Telecom PARIS
Post Masters' Degree in A.I.

IA705 – Multi-Modal Dialogue
Author: Souhail OUMAMA

Review of the article:
Region-based facial representation for real-time Action Units intensity detection across datasets

Abstract

The paper we will be reviewing proposes a method for real-time detection of AUs intensity in terms of the Facial Action Coding System Scale.
It is grounded on a novel and robust based facial representation strategy, for which features are registered from a different region of interest depending on the AU considered.
We will be conducting our review of this article by explaining the main topic the paper deals with, and developing on the approach it uses to address the scientific issue it presents.
As the authors present their work, they first introduce the matter at hand, and explain the aim of their study.

They then perform a review of the state of the art to establish the standard they will be complying to.
After laying the ground, they introduce how their approach works, which is mainly achieved by combining Histogram of Gradients descriptors with linear kernel Support Vector Machines.
The next step in the process is to evaluate the proposed approach, and compare the results to existing papers published in academia.
The article concludes by outlying which improvements to the approach the authors will be focusing their upcoming work on.

Introduction

Facial expressions are the most versatile tool in the human communication toolbelt, as the complexity of its messages allows for powerful ways that humans communicate and understand each other's psyche.
There are two main approaches to measuring facial expressions: the *message-based* approach and the *sign-based* approach.

The most common sign-based approach is the Facial Action Coding System (FACS), which breaks down facial expressions into small anatomical components called action units (AU), that in turn can be graded for changes in intensity using a 5-level scale ranging from "A" (trace) to "E" (maximum).
FACS is well suited for automatic facial expression recognition, as most existing methods focus on detecting AUs and, optionally, map them to targeted emotions as a second step.
However, almost all research focus on binary AU detection (present or absent), instead of mapping their full range intensity estimation. This ends up losing part of the descriptive power of FACS and valuable information about the meaning and dynamics of facial expressions.

The main contributions of the paper at hand are the novel method it proposes for AU intensity detection.
The unprecedented feat it has accomplished is being able to work in real-time while maintaining or even outperforming state-of- the-art approaches.

Previous work on AU intensity detection

Before delving into the workings of the new method, the article outlines the fact that, in literature today, even though there are many visual datasets which are annotated in terms of presence or absence of the AU (or message base descriptors), really few databases that include AU intensity labels according to the FACS A–B–C–D–E intensity scale. Which explains why the main databases considered in this study are *DISFA BP4D* and *SP* datasets.
Having the initial datasets, the article moves on to explain how Detecting AUs intensity operates. It could be split into two main steps:

1. **extracting the most appropriate and representative set of features from the face**

In the overwhelming majority of previous research, there seemed to be an overall preference for *Appearance-Based Approaches* to *Geometry-Based Approaches* (e.g. facial angles or distances). This could be explained by the fact that the latter are vulnerable to errors in the facial landmarks' extraction task, because the smallest variations in intensities have to be registered.

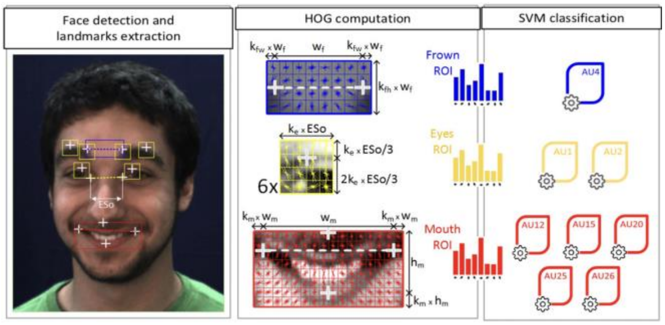
2. **use these features as inputs for a prediction model.**

Most existing research is 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".

Proposed AU intensity detection method description

As stated earlier, the main characteristics the proposed approach presents are it operating in real-time, which is an end-goal for most facial expression detection algorithms. The other one is its capability to perform well across as many datasets as possible (here being the 3 considered datasets).

The method revolves around three pillars, which are illustrated in the following figure.
First is facial segmentation from the whole input image, from which a set of facial landmarks are extracted. Using the positions of these facial landmarks, three specific Regions of Interest (ROIs), representing the most expressive features of the human face: the mouth, the eyes and the frown region, each having their own AU.
The second step is computing HOG (Histogram Of Gradients) features separately from each of them.
The final phase is classifying each AU intensity by different pre-trained SVM models, that use as input their corresponding ROI features.



1. *ROI-based facial feature extraction*

The feature extraction task starts with face segmentation, using OpenCV or the Haar Cascade Algorithm. Twelve facial landmarks (4 in the eyebrows, 4 in the eyes, 1 in the nose and 3 in the mouth) are extracted. These facial points are then for defining the 3 rectangular ROIs where the HOG features are computed. HOG descriptors are able to capture shape and texture information from gradients by breaking an image region into smaller connected regions, and computing a histogram of edge orientation for each.
HOG has the advantage of being robust to illumination changes and fast to compute, but it is not rotation invariant and thus requires an accurate alignment of the facial region before it is computed.
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 the various facial landmarks.

2. *Classification models training and benchmarking procedure*

The model was trained by building one linear kernel SVM model per AU by using LIBSVM library. The 6 intensity categories ("N", "A", "B", "C", "D" and "E") were classified via a one-vs-all strategy. DISFA database was used for training and testing the models, and data imbalance between intensity classes was reduced by combining under-sampling 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).
The target of the competition was to optimize Intraclass Correlation Coefficient ICC(3,1), used to measure the correlation or conformity of data with multiple targets, commonly used in behavioral research to quantify agreement between different raters.

The optimization of ICC(3,1) was preferred to that of the popular F1score because it is the weighted average of per-class precision and recall, and thus treats any confusion between classes similarly.
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.

Analysis of the results and performance (Solely on the DISFA dataset)

1. *Classification results*

The first remark the authors made is that 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). And, accordingly, lower ICC(3,1) performances correspond to the AUs with less available training samples: AU15 (0.63) and AU20 (0.62). This shows that the model has been able to capture the slight nuances between labels it had enough data to train on, and perfectly demonstrates the effects of class imbalance.
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 authors compare their models to five recent state-of-the-art approaches, under the same conditions (training and validating models with DISFA database, provide ICC(3,1) and/or FM 1 performance)
The values obtained using this approach are only outperformed for some AUs by the works which make use of a costly SVM with Radial Basis Function kernel and therefore lacks real-time capabilities.

2. *Computational performance*

With real-time capabilities being the main focus of the study, it was essential that the authors detail the performance of their algorithm in areas where fast responses are crucial.
Videos taken under the same conditions as the DISFA database (1024x768 pixels at 20 fps) could be analyzed in real time without frame loss with a mid-range processor, while videos similar to those in the BP4D database (1040x1392 pixels at 25 fps) work with minor frame loss with a high-end processor.

An analysis of the computation times indicates that the we phase of the pipeline is face detection, by several orders of magnitude depending on the next task we compare it to. However, for good quality videos, it is performed once for the first frame and then the face points are tracked, so it has no impact on the overall computation time. The second longest process is the cropping, alignment, and rescaling required to compute the HOG features, which is proportional to the resolution of the image.

Such metrics have not been reported in most previous AU intensity detection works, which makes comparison difficult. But the studies mentioned in the article can serve as a baseline to demonstrate that speed is competitive with other state-of-the-art approaches.

Analysis of the results and performance (Cross-Dataset Experiments)

Having established the performances of the algorithm on a fixed dataset, the author went to test their adaptability and robustness, with an in-depth cross-dataset evaluation on DISFA, BP4D and SP databases.
This proved to be 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. It should be mentioned that all the AU12 models have been built following the method presented, both regarding features extraction and training procedure. The study outlined four kinds of cross-dataset testing conditions:

a. Within-corpus testing (WCT): the model is both machine-learned and tested using AU12 samples from one of the corpora.
b. Off-corpus testing (OCT): the model is first machine-learned using one corpus and, subsequently, tested on AU12 samples from another corpus.
c. 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.
d. Integrated off-corpus testing (IOCT): two different corpora are merged into one single corpus, but testing is performed using samples of a third dataset.

After conducting the presented experiments, the authors come to the following conclusions:
Overall, within-corpus tests achieve better classification results.
When the SP dataset is used as a test set, there is a substantial decrease in performance for each test condition, which is not surprising when differences in participant age and context are considered.
In addition to the social differences, from a technical perspective, the resolution of the SP samples is significantly lower (352x240 pixels) and likely led to a decrease in the HOG feature descriptor power
After digging deeper into the numbers using Manifold Learning techniques, the authors also realized that even though the inter-rater agreement is high for each individual database, the inter-dataset agreement is not so accurate.

Conclusions and future work

The paper concludes with a summary of its findings, which presents a real-time AU intensity detection method, the novelty of which lies in the proposed facial representation strategy: features were computed from a different ROI (mouth, eyes or frown) depending on the considered AU.
Real-time performance of 20-30 fps was achieved using HOG as feature descriptor and linear kernel SVM as classification mechanism. Eight different AU intensity detection models were built according to the proposed method and validated by cross-checking via the DISFA database, achieving similar or even better results than previous approaches without real-time capabilities.
Further evaluation of the datasets using AU12 samples from DISFA, BP4D, and SP databases validated the suitability of the proposed facial representation, proving that the AU12 model generalizes better to new datasets and contexts.
Therefore, reducing co-occurrence learning by separating the ROIs from the muscular face led to a more adaptable system for new datasets.

Critical analysis of the adopted methodology

The methodology adopted by the authors of the article is pretty interesting in the sense that is extends on previous works, by adding a very important feature, that forwards the field as a whole, which is live-time detection. The article, per say, did not engage the research community as a whole, but it was one of the very first ones to tackle this issue, and has opened the way for other studies of the same matter. As an example, we can cite "Real-time Action Unit Intensity Detection" (Saurabh Hinduja & Shaun Canavan, 2020). The authors themselves did not further extend their work, as they both moved to, related, yet unrelated research domains.

Another advantage of this approach is that is doesn't need an initial Neutral expression to take off from, but instead could analyze any video (as long as it is under the same conditions as the training datasets), and immediately classify the identified AU. This fact also highlights one of the study's shortcomings, which is its high reliance on lab-condition images for it to work best. And as we all know, real-life detection requires a very high generalization capability, as images quality is very random, depending on the setting. Moreover, cross-dataset capacities are to be taken with a grain of salt, as all results depend on one sole AU. More work could have been put into the creation of a bigger dataset comprising of more AUs, which would have enabled the algorithm to generalize better.

The authors themselves, at the end of the paper, introduce a method (Transfer Learning) that 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 contribute to extending cross-dataset study to more AUs. The authors indicated this was their future line of research.

Another method to improve the cross-dataset results, by using the same idea of extending to more AUs would be to address the issue that made us drop the other AUs in the first place. We could think of augmenting the datasets in order to include more pictures of other AUs, in order to try to balance the class imbalance that affects the datasets. Doing so would also allow us to deal with the rotation invariance from which suffers the HOG descriptor. By adding small rotations to our already existing images, we could make the algorithm robust to such changes.

Augmenting data by changing luminosity and intensity could also be another idea, but is less useful as the HOG is already robust to that.

As a conclusion, I would like to address the fact that nowadays, AU intensity detection is of great interest to many different fields, and has been for quite some years. And when the day comes, multimodal dialogue will find itself in a unique position, in the center of most AI disciplines, as it deals with most state-of-the-art approaches in different domains, enabling an interoperability between disciplines.

This is exactly the reason why I chose this article, as Artificial Intelligence is working towards a General AI that could make use of all the modalities of the human communication system, with all its components, to help the computer understand human dialogue and interactions better. This would allow the AI to first mimic it to its best extent, and maybe in the long future, use these modalities as well as a human does. And in order to do so, it needs to understand human emotions, and the best way to achieve that is through facial expressions, as stated earlier.