Image and Vision Computing 32 (2014) 476–486

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

Automatic expression spotting in videos ☆

Matthew Shreve *, Jesse Brizzi, Sergiy Fefilatyev, Timur Luguev, Dmitry Goldgof, Sudeep Sarkar

Department of Computer Science and Engineering, University of South Florida, Tampa, FL, USA

ARTICLE INFO

Article history:
Received 12 February 2013
Received in revised form 18 February 2014
Accepted 30 April 2014
Available online 9 May 2014

Keywords:
Expression spotting
Macro-expressions
Micro-expressions

ABSTRACT

In this paper, we propose a novel solution for the problem of segmenting macro- and micro-expression frames (or retrieving the expression intervals) in video sequences, which is a prior step for many expression recognition algorithms. The proposed method exploits the non-rigid facial motion that occurs during facial expressions by capturing the optical strain corresponding to the elastic deformation of facial skin tissue. The method is capable of spotting both macro-expressions which are typically associated with expressed emotions and rapid micro-expressions which are typically associated with semi-suppressed macro-expressions. We test our algorithm on several datasets, including a newly released hour-long video with two subjects recorded in a natural setting that includes spontaneous facial expressions. We also report results on a dataset that contains 75 feigned macro-expressions and 37 feigned micro-expressions. We achieve over a 75% true positive rate with a 1% false positive rate for macro-expressions, and a nearly 80% true positive rate for spotting micro-expressions with a .3% false positive rate.

1. Introduction

Accurately and automatically spotting frames containing facial expressions in videos is often a prior step required for high-level facial analysis, including identifying emotional response, gestures, and human identification. In many papers, this is a manual pre-processing step, or it is assumed that the data consists of a single facial expression sequence. We address this problem using an optical strain based method that is capable of automatically spotting expressions in video sequences.

In this paper we do not address the problem of identifying expressions, but only solve the expression segmentation problem (see Fig. 1). Since our method is based on the non-rigid motion of the face, and not on pre-defined expression models, we are able to capture a large variety of facial motion corresponding to facial expressions. In other words, while some traditional techniques are capable of recognizing pre-defined expressions (such as recognizing when a person smiles or shows surprise in a video sequence), it is not possible for these types of methods to recognize expressions for which the algorithm has not been trained. We propose a novel expression spotting method that can be used to locate and distinguish between two broad classes of expressions. First are macro-expressions, which are generally characterized as occurring several seconds over several regions of the face. The second class of expressions typically occurs rapidly and in a single region of the face, or micro-expressions.

The method presented in this paper represents our complete work on expression spotting. Earlier ideas related to this method were documented in [1], with some further results reported in [2]. Some of our early micro-expression work was reported in [3] and also in a medical application [4]. We have developed a completely new algorithm compared with our prior work. Specifically, we have included more robust face tracking, a new masking technique, and a new peak detection algorithm. We show the performance of the algorithm at several scaled resolutions. We give results on more challenging datasets, including longer videos that contain a mixture of both macro- and micro-expressions during a single sequence, as well as spontaneous (genuine) facial expressions.

The method consists of the following steps: (i) the face and eyes are detected in all frames of the video sequence. These coordinates are then used to segment the face in to several regions; (ii) the non-rigid motion is estimated using an optical flow based method over several frames, for each region; (iii) a masking technique is used that removes erroneous flow estimation caused by blinking the eyes or the opening and closing of the mouth; (iv) optical strain maps are calculated over each pair of frames to generate strain maps, which are then summed to generate strain magnitude for four different regions on the face; (v) lastly, a peak detection method is used to locate intervals that correspond with macro-expressions [5], and the remaining intervals are then analyzed for micro-expressions.

It is important to note that while many papers address the problem of expression spotting, the definition of spotting is not always consistent. It may also be referred to as expression detection, or in the case of genuine expressions, spontaneous [6] or authentic expression analysis [7]. In some papers, this refers to determining if a pre-segmented

☆ This paper has been recommended for acceptance by Georgios Tzimiropoulos.

* Corresponding author at: 4202 E. Fowler Avenue, University of South Florida, ENB-118, Tampa, FL 33620-5399, USA. Tel.: +1 813 974 3652; fax: +1 813 974 5456.
E-mail address: mshreve@mail.usf.edu (M. Shreve).

http://dx.doi.org/10.1016/j.imavis.2014.04.010
0262-8856/

Fig. 1. The scope of this paper is seen in red (dashed circle).

group of frames does or does not contain an expression. For example Zeng et al. [6] propose a single classification method for detecting spontaneous expressions (emotion or non-emotion) by training a Support Vector Data Description (SVDD) on several examples of emotion data. Then, test segments containing roughly an equal number of frames are classified with a single binary decision (hence chance is 50%). The same type of experimental setup can be found in [8] for macro-expressions, and in Pfister et al. [9] and Polikovsky et al. [10] for micro-expressions. In [9], local spatio-temporal features are extracted from a high-speed video sequence (100 frame/s) and then performance is measured using several classifiers. Similarly in [10], a high speed camera (300 frames/s) is used to capture rapid micro-expressions. In their work, they use 3-D gradient histograms and the Facial Action Coding System (FACS) to spot the 4 states of micro-expressions: onset, apex, apex offset, and apex neutral.

A single frame approach using Gabor filters and GentleSVM is used by Wu et al. [11] Their work is based on the assumption that the appearance of micro-expressions completely resemble macro-expressions, and thus they reduce the entire micro-expression problem to the temporal duration of macro-expressions. While this definition may fit a subset of micro-expressions, we instead use the general categorization found in [12] that defines them as a suppression of macro-expressions. Hence, they are often distorted or only fractional representations of macro-expressions [13].

Some papers that address the problem of automatic expression analysis do so only for pre-trained examples, or in other words, they are capable of recognizing a subset of expressions in uncut videos. For instance, a dynamic approach can be found in the work by Sung et al. [14], the authors use generalized discriminant analysis for recognizing a subject randomly expressing four different facial expressions (neutral, happy, surprised, angry) in roughly 30 second videos. Another example can be found in [15] where seven expressions (neutral, sadness, anger, joy, fear, disgust, surprise) are automatically recognized in videos. A method that can detect several more affective states can be found in [16].

Static (single frame) approaches that model a subset of macro-expressions are also found extensively in the literature, with perhaps local binary patterns performing among the best [17]. In the work by Ruiz-Hernandez and Pietikainen [18], a novel LBP encoding technique is proposed that uses a re-parametrization of the second order Gaussian jet. Similar to [9] they do not perform spotting, but test on sequences that each contains a single expression.

A popular method for describing several types of expressions on the face is the Facial Action Coding Systems (FACSs). In this system, an action unit label is given to different types of facial motion, and the activation of one or more action units can be used to describe a facial expression. There are several works in the literature that automatically detect action units, as well as provide a measure of intensity [19,20]. While there is a similarity between detecting the activation of an action unit and detecting expressions, the two are not equivalent. For instance, in the DISFA dataset [21], every action unit is a measure of several parts of expressions, but not fully reduced; in other words, there are still types of expressions, especially micro-expressions, that are not represented by the labeled action units. Of course, it is possible to find an action unit to describe every type of motion on the face, but then training would be required on each of them. Hence, the main contribution of our work is that individual training of for each type of motion, or action unit, is not needed. The goal of our approach is to successfully detect any type of motion that causes strain, or deformation on the face.

In the work by Zhou et al. [22], the authors propose to use Aligned Cluster Analysis on points obtained using FACS. In their work, they are able to identify differing types of spontaneous facial expressions, although it is dependent on a manually defined number of expression clusters, and performance on micro-expression detection is not given. Another unsupervised approach is given by Liwicki et al. [23]. In this approach, an online temporal video segmentation technique is described that uses a subspace learning method. While it has been used to successfully segment several macro-expressions from a video sequence, it assumes that changes in a scene (or signal) occur slowly, thus it does not appear to be suited for rapidmicro-expressions which can occur in a little as 2–3 frames.

Similarly to the unsupervised methods, we do not rely on previously trained models of expressions. However, we want to emphasize a few key highlights of our method: (i) we rely on the fundamental dynamics of facial expressions, in that they cause the non-rigid deformation of the facial skin tissue. Hence, our method is naturally suited to spot all expressions that cause facial skin deformation; (ii) we detect facial expression over the entire video sequence, without any manual temporal pre-segmentation (however we do provide results for an experiment that is formatted similarly to [9] and [18] who both test on the SMIC corpus, so performance can be compared); and (iii) we are not aware of any other method that has yet been proposed that detects both macro- and micro-expressions. Finally, for clarity, we propose that expression spotting refer to the temporal segmentation of an entire video into segments that only contain the frames of each expression.

To further place our method into perspective, we find the categorization of motion-based methods useful for expression spotting. In general, they have been organized [24] into three types, namely: point-model, holistic, or some mixture of these two. Point-model approaches track several key points on the face over time. The interplay of these points can then be used to recognize the expression [25]. However, while these algorithms may be sufficient for large macro-expressions, the average 2–3 pixel "jittering" in frame to frame tracking often suppresses the nearly equally small movement observed in micro-expressions. Alternatively, holistic approaches track all points on the face [26], and hence become more suitable for detecting a larger variety of possible facial expressions, including small motion.

Our method fits into the last category, i.e., it uses both a point-model and holistic approach. Our approach segments the face into several regions based on several detected landmarks. Then, within these segmented regions we use a holistic optical flow method that densely tracks each point on the face. Hence, we hope to have minimized the drawback of approaches in the first category, while taking advantage of the potential robustness associated with methods in the second category.

2. Background

Expressions are generally believed to be the physiological response to an internal emotional state. While there does appear to be a universality for some expressions (such as happiness, sadness, surprise, disgust, and anger) there are a much larger number of possible expressions possible, as well as large inter- and intra-variability between subjects for the same expression. For instance in Fig. 2 there some expressions we may immediately recognize (such as anger in column d), however some other expressions may be harder to recognize, or in fact may not be recognized without context. Hence, the goal is of this paper is not to present a method which only spots pre-defined expressions, but to spot segments containing any possible type of facial expression that involves the strain (or deformation) of facial skin tissue.

2.1. Macro-expressions

Macro-expressions typically last 3/4th of a second to 2 s (roughly 24–60 frames) [13]. There are 6 universal expressions: happiness,