Pattern Recognition 51 (2016) 187–196

Contents lists available at ScienceDirect

Pattern Recognition

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

Task-dependent multi-task multiple kernel learning for facial action unit detection

Xiao Zhang, Mohammad H. Mahoor *

University of Denver, 2390 S York Street, CMK 308, Denver, CO 80210, USA

ARTICLE INFO

Article history:
Received 27 January 2015
Received in revised form
26 August 2015
Accepted 29 August 2015
Available online 25 September 2015

Keywords:
Facial action unit detection
Multi-task multiple kernel learning
Support vector machines

ABSTRACT

Facial action unit (AU) detection from images and videos is a challenging research topic and has attracted great attention in the past few years. This paper presents a novel method, task-dependent multi-task multiple kernel learning (TD-MTMKL), to jointly detect the absence and presence of multiple AUs. TD-MTMKL attempts to learn an optimal kernel combination from a given set of basis kernels for each involved AU and obtain a finer depiction of AU relations through kernel combination weights. In other words, AU detection is solved as a multi-task multiple kernel learning problem, where AU relations are encoded via their SVM discriminative hyperplanes and kernel combination weights. The kernel learning increases the discriminant power of the classifier by fusing different types of facial feature representations with multiple kernels. Specifically, based on the TD-MTMKL method proposed in this paper, co-occurrence AUs share the same kernel weights while AUs with weak co-occurrence relations may employ distinct sets of kernels. Such "task-dependent" kernel learning framework seeks a trade-off between capturing commonalities and adapting to variations in modeling AU relations. Our experiments on the CK+ and DISFA databases show that our method achieved encouraging detection results of both post and spontaneous AUs compared to the state-of-the-art methods.

1. Introduction

In the past few decades, designing automated facial action unit detection systems has become an attractive and challenging research topic in the field of pattern recognition, computer vision and robotics. The well-known Facial Action Coding System (FACS) [1,2] defines 44 AUs to capture all visually detectable muscle movements in human faces, and also describes the intrinsic relationships among AUs such as simultaneous presence (co-occurrence). For examples, AU4 (brow lowerer) is usually co-occurred with AU1 (inner brow raiser) and AU2 (outer brow raiser) to generate negative expressions such as "fear" and "sadness". Whereas AU6 (cheek raiser) is usually co-occurred with AU12 (lip corner puller) in the case of Duchenne smile [3]. These AU inter-relationships are by their nature good resources for achieving an accurate understanding of human facial activities. However, almost all the existing AU detection approaches in the literature, based on either static [4–6] or dynamic modeling techniques [7–10], were proposed to recognize AUs or certain AU combinations separately without considering their inter-relations.

In this paper, we propose to employ the idea of multi-task learning (MTL) to simultaneously detect a set of AUs by exploiting their inter-relations. In this framework, the detection of each AU is considered as a task. Our view is upon the fact that there exist commonalities among the classification tasks for multiple AUs. One instance of these commonalities as shown in Fig. 1 can be that the same set of training data is usually shared and commonly used to learn the SVM hyperplanes for detecting different AUs. Another instance can be that there exists a main task among multiple AU detection tasks, which is to distinguish between the neutral faces and the occurrences of AUs. However, in humans' social interactions, similar emotions can be exhibited differently by subjects either via a single AU or a combination of AUs. Even an individual may show various combinations of AUs for demonstrating the same emotions, such as the difference between Duchenne and polite smiles. These imply that the AU relations defined in FACS are not always fixed, or at least the degree of the relations among AUs are not uniform. Thus, when detecting a set of AUs that is usually co-occurred in specific emotions, it is essential for the system to not only determine the commonalities across multiple AU detection tasks but also adapt to the task differences, or say, diversities.

This paper presents a novel task-dependent multi-task multiple kernel learning (TD-MTMKL) that learns an optimal kernel combination from a given set of basis kernels for each involved task and obtains a finer depiction of task relations through kernel

* Corresponding author. Tel.: +1 303 871 3745; fax: +1 303 871 2194.
E-mail address: mmahoor@du.edu (M.H. Mahoor).

http://dx.doi.org/10.1016/j.patcog.2015.08.026
0031-3203/

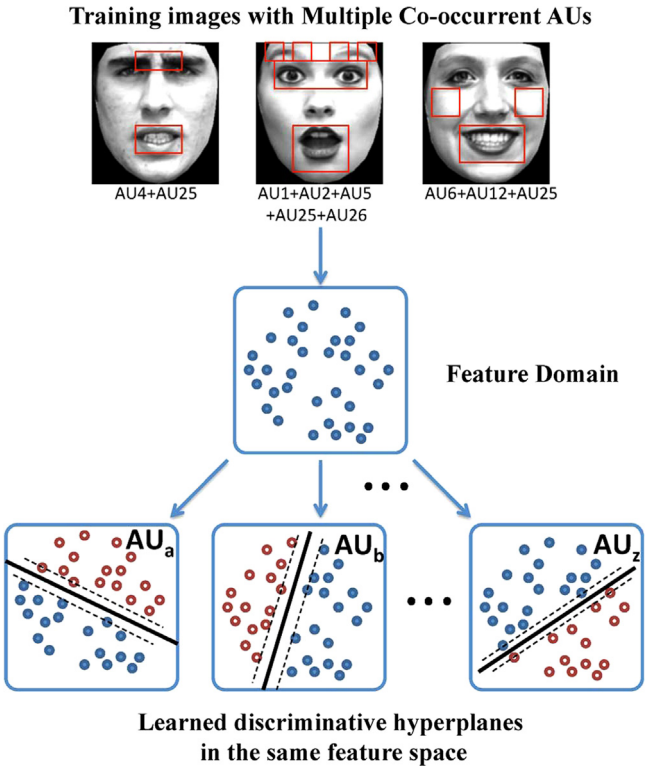

Fig. 1. SVM classifier training in common AU detection systems.

combination weights. In our TD-MTMKL, highly related tasks share the same kernel weights while non-related tasks may employ distinct sets of kernels. This "task-dependent" characteristic seeks to capture the AU commonalities through MTL meanwhile adapt to AU variations via kernel learning. By doing this, our proposed method can incorporate the benefits of both MTL and MKL, and identify local distributions in the training data from all AU detection tasks. The experimental results on the extended Cohn–Kanade (CK+) database [11] and the Denver Intensity of Spontaneous Facial Action (DISFA) database [12] confirm that our proposed TD-MTMKL approach is a powerful technique for automated AU detection.

The rest part of this paper is divided as follows. Section 2 reviews the previous work on facial expression analysis and multi-task learning, and also describes their differences from our method. Section 3 formulates our proposed TD-MTMKL algorithm, and presents the modeling of AU inter-relations. Section 4 discusses the experimental results on public face databases, and also shows the comparison with the state-of-the-art AU detection methods. Finally, Section 5 concludes the paper and presents our future work.

2. Related work

In the past few decades, good progresses have been made in automatic facial expression analysis. For a comprehensive study on the state-of-the-art methods, we refer our readers to these survey papers [13,14]. Recently, the Facial Expression Recognition and Analysis Challenge (FERA 2011) [15] outlined the face database used and the benchmark framework as well as the performance evaluation protocol for automated recognition of facial AU and basic expressions. In this section, we review several works on two major aspects of this challenge including facial feature representation and AU detector design. We also review the related work to our proposed TD-MTMKL algorithm from the aspects of methodologies and the applications in AU detection.

2.1. Facial feature representation

In automatic facial expression analysis systems, facial images or video frames are registered and then represented by a set of features. Usually, facial landmarks are annotated and utilized for image registration such as in [16,17]. There are three categories of features that are commonly used for facial image representation including geometric features, appearance features, and a combination of them.

Geometric features model the geometry of human faces by extracting the locations and shapes of facial components. Pantic et al. [4,18] tracked a set of facial landmarks around eyebrows, eyes, nose, mouth, and chin as the characteristic points to capture the geometric information of these facial organs. In [19], Chang et al. trained an active shape model (ASM) [20] for feature representation. 58 fiducial points were utilized to avoid incorrect matching caused by non-linear image variations. The famous active appearance models (AAM) [21] extended ASM by jointly extracting the geometric and the appearance information from human faces, and are widely used for facial expression recognition [22,23]. However, both ASM and AAM rely on accurate matching results of their geometric face models, where the model training phases are usually time consuming and sometimes need manual works.

Appearance features represent textures of facial images including wrinkles, bulges and furrows in expressive activities. Gabor wavelet analysis [24,25] is one of the first features applied for describing the variations of facial appearance in either the entire face or specific face regions. However, due to the high computation and memory complexity in extracting Gabor-wavelet features, local binary pattern (LBP) operator [26] was introduced as an effective appearance feature for facial image analysis [27,28]. The tolerance against illumination variations across images and the computational simplicity are the two major advantages of LBP features. The authors of [16] comprehensively studied the effect of applying LBP features for facial expression recognition. In their work, better experimental results were obtained compared to Gabor features. In [29], the authors proposed a temporal extension of the canonical LBP feature and named it LBP-TOP. This feature can represent the dynamic appearance information between consecutive video frames. Histogram of oriented gradients (HOG) is another famous appearance feature that is firstly proposed in [30] for pedestrian detection. The HOG feature counts the occurrences of gradient orientations in localized portions of an image, and is designed to keep invariant to geometric transformations. Recent works [31,32] applied and revised HOG features to capture appearance information and shape orientations of facial images for expression analysis.

Our previous works [33–35] compare the performance of utilizing one type of features over multiple features for recognizing basic facial expressions and action units, and show that a single feature is insufficient to provide a robust facial image representation. As comprehensively stated and proven in [35], one type of facial features (e.g. HOG, LBPH, Gabor) may not be distinguishable for representing all expressions whereas using another type of features may produce better results in some other expressions. Facial expressions and actions are depicted by wrinkles, edges, and bulges (both appearance/texture and shape variations). Usually HoG operators are more sensitive to edges and shape variations while LBPH filters are more suitable for representing texture and appearance patterns. Also, various features have different distributions, and hence it is necessary to take advantage of multiple facial features to increase the discriminative