerical value from 1 to 5 and AU absence is assigned 0. The regression model is trained to provide an intensity output on a continuous dimension, and this output is then re-discretized to the nearest integer to obtain its FACS value. Regression models are popular in the literature and can be found in many different versions: epsilon Support Vector Regression [36, 41, 45], hierarchical Partial Least Squares regression [12], Logistic Regression [3] and others [31, 34]. However, this mapping of intensities to ordinal values deviates in concept from the non-metric definition of AU intensity in the FACS manual, wherein each intensity level spreads over a different range of appearance changes (c.f. Fig. 1). Besides the type of model chosen, interestingly, Girard et al. [14] started to pinpoint that other factors closely related to prediction, such as the social context of the database used for training or the choice of the dimensionality reduction technique applied to input facial features, also play a crucial role in the final performance of the system.

2.3 Proposed approach

Our system has two main characteristics: (1) it works in real-time and (2) it is built to generalize as much as possible across datasets. Given their real-time capabilities and previous promising results, the method for AU intensity detection is based on the combination of HOG features [9] and linear kernel multiclass SVM [4]. To mitigate the effect of

facial landmarks extraction inaccuracy and the learning of co-occurrences between AUs, and thus maximize characteristic (2), features are computed from different anatomically based facial ROIs for each AU. This method is used to build one intensity detection model for each of the eight targeted AUs and thorough performance metrics, including results from cross-validation and cross-dataset experiments, are presented and discussed. DISFA is used in the first phase of the work, as it is the most comprehensive database for the AUs considered. The cross-dataset study (second phase) also involves BP4D and SP, and focuses on AU12 which is the only targeted AU common to the three databases. The latter study aims, on the one hand, to demonstrate the adaptation capability of our models to new datasets and, on the other hand, to analyze the impact of context on AUs intensity detection.

3 AU intensity detection method description

The AUs intensity detection pipeline proposed in this work is shown in Fig. 3. The system is implemented in C++ for real-time purposes and involves three main steps: firstly the face is segmented from the whole input image and a set of facial landmarks are extracted; on the basis of facial landmarks positions, three specific ROIs are defined (mouth ROI, frown ROI and eyes ROI) and HOG features are computed separately from each one of them; finally classification of each AU intensity is performed by one different pre-trained SVM model using as input its corresponding ROI features. Section 3.1 describes the first two steps while Sect. 3.2 details how models were built.

3.1 ROI-based facial feature extraction

The feature extraction procedure starts with face segmentation, which is performed by means of the OpenCV version of Viola and Jones Haar Cascade algorithm [44]. After that, twelve facial landmarks (four in the eyebrows, four in the eyes, one in the nose and three in the mouth, see white crosses in Fig. 3) are extracted thanks to the Supervised Descent Method provided by Intraface library [47]. These facial points are the reference for the definition of the three rectangular ROIs where HOG features are computed.

HOG descriptor is able to capture shape and texture information from gradients by breaking an image region into smaller connected regions, called cells, and computing a histogram of edge orientation for each cell [9]. To account for changes in illumination and contrast, the gradient strengths must be locally normalized, which requires grouping the cells together into larger, spatially connected blocks. The final HOG descriptor of the region is then the concatenated vector of the components of the normalized cell histograms from all of the blocks. These blocks typically overlap, meaning that each cell contributes more than once to the final descriptor. HOG has the advantage of being robust to changes in illumination and computationally fast (e.g. its cost is $^1/_5$ compared to SIFT [33]), but is not invariant to rotations and therefore requires an accurate alignment of the facial region before being computed.

The common trend in the literature is to perform the alignment of the whole face region by using the angle formed by a couple of facial landmarks, generally both eyes centers, to cancel rotations [43]. However, even small registration errors during landmarks extraction may cause that

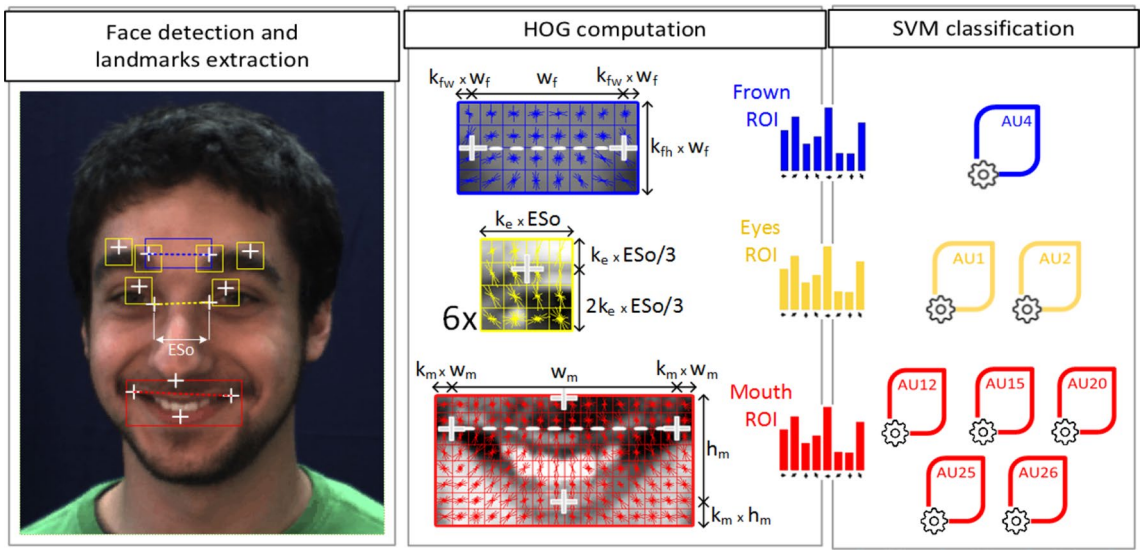

Fig. 3 Proposed pipeline for AUs intensity detection. *White crosses* represent facial landmarks extracted from the face, while *dashed lines* link landmarks used for each ROI alignment

regions surrounding the points used to align are correctly transformed, while other more distant (such as mouth) are dramatically affected. To avoid this problem, each of the three facial ROIs used in this work has its own coordinate system and is aligned with regard to different facial landmarks. More precisely, the parameters defined for each ROI are the following:

- *Frown ROI* (used for AU4 model). This ROI is located around inner eyebrows landmarks, which are used for alignment. The ROI width is established as the distance between both landmarks in the horizontal axis, plus an extra k_{fw} margin of 5%; ROI height is defined as the $k_{fh} = 25\%$ of its width. The region is re-scaled to fit 64×32 pixels, with 8×8 cells, 16×16 blocks overlapped at 50% and 9 bins for gradients orientations. The final HOG descriptor dimension is therefore 756.
- *Eyes ROI* (used for AU1 and AU2 models). This ROI is made up of six square patches located around inner eyebrows, middle eyebrows and upper eyes landmarks. Patches size is equal to the $k_e = 20\%$ of the distance between inner eyes corners ("ESo" in Fig. 3), which are also used for ROI alignment. Each patch is scaled to 16×16 pixels and a HOG descriptor is computed from it by using 4×4 cells, 8×8 blocks with 50% overlap and 9 orientation bins. The six concatenated descriptors lead to a final HOG vector of 1944 features.
- *Mouth ROI* (used for AU12, AU15, AU20, AU25 and AU26 models). This ROI is bounded by the nose center, the two lip corners and the lower lip, plus an extra k_m margin assigned to 5% both for width and height. Alignment is done with regard to lip corners positions. The region is re-scaled to a size of 128×64 pixels. Cells are defined as 8×8, blocks as 16×16 with an overlap of 50% and 8 bins are used for gradients orientations, leading to a total HOG dimension of 3360 features.

For each ROI, corresponding HOG parameters and constants values (window size, cell size, block size, block overlap percent, number of bins, k_{fw}, k_{fh}, k_e and k_m) were selected on the basis of a grid-search benchmarking process deeper explained in the following section.

3.2 Models training and benchmarking procedure

One linear kernel SVM model was built per AU by using LIBSVM library [5]. Multiclass classification into six intensity categories ("N", "A", "B", "C", "D" and "E") was tackled through a one-versus-all strategy. Unlike most previous approaches [15, 31, 34, 41], any dimensionality reduction algorithm was applied to features since recent surveys on facial expression recognition have started to question the suitability of using such techniques [35]. Specifically, for the AU intensity detection task, it was found that the success of dimensionality reduction techniques is highly dependent on the database used [14], which is not desirable for a system aiming to generalize well across datasets.

DISFA database was used for training and testing the models. Overall 0.2% of the samples were untrackable, mostly due to occlusions or extreme out-of-plane rotations, and were removed. From the remaining samples, data imbalance between intensity classes was reduced as much as possible, as frequency of occurrences was highly skewed toward lower intensities. Previous research has demonstrated that datasets need to be carefully balanced to achieve a good performance in AU recognition systems [2, 46]. To balance the data, under-sampling was combined with cost-sensitive learning techniques. Both for training and testing data, the number of neutral frames (class "N" samples) was balanced with the second most frequent intensity level for each AU (up to 4000 samples per class). For example, the total number of samples per intensity class considered for AU4 (c.f. Table 2) was: 4000 samples for classes "N", "A", "B", "C" and "D", and 1383 samples for class "E". Then, SVM weights were assigned proportionally to the number of samples in each class. For AU15 and AU20 class "E" was not considered, as DISFA does not provide any highest intensity sample, and therefore the classification problem was reduced to five categories.

Regarding testing and benchmarking issues, the C value for the SVM soft margin cost function, HOG parameters (window size, cell size, block size, block overlap percent and number of bins) and ROIs constants (k_{fw}, k_{fh}, k_e and k_m) were optimized using a grid-search procedure. A 10-fold cross-validation strategy was applied in which no subject overlap between training and testing sets was ensured. The target of the competition was to optimize Intraclass Correlation Coefficient *ICC(3,1)*. ICC is a 0-to-1 measure of correlation or conformity of data with multiple targets, commonly used in behavioral research to quantify agreement between different raters. Depending on how the ratings are obtained, different types of this score should be used [37]. As pointed out by [34] the most appropriate version of ICC for AUs intensity detection is *ICC(3,1)*, which is based on a Mixed Model ANOVA, with J judges, treated as fixed effects, and N targets, considered as random effects. In that case $J = 2$ (the ground truth and predicted values) and N is the total number of test samples.

The optimization of *ICC(3,1)* was preferred to that of the popular macro-F1 (F_1^M) score. The reason is that the latter metric, which is the weighted average of per-class precision and recall, treats any confusion between classes as the same type of error. It does not take into account that FACS intensity levels follow a conceptual ordering and thus that, for instance, predicting label "D" for a true label "E" is less grave than predicting "A" for the true label "E". On the

contrary, *ICC(3,1)* considers the difference between large and small errors among judges and inconsistencies in judgments, and therefore turns out a more interesting reference of performance.

4 Within-DISFA results and performance

4.1 Classification results

The confusion matrix, *ICC(3,1)* and F_1^M scores obtained after the benchmarking procedure on DISFA samples for each AU intensity model are provided in Fig. 4. The best *ICC(3,1)* scores are found for the AUs that occur more frequently in the database: AU12 (0.84) and AU25 (0.83). Accordingly, lower *ICC(3,1)* performances correspond to the AUs with less available training samples: AU15 (0.63) and AU20 (0.62). Regarding per-class accuracy, it can be depicted from confusion matrices that it is in general significantly higher for classes "N" and "E" (extreme classes) than for middle classes ("A", "B", "C" and "D"), which are frequently confused with their neighboring categories. The latter fact strongly impacts F_1^M values, but *ICC(3,1)* scores are still promising as most confusions are located around matrices diagonals.

The comparison of our results to previous works is not always rational for several reasons, such as that database partitions are built in different ways, some models are evaluated with subject overlap between training and test sets [3, 14], or because confusions matrices, *ICC(3,1)* and F_1^M scores are not given. Figure 5 compares our models to five recent state-of-the-art approaches. These approaches have all been evaluated under similar conditions as ours, i.e., by training and validating models with DISFA database, and provide *ICC(3,1)* and/or F_1^M performance metrics. Thus, the direct comparison of results is reasonable. As it can be seen, our *ICC(3,1)* values are only outperformed for some AUs by the works by Mavadati et al. [28] and Li et al. [23]. However, the former makes use of a costly SVM with Radial Basis Function kernel and therefore lacks real-time capabilities. Moreover, none of them has been tested in cross-dataset conditions and thus—contrary to our approach, as it will be further demonstrated—their good performance is not ensured for new datasets. Regarding F_1^M scores, we improve for every AU the results by Rudovic et al. [34] which is to our knowledge the only baseline reference available for this metric.

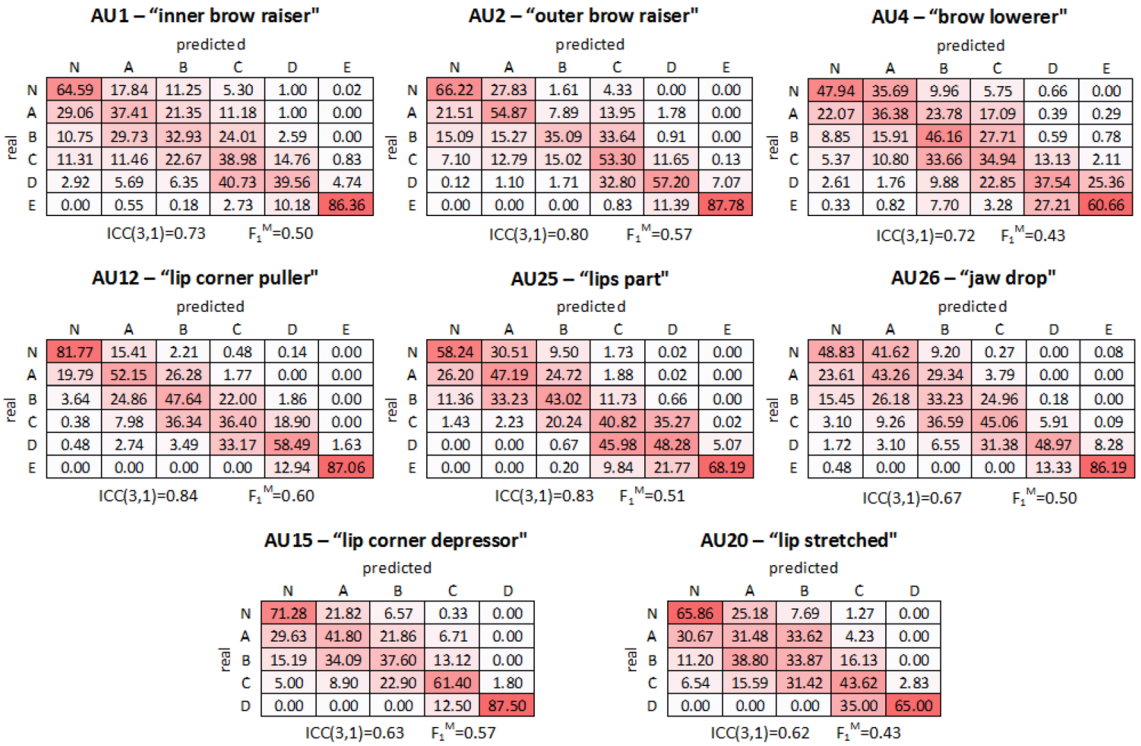

Fig. 4 Confusion matrices computed from the true and predicted intensity labels of the DISFA database. Matrix elements have been normalized by each intensity class size. Results were obtained by applying a 10-fold cross-validation strategy ensuring no subject overlap between training and testing sets

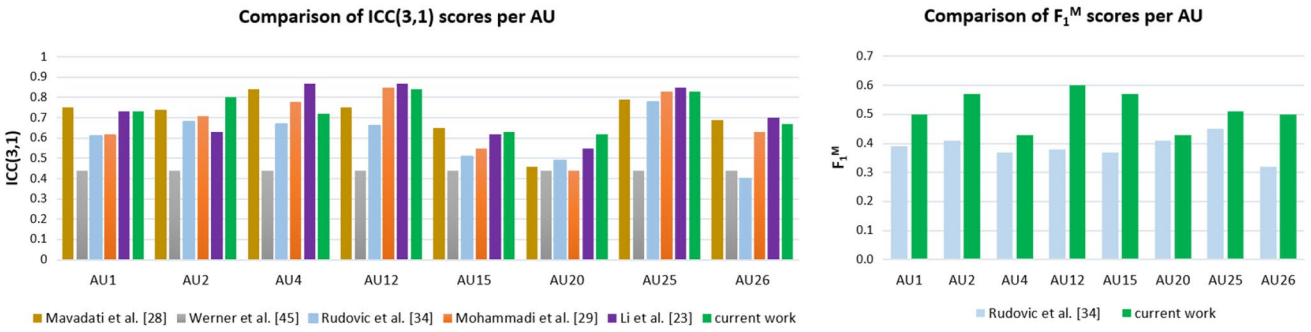

Fig. 5 Comparison of the proposed models with recent state-of-the-art approaches making use of DISFA database for AU intensity detection

4.2 Computational performance

Real-time facial expression recognition is crucial for certain application domains where fast reactions are needed. Figure 6 depicts computational performance metrics for our whole system, i.e. for per frame detection of the eight AUs intensities. Mean speeds of 20–30 frames per second (fps) are achieved with mid-range processors Intel Core i5 @ 1.80 GHz and Intel Core i7 @ 2.4 GHz, for commonly used resolutions of 780 × 580 to 1280 × 720 pixels. Performance speeds up to 30–45 fps when a high-end Intel Core i5 @ 3.20 GHz processor is used. Thus, videos taken in the same conditions as DISFA dataset (1024 × 768 pixels at 20 fps) could be analyzed in real-time without frame drops with a mid-range processor, while videos similar to those from the BP4D database (1040 × 1392 pixels at 25 fps) could be processed with minor frame drops with a high-end processor.

Interestingly, it can also be observed that the most costly phase of the pipeline is face detection (marked with * in the figure). However, for good quality videos, it is just performed once for the first frame—then facial points are tracked—and it therefore does not impact the overall computation time. The second heaviest process is the previous cropping, alignment and re-scaling necessary to HOG features computation, whose cost is proportional to image resolution.

Such metrics have not been reported in most previous AU intensity detection works, which makes comparison difficult. One reference can be found in the study by Li et al. [23], where each AU intensity detector works at 7.1 fps in a 2.4 GHz Intel Core2 PC. Also, the work by Baltrušaitis et al. [2] reports the detection of four AUs intensities using a dual core 3 Ghz *Intel i3* processor at 20–30 fps and 5–10 fps, for resolutions of 780 × 580 and 1040 × 1392 pixels, respectively. Regarding other kind of facial expression recognition works: the state-of-the-art Computer Expression Recognition Toolbox (CERT) [24] performs nineteen AUs binary detection at 10 fps for 320 × 240 pixels videos; the recent system by Agarwal and Umer [1] detects two basic facial emotions per second; and Suk and Prabhakaran [39] recognize 6 basic emotions at 2.4 fps in a Samsung Galaxy S3 mobile device. The

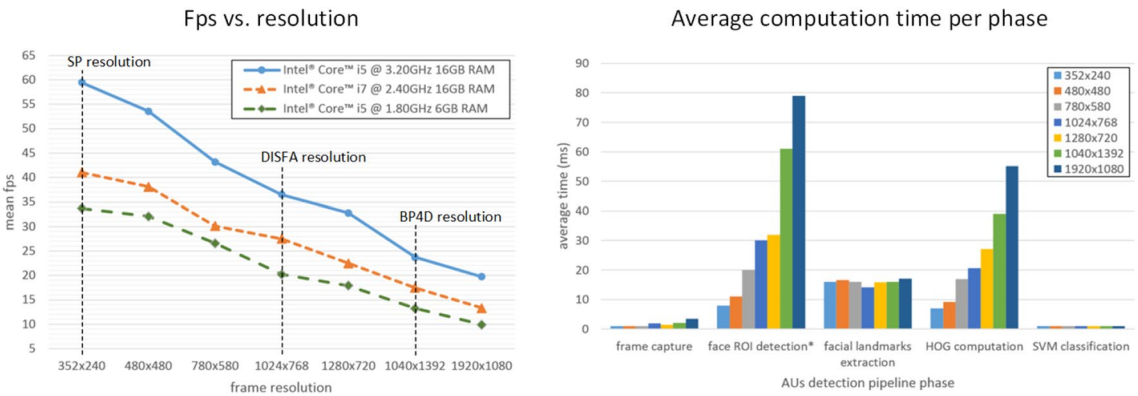

Fig. 6 Computational performance of the system. *Left* mean frames per second (fps) obtained, depending on the frame resolution and processor used. *Right* average computation time obtained with the *Intel Core i7* processor, per phase of the AUs intensity detection pipeline and for different resolutions. Note that face detection (marked with *) is only performed when the facial tracker initializes or facial tracking is lost, and therefore does not dramatically impact overall computational time

aforementioned studies can serve as a baseline to demonstrate that our speed is competitive.

5 Cross-dataset experiments

5.1 Experiments description

To test the adaptability and robustness of the models, an in-depth cross-dataset evaluation has been performed by means of DISFA, BP4D and SP databases. As pointed out in Sect. 2.1, this is challenging mainly due to differences in terms of participants demographics, FACS annotators involved, quality of samples and context, which may affect the frequency, intensity and co-occurrence of AUs. From all the AUs taken into consideration in this study, AU12 is the only common to the three databases and with a number of samples allowing to build balanced cross-dataset partitions (c.f. Table 2). In this section we therefore focus on this AU, and four kinds of cross-dataset testing conditions are presented:

1. Within-corpus testing (WCT): the model is both machine-learned and tested using AU12 samples from one of the corpus.
2. Off-corpus testing (OCT): the model is first machine-learned using one corpus and, subsequently, tested on AU12 samples from another corpus.
3. Integrated within-corpus testing (IWCT): it involves merging two different corpora into one single corpus, and then performing within-corpus testing on the resulting corpus.
4. Integrated off-corpus testing (IOCT): two different corpora are merged into one single corpus, but testing is performed using samples of a third dataset.

All the AU12 models used in the aforementioned experiments have been built following the method presented in Sect. 3, both regarding features extraction and training procedure. The SP dataset has not been used for training in OCT, IWCT and IOCT conditions, as the amount of annotated samples is much lower compared to DISFA and BP4D (c.f. Table 2), and therefore training sets would have been difficult to balance.

5.2 Results

Table 3 shows *ICC(3,1)*, F_1^M, mean and per-class accuracy scores obtained in the different cross-dataset experiments for AU12 intensity detection. Confusion matrices resulting from OCT experiments are also provided in Fig. 7. As expected, within-corpus tests obtain better classification results. When SP is used as testing set, there is a substantial decrease in performance for every testing condition, which is not surprising either taking into account differences in age of participants and context. Besides social differences, also from a technical point of view SP samples resolution is critically lower (352×240 pixels) and has probably lead to decrease HOG features descriptor power.

As happened with within-DISFA testing (c.f. Sect. 4.1), accuracy is significantly higher for classes "N" and "E" than for middle classes in every within-corpus experiment. This problem with middle classes has even more impact in the off-corpus scenario. For the pair DISFA-BP4D, despite the drop in F_1^M score, *ICC(3,1)* values are still high in off-corpus experiments (0.81 and 0.74 respectively), which means that confusions are coherent and do not imply large intensity jumps. This is not the case when SP samples are used as testing set: confusion matrices are strongly left-shifted and middle intensity classes are too frequently confused with "N" and "A" classes, leading to much lower *ICC(3,1)* scores (0.49 and 0.50 respectively).

Regarding integrated within- and off-corpus tests, it is interesting to notice that neither *ICC(3,1)* nor accuracy scores increase with regard to their single corpus counterparts. For example, in the within-corpus scenario *ICC(3,1)*

Table 3 Results obtained for AU12 intensity detection in the different cross-dataset experiments

Condition	Training	Testing	ICC(3,1)	F_1^M	Mean Acc.	Acc. "N"	Acc. "A"	Acc. "B"	Acc. "C"	Acc. "D"	Acc. "E"
WCT	DISFA	DISFA	0.84	0.60	60.59	81.77	52.15	47.64	36.40	58.49	87.06
	BP4D	BP4D	0.85	0.40	40.11	63.12	34.64	32.68	29.32	32.02	48.85
	SP	SP	0.59	0.51	61.28	63.26	47.94	34.37	55.21	76.89	90.00
OCT	DISFA	BP4D	0.81	0.32	36.23	83.83	22.46	40.15	47.10	21.50	2.36
	DISFA	SP	0.49	0.40	39.23	85.99	21.45	23.74	17.33	20.22	66.67
	BP4D	DISFA	0.74	0.27	35.60	54.50	10.11	34.07	31.90	17.34	65.70
	BP4D	SP	0.50	0.26	27.92	64.14	18.37	33.89	22.34	20.44	8.33
IWCT	BP4D + DISFA	BP4D + DISFA	0.82	0.38	38.35	70.67	29.75	34.01	23.34	24.53	47.80
IOCT	BP4D + DISFA	SP	0.49	0.38	38.77	84.08	17.86	31.92	9.11	23.00	66.67

Acc. is class accuracy, defined as the proportion of the total number of predictions that were correct

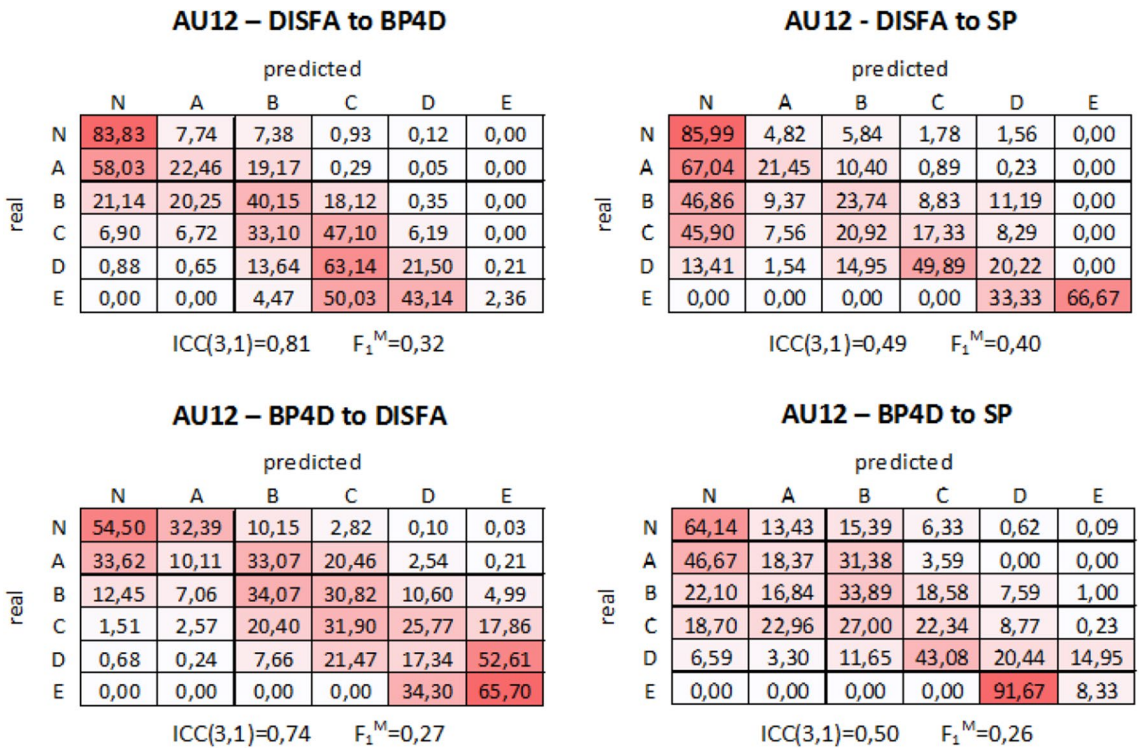

Fig. 7 Normalized confusion matrices obtained for AU12 intensity detection in off-corpus testing (OCT). *Left column* OCT experiments related to the databases pair DISFA-BP4D. *Right column* OCT experiments where SP is used as testing set

is 0.84 for DISFA and 0.85 for BP4D, while the integration of both datasets (DISFA+BP4D) achieves 0.82. Similarly, in the off-corpus experiment where BP4D is used for training and SP for testing the *ICC(3,1)* score is 0.50; however if DISFA is added to the training set, this value drops to 0.49.

There are few AU intensity detection cross-dataset experiments in the literature to which compare obtained results. Baltrušaitis et al. [2] recently performed tests involving BP4D and DISFA databases. Their system extracts HOG features from the whole facial area which are then classified using Support Vector Regression. They reported similar correlation coefficients for AU12: 0.86 for BP4D within-corpus testing and 0.83 for DISFA to BP4D off-corpus testing. Interestingly, they also found that integrated within-corpus experiments did not improve performances. In what concerns works involving the very different pain context SP database, it seems that the proposed methodology outperforms existing approaches in relation to cross-dataset adaptation. For instance, Werner et al. [45] report an *ICC(3,1)* of 0.22 (far below our 0.49) with linear kernel Support Vector Regressor models trained on DISFA and tested with SP. For the same experiment, Rudovic et al. [34] propose a Conditional Ordinal Random Field model that achieves a F_1^M score of 0.27 (lower than our 0.40) and a slightly higher *ICC(3,1)* value of 0.55; however, their system explicitly sets "subject" and "context" as parameters learned from initial frames assumed neutral.

5.3 Discussion through manifold learning techniques

Results presented in Sect. 5.2 suggest that, even though the inter-rater agreement is high for each individual database, the inter-dataset agreement is not so accurate. This hypothesis is especially reinforced by the fact that integrated corpus experiments did not improve performances. A reasonable explanation could be that raters of a given dataset agreed on a common baseline to annotate samples intensities within their range of facial appearance changes, but that this baseline is not strictly the same across databases (although FACS is followed in any case).

In order to visually interpret and further analyze the proximity of AU12 intensity classes across datasets in the feature space, we apply Manifold Learning using the Multi-dimensional Scaling (MDS) algorithm implemented in the *Scikit-Learn* toolkit [32]. MDS is especially appropriate for this purpose as it seeks a low-dimensional representation of the data in which the distances respect well the distances in the original high-dimensional space. Figure 8 shows the results of the different Manifold Learning experiments carried-out: top figures are individual corpus representations in the embedded feature space, while bottom figures are integrated

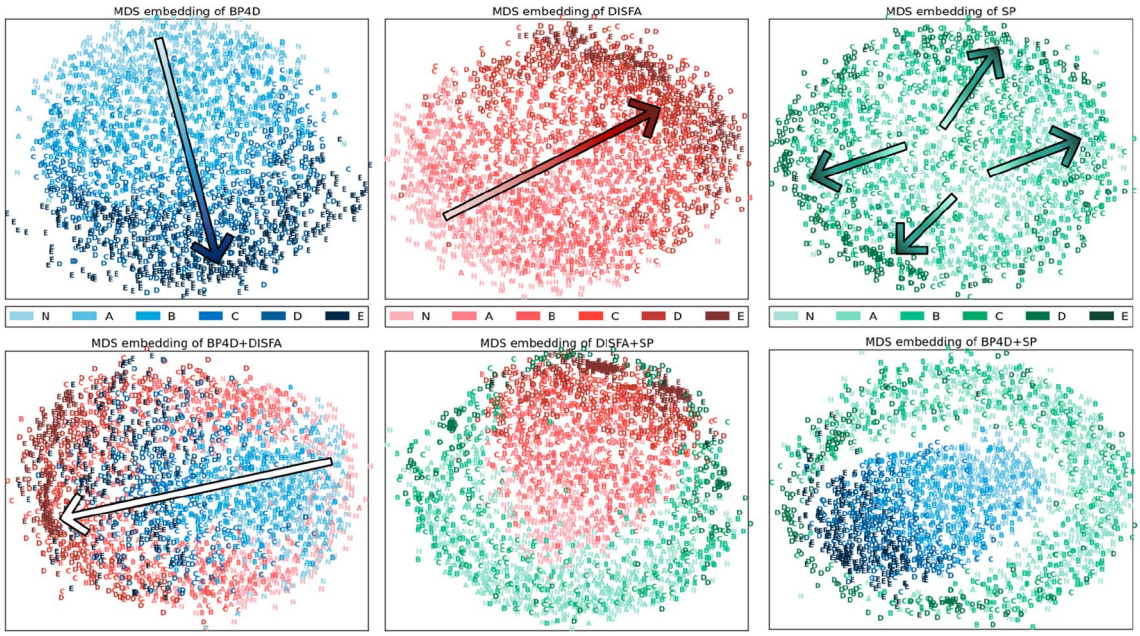

Fig. 8 Results from MDS Manifold Learning experiments. *Top* single corpus representations in the embedded feature space; *bottom* integrated corpus representations. *Graded-blue scale* is used for BP4D samples, *red tones* correspond to DISFA and *green colors* represent SP (color figure online)

corpus representations (BP4D+DISFA, DISFA+SP and BP4D+SP).

There is a clear similarity in the representations obtained for BP4D and DISFA, where "N" samples are located in one side of the embedded feature space and classes progressively grow in orderly fashion until arriving to "E" samples at the diagonally opposite side of the graphic (indicated by an arrow in the figures). This tendency is also repeated in the integrated BP4D+DISFA case, with a clear overlap between classes of both databases. However, it is interesting to notice that this overlap is shifted by one class, especially for middle-high intensity categories (i.e. BP4D "E" samples are overlapped with DISFA "D" samples, BP4D "D" samples are close to DISFA "C" samples, and so on). Nevertheless, Manifold Learning demonstrates that BP4D and DISFA share a reasonably close annotation baseline, which explains the high *ICC(3,1)* performances obtained in cross-dataset tests.

The single corpus representation of SP has a totally different shape: "N" samples are situated in the center of the embedded feature space and intensity classes grow up in all directions in a concentric fashion. Consequently, in integrated representations with BP4D and DISFA, there is almost any or very little overlap between samples of the different corpora (excepting for some "N" and "E" samples that are fairly close). The faces of pain together with smile, which imply e.g. the co-occurence of AU10 "upper lip raiser" with AU12 [26], have therefore strongly impacted the annotation baseline with regard to other contexts and thus classification results.

6 Conclusions and future work

This paper presented a method for the real-time detection of AUs intensity. Its main novelty relies on the proposed facial representation strategy: features were computed from a different ROI (mouth, eyes or frown) depending on the AU considered. Real-time performances of 20–30 fps were achieved by using HOG as features descriptor and linear kernel SVM as classification mechanism. The ROI-based strategy allowed to obtain better registered HOG descriptors, locally aligned with respect to the most appropriate facial landmarks, and to reduce the effect of AUs co-occurrences learning. Eight different AU intensity detection models were built following this method and cross-validated by means of the DISFA database, obtaining similar or even better results than previous approaches without real-time capabilities. An in-depth cross-dataset evaluation through AU12 samples from DISFA, BP4D and SP databases further validated the proposed facial representation suitability, demonstrating that our AU12 model generalizes better to new datasets and contexts. Therefore, the fact of reducing co-occurrences learning by separating muscular facial ROIs has lead to a system more adaptable to new datasets. This study also provided novel insights about existing public corpora: high

inter-dataset agreement is not ensured and annotations baselines strongly depend on the context.

Nowadays, AU intensity detection starts to be of great interest to many different fields, sometimes with commercial purposes requiring high accuracy. As AUs intensity annotation is tedious and can only be performed by FACS experts (e.g. it cannot be crowdsourced), the combination of existing databases is a valuable mean to build more general and robust systems but, as demonstrated in this paper, has to be faced carefully. Transfer Learning techniques could help to take advantage of learning in a new database through the transfer of knowledge from another that has already been learned. This idea could also contribute to extend our cross-dataset study to more AUs, which is our future research line.

Acknowledgements This research has been supported by the Laboratory of Excellence SMART (ANR-11-LABX-65) supported by French State funds managed by the ANR within the Investissements d'Avenir programme (ANR-11-IDEX-0004-02).

References

1. Agarwal S, Umer S (2015) Mp-feg: media player controlled by facial expressions and gestures. In: 5th National conference on computer vision, pattern recognition, image processing and graphics, pp 1–4
2. Baltrušaitis T, Mahmoud M, Robinson P (2015) Cross-dataset learning and person-specific normalisation for automatic action unit detection. In: 11th IEEE international conference and workshops on automatic face and gesture recognition, vol 6, pp 1–6
3. Bingol D, Celik T, Omlin CW, Vadapalli HB (2014) Facial action unit intensity estimation using rotation invariant features and regression analysis. In: 2014 IEEE international conference on image processing, pp 1381–1385
4. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: 5th Annual workshop on computational learning theory, pp 144–152. ACM
5. Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
6. Chew SW, Lucey P, Lucey S, Saragih J, Cohn JF, Matthews I, Sridharan S (2012) In the pursuit of effective affective computing: the relationship between features and registration. IEEE Trans Syst Man Cybern B Cybern 42(4):1006–1016
7. Chu WS, Torre F, Cohn J (2013) Selective transfer machine for personalized facial action unit detection. In: IEEE conference on computer vision and pattern recognition, pp 3515–3522
8. Cohn JF, Ambadar Z, Ekman P (2007) Observer-based measurement of facial expression with the facial action coding system. The handbook of emotion elicitation and assessment, pp 203–221
9. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. IEEE Comput Soc Conf Comput Vis Pattern Recognit 1:886–893
10. De Moor K, Mazza F, Hupont I, Quintero MR, Mäki T, Varela M (2014) Chamber QoE: a multi-instrumental approach to explore affective aspects in relation to quality of experience. In: IS&T/SPIE electronic imaging. International society for optics and photonics
11. Flores VC (2005) Artnatomy. www.artnatomia.net
12. Gehrig T, Al-Halah Z, Ekenel HK, Stiefelhagen R (2015) Action unit intensity estimation using hierarchical partial least squares. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition, vol 1, pp 1–6
13. Girard JM, Cohn JF, Mahoor MH, Mavadati S, Rosenwald DP (2013) Social risk and depression: evidence from manual and automatic facial expression analysis. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition, pp 1–8
14. Girard JM, Cohn JF, De la Torre F (2015) Estimating smile intensity: a better way. Pattern Recogn Lett 66:13–21
15. Gonzalez I, Verhelst W, Oveneke M, Sahli H, Jiang D (2015) Framework for combination aware au intensity recognition. In: 2015 International conference on affective computing and intelligent interaction, pp 602–608
16. Hager JC, Ekman P, Friesen WV (2002) Facial action coding system. A Human Face, Salt Lake City
17. Happy S, George A, Routray A (2012) A real time facial expression classification system using local binary patterns. In: 4th IEEE international conference on intelligent human computer interaction, pp 1–5
18. Hupont I, Cerezo E, Baldassarri S (2008) Facial emotional classifier for natural interaction. ELCVIA Electron Lett Comput Vis Image Anal 7(4):1–12
19. Jiang B, Martinez B, Valstar MF, Pantic M (2014) Decision level fusion of domain specific regions for facial action recognition. In: 2014 22nd international conference on pattern recognition, pp 1776–1781
20. Jiang B, Valstar MF, Pantic M (2011) Action unit detection using sparse appearance descriptors in space-time video volumes. In: 2011 IEEE international conference on automatic face and gesture recognition and workshops, pp 314–321
21. Kaltwang S, Todorovic S, Pantic M (2016) Doubly sparse relevance vector machine for continuous facial behavior estimation. IEEE Trans Pattern Anal Mach Intell 38(9):1748–1761
22. Li Y, Mavadati SM, Mahoor MH, Ji Q (2013) A unified probabilistic framework for measuring the intensity of spontaneous facial action units. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition, pp 1–7
23. Li Y, Mavadati SM, Mahoor MH, Zhao Y, Ji Q (2015) Measuring the intensity of spontaneous facial action units with dynamic Bayesian network. Pattern Recogn 48(11):3417–3427
24. Littlewort G, Whitehill J, Wu T, Fasel I, Frank M, Movellan J, Bartlett M (2011) The computer expression recognition toolbox (CERT). In: 2011 IEEE international conference on automatic face and gesture recognition and workshops, pp 298–305
25. Littlewort GC, Bartlett MS, Lee K (2009) Automatic coding of facial expressions displayed during posed and genuine pain. Image Vis Comput 27(12):1797–1803
26. Lucey P, Cohn JF, Prkachin KM, Solomon PE, Matthews I (2011) Painful data: the UNBC-McMaster shoulder pain expression archive database. In: 2011 IEEE international conference on automatic face and gesture recognition and workshops, pp 57–64
27. Mahoor MH, Cadavid S, Messinger DS, Cohn, JF (2009) A framework for automated measurement of the intensity of non-posed facial action units. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 74–80
28. Mavadati SM, Mahoor MH, Bartlett K, Trinh P, Cohn JF (2013) DISFA: a spontaneous facial action intensity database. IEEE Trans Affect Comput 4(2):151–160
29. Mohammadi MR, Fatemizadeh E, Mahoor MH (2016) Intensity estimation of spontaneous facial action units based on their sparsity properties. IEEE Trans Cybern 46(3):817–826
30. Nakamura T, Maejima A, Morishima S (2014) Driver drowsiness estimation from facial expression features computer vision feature investigation using a CG model. In: 2014 IEEE international

conference on computer vision theory and applications, vol 2, pp 207–214
31. Nicolle J, Bailly K, Chetouani M (2015) Facial action unit intensity prediction via hard multi-task metric learning for kernel regression. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition, vol 6, pp 1–6
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
33. Ren H, Li ZN (2014) Age estimation based on complexity-aware features. In: Computer vision–ACCV 2014, pp 115–128. Springer
34. Rudovic O, Pavlovic V, Pantic M (2015) Context-sensitive dynamic ordinal regression for intensity estimation of facial action units. IEEE Trans Pattern Anal Mach Intell 37(5):944–958
35. Sariyanidi E, Gunes H, Cavallaro A (2015) Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans Pattern Anal Mach Intell 37(6):1113–1133
36. Savran A, Sankur B, Bilge MT (2012) Regression-based intensity estimation of facial action units. Image Vis Comput 30(10):774–784
37. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86(2):420
38. Suk M, Prabhakaran B (2014) Real-time mobile facial expression recognition system–a case study. In: 2014 IEEE conference on computer vision and pattern recognition workshops, pp 132–137
39. Suk M, Prabhakaran B (2014) Real-time mobile facial expression recognition system-a case study. In: IEEE conference on computer vision and pattern recognition workshops, pp 132–137
40. Tian Yl, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell 23(2):97–115
41. Valstar MF, Almaev T, Girard JM, McKeown G, Mehu M, Yin L, Pantic M, Cohn JF (2015) FERA 2015-second facial expression recognition and analysis challenge. In: 2015 11th IEEE international conference and workshops on automatic face and gesture recognition, vol 6, pp 1–8
42. Valstar MF, Pantic M (2012) Fully automatic recognition of the temporal phases of facial actions. IEEE Trans Syst Man Cybern B Cybern 42(1):28–43
43. Velusamy S, Gopalakrishnan V, Anand B, Moogi P, Pandey B (2013) Improved feature representation for robust facial action unit detection. In: IEEE consumer communications and networking conference, pp 681–684
44. Viola P, Jones MJ (2004) Robust real-time face detection. Int J Comput Vis 57(2):137–154
45. Werner P, Saxen F, Al-Hamadi A (2015) Handling data imbalance in automatic facial action intensity estimation. FERA, p 26
46. Wu T, Butko NJ, Ruvolo P, Whitehill J, Bartlett MS, Movellan JR (2011) Action unit recognition transfer across datasets. In: 2011 IEEE international conference on automatic face and gesture recognition and workshops, pp 889–896
47. Xiong X, Torre F (2013) Supervised descent method and its applications to face alignment. In: IEEE conference on computer vision and pattern recognition, pp 532–539
48. Zhang X, Yin L, Cohn JF, Canavan S, Reale M, Horowitz A, Liu P, Girard JM (2014) BP4D-spontaneous: a high-resolution spontaneous 3D dynamic facial expression database. Image Vis Comput 32(10):692–706