Computer Vision and Image Understanding 147 (2016) 95–101

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Learning prototypes and similes on Grassmann manifold for spontaneous expression recognition

Mengyi Liu, Ruiping Wang, Shiguang Shan*, Xilin Chen

Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China

ARTICLE INFO

Article history:
Received 28 March 2015
Accepted 17 August 2015

Keywords:
Expression prototype
Simile representation
Grassmann manifold
Spontaneous expression recognition

ABSTRACT

Video-based spontaneous expression recognition is a challenging task due to the large inter-personal variations of both the expressing manners and the executing rates for the same expression category. One of the key is to explore robust representation method which can effectively capture the facial variations as well as alleviate the influence of personalities. In this paper, we propose to learn a kind of typical patterns that can be commonly shared by different subjects when performing expressions, namely "prototypes". Specifically, we first apply a statistical model (i.e. linear subspace) on facial regions to generate the specific expression patterns for each video. Then a clustering algorithm is employed on all these expression patterns and the cluster means are regarded as the "prototypes". Accordingly, we further design "simile" features to measure the similarities of personal specific patterns to our learned "prototypes". Both techniques are conducted on Grassmann manifold, which can enrich the feature encoding manners and better reveal the data structure by introducing intrinsic geodesics. Extensive experiments are conducted on both posed and spontaneous expression databases. All results show that our method outperforms the state-of-the-art and also possesses good transferable ability under cross-database scenario.

1. Introduction

In recent years, facial expression recognition has become a popular research field due to its wide applications in many areas such as biometrics, psychological analysis, human-computer interaction, and so on. In the early stage, many works have been done to classify human posed expressions in static images [1]. However, as facial expression can be viewed as a sequentially dynamic process, it is natural and proved to be more effective to be recognized from video clips [2–5]. For spontaneous expression recognition in video, one of the main challenges is the large inter-personal variations of expressing manners and executing rates for the same expression category. The key issue to cope with the challenge is to develop a more robust representation for facial expression, which can better capture the subtle facial variations as well as alleviate the influence of personalities in performing expression.

According to the theory from physiology and psychology, facial expressions are the outcome of facial muscle motions over various time intervals. When captured by cameras, an observed expression can be decomposed into a set of local appearance variations produced by the motions occurring in different facial regions. In spite of the large inter-personal variations, there still exist some typical motion patterns, that can be commonly shared by different subjects in performing expressions. The similar idea is also reflected in a pioneering work Facial Action Coding System (FACS) [6], where a number of Action Units (AU) are manually defined to describe some emotion-related facial actions aroused by muscle motions. Then each expression is represented by the existence of these AUs in a binary coding manner.

In light of such theory, we propose to explore a batch of commonly shared typical patterns, i.e. "prototypes", using data-driven approach, and then design a prototype-based encoding manner to generate the feature representation for each sample. An schema of our basic idea is illustrated in Fig. 1. Specifically, we first apply a statistical model (i.e. linear subspace) on facial regions to model the local variations of local patterns, which can generate the specific expression patterns for each video sample. Then a clustering algorithm is employed on all these expression patterns, and each cluster mean can be regarded as a "prototype", which integrates the common properties of the samples assigned to this cluster. Note that, all of the original patterns and the learned "prototypes" are represented as linear (orthogonal) subspaces lying on Grassmann manifold, thus intrinsic geodesic distance [7] and Karcher means [8] are employed in this procedure for accurate estimation. To obtain the unified prototype-based representation, we further design "simile" features to measure the similarities of personal specific patterns to our learned "prototypes" on Grassmann manifold. The idea is derived from [9] for face verification, which

[*] Corresponding author. fax: +86 10 6260 0548.
E-mail address: sgshan@ict.ac.cn (S. Shan).

http://dx.doi.org/10.1016/j.cviu.2015.08.006
1077-3142/© 2015 Elsevier Inc. All rights reserved.

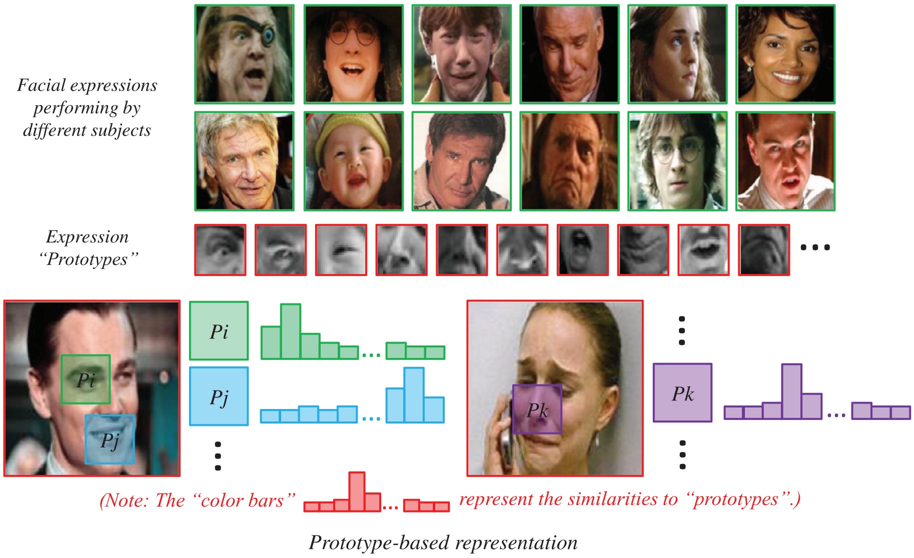
Facial expressions performing by different subjects
Expression "Prototypes"
...
Pi
Pj
Pk
(Note: The "color bars" represent the similarities to "prototypes".)
Prototype-based representation

Fig. 1. An schema of our basic idea (best viewed in color).

assumed that an unseen face can be described as having a mouth that *looks like* A's and a nose that *looks like* B's, where A and B are individuals in the reference set. In our method, the static facial attributes are replaced by some dynamic variation manners of facial regions when performing expressions. However, different from [9] which introduced an auxiliary reference set, we measure the similarities referring to the "prototypes" directly explored from the data, thus brings favorable robustness against the bias in the construction of reference set.

The main contributions of this paper are summarized as follows: (1) We propose a novel approach for modeling expression patterns and learning "prototypes" using statistical model on Grassmann manifold; (2) "Similes" are designed to explore the relations among common prototypes and specific patterns, which provides a new viewpoint to analyze the generality and specificity in the manner of human performing spontaneous expressions. (3) Comprehensive experiments are conducted with different parameter settings. The transferable ability of prototypes and similes are further discussed in cross-database test.

The rest of this paper is structured as follows. Section 2 reviews several most related work for video-based facial expression recognition. Section 3 introduces the essential components of our proposed method, including facial expression patterns, prototypes learning, and simile representation. In Section 4, we provide comprehensive evaluations on the whole framework as well as discussing the important parameters. Finally, we conclude the work and discuss possible future efforts in Section 5.

2. Related works

For video-based facial expression recognition, there is always strong interest in modeling the temporal dynamics among video frames. The mainstream approaches of dynamic representation are based on local spatial-temporal descriptors. For example, Yang et al. [3] designed Dynamic Binary Patterns (DBP) mapping based on Haar-like features. Zhao et al. [2] proposed LBP-TOP to extract the spatial-temporal information from three orthogonal planes (i.e. X–Y, X–T, Y–T) in image volumes. Hayat et al. [10] conducted a comprehensive evaluation based on various descriptors, e.g. HOG3D [11], HOG/HOF [12], 3D SIFT [13], using bag of features framework for facial expression recognition. All these hand-crafted methods possess favorable computational efficiency and generalization ability due to the independency of data.

Another line of methods attempt to explore the specific characteristics in expression evolution using dynamic graphic models. For instance, Shang et al. [14] employed a non-parametric discriminant Hidden Markov Model (HMM) on tracked facial features to for dynamic expressions modeling. Jain et al. [15] proposed to model the temporal variations within facial shapes using Latent-Dynamic Conditional Random Fields (LDCRFs), which can obtain the entire video prediction and continuously frame labels simultaneously. Wang et al. [4] proposed Interval Temporal Bayesian Networks (ITBN) to represent the spatial dependencies among primary facial actions as well as the variety of time-constrained relations, which characterize the complex activities both spatially and temporally. Although these schemes can better reveal the intrinsic principles of facial expressions, the optimization requires lots of domain knowledge and large computational cost.

More recently, statistical models were employed to encode the appearance variations occurring in dynamic facial expressions, which proved to be more effective when dealing with real-world data [16,17]. In [16], linear subspace were applied on the feature set of successive image frames to model the facial feature variations during the temporal evolution of expression. For more robust modeling, [17] integrated three different types of statistics into the framework, i.e. linear subspace, covariance matrix, gaussian distribution, to model the feature variations from different perspectives. As these statistical models all reside on Riemannian manifold, intrinsic geodesics or extrinsic kernel methods were exploited to perform representation and classification on Riemannian manifold instead of traditional Euclidean space. A common property of these two methods is to model the facial variation globally, which could be easily affected by misalignment or partial occlusion. Since localization has been proved to be more natural and effective when processing face-related application, in this paper, we focus on modeling local variation based on statistical models and extracting generic local dynamic patterns to unify the description of facial expressions performed by different subjects. We believe that this scheme can make balance between the generalization ability by considering repeatable local features and the modality specificity by learning directly from data.

3. The proposed method

In this section, we introduce the proposed method in three stages: facial expression patterns generation, prototypes learning, and simile representation.