Pattern Recognition 48 (2015) 3417–3427

Contents lists available at ScienceDirect

Pattern Recognition

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

Measuring the intensity of spontaneous facial action units with dynamic Bayesian network

Yongqiang Li [a,*], S. Mohammad Mavadati [b], Mohammad H. Mahoor [b], Yongping Zhao [a], Qiang Ji [c]

[a] School of Electrical Engineering and Automation, Harbin Institute of Technology, Harbin 15001, China
[b] Department of Electrical and Computer Engineering, University of Denver, Denver, CO 80208, USA
[c] Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

ARTICLE INFO

Article history:
Received 27 March 2014
Received in revised form
16 March 2015
Accepted 20 April 2015
Available online 29 April 2015

Keywords:
Spontaneous facial expression
AU intensity
Dynamic Bayesian Network
FACS
DISFA database

ABSTRACT

Automatic facial expression analysis has received great attention in different applications over the last two decades. Facial Action Coding System (FACS), which describes all possible facial expressions based on a set of facial muscle movements called Action Unit (AU), has been used extensively to model and analyze facial expressions. FACS describes methods for coding the intensity of AUs, and AU intensity measurement is important in some studies in behavioral science and developmental psychology. However, in majority of the existing studies in the area of facial expression recognition, the focus has been on basic expression recognition or facial action unit detection. There are very few investigations on measuring the intensity of spontaneous facial actions. In addition, the few studies on AU intensity recognition usually try to measure the intensity of facial actions statically and individually, ignoring the dependencies among multilevel AU intensities as well as the temporal information. However, these spatiotemporal interactions among facial actions are crucial for understanding and analyzing spontaneous facial expressions, since these coherent, coordinated, and synchronized interactions are that produce a meaningful facial display. In this paper, we propose a framework based on Dynamic Bayesian Network (DBN) to systematically model the dynamic and semantic relationships among multilevel AU intensities. Given the extracted image observations, the AU intensity recognition is accomplished through probabilistic inference by systematically integrating the image observations with the proposed DBN model. Experiments on Denver Intensity of Spontaneous Facial Action (DISFA) database demonstrate the superiority of our method over single image-driven methods in AU intensity measurement.

1. Introduction

Facial expression is one of the most common nonverbal communication media that individuals use in their daily social interactions. Analyzing facial expression will provide powerful information to describe the emotional states and psychological patterns of individuals. In the last two decades automatic facial expression recognition has gained more attention in several applications in developmental psychology, social robotics, affective online tutoring environment and intelligent Human–Computer Interaction (HCI) design [1,2].

Facial Action Coding System (FACS) is one of the well-known approaches for describing and analyzing facial expressions [3]. FACS describes all possible facial expressions based on a set of anatomical facial muscle movements, called *Action Unit (AU)*. For instance, AU12 or *lip corner puller* specifies contractions that occur on the face by Orbicularis oculi muscle [3]. FACS is also capable of representing the dynamics of every facial behavior by annotating the intensity of each AU in a five ordinal scale (i.e. scales A–E that indicate the barely visible to maximum intensity of each AU). AU intensity can describe the occurrence of spontaneous facial expressions in more detail. The general relationship between the scale of evidence and the A–B–C–D–E intensity scoring, as well as some AU samples is illustrated in Fig. 1. Generally, the A level refers to a *trace* of the action; B, *slight* evidence; C, *marked* or *pronounced*; D, *severe* or *extreme*; and E, *maximum* evidence. For example, “AU12B” indicates AU12 with a B intensity level. Manual FACS coding is an intensive and time consuming task and designing an automatic system which can specify the list of AUs and their intensities would help the community to analyze spontaneous facial behaviors accurately and efficiently.

* Corresponding author. Tel.: +86 15245104277.
E-mail address: liyongqiang@hit.edu.cn (Y. Li).

http://dx.doi.org/10.1016/j.patcog.2015.04.022
0031-3203/

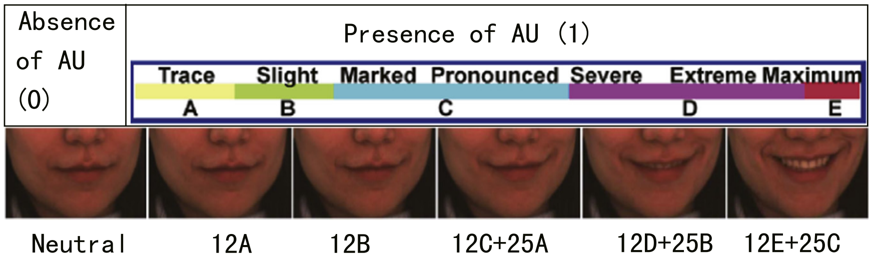

Fig. 1. Relation between the scale of evidence and intensity scores for facial action units [3].

Majority of the existing literature has been focusing on two types of facial expression studies. The first category is concerned with analyzing and classifying prototypic facial expressions (aka as six basic expressions: happy, sad, disgust, anger, surprise, fear). These studies are mostly designed to recognize the basic expressions that can represent the human emotions. These expressions of emotions are known to be similar among different cultures [39]. The second category of facial behavior analysis specifies expressions by a set of AUs where the goal is to represent and recognize facial AUs defined by FACS. The latter approach can comprehensively describe a wider range of facial expressions. AU-based analyzers are also capable for representing the prototypic facial expressions as a combination of AUs. For instance 'fear' can be represented by combination of AU1, 2, 4, 5, and 25 [3].

Most of the existing studies have been focused on prototypic facial expressions and detecting the occurrence of AUs in posed facial expressions. In many real world application we need to analyze spontaneous facial expression, such as categorizing pain related facial expressions [7] and measuring the engagement of students for online tutoring applications. Spontaneous facial behavior analysis can be very challenging due to several factors, such as out-of-plane head motion and different poses, subtle facial expressions and intra-subject variability in dynamics and timing of different facial actions. In addition it has been shown that the dynamics and patterns of spontaneous facial expressions can be very different from the posed ones.

Analyzing spontaneous facial expressions, especially for intensity measurement, is not as robust and accurate as the posed one, because of the aforementioned challenging factors. In early works for automatic AU intensity measurement, Bartlett et al. [13] measured the intensity of AUs in posed and spontaneous facial expressions by using Gabor wavelet and support vector machines. The mean correlation with human-coded intensity of their automated face recognition system for posed and spontaneous facial behavior is 0.63 and 0.3, respectively. These quantitative results demonstrate that recognizing spontaneous expressions is more challenging than posed expressions.

In the area of spontaneous facial actions recognition, there are very few works on detecting or measuring the intensity of spontaneous facial actions [5,15]. To the best of the authors' knowledge, most of the current studies, including [5,15,14], analyze spontaneous facial actions statically and individually. In other words, the dependencies among multilevel AU intensities as well as the temporal information are ignored. The semantic and dynamic relationships among facial actions are crucial for understanding and analyzing spontaneous expression. In fact, the coordination and synchronized spatiotemporal interactions between facial actions produce a meaningful facial expression. Tong et al. [25] employed dynamic Bayesian Network (DBN) to model the dependencies among AUs and achieved improvement over single image-driven methods, especially for recognizing AUs that are difficult to detect but have strong relationships with other AUs. However, their work [25] focuses on AU detection of posed expression.

Following the idea in [25], in this paper, we introduce a framework based on DBN to systematically model the spatiotemporal dependencies among multi-AU intensity levels in multiple frames, in order to measure the intensity of spontaneous facial actions. The proposed probabilistic framework is capable of recognizing multi-level AU intensities in spontaneous facial expressions. Denver Intensity of Spontaneous Facial Action (DISFA) database [9] is employed in this study. DISFA database is publicly available for analyzing the AU intensities and their dynamics. For every frame in DISFA, the intensity of every AU within the scale of 0 (absence of an AU) to 5 (maximum intensity) has been provided. To demonstrate the effectiveness of the proposed model, rigorous experiments are performed on DISFA database. The experimental results as well as the detailed analysis on the improvements are reported in this paper.

2. Related works

Given the significant role of faces in human's emotional and social life, automating the analysis of facial expression has gained great attention in both academia and industry. An automated facial expression recognition system usually consists of two key stages: feature extraction and machine learning algorithm design for classification. Commonly used features that represent facial gestures or facial movements include optical flow [35,41], explicit feature measurement (e.g., length of wrinkles and degree of eye opening) [42], Haar features [43], Local Binary Patterns (LBP) features [44,45], independent component analysis (ICA) [46], feature points [36], 3D geometries features [54,55], and Gabor wavelets [4]. Given the extracted facial features, facial actions are identified using machine learning methods such as Neural Networks [42], Support Vector Machines (SVM) [5], rule-based approaches [47], AdaBoost classifiers, and Sparse Representation (SR) classifiers [48].

The recent attempt of facial action intensity estimation has received increasing interest due to the semantic richness of the predictions. Bartlett et al. [49] estimated the intensity of action units by using the distance of a test example to the SVM separating hyperplane, while Hamm et al. [50] used the confidence of the decision obtained from AdaBoost. Multi-class classifier is a natural choice for this problem. For example, Mahoor et al. [5] utilized AAM features in conjunction with six one-vs-all SVM classifiers to automatically measure the intensity of AU6 and AU12 in videos captured from infant–mother communications. More recently, Mavadati et al. presented the DISFA database in [9], which is fully coded by six level AU intensity and employed Local Binary Pattern histogram (LBPH) features, HOG features, and Gabor features followed by multi-class SVM classification for AU intensity measurement. Besides multi-class classifiers, regressor is also commonly used for the task of continuous expression estimation. For instance, Jeni et al. [51,56] and Savran et al. [52] applied Support Vector Regression (SVR) for prediction, while Kaltwang et al. [53] used Relevance Vector Regression (RVR) instead. Both methods, SVR and RVR, are extensions to regression of SVM, although RVR yields a probabilistic output.

Most of the current works recognize facial actions individually and statically. However, due to the richness, ambiguity, and