Intensity Score for Facial Actions Detection in Near-Frontal-View Face Sequences

Moi Hoon Yap  
School of Computing, Mathematics, and Digital Technology  
Manchester Metropolitan University  
John Dalton Building, Chester Street  
Manchester, M1 5GD, UK  
M.Yap@mmu.ac.uk

Hassan Ugail  
Centre for Visual Computing  
University of Bradford  
Bradford, BD7 1DP, UK  
H.Ugail@bradford.ac.uk

Reyer Zwiggelaar  
Department of Computer Science  
Aberystwyth University  
Aberystwyth, SY23 2AX, UK  
rrz@aber.ac.uk

Abstract—this paper proposes a method to detect the facial Action Units (AUs) and introduce an automatic measurement in predicting the intensity scores of each AU in near-to-frontal face image sequences. To our knowledge, this is the first attempt in computer vision to automate the intensity scores of facial expression research. First, the facial feature points are detected by using Gabor feature based boosted classifiers and the movement of each point is tracked by optical flow. Then, we introduce a set of distance measurement for the feature points and analyze the distances by using Sequence Analysis. Further, we exploit the sequence partition to predict the possible temporal segments in the sequence. We found there is a relationship between the intensity scores and the partition threshold, which we automated the process of threshold selection in our work. We tested the proposed prototype with our in-house dataset and MMI database. Finally, we discuss the result, the possibilities of further research, and the next challenges for computer vision scientist in facial actions detection.

Keywords—Facial Action; FACS; intensity score; facial expressions; temporal segments

I. INTRODUCTION

Human face is explained as the major communicative outputs and major sensory inputs [1]. According to Kong et al [2], the demand and research activities in machine recognition of human faces have increased significantly over the past 30 years. Such statement can be supported by the facts that a lot of past researches [3-5] and on-going researches [6] are trying to reveal the face information to provide clues/cues for entertainment, security and other technology development.

In psychology, it has a long history in facial behavioral analysis and measurement. The study of facial expressions has been conducted in last century [7, 8]. Facial Action Coding System (FACS) is less time consuming and cost saving compare to other facial behavioral measurement methods [9]. A few experiments have shown the popularity of FACS in research, for instance, Ekman et. al. [9] used FACS in predicting subjects’ emotional experience while watching the films. Also, Kunz et. al. [10] has utilized FACS in pain assessment for dementia diagnostics.

In computer vision, the considerable progress has been made in automated facial expression analysis from digital video input within the past decade [4, 11]. Currently, there are a lot of researchers involved in the automatic recognition of AUs temporal segments. The existing approaches to facial expression analysis include geometric approach [12, 13], appearance-based approach [14], and Dynamic-Texture based approach [6]. However in the existing systems of automatic recognition of AUs temporal segments [6, 12, 13, 15], none of them make an attempt in recognising or predicting the intensity scores in the AUs. The novelties in this work are the implementation of sequence analysis in AUs detection and the introduction of an automatic method in intensity scores prediction. Intensity scores in facial expression are important in measuring the facial responses. Kunz et. al. [10] has demonstrated the contribution of the intensity scores in pain diagnostics for dementia research. They showed that the facial expression of pain, with intensity stimulation, has the potential to serve as alternative pain assessment tool in demented patients [10].

For clarity of presentation this paper has been divided into 5 sub-sections. Apart from this section which provides the reader an introduction to the problem domain and setting out the objectives of the proposed research, Section-2 introduces the background and research motivation. Section-3 proposes a methodology to detect the Action Units and intensity scores. Section-4 provides details of results and a comprehensive analysis. Finally section 5 provides the conclusions with an insight to possibilities of further research.

II. BACKGROUND AND MOTIVATION

The motivation for this work arises from a research project funded by UK Engineering Physical Sciences Research Council under the theme of the facial analysis for
real-time profiling project. The aim of the project is to provide a real-time dynamic passive profiling technique which will assist as a decision aid to Border Control Agencies, which has the potential to improve hit rates. The project is intended to combine and build on several research areas, which include multi-modality face and eye-motion tracking, eye-motion related to intent, dynamic thermal/visible face information related to intent, statistical shape and appearance modeling and face modeling and recognition.

Another motivation, like any other computer vision scientists, we attempt to automate the recognition and scoring of AUs. Face is the prominent visual object, and the expression is the basic mode of nonverbal communication among people [16]. It offers non-intrusive, perhaps the most natural way of authentication [2] – intuitive and does not stop user’s activities. Research in measuring the face has been impeded by the problems of devising an adequate technique in distinguishing all possible visual facial movements. In psychology and human vision, Ekman et. al. introduced a way to recognize and score the AUs [17, 18]. This is not only a breakthrough in human vision and psychology, but it has motivated researchers from other domains (especially in computer vision) to work in this area. FACS is an extension of Facial Affect Scoring Technique (FAST) to distinguish all possible visually distinguishable facial movements [18]. It was built based on the neuroanatomy of facial behavior on the evolution of the nervous system and parallel elaborations of facial musculature [9]. There are three measurements in FACS: Type, Intensity, and Timing. The face is neutral when evidence of any specific AU is absent. Type refers to the action type, for instance, inner brow raise. Intensity refers to the magnitude of the appearance change. Timing refers to the duration of the movement. The five-point ordinal scale (A-B-C-D-E) is used to rank the present of the AU [18]:

- A: trace of the action
- B: slight evidence
- C: marked or pronounced
- D: severe or extreme
- E: maximum evidence

The following section explains the methodology used in AUs detection and prediction of the intensity scoring.

III. METHODOLOGY

This section explains the methodology of our project, which includes facial feature point detection, facial feature point tracking, facial actions and analysis, and intensity scores measurement.

A. Facial Feature Point Detection

Among the facial feature point detection algorithms, Gabor feature based boosted classifiers is most robust at the time we write this paper. It is known that Gabor filters remove most of the variability in image in different illumination and contrast, it also robust against small shift and deformation [19]. In our work, we adopted the algorithm introduced by Vukadinovic and Pantic [19], which defines 20 facial points in images of neutral faces. This algorithm implemented Viola and Jones face detector [20] and divided the detected face into 20 regions of interest. Then, each of the regions is examined further to predict the location of the facial feature points. They used the individual feature patch templates to detect points in the relevant region of interest. The feature models are GentleBoost templates built from the gray level intensities and Gabor wavelet features [19]. Figure 1 illustrates the implementation of Gabor feature based boosted classifier on our in-house data. For further details of the Gabor Feature Based Boosted Classifiers in facial feature point detection, please refer to Vukadinovic and Pantic [19]. The executable version of the algorithm is freely available for research purpose from Pantic’s personal website [21].

B. Facial Point Tracking

In early tracking system, for e.g. in [22-24], feature matching was carried out from one frame to the next using optical flow computations. This has caused a problem which resulting in drifting errors accumulating over long image sequences. Face motion produces optical flow in image. The optical flow approach has the advantage of not requiring a feature detection stage of processing. Optical flow is the visible result of movement and is expressed in terms of velocity, it is a direct representation of facial action [24, 25]. Muscles actions can be directly observed in image sequence as optical flow, which is calculated by facial features and skin deformation [26]. Dense flow information was computed from images sequences of facial expression with Horn et al’s [27] basic gradient algorithm, which is then reduced by taking the average length of directional components in the major directions of muscle contraction. Several muscle windows are located manually to define each muscle group using feature points as references. The muscle action derived from the muscle (group) model can be associated with several AUs of the FACS. In Mase’s [25] experiment, he managed to identify 19 out of 22 test data in recognizing four expressions (happiness, anger, surprise and disgust) in motion. Due to the history and benefits of optical flow, we have adopted optical flow in our prototype.
C. Facial Actions and Analysis

Figure 2 shows the labeling of the facial feature points defined by Vukadinovic & Pantic [19]. They defined facial feature points as the corners of the eyes, corners of the eyebrows, corners and outer mid points of the lips, corners of the nostrils, tip of the nose, and the tip of the chin [19]. For instance, Point 19 shows the location of nose tip. The following explain the definition of each parameter used in facial action analysis:

![Figure 2. Labeling of the facial feature points](image)

In facial action analysis, we assume each action consists of four temporal segments: Neutral, Onset, Apex, and Offset [18]. Neutral is the condition of expressionless face. Onset is the duration of the start frame of the action to apex. Apex is the duration of the greatest excursion of that action [18]. Offset is the duration of the end of Apex to Neutral. Figure 3 describes the temporal segments for the subject who expressed a happy expression. The first frame and the last frame of the sequence are the neutral state of the subject. The corner of the mouth start to pull from frame 15 to frame 25, where greatest pulling happens and the apex of this action is achieved. The subject holds the expression for 53 frames, at frame 78, she starts to resume to neutral state - the offset is from frame 78 to frame 88. The duration of temporal segments is very important in differentiate a blink from closed eyes, and possibly, in the future, to decide the micro-expressions.

D. Intensity Scores

Due to the limitation of current research in computer vision for facial features detection/tracking, we reduced the five scales into three scales. We regroup the scoring into:

- **Slight**: which consist of trace of an action and slight evidence, category A and B in FACS intensity scoring
- **Marked**: marked or pronounced, category C in FACS intensity scoring
- **Extreme**: which consist of severe or extreme, and maximum evidence, i.e. category D and E in FACS intensity scoring

In Table I, parameters for face action units recognition are listed.

<table>
<thead>
<tr>
<th>Action Unit (AU)</th>
<th>Name of AU</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU0</td>
<td>Neutral</td>
<td></td>
</tr>
<tr>
<td>AU1</td>
<td>Inner Brow Raiser</td>
<td>BbrR AND BbrL, increase</td>
</tr>
<tr>
<td>AU2</td>
<td>Outer Brow Raiser</td>
<td>BbrR and BbrL, increase</td>
</tr>
<tr>
<td>AU4</td>
<td>Brow Lowerer</td>
<td>BbrDist, BbrL, BbrR, BbrL, decrease</td>
</tr>
<tr>
<td>AU5</td>
<td>Upper Lid Raiser</td>
<td>EerL &amp; EerR, decrease</td>
</tr>
<tr>
<td>AU6</td>
<td>Check Raiser and Lid</td>
<td>EerL &amp; EerR, decrease</td>
</tr>
<tr>
<td>AU7</td>
<td>Lid Tightener</td>
<td>EerL &amp; EerR, decrease, EerL &amp; EerR, decrease</td>
</tr>
<tr>
<td>AU10</td>
<td>Upper Lip Raiser</td>
<td>MbrL decreases, MbrR increases</td>
</tr>
<tr>
<td>AU12</td>
<td>Lip Corner Puller</td>
<td>LbrL &amp; LbrR, decrease</td>
</tr>
<tr>
<td>AU15</td>
<td>Lip Corner Depressor</td>
<td>LbrL &amp; LbrR, increase, MbrL increases</td>
</tr>
<tr>
<td>AU25</td>
<td>Lips Part</td>
<td>MbrR &amp; MbrL increase, EerL &amp; EerR, increase</td>
</tr>
<tr>
<td>AU45</td>
<td>Blink</td>
<td>Only AU45 is available, no other intensity scores</td>
</tr>
<tr>
<td>AU43</td>
<td>Eyes are closed completely</td>
<td>EerL &amp; EerR = 0</td>
</tr>
<tr>
<td>AU46</td>
<td>Wink</td>
<td>Unilateral Blink (AU45), Extreme changes in EerL or EerR, but not both</td>
</tr>
</tbody>
</table>

![Figure 3. Analysis of temporal segments (onset, apex, and offset) in a sequence of frames](image)
Each of the movement tracking of the points will generate a sequence/series as show in figure 3. In order to automatically detect the hill or valley pattern in the sequence, we implement Sequence Partition [28] to split a sequence into equivalency classes. The categorical predictor is using a quadratic algorithm for splitting a set into one or more equivalency classes and the function returns the number of equivalency classes [29]. The comparison function is based on Euclidean distance.

In order to obtain an automatic intensity score, we make some assumptions: First, we expect the number of classes should be in between 3 and 15. Second, we predict the threshold’s range is from 4 to 12. By default, the threshold is set to 7. The value of 7 represents the Marked/Pronounced of the facial action, the value from 4 to below 7 represent Slight of the facial action, and the value above 7 to 12 represents Extreme of the facial action. We programmed the prototype to automatically change the value of the threshold and stop when the number of classes is fall in between 3 and 15. Once we have the threshold fixed, we compare the classes to detect the hill and valley patterns.

We will further discuss and demonstrate the meaning of the intensity scores in Section 4.

IV. RESULT AND ANALYSIS

To explain the full functionality of our prototype, we illustrate the process by using a short clip from our in-house database. Figure 4 is a sequence of images from our in-house dataset, with the surprise expression.

Figure 4. A sequence of images from a short clip of our in-house database

From this video, we observe Slight, Marked, and Extreme facial actions on the face. For instance, Slight for the distance between upper lip to nose tip (MouthUp), Marked for the distance between right inner brow and nose tip (BInR), and Extreme for the Mouth size vertically (MSizeV). Some of the screen shots of results of the AUs detection and automated intensity scoring are illustrated in figure 5. Figure 5(a) shows an increase distance between Upper Lip from nose with intensity 4, this can be interpreted as a Slight movement of the upper lip. Figure 5(b) shows an increase distance between the eye brows with intensity 7, we interpreted this as a Marked increase in the brows distance. Figure 5(c) shows the increment of the vertical size of the mouth with intensity 10. This can be interpreted as an Extreme lips part. In figure 5(d), the distance of the lower lip from the nose tip is increase with intensity 12. This figure shows an Extreme lower lip moving away from nose tip, adding the confidence of Extreme lip parts.

Figure 5. Screen Shots of results of the Aus detection and automated intensity scoring.

In addition, we also detect the wink and blink of the eyes by checking the size of the eyes. The distance of the eye lids is close to 0 if a wink or a blink is happened. For example, in this case, a blink is detected in this short clip at frame 43 for duration of 2 frames, as shown in figure 6. The prototype also generated a detailed report on the analysis at the end of the analysis, as in figure 6. From the report, we conclude that the short clip consists of Slight AU10, Marked AU1 and AU2, Extreme AU25, and AU46.

We have tested the prototype with MMI database [30]. From the database, we managed to detect AUs listed in Table I. The result is quite promising and the prediction is sound. In the case of miss detection/ failure in recognition, it is always caused by the error in computer vision algorithms, it is, detection errors and tracking errors.
V. CONCLUSION

In this project, we take the first step in the attempt to predict the intensity scores of the Facial Action Units. Also, we introduce sequence analysis in automatically recognizing AUs and its’ intensity scores. We believe that our work will motivate a lot of researchers in looking into the effort of intensity scoring automation and AUs recognition. The inaccuracy of the computer vision algorithms had a great impact on the facial action analysis. We will investigate in improving the facial feature tracking algorithm by replacing optical flow with particle filtering [31, 32]. The current prototype has a limitation: it only works on short clips (100 frames). In future, we will extend our work to long video clips analysis and recognize the AUs from the speech. In practical research, the subjects are normally facing an interview. It is a challenge for the FACS coder and researcher to recognize the AUs from the speech. On the other hand, we also plan to extend our work on the recognition of other AUs and the combination of AUs.

ACKNOWLEDGMENT

This work is supported by EPSRC grant on “Facial Analysis for Real-Time Profiling” (EP/G004137/1). The authors would like to thank Pantic et al for providing the MMI database.

REFERENCES


Figure 6. A Report on AUs recognition and intensity scores.


[31] I. Patras, and M. Pantic, “Particle Filtering with Factorized Likelihoods for Tracking Facial Features.”