Pattern Recognition 61 (2017) 78–91

Contents lists available at ScienceDirect

Pattern Recognition

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

Expression-assisted facial action unit recognition under incomplete AU annotation

Shangfei Wang [a,*], Quan Gan [a], Qiang Ji [b]

[a] Key Lab of Computing and Communication Software of Anhui Province, School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, PR China

[b] Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180, USA

ARTICLE INFO

Article history:
Received 15 March 2016
Received in revised form
9 July 2016
Accepted 18 July 2016
Available online 21 July 2016

Keywords:
AU recognition
Incomplete annotation
Bayesian network
Expression

ABSTRACT

Facial action unit (AU) recognition is an important task for facial expression analysis. Traditional AU recognition methods typically include a supervised training, where the AU annotated training images are needed. AU annotation is a time consuming, expensive, and error prone process. While AU is hard to annotate, facial expression is relatively easy to label. To take advantage of this, we introduce a new learning method that trains an AU classifier using images with incomplete AU annotation but with complete expression labels. The goal is to use expression labels as hidden knowledge to complement the missing AU labels. Towards this goal, we propose to construct a Bayesian network (BN) to capture the relationships among facial expressions and AUs. Structural expectation maximization (SEM) is used to learn the structure and parameters of the BN when the AU labels are missing. Given the learned BNs and measurements of AUs and expression, we can then perform AU recognition within the BN through a probabilistic inference. Experimental results on the CK+, ISL and BP4D-Spontaneous databases demonstrate the effectiveness of our method for both AU classification and AU intensity estimation.

1. Introduction

Facial expression recognition has attracted increasing attention due to its wide applications in human–computer interaction [1]. There are two kinds of descriptors of expressions: expression category and AUs [2]. The former describes facial behavior globally, and the latter represents facial muscle actions locally. To recognize AUs and expressions, a large number of annotated training images are required. In general, AU annotation is more expensive and harder than expression annotation, since expression is global and easier to recognize, while AUs are local and subtle, and harder to recognize. Furthermore, the number of AUs for an image is usually larger than that of expressions. Therefore, the AUs should be labeled by qualified facial action coding system (FACS) experts. Current research [3–6] reveals that some AUs are obvious and easy to be annotated, while others are subtle and hard to annotate. This phenomenon not only increases the difficulty of AU annotation, but also makes the AU labels error prone. Compared with AU annotation, expressions are much easier to annotate, and can be labeled with great accuracy.

In this work, we try to design an AU recognition method with the assistance of expression labels under the incomplete AU labeling. Specifically, during training, instead of trying to label every AU in each image as being done by the existing AU recognition methods, we only label AUs that can be labeled confidently and leave those difficult and subtle AUs unlabeled. In addition, we provide expression label for each image. Using such annotated images, we then train an AU recognition algorithm by leveraging on the relationships among AUs and the knowledge of the expressions. To take advantage of the available expression labels during training, we propose to construct a BN to systematically capture the dependencies among AUs and expressions. The nodes of the BN represent the AUs and expressions. The links and their parameters capture the probabilistic relations among AUs and expressions. Since some AU labels are missing for some training images, structural expectation maximization (SEM) is adopted to learn the structure and parameters of the BN. Given the learned BN, we can infer the AUs by combining the AU-expression relationships encoded in the BN and the AU measurements. The experimental results on the CK+ database show that, with complete annotation, our method outperforms the state of the art model-based and image-driven AU classification methods; with incomplete annotation, our method performs much better than state of the art AU classification methods. The experimental results on the ISL database demonstrate the cross-database generalization

* Corresponding author.
E-mail addresses: sfwang@ustc.edu.cn (S. Wang), gqquan@mail.ustc.edu.cn (Q. Gan), qji@ecse.rpi.edu (Q. Ji).

http://dx.doi.org/10.1016/j.patcog.2016.07.028
0031-3203/

ability of our method for AU classification. Furthermore, the experimental results on the BP4D-Spontaneous databases demonstrate that for AU intensity estimation, our method outperforms current model-based and image-driven AU intensity estimation methods under both complete and incomplete annotation.

2. Related work

Usually, several AUs can be present at the same image or image sequence. Thus, AU recognition can be formulated as a multi-label classification problem. Due to the large number of possible label sets, multi-label classification is rather challenging. Successfully exploiting the dependencies inherent in multiple labels is the key to facilitate the learning process. Accounting for dependencies among AUs, present AU recognition research can be divided into three groups.

The first group recognizes each AU individually and directly from images or sequences [7,8]. They are referred to as image-based AU recognition methods. Valstar and Pantic [7] proposed an automatic method to detect 22 AUs. They first detected and tracked 20 facial points, and then used a combination of GentleBoost, support vector machines (SVMs), and hidden Markov models as a classifier. van der Maaten and Hendriks [8] adopted AAM features and linear chain conditional random field to detect the presence of AUs. These works treat recognition of each AU individually as one-vs.-all scheme, ignoring the dependencies among AUs. However, multiple AUs can appear together and thus there exist dependencies among them. The AU relationships may help AU recognition.

The second group recognizes fixed AU combinations. One approach regards the AU combination as a new AU. For example, Littlewort et al. [9] analyzed the AU combinations of 1+2, 2+4, 1+4 and 1+2+4 using a linear SVM with Gabor features. Lucey et al. [10] used SVM and nearest neighbor to detect a few combinations of AUs (i.e. 1, 1+2, 4, 5) with active appearance model (AAM) features. The other approach integrates AU relations existing in AU labels into AU classifiers. Zhang and Mahoor [11] first proposed a hierarchical model to group multiple AUs into several fixed groups based on AU co-occurrences existing in AU labels and facial regions. Then, each AU recognition is regarded as a task, and AUs in the same groups share the same kernel. A multi-task multiple kernel learning is used to learn AU classifiers simultaneously. Zhao et al. [12] selected a sparse subset of facial patches and learned multiple AU classifiers simultaneously under the constraints of group sparsity and local AU relations (i.e. positive correlation and negative competition). Although the local dependencies among AUs have been exploited in these works, the combinations are manually determined and fixed. Thus, it is only feasible for a few combinations and is hard to detect thousands of possible combinations.

The third group explicitly exploits the co-existent and mutually exclusive relations among AUs from target labels. They are referred to as model-based AU recognition methods. Tong et al. [13,14] used Gabor features and SVM to recognize each AU first, then they model the relations among AU labels by dynamic Bayesian network (DBN). Eleftheriadis et al. [15] proposed a multi-conditional latent variable model to combine global label dependencies into latent space and classifier learning. The image features are projected onto the latent space, which is regularized by constraints, encoding local and global co-occurrence dependencies among AU labels. Then, multiple AU classifiers are learned simultaneously on the manifold. Both work assumes complete AU labeling and only involves AUs without using expressions. Wang et al. [16] proposed a hierarchical model to integrate the low-level image measurements with the high-level AU semantical relationships for AU recognition. A restricted Boltzmann machine (RBM) is used to capture higher-order AU interactions, and a 3-way RBM is further developed to capture related factors such as the facial expressions to achieve better characterization of the AU relations. Although their model can capture high order AU relationships, it cannot effectively handle missing labels. Therefore, all the model-based AU recognition methods require completely annotated images.

Due to the difficulties of collecting data with AU intensity values and the limited available database, current AU analyses mainly focus on AU occurrence detection, and few work measures the intensity of AUs. Furthermore, among the very few existing AU intensity estimation work, most work, such as [17–23], measures the intensity of each AU independently. They do not make use of the intensity dependencies that are crucial for analyzing AUs. In this paper, we refer to these methods as image-driven intensity estimation methods. It is only recently that two works have considered AU relations for AU intensity estimation. Li et al. [24] proposed using DBN to model AU relationships for measuring their intensities. In order to estimate the intensity of AUs present in a region of the upper face, Sandbach et al. [25] adopted Markov random field structures to model AU combination priors. Similar to Tong et al.'s [13,14] work, both works assume complete AU intensity labeling and only involve AUs without using expressions. They are referred to as model-based AU intensity estimation methods. Therefore, all the model-based AU intensity estimation methods [24,25] require completely AU-annotated images without using expressions.

To the best of our knowledge, there is little reported work that recognizes AUs or estimates AU intensities with the assistance of expressions [16,26], although there exist a few works considering the relations among expressions and AUs to help expression recognition or to jointly recognize AUs and expressions [27]. For example, Pantic and Rothkrantz [28] summarized the production rules of expressions from AUs using the AUs-coded descriptions of the six basic emotional expressions given by Ekman and Friesen [2]. Velusamy et al. regarded AU to expression mapping as a problem of approximate string matching, and they adopted a learned statistical relationship among AUs and expressions to build template strings of AUs for six basic expressions [5]. Zhang and Ji proposed to use DBNs to model the probabilistic relations of facial expressions to the complex combination of facial AUs and temporal behaviors of facial expressions [6]. Li et al. [27] introduced a dynamic model to capture the relationships among AUs, expressions and facial feature points. The model was used to perform AU and expression recognition as well as facial feature tracking.

Current AU recognition and AU intensity estimation methods require complete AU label assignments. However, AU analyses with incomplete AU label assignments are frequently encountered in realistic scenarios, due to the large number of AUs and difficulty in manual AU annotation. Till now, little research has addressed the challenge of AU analyses with incomplete AU labels. While AUs are hard to annotate, facial expression is relatively easy to label. The available expression labels during training and the dependencies among expression and AUs may be useful for AU recognition and AU intensity estimation. Thus, in the paper, we construct AU classifiers with the ability of learning and inferring from incomplete AU annotations with the help of ground truth of expression knowledge that is available during training only.

Compared with related work, the main contribution of this work lies in the introduction of a probabilistic framework to use the ground truth of expression labels available during training to help train an improved AU classifier under incomplete AU annotation. In addition, we formulate AU detection as a multi-label classification problem.