Engineering Applications of Artificial Intelligence 64 (2017) 25–32

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

3D facial expression recognition using kernel methods on Riemannian manifold

Walid Hariri [a,b,*], Hedi Tabia [a], Nadir Farah [b], Abdallah Benouareth [b], David Declercq [a]

[a] ETIS, UMR 8051, ENSEA, Université Paris Seine, Université Cergy-Pontoise, CNRS, Cergy, France
[b] Labged Laboratory, Badji Mokhtar Annaba University, B.P.12, Annaba 23000, Algeria

A R T I C L E I N F O

Keywords:
Covariance matrix
Manifold
Facial expression
Kernel-SVM

A B S T R A C T

Automatic human Facial Expressions Recognition (FER) is becoming of increased interest. FER finds its applications in many emerging areas such as affective computing and intelligent human computer interaction. Most of the existing work on FER has been done using 2D data which suffers from inherent problems of illumination changes and pose variations. With the development of 3D image capturing technologies, the acquisition of 3D data is becoming a more feasible task. The 3D data brings a more effective solution in addressing the issues raised by its 2D counterpart. State-of-the-art 3D FER methods are often based on a single descriptor which may fail to handle the large inter-class and intra-class variability of the human facial expressions. In this work, we explore, for the first time, the usage of covariance matrices of descriptors, instead of the descriptors themselves, in 3D FER. Since covariance matrices are elements of the non-linear manifold of Symmetric Positive Definite (SPD) matrices, we particularly look at the application of manifold-based classification to the problem of 3D FER. We evaluate the performance of the proposed framework on the BU-3DFE and the Bosphorus datasets, and demonstrate its superiority compared to the state-of-the-art methods.

1. Introduction

Facial Expression Recognition (FER) has emerged as an active research field in several areas, such as human–machine interaction, facial animation, and robotics. The recognition of the movements of the eyes, mouth, and facial muscles has attracted a great amount of researchers in the past decade. Detailed surveys of previous work can be found in Fang et al. (2011, 2012); Zeng et al. (2009); Sandbach et al. (2012); Shan et al. (2009); Tian et al. (2003). Most of these previous works were developed for 2D data (Fasel and Luettin, 2003; Pantic and Rothkrantz, 2000; Zeng et al., 2009; Ilbeygi and Shah-Hosseini, 2012; Mahersia and Hamrouni, 2015; Chakrabarty et al., 2013). Although the remarkable performance achieved, most of these works are still sensitive to many variations, particularly illumination and pose. Recent progress in 3D acquisition techniques, on the other hand, has provided a new alternative to overcome these issues (Yin et al., 2006). 3D data bring additional information which are more robust to illumination (Al-Osaimi et al., 2012; Patil et al., 2015) and pose changes (Ocegueda et al., 2013).

1.1. Related works

3D face datasets become more and more available, providing the worldwide researchers of face and FER community a large scale data for training and evaluating their approaches. Different approaches were proposed to address the problem of 3D FER. Most of these works focus on recognizing six basic expressions; anger (AN), fear (FE), disgust (DI), sadness (SA), happiness (HA) and surprise (SU) (Ekman and Friesen, 1971). Proposed methods often use 3D local features which capture the geometrical and topological properties of the face expression (Bowyer et al., 2006; Zhao et al., 2011; Tabia et al., 2011; Mohammadzade and Hatzinakos, 2013). One of the main strengths of local features is their flexibility in terms of type of analysis that can be performed with. Wang et al. (2006) proposed to extract geometric based features to describe facial expressions. These features have been estimated using the principle curvature information calculated on the 3D triangulated mesh model of a face. A linear discriminant analysis classifier has been used for features classification. Shao et al. (2015) proposed to learn sparse features from spatio-temporal local cuboids extracted from the face. They applied conditional random fields classifier to train and classify the expressions. Other methods use distance-based features between certain

* Corresponding author at: ETIS, UMR 8051, ENSEA, Université Paris Seine, Université Cergy-Pontoise, CNRS, Cergy, France. Tel.: +33 785 06 83 98, +213 669 28 76 44.

http://dx.doi.org/10.1016/j.engappai.2017.05.009
Received 1 December 2016; Received in revised form 3 April 2017; Accepted 15 May 2017
0952-1976/

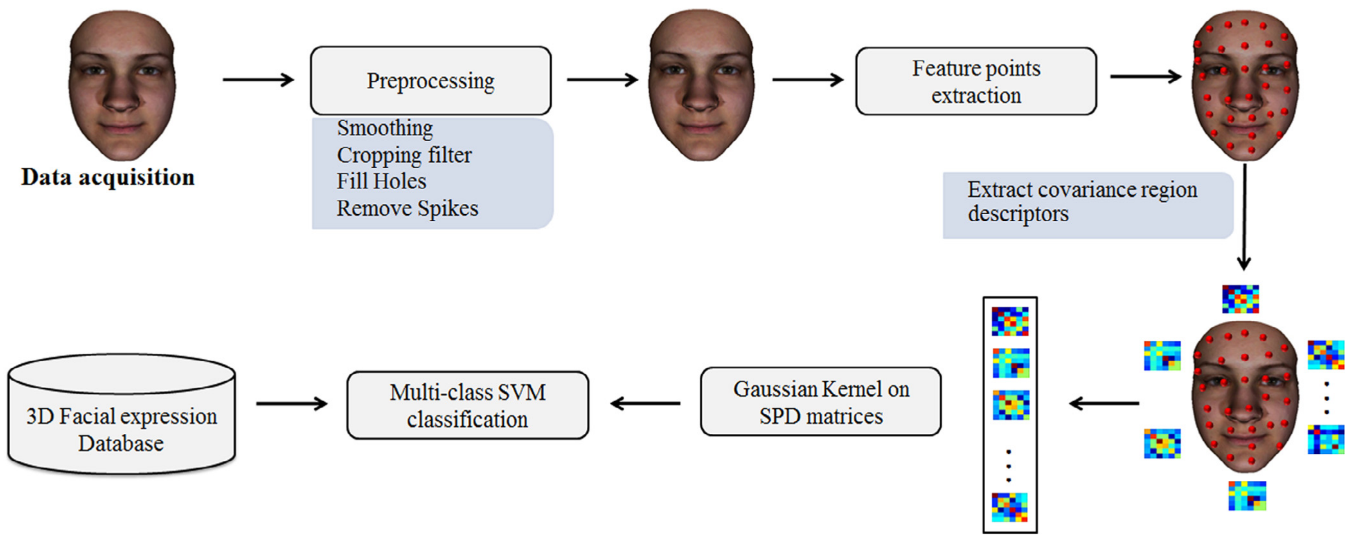

Fig. 1. Overview of the proposed method.

facial landmarks, from a neutral face. Next they compute the changes due to facial deformation (Soyel and Demirel, 2008, 2010; Tang and Huang, 2008; Li et al., 2010; Tekguc et al., 2009; Sha et al., 2011; Srivastava and Roy, 2009). For instance, Soyel and Demirel (2008, 2010) used distance vectors computed between landmarks on the 3D face to describe facial features, and used probabilistic neural network for expression classification.

The use of local features in 3D facial expression recognition, however, has several limitations. For instance, 3D face expressions often exhibit large inter-class and intra-class variability that cannot be captured with a single feature type. This triggers the need for combining different modalities or feature types. However, different shape features often have different dimensions, scales and variation range, which makes their aggregation difficult without normalizing or using blending weights.

Covariance matrices have been successfully used in the literature for object detection and tracking (Tuzel et al., 2006, 2008), shape retrieval (Tabia et al., 2014; Tabia and Laga, 2015), face recognition (Krizaj et al., 2013; Hariri et al., 2016), and image set classification (Wang et al., 2012). The use of covariance matrices has several advantages. First, they provide a natural way for fusing multi-modal features, eventually of different dimensions, without normalization or joint distribution estimation. Second, covariance matrices extracted from different regions have the same size, which is significantly compact compared to the features themselves and to their statistics. This enables comparing any regions without being restricted to a constant window size or specific feature dimension.

Covariance matrices, however, lie on the manifold of Symmetric Positive Definite (SPD) tensors Sym_d^+, a special type of Riemannian manifolds. Therefore, matching with covariance matrices requires the computation of geodesic distances on the manifold using proper metrics.

Moreover, Sym_d^+ manifold has a non-linear structure which makes impossible the use of many conventional classification algorithms such as the Support Vector Machines (SVM). Non-linear mappings to Riemannian manifolds (Tuzel et al., 2008; Yun et al., 2013; Yun and Gu, 2016; Harandi et al., 2014a) or the reproducing kernel Hilbert space (Harandi et al., 2014b; Jayasumana et al., 2013; Khan and Gu, 2014) are, therefore, used to obtain vector spaces, in which the metrics for machine learning methods are defined.

In this paper, we look at the application of manifold-based classification of covariance matrices to the problem of 3D FER. We represent each 3D face surface using a set of unordered covariance descriptors. We then use kernel functions which map covariance matrices into a high dimensional Hilbert space. This hence enables to use conventional classification algorithms on such non-linear valued data. We evaluate the performance of the proposed framework on the BU-3DFE and the Bosphorus datasets, and demonstrate its superiority compared to the state-of-the-art methods. The rest of the paper is organized as follows, the proposed method is detailed in Section 1.2. In Section 2, we present the covariance matrices on 3D face. Section 3 describes the Riemannian geometry of SPD matrices. In Section 4, we explain how to classify facial expressions on manifold. Experimental results are presented in Section 5. Conclusions end the paper.

1.2. Method overview

Fig. 1 presents an overview of the proposed method. Since raw 3D face surfaces may contain imperfections such as holes, spikes, and include some undesired parts (clothes, neck, ears, hair, etc.), we successively apply a set of preprocessing filters; including (1) a smoothing filter which reduces spikes, (2) a cropping filter which returns the central part of the face surface, (3) a filling holes filter, and (4) a median filter which removes spikes.

Once the 3D face mesh has been preprocessed, we uniformly select m feature points over the whole 3D surface. The feature points are the center of m patches from a paving of the face. Each point has a region of influence, which we characterize by the covariance of its geometric features instead of directly using the features themselves. Each feature captures some properties of the local geometry. They can be of different type, dimension or scale.

Covariance matrices lie on the (Sym_d^+) which lacks Euclidean structures such as norm and inner product. This makes impossible the application of conventional clustering algorithms in their original forms. In this paper, we take advantages of recent works on kernel methods on manifold-valued data (Harandi et al., 2014a; Tabia and Laga, 2015; Yun and Gu, 2016) and explore, for the first time, their usage in 3D FER. Since covariance matrices are considered in this work as local descriptors, we propose to apply the SVM algorithm to this local representation. For this end, we build a global kernel function so that one can compare two 3D facial expressions by using the covariance descriptors.

The proposed 3D FER method has been evaluated on the two well known datasets, namely the BU-3DFE and the Bosphorus. Results demonstrate the superiority of the proposed method compared to the state-of-the-art ones.

2. Covariance matrices on 3D face

Let $\mathcal{P} = \{p_i, i = 1 \dots m\}$ be the set of patches extracted from a 3D face. Each patch p_i defines a region around a feature point $p_i = (x_i, y_i, z_i)^t$.