Image and Vision Computing 51 (2016) 36–48

Contents lists available at ScienceDirect

Image and Vision Computing

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

Cascade of Tasks for facial expression analysis ☆

Xiaoyu Ding[a,*], Wen-Sheng Chu[b], Fernando De la Torre[b], Jeffery F. Cohn[b,c], Qiao Wang[a]

[a] School of Information Science and Engineering, Southeast University, Nanjing, China
[b] Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, United States
[c] Department of Psychology, University of Pittsburgh, Pittsburgh, PA 15260, United States

ARTICLE INFO

Article history:
Received 21 February 2014
Received in revised form 28 August 2015
Accepted 22 March 2016
Available online 31 March 2016

Keywords:
Automated facial expression analysis
Action unit detection
FACS

ABSTRACT

Automatic facial action unit (AU) detection from video is a long-standing problem in facial expression analysis. Existing work typically poses AU detection as a classification problem between frames or segments of positive and negative examples, and emphasizes the use of different features or classifiers. In this paper, we propose a novel AU event detection method, Cascade of Tasks (CoT), which combines the use of different tasks (*i.e.*, frame-level detection, segment-level detection and transition detection). We train CoT sequentially embracing diversity to ensure robustness and generalization to unseen data. Unlike conventional frame-based metrics that evaluate frames independently, we propose a new event-based metric to evaluate detection performance at the event-level. The event-based metric measures the ratio of correctly detected AU events instead of frames. We show how the CoT method consistently outperforms state-of-the-art approaches in both frame-based and event-based metrics, across four datasets that differ in complexity: CK+, FERA, RU-FACS and GFT.

1. Introduction

Facial expressions convey varied and nuanced meanings. Small variations in the timing and packaging of smiles, for instance, can communicate politeness, enjoyment, embarrassment, or social discomfort [1,2]. To analyze information communicated by facial expressions, Ekman and Friesen proposed the Facial Action Coding System (FACS) [3]. FACS describes facial activity in terms of anatomically based action units, which can occur alone or combine to represent all possible facial expressions. Action units (AUs) have a temporal envelope that minimally include an onset (start) and offset (stop) and may include change in intensity. Researchers have defined 33 to 44 AUs, depending on FACS version [4].

In computer vision, automated AU detection has become an active area of research [6–15] and has been applied to marketing, mental health, instructional technology, and media arts [16–20]. Despite its descriptive power [5], automatic AU detection is challenging: non-frontal pose and moderate to large head motion complicate facial image registration; the temporal scale of facial actions varies considerably; individual differences occur in shape and appearance of facial features; and many facial actions are inherently subtle. Due to the thousands of possible combinations of 30-40 or more AUs, detection typically is decomposed to a binary classification problem on each AU.

Existing AU detection methods broadly fall into one of three types: frame-level detection, segment-level detection, and transition detection. Frame-level detection independently evaluates each video frame for the occurrence of one or more AUs [8,11,13,21–24]. Segment-level detection seeks to detect contiguous occurrences of AU that ideally map onto what manual FACS coders perceive as an event [12,25–27]. Transition detection seeks to detect the onset and offset of each segment, or event [28]. See [29,30] for recent surveys.

Most approaches to AU detection are frame-level detectors, which consider each video frame as independent. Because this assumption ignores the inherent auto-correlation of behavioral data, detection tends to be noisy with classifiers firing on and off in proximal frames. By contrast, human observers do not evaluate video frames individually. Rather, they perceive AU as *events* that have a beginning (onset), an end (offset), and a certain duration. Consequently, manual FACS coding requires significant effort to first perceive an AU event and then identify its precise onset and offset. To identify such events, researchers rely on segment-level detection. Often, it is relatively easy to detect the temporal segment in the middle of an AU event with high intensity or large facial movement, yet the transition points between AU inactivation and activation are more subtle and difficult to detect. We seek to automatically detect

☆ This paper has been recommended for acceptance by Vladimir Pavlovic.

* Corresponding author. Tel.: +1 412 999 8605.
E-mail addresses: leonxd@andrew.cmu.edu (X. Ding), wschu@cmu.edu (W. Chu), ftorre@cs.cmu.edu (F. De la Torre), jeffcohn@pitt.edu (J. Cohn), qiaowang@seu.edu.cn (Q. Wang).

http://dx.doi.org/10.1016/j.imavis.2016.03.008
0262-8856/

AU events, including onsets and offsets, with high fidelity to human perception.

To achieve this goal, we propose a Cascade of Tasks (CoT). CoT detects AU events including their onsets and offsets, by sequentially integrating the three AU detection tasks: frame-level detection, segment-level detection, and detection of onsets and offsets. Fig. 1 illustrates the CoT process. The first task detects AU at the frame-level. The results of this task tend to be noisy, or less reliable, because it fails to exploit the temporal dependencies among proximal frames.

The second task combines the output of the frame-level detection with new segment-level features with a segment-based classifier (see Fig. 1 second row). The segment-level detector gives a rough location of the AU event and reduces the frame-level false positives, but is imprecise in the boundaries (*i.e.*, onset and offset). The third task addresses this problem. By integrating the three tasks, CoT provides a more robust and precise detection of AUs than previous approaches.

Our contributions are two-fold. 1) To the best of our knowledge, CoT is the first approach to integrate multiple *tasks* for AU detection. Most other algorithms for AU detection emphasize different features or a classifier, or combine them with ensemble-type methods to solve a single task. However, our approach combines different tasks.

2) CoT fully recovers AU events instead of isolated AU frames or incorrectly parsed segments.

To evaluate AU detection performance at event-level, we propose a new event-based metric, as opposed to conventional frame-based metrics that evaluate frames independently.

2. Previous work

We broadly categorize AU detection approaches into three types of *task*: frame-level detection, segment-level detection and transition detection. These approaches largely differ on the methods for registration, feature representation, and classifier learning. Here we review recent work on AU detection. Refs. [6,7,29–31] offer more complete surveys.

The first AU detection challenge (FERA) [7] indicates that most approaches, including the winning one, were frame-based. Frame-level methods detect AU occurrences in individual frames by extracting geometric or appearance features to represent each frame, which are then fed into static classifiers (*e.g.*, SVM [8,32] or AdaBoost [11,13]). Geometric features contain information of facial feature shapes, including landmark locations [22,32,33] and geometry of facial components [34]; appearance features capture texture changes of the face, such as wrinkles and furrows, and can be typically represented by Gabor [11,35], LBP [24,36,37] and DAISY/SIFT descriptors [13]. A notable trend in this area is fusing various features/classifiers to generate more accurate and robust results [38,39]. For example, Tariq et al. [40] concatenated image features, including SIFT, Hierarchical Gaussianization and optical flow, as input to a SVM classifier. Later, Tariq et al. [9] used a log sum model to fuse the outputs of classifiers trained separately with different low-level image features.

In their study of multilayer architectures of texture-based image feature descriptors (filters), Wu et al. [21] showed that adding a second layer of nonlinear filters consistently improved performance. This approach represents a special way to fuse feature descriptors. Almaev and Valstar [41] proposed a temporal extension to the multi-layer appearance features (LGBP-TOP). More recently, Jiang et al. [42] proposed a decision-level fusion strategy to combine region-level classifiers. First, domain knowledge regarding FACS AU definition is used to define a face region. Second, a region-specific classifier is trained for each region. Finally, a weighted sum combines outputs of these classifiers.

Segment-level approaches seek to incorporate temporal information of facial action, and to detect AU as a set of contiguous frames. To capture temporal information, dynamic features have been used to measure motions on a face [43,44], such as raising mouth corners. Recent work on exploiting dynamic features includes bag of words [12] and temporal extensions to LBP, LGBP and LPQ [20,23,37,41,45]. Another approach models the AU state change over time using temporal classifiers or models. For example, Chang et al. [25] use hidden conditional random fields to link the AU state with underlying emotions in facial expression sequences. At each time

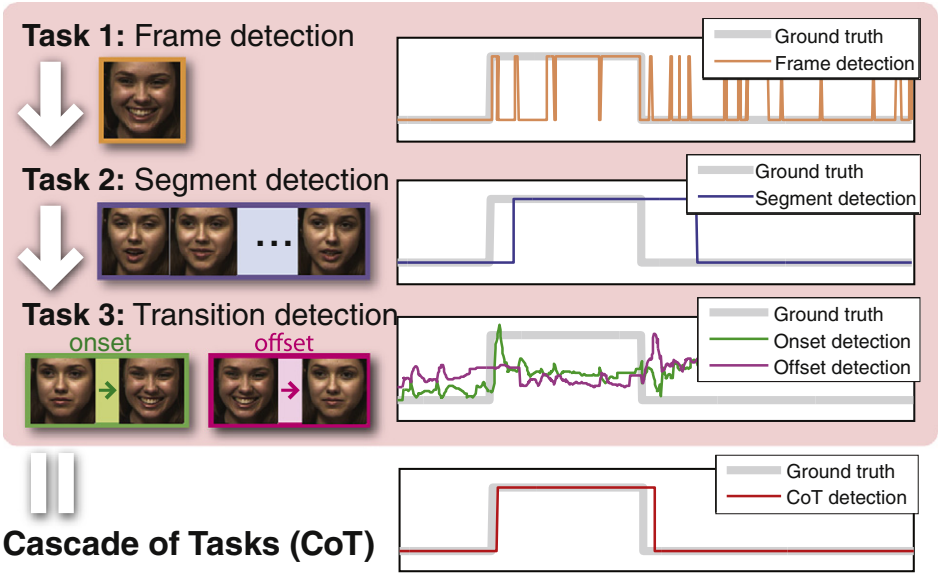

Fig. 1. Detection of AU 12 (smile) from its onset to offset using our proposed CoT method. In the plots to the right above, thick gray lines indicate ground truth and thin lines indicate prediction results. First, CoT detects AU 12 in individual frames (Task 1). Because this step assumes that individual frames are independent, it is prone to error. Next, CoT uses the responses of the frame-level detector and segment-based features to detect a segment for AU 12 (Task 2). Finally, CoT more precisely estimates the onset and offset frames by learning transition detectors (Task 3).