Computer Vision and Image Understanding 140 (2015) 83–92

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

An efficient multimodal 2D + 3D feature-based approach to automatic facial expression recognition ☆

Huibin Li [a], Huaxiong Ding [b], Di Huang [c,*], Yunhong Wang [c], Xi Zhao [d], Jean-Marie Morvan [e,f], Liming Chen [b]

[a] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China
[b] Ecole Centrale de Lyon, LIRIS UMR5205, Lyon, France
[c] State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, China
[d] School of Management, Xi'an Jiaotong University, Xi'an, China
[e] Université Lyon 1, Institut Camille Jordan, Lyon, France
[f] GMSV Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

ARTICLE INFO

Article history:
Received 14 June 2014
Accepted 9 July 2015
Available online 29 July 2015

Keywords:
Facial expression recognition
Local texture descriptor
Local shape descriptor
Multimodal fusion

ABSTRACT

We present a fully automatic multimodal 2D + 3D feature-based facial expression recognition approach and demonstrate its performance on the BU–3DFE database. Our approach combines multi-order gradient-based local texture and shape descriptors in order to achieve efficiency and robustness. First, a large set of fiducial facial landmarks of 2D face images along with their 3D face scans are localized using a novel algorithm namely incremental Parallel Cascade of Linear Regression (iPar–CLR). Then, a novel Histogram of Second Order Gradients (HSOG) based local image descriptor in conjunction with the widely used first-order gradient based SIFT descriptor are used to describe the local texture around each 2D landmark. Similarly, the local geometry around each 3D landmark is described by two novel local shape descriptors constructed using the first-order and the second-order surface differential geometry quantities, i.e., Histogram of mesh Gradients (meshHOG) and Histogram of mesh Shape index (curvature quantization, meshHOS). Finally, the Support Vector Machine (SVM) based recognition results of all 2D and 3D descriptors are fused at both feature-level and score-level to further improve the accuracy. Comprehensive experimental results demonstrate that there exist impressive complementary characteristics between the 2D and 3D descriptors. We use the BU–3DFE benchmark to compare our approach to the state-of-the-art ones. Our multimodal feature-based approach outperforms the others by achieving an average recognition accuracy of 86.32%. Moreover, a good generalization ability is shown on the Bosphorus database.

1. Introduction

Affect recognition aims to determine an individual's emotion by detecting and measuring the emotion related physiological (e.g., bodily symptoms), psychological (e.g., feelings) or behavioral (e.g., facial expression) characteristics [1,2]. As an easily detectable, collectible, and measurable emotion component, facial expression is ideal for affect recognition and for human-computer interaction related applications [3]. However, Facial Expression Recognition (FER) is a very challenging problem mainly because of the diversity and hybridity of human expressions among different subjects in different cultures, genders and contexts.

In the past decades, a large number of FER approaches have been proposed. They can be categorized from three perspectives, namely the data modality, expression granularity, and temporal dynamics. From the first perspective, they are classified into 1) 2D FER (which uses 2D gray or color face images), 2) 3D FER (which uses 3D range images, point clouds, or meshes of faces), and 3) multimodal 2D + 3D FER (which uses both 2D and 3D facial data). From the second perspective, they are divided into 1) six basic facial expression (i.e., anger, disgust, fear, happiness, sadness, and surprise) recognition, 2) facial Action Unit (AU, e.g., brow raiser, lip tightener, and mouth stretch) detection and recognition. From the third perspective, they are categorized into static (still images) and dynamic (image sequences) FER. In this paper, we focus on the problem of recognizing the six basic facial expressions using multimodal 2D + 3D static images.

☆ This paper has been recommended for acceptance by Nikos Paragios.

* Corresponding author.
E-mail addresses: huibinli@mail.xjtu.edu.cn (H. Li), huaxiong.ding@ec-lyon.fr (H. Ding), dhuang@buaa.edu.cn (D. Huang), yhwang@buaa.edu.cn (Y. Wang), zhaoxi1@mail.xjtu.edu.cn (X. Zhao), morvan@math.univ-lyon1.fr (J.-M. Morvan), liming.chen@ec-lyon.fr (L. Chen).

http://dx.doi.org/10.1016/j.cviu.2015.07.005
1077-3142/

Appearance-based 2D FER has been widely investigated since 1990s [3]. The main research topics lie in three aspects: face detection, expression related feature extraction and classification. Comprehensive surveys of 2D FER approaches are given in [3,4]. They are mainly classified into two categories, i.e., template-based and feature-based [4]. Template-based approaches usually fit a holistic face model to the input image or track it in the input image sequence. Active appearance model [5], point distribution model [6], mixture of probabilistic PCA [7], and topographic modeling [8] are some typical examples. Feature-based approaches generally localize the features of an analytic face model in the input image or track them in the input sequence. Gabor wavelets [9] and Local Binary Patterns (LBP) [10] based face representations are two popular representatives. Although considerable advancements have been achieved, 2D FER is still very challenging mainly due to its sensitivity to illumination, pose variations, and possible occlusions [3,4].

Recently, with the rapid development of 3D imaging and scanning technologies, it becomes more and more popular to capture 3D face scans. Comparing with 2D face images, 3D face scans contain precise geometric shape information of facial surfaces, which is robust to illumination and pose variations, but more sensitive to facial expression changes. Thus, shape-based 3D FER has attracted increasing attentions. Similar to 2D, 3D FER approaches can also be categorized into template-based and feature-based. Template-based approaches usually build a parametric deformable face model first, and then extract the model parameters as expression features for recognition. 3D morphable model [11], bilinear deformable model [12], shape deformation model [13], and statistical feature model [14] are some famous examples. The main drawback of template-based approaches lies in that they require to establish one-to-one correspondence between 3D face scans, which is still a very challenging issue. Meanwhile, time consuming procedures like dense 3D face registration and model fitting are indispensable. Feature-based approaches generally extract 3D expression cues around facial landmarks using different facial surface geometric or differential quantities. For example, the distances between 3D facial landmarks are widely used in [15–17], and [18]. Moreover, 3D facial curves [19], facial geometry images and normal maps [20,21] facial conformal images [22], facial surface normal [23,24] and curvatures [23–25], and local depth-SIFT features [26] are some popular expression features. Feature-based approaches generally perform better than template-based ones. However, the bottleneck of feature-based approaches lies in accurate and robust 3D facial landmark localization, which is still a very difficult task [27]. More detailed surveys of 3D facial expression recognition are given in [28,29].

Although the effectiveness of multimodal 2D + 3D face recognition has been well presented as in [30,31], the investigation of multimodal 2D + 3D FER is very limited. Wang et al. [25] compared the FER accuracy of 3D primitive surface feature distribution based approach with 2D Gabor-wavelet and Topographic Context based ones on the BU–3DFE database, and found that 3D shape based approach is superior to 2D ones, especially for non-frontal faces. However, the effectiveness of combing 3D and 2D approaches was not discussed. Zhao et al. [14] used both 2D features (RGB values and LBP) and 3D features (3D coordinates and shape index values) in the 3D statistical feature model for prototypical expression recognition. But the results using only 2D features or 3D features were not reported, and thus the complementarity between 2D and 3D features was also not studied. In [32], the authors used both 2D and 3D dynamic data for real-time facial action and expression recognition. More precisely, they first extended the active shape model to handle 3D data for facial feature tracking. Then, they extracted numerous geometric measurements (e.g., the distances between landmarks and the boundary shape of lips) and surface deformation measurements (e.g., image gradient and surface curvature descriptors). Finally, the Rule Classifier was used for recognizing a subset of 11 important AUs and 4 facial expressions (i.e., happy, sad, surprise, disgust) on a dataset consisting of 832 sequences of 52 participants. Their experimental results demonstrated that the proposed 2D+3D algorithm performed much better than the 2D appearance-based algorithm (i.e., 2D ASM + Gabor filters + LDA) for recognizing the four facial expressions. This is a very illuminating approach for 2D+3D multimodal FER. However, they did not report the performance of each modality under their own framework. The importance of each modality is still unclear. Savran et al. [33] utilized multimodal 2D + 3D face data for facial AU detection. They found that 3D data generally perform better than 2D data, especially for lower AUs. Moreover, the fusion of two modalities can improve the detection rates from 93.5% (2D) and 95.4% (3D) to 97.1% (2D+3D). Except for facial AU detection and expression recognition, Wang et al. [34] quantified facial expression abnormality in Schizophrenia by combining 2D and 3D features. Their experimental results demonstrated that the combined features better characterized facial expressions than either individual 3D geometric or 2D texture features.

The above studies have preliminarily proved the fact that the combination of 2D and 3D data is better than either of the single 2D or 3D modality for expression characterization and AU detection, but deep analysis of the superiority for multimodal 2D+3D FER is still missing. An advantage of using 2D data is that it can be used to accurately localize a large set of facial landmarks on 2D face images and further on their 3D face scans due to the 2D–3D correspondence, which is the first contribution of this paper. More precisely, we propose to explore the incremental Parallel Cascade of Linear Regression (iPar–CLR) algorithm [35] to automatically localize 49 landmarks for each 2D face image and its corresponding 3D mesh scan. This large set of expression related landmarks are then used for extracting local texture and shape descriptors for expression classification. To the best of our knowledge, this is the first work which uses such large number of automatically detected landmarks for 2D and 3D multimodal FER. In contrast, the majority of existing feature-based 3D FER approaches reported their results on the BU–3DFE benchmark based on a large set of (typically 83) 3D facial landmarks manually localized by the database providers [15–19,23,25,26]. Therefore, the proposed framework presents a promising way to these landmark-based approaches so that they can be made automatic using the iPar–CLR algorithm in 2D and 3D multimodal face space.

The second contribution of this paper is that a novel second-order image gradient based local texture descriptor (HSOG), a novel first-order mesh gradient (i.e., surface normal) based local shape descriptor (meshHOG), as well as a second-order mesh gradient (i.e., surface curvature) based local shape descriptor (meshHOS) are adapted in FER to comprehensively encode the expression variations in both the 2D and 3D modalities. According to our previous work [36], most of existing popular local image descriptors, such as HOG, LBP, and SIFT, only employ the first-order gradient information related to the slope and the elasticity, i.e., length, area, etc. when the image is regarded as a surface, and thereby partially characterize its geometric properties. By contrast, HSOG captures the curvature related cues of the surface, i.e., cliffs, ridges, summits, valleys, basins, and so on. Thus, HSOG can be applied to describe facial expression deformations (e.g., mouth stretch, lip stretcher, brow raiser). Moreover, in that paper, it was also demonstrated that HSOG outperformed the first-order gradient based local image descriptors (i.e., HOG, LBP, SIFT) when there were not severe scale variations, as in the applications of local image matching and scene classification. In this paper, we give another evidence of the effectiveness and generalization ability of HSOG for FER. Similarly, as general local shape descriptors, meshHOG and meshHOS provide a compact description of the facial surface normal and curvature information, and they have proved very efficient for 3D face identification in our previous works [37,38]. In this paper, we interested in exploring their generalization abilities in 3D FER.