Pattern Recognition ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Pattern Recognition

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

Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos

Seung Ho Lee, Wissam J. Baddar, Yong Man Ro *

Image and Video Systems Lab, Dept. of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Yuseong-Gu, Daejeon 305-701, Republic of Korea

ARTICLE INFO

Article history:
Received 19 May 2015
Received in revised form
14 December 2015
Accepted 25 December 2015

Keywords:
Facial expression recognition (FER)
Collaborative expression representation
Subject-independent FER
Peak expression face
Intra class variation (ICV) face

ABSTRACT

This paper proposes a facial expression recognition (FER) method in videos. The proposed method automatically selects the peak expression face from a video sequence using closeness of the face to the neutral expression. The severely non-frontal faces and poorly aligned faces are discarded in advance to eliminate their negative effects on the peak expression face selection and FER. To reduce the effect of the facial identity in the feature extraction, we compute difference information between the peak expression face and its intra class variation (ICV) face. An ICV face is generated by combining the training faces of an expression class and looks similar to the peak expression face in identity. Because the difference information is defined as the distances of locally pooled texture features between the two faces, the feature extraction is robust to face rotation and mis-alignment. Results show that the proposed method is practical with videos containing spontaneous facial expressions and pose variations.

1. Introduction

Thanks to the increasing availability of computers and powerful electronic devices, the human computing needs to develop more human-centered user interfaces which respond quickly to naturally occurring human communication [1]. An important functionality of such interfaces is to understand emotions represented by facial expressions [1]. Facial expressions are the most natural and effective tools which allow humans to communicate their emotions, to express their intentions, and to interact with each other [2]. The aforementioned reasons have been emphasizing the importance of automatic facial expression recognition (FER) and justifying the great interest this research topic has attracted in the past few years [2].

Several research efforts have been made regarding automatic FER. In general, existing FER methods can be classified into geometric-based methods and appearance-based methods [3,4]. In the beginning, most of the FER methods were based on facial landmark localization and face geometry (which are called geometric-based methods). These methods use the characteristics of a face such as the shape and locations of facial components (including mouth, eyes, eyebrows and nose), the distances between pairs of facial landmark points or the velocities of particular facial landmark points [3,5]. It has been reported that geometric features can provide sufficient information to achieve accurate FER [6]. However, one crucial limitation these methods is that they suffer from mis-alignment of face due to inaccurate detection or tracking of facial landmark points under challenging image conditions (e.g., occlusion, low-resolution image, illumination change etc.) [7,8,11]. Mis-alignment of face unavoidably leads to degradation of the feature extraction for FER. Another limitation is that many of them require a neutral face of the corresponding subject for the normalization of a query face (i.e., eliminating the effect of facial identity [9–13]), or for the initialization of a facial landmark tracker (e.g., [35]). However, a neutral face of a subject is not always available in real applications [4,32]. The authors in [14] proposed neutral-independent geometric features for FER without using neutral face. In this method, the locations of eight facial landmark points, and the six distances between the facial landmark points were used as features for FER [14]. However, such features (e.g., distances between facial feature points) can also be used for face recognition (i.e., identifying subject) [15], which means that geometric features may be different across subjects and not fully subject-independent. Indeed, as is seen in Table 7, this method shows a relatively poor FER performance under subject-independent recognition.

* Corresponding author. Tel.: +84 42 350 3494; fax: +82 42 350 7619.
E-mail addresses: leesh09@kaist.ac.kr (S.H. Lee), wisam.baddar@kaist.ac.kr (W.J. Baddar), ymro@kaist.ac.kr (Y.M. Ro).

http://dx.doi.org/10.1016/j.patcog.2015.12.016
0031-3203/

Appearance-based methods aim to capture the changes in face textures such as those created by wrinkles and bulges [3,5]. These methods apply image filters (such as Gabor wavelets [16]) to the whole face or, specific face regions, or local patches around some facial components [5]. Most appearance-based FER methods that rely on static appearance in still images were investigated either using still image dataset (e.g., [17]) or manually selected peak expression faces from video sequences (e.g., [7,18,19,32,43–45]). To exploit the temporal dynamic information present in a video sequence, there have been some methods which have used a spatio-temporal appearance descriptor such as Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [20,21], Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) [5,21], Spatio-Temporal Local Monogenic Binary Pattern (STLMBP) [34] (that exploits the local phase and the local orientation of an image in addition to the local magnitude of LBP) and so on. To exploit both spatial and temporal discriminative information, a texture operation is independently applied to each of the three orthogonal planes (XY plane: spatial appearance, XT and YT planes: appearances of facial expression changes over time [20]) in a video volume. Note that the video volume can be of the whole face or of a local region of the face. However, there exist two main limitations of the spatio-temporal appearance descriptors. First, temporal motion appearance in XT or YT plane could negatively affect the FER when faces within a sequence are temporally not consecutive. Specifically, if a number of faces in the sequence are not detected by a face detector or tracker, some parts of the appearance associated with the facial expression change may be lost. In this case, the discriminative capability of the spatio-temporal descriptor will be degraded due to the loss of the information. Second, the appearance of the XY plane could be dependent on the facial identity which is not desirable for subject-independent FER.

It has been known that combining geometric and appearance features is better than using only geometric or appearance feature for facial expression recognition (FER) [4]. This hybrid representation is able to incorporate local pixel variation pattern (related to face texture information) while exploiting face geometry information at a global level [56]. There have been an increasing number of FER methods which make use of hybrid feature extraction. To generate a hybrid feature, the FER methods in [12] and [56] simply concatenated the appearance features and geometric features. However, one could not ensure that every appearance or geometric feature is helpful for classification. To address this issue, feature selection techniques have been adopted for more discriminative hybrid features [57,58]. In [57] Adaboost [59] was used to select a set of discriminative geometric and appearance features for recognizing facial action units. In [58], the backward elimination method was used to select the discriminative features. The main idea was that, out of all features extracted, a feature to make the quadratic mutual information [60] between the remaining feature set and an emotion category maximized was discarded [58]. In [56] and [58], the geometric and appearance features were extracted directly from a face which contained both identity and expression information. Thus, this might not be optimal for subject-independent FER scenarios due to the confusion between the identity and expression. In [12] and [57], a neutral face was used to eliminate the effect of the facial identity during the appearance and geometric feature extractions. However, the main limitation of these methods was that a neutral face of a subject was not always available in real applications.

Instead of exploiting hand-crafted descriptors, some research efforts have been dedicated to learn semantic expression features using deep learning [45,65,66]. In [65], a deep architecture based FER method was proposed, which was inspired by facial action coding system (FACS). In this method, a convolutional neural network (CNN) was used to generate an over-complete representation of expression specific appearance variation [65]. For simulating specific AUs, groups of local patches (called AU-aware receptive fields (AURFs)) were selected [65]. Restricted Boltzmann machines [70] were used to extract high-level features of the AURFs, which were concatenated to construct the final hierarchical feature for FER [65]. Different from the methods [45,65] that relied on a still image (i.e., peak expression face), 3D CNN was applied to a face image sequence in [66] to exploit facial dynamics for FER. A 3D CNN was achieved by convolving the 3D kernels on the cube constructed by face images. However, these methods were evaluated using manual preprocessing (peak expression face selection in [45,65] or manual face alignment [65,66]). Thus, their feasibilities in fully automatic FER applications were not verified.

In this paper, we propose a new practical FER in video sequences aiming to overcome three following difficulties. First, we may have incomplete temporal dynamic (such as temporal discontinuity) in a video sequence due to face detection or tracking error. Second, normalization using a neutral face from the same subject is required for performing a subject-independent FER, which is not always satisfied. Third, subtle mis-alignment of a face image may occur due to the incorrect landmark detection [55]. Many facial expression recognition (FER) methods were evaluated using manually aligned faces [7,8,32,34,45]. However, the manual alignment of face is far from the realistic. In the context of a fully automatic FER method, faces are not always perfectly aligned [61]. Experimental results in [61] showed that FER performance could be severely degraded even in small degrees of perturbation on landmarks (e.g., 3% perturbation of the eye distance). The main contributions of the paper are threefold:

1) We propose a new feature extraction method (called collaborative expression representation (CER)) where the peak expression face of a video sequence and an artificially generated face collaboratively represent an expression related facial appearance. This artificial face image is called Intra class variation (ICV) face and aims to eliminate the intra class variation due to the facial identity appearance on a peak expression face image. An ICV face is generated by combining the training face images of an expression class and is characterized to be similar to the facial identity appearance of the peak expression face [32]. To reduce the appearance related to facial identity, the proposed CER computes the distances of the locally pooled texture features between the peak expression face and the ICV face. The CER does not need prior knowledge of the query subject's neutral state. In addition, due to the differences of locally pooled texture features, the representation is robust to subtle rotation and mis-alignment of the peak expression face.

2) We propose a method to select the peak expression face from a video sequence while discarding the degraded faces. Herein, degraded face refers to face image with low quality in terms of face alignment and/or frontal pose degree. In terms of face alignment, incorrectly scaled or rotated face images [55] due to incorrect landmark detection are regarded as degraded faces. In terms of frontal pose degree, face images with out-of-plane face rotations (rotations in yaw or pitch [52]) are regarded as degraded faces. Among the properly aligned and frontal faces, we select the most expressive (peak expression) face image.

3) We propose a sparsity based weighting scheme that aims to fuse the complementary CERs derived by using a given peak expression face and its multiple ICV face images. We make use of an assumption that a sparse solution with higher sparsity provides more discriminative information for classification [23,30]. Using this assumption, the effects of more discriminative CERs can be emphasized while the effects of less discriminative ones can be suppressed. The weighting scheme can be practically used because it is adaptive to the appearance of a given peak expression face image rather than pre-training using fixed training face images.

Please cite this article as: S.H. Lee, et al., Collaborative expression representation using peak expression and intra class variation face images for practical..., Pattern Recognition (2016), http://dx.doi.org/10.1016/j.patcog.2015.12.016