Pattern Recognition 57 (2016) 21–30

Contents lists available at ScienceDirect

Pattern Recognition

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

A Grassmann framework for 4D facial shape analysis

Taleb Alashkar [a], Boulbaba Ben Amor [a,*], Mohamed Daoudi [a], Stefano Berretti [b]

[a] Institute Mines-Télécom/Télécom Lille; CRIStAL (UMR CNRS 9189), France
[b] Department of Information Engineering, University of Florence, Italy

ARTICLE INFO

Article history:
Received 31 December 2015
Received in revised form
5 March 2016
Accepted 9 March 2016
Available online 25 March 2016

Keywords:
4D face recognition
Curvature-maps
Grassmann manifold
Dictionary learning
Sparse coding

ABSTRACT

In this paper, we investigate the contribution of dynamic evolution of 3D faces to identity recognition. To this end, we adopt a subspace representation of the flow of curvature-maps computed on 3D facial frames of a sequence, after normalizing their pose. Such representation allows us to embody the shape as well as its temporal evolution within the same subspace representation. Dictionary learning and sparse coding over the space of fixed-dimensional subspaces, called Grassmann manifold, have been used to perform face recognition. We have conducted extensive experiments on the BU-4DFE dataset. The obtained results of the proposed approach provide promising results.

1. Introduction

In recent years automatic face analysis has attracted increasing interest in the field of computer vision and pattern recognition due to its inherent challenges and its potential in a wide spectrum of applications, including security surveillance [1,2] and diagnostic of facial pathology [3]. Despite the great progress, 2D face analysis approaches that depend on color or gray-scale image analysis, still suffer from illumination and pose variations, which often occur in real-world conditions. With the rapid innovation of 3D cameras, the 3D shape is regarded as a promising alternative to achieve robust face analysis [4,5]. Very recently, the advent of 4D imaging systems capable of acquiring temporal sequences of 3D scans (i.e., 4D is regarded as 3D over the time) made possible comprehensive face analysis by introducing the temporal dimension, where the temporal behavior of 3D faces is captured by adjacent frames [6,7]. Note that such temporal information is crucial for analyzing the facial deformations. Despite the large amount of work on static and dynamic 3D facial scans analysis, temporal modeling is still almost unexplored for identity recognition. Moving from shape analysis of static 3D faces to dynamic faces (4D faces) gives rise to new challenges related to the nature of the data and the processing time – which static and dynamic shape representations are most suited to 4D face analysis? How the temporal dimension can contribute to face analysis? Is it possible to compute statistical summaries on dynamic 3D faces? From a perspective of face classification, which relevant features and classification algorithms can be used?

In this paper, we aim to answer the above questions by proposing a comprehensive framework for modeling and analyzing 3D facial sequences (4D faces), with an experimental illustration in face recognition from 4D sequences.

Recently, works addressing face analysis from temporal sequences of 3D scans start to appear in the literature, encouraged by the advancement in 3D sensors' technology, with some of them restricted to RGB-D Kinect-like sensors. In [8], Berretti et al. investigated the impact of 3D facial scans' resolution on the recognition rate by building super resolution 3D models from consumer depth camera frames. Experimental studies using the new 3D super resolution method validate the increase of recognition performance with the reconstructed higher resolution models. Hsu et al. [9] showed that incorporating depth images of the subjects in the gallery can improve the recognition rate, especially in the case of pose variations, even though there are only 2D still images in the testing. In the last few years, some works addressed face recognition from dynamic sequences of 3D face scans as well like in [6], where Sun et al. proposed a 4D-HMM based approach. In this work, a 3D dynamic spatio-temporal face recognition framework is derived by computing a local descriptor based on the curvature values at vertices of 3D faces. Spatial and temporal HMM are used for the recognition process, using 22 landmarks manually annotated and tracked over time. As an important achievement of this work, it is also evidenced that 3D face dynamics provides better results than 2D videos and 3D static scans.

* Corresponding author.

http://dx.doi.org/10.1016/j.patcog.2016.03.013
0031-3203/

Subspace representation for dynamic facial information either for image sets or for image sequences (videos) showed a great success. Shigenaka et al. [10] proposed a Grassmann distance mutual subspace method (GD-MSM) and Grassmann Kernel Support Vector Machine (GK-SVM) comparison study for the face recognition problem from a mobile 2D video database. In [11], Lui et al. proposed a geodesic distance based algorithm for face recognition from 2D image sets. Turaga et al. [12] presented a statistical method for video based face recognition. These methods use subspace-based models and tools from Riemannian geometry of the Grassmann manifold. Intrinsic and extrinsic statistics are derived for maximum-likelihood classification applications. More recently, Huang et al. [13] proposed learning projection distance on Grassmann manifold for face recognition from image sets. In this work, an improved recognition is obtained by representing every image set using a Gaussian distribution over the manifold.

Sparse representation and dictionary learning attracted a lot of attention recently, due to their success in many computer vision problems. In [14], a sparse coding framework was presented for face recognition from still images. In this work, Wright et al. showed that using sparse coding the role of feature extraction on the performance is not so important, and the sparse coding is more tolerant with face occlusion. Yang et al. [15] proposed a robust sparse coding (RSC) approach for face recognition. In this work, the sparse coding problem is solved as a constrained robust regression, which makes the recognition more robust against occlusion, change of lighting and expression variation in still images. Elhamifar et al. [16] presented the Sparse Subspace Clustering (SSC) algorithm that classifies linear subspaces after finding their sparse coding. A generalization of sparse coding and dictionary learning was proposed by Xie et al. [17], which permits its application on subspace data representations that do not have a linear structure, like the Riemannian manifold. Mapping points from a non-linear manifold to tangent spaces shows good classification results on texture and medical images' classification.

In [18], Harandi et al. proposed an extrinsic solution to combine sparse coding and dictionary learning with nonlinear subspaces, like the Grassmann manifold. Embedding the Grassmann manifold into the symmetric matrices' sub-manifold makes the sparse coding on the induced manifold possible, faster, and more coherent than intrinsic embedding on one or more tangent spaces. Application to 2D video face datasets shows the efficiency of this approach against other learning solutions.

2. Methodology and contributions

In this paper, we investigate the contribution of 3D face dynamics in face recognition. To this end, after a preprocessing step, we compute surface curvature from each 3D static mesh of a sequence, and project it to a 2D map (call edcurvature map). A sequence of curvature maps is then cast to a matrix form by re shaping the 2D maps to column vectors, Singular Value Decomposition (SVD) is used to reduce the subspace spanned by the matrix to that of the first k-singular-vectors, which in turn is regarded as a point on a Grassmann manifold. Recognition using extrinsic methods based on sparse coding and dictionary learning on the manifold achieved the best performance. An overview of the proposed approach is shown in Fig. 1.

In summary, the main contributions of this paper are:

- A fully automatic and computationally cheap face recognition approach using 4D data. To the best of our knowledge, this is the first study in the literature, which brings the subspace modeling methodology with advanced geometric and learning tools to 3D face sequences.
- An in-depth investigation of the contribution of the 3D shape dynamics to face recognition.
- An extensive experimental analysis, involving the BU-4DFE dataset and three classification schemes based on intrinsic and extrinsic methods on the manifold.

The rest of the paper is organized as follows: in Section 3, the methodology of modeling 4D faces on Grassmann manifold as well as essential elements on the geometry of these manifolds is presented; Section 4 discusses sparse representation and dictionary learning on the Grassmann manifold; our 3D dynamic face recognition framework is presented in Section 5; Experimental results and their discussion are given in Section 6; finally, our conclusions and future work are drawn in Section 7.

3. Modeling sequences of 3D faces on Grassmann manifold

The idea of modeling multiple-instances of visual data, like set of images or video sequences, as linear subspaces for classification and recognition tasks has revealed its efficiency in many computer vision problems [12,19,20]. This compact low-dimensional data representation has the main advantage in its robustness against noise or missing parts in the original data. Besides, the availability of computational tools from differential geometry makes working on non-linear data (e.g., the space of k-dimensional subspaces) possible, and allows managing the non-Euclidean nature of these spaces. Accordingly, in this work, we adopt the subspace representation solution for analyzing 4D facial sequences. To our knowledge, this is one of the first investigations on modeling the temporal evolution of 3D facial shapes with application to face recognition. Studying the effects of these two aspects together is still an open problem in computer vision applications.

In the remaining of this section, we will describe the static 3D shape representation using mean curvature computed on 3D facial surfaces as well as the associated subspace representation to capture their temporal dynamics (Section 3.1). In addition, since the subspace learning approach that we propose lies on the Grassmann manifold, we will also recall essential background on its geometry, and related definitions including metrics and distances (Section 3.2) and sample mean computation (Section 3.3).

3.1. Static and dynamic 3D shape representation

In the proposed solution, we consider 3D scans of the face acquired continuously via a dynamic 3D scanner (3D plus time, also called 4D), thus producing a temporal 3D sequence with the dynamic evolution of the 3D face. Using these data, the proposed approach is designed to exploit the spatio-temporal information. To achieve this goal, a subspace modeling technique is applied as follows: (i) the 3D scans are preprocessed by cropping the facial region from the rest of the scan, then pose normalization, denoising via smoothing, and holes filling are performed; (ii) the mean curvature on 3D surfaces is computed, so that a flow of curvature-maps is produced by projection; (iii) the k-SVD orthogonalization procedure is applied to subsequences of the curvature-maps, so as to obtain an orthonormal basis spanning an optimized subspace. This subspace represents an element on the Grassmannian manifold $\mathcal{G}_k(\mathbb{R}^n)$, being n the dimension of curvature maps.

The shape information of every 3D scan is captured first by computing, as 3D local descriptor, the mean curvature $H=(k_1+k_2)/2$, where k_1 and k_2 are the two principle curvatures. The mean curvature values are computed at every vertex, then they are visualized and saved as a 2D map using a blue-red color

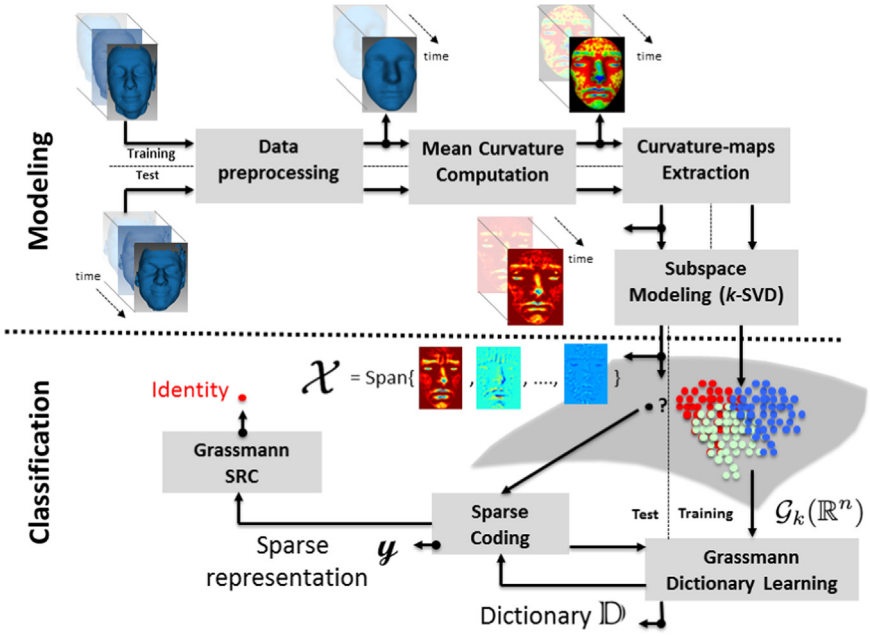

Fig. 1. Overview of the proposed approach: top – modeling the shape and its dynamics using a subspace representation; bottom – classification of space representations using the proposed Sparse Representation based Classification (SRC) algorithm.

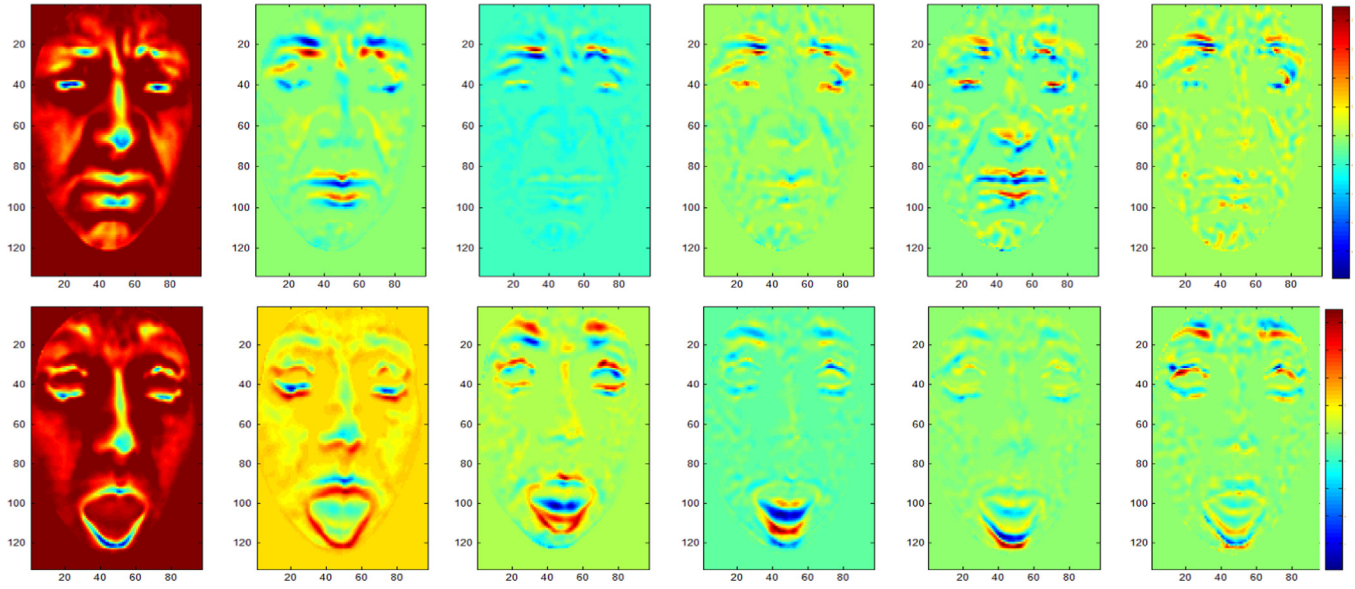

Fig. 2. Visual illustration of two subspaces using their singular vectors derived by SVD *orthogonalization* on sequences of size 50 frames (*Angry*, top row – *Surprise*, bottom row). In the plots, colors ranging from blue to red reflect increasing values of the curvature. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

scale.[1] The main motivation in using this descriptor is its ability to capture the local facial shape and the non-rigid deformation, and also its invariance against rotation, scale and mesh resolution.

Fig. 2 shows, as color maps, the matrices representing the subspaces computed from two different 3D facial sequences. It can be appreciated that a subspace (k-first dominant left singular vectors of the original matrix of data) can be viewed as the mean shape computed over the subsequence (leftmost images), followed by the dominant deformations (remaining images on the right). These deformation images change according to the expression exhibited by the face (*Angry* in the first row, and *Surprise* in the second). We note histogram equalization is used here (except for the images on the left column) to highlight the location of the deformation areas, using cold to warm colors. Colors in between reflect the most stable areas of the curvature-maps over the 3D video.

3.2. Distances on the manifold

The idea of using the Grassmann manifold representation is that a subsequence of 3D scans can be cast to a matrix representation, and thus mapped to a unique point on the manifold. In this way, computing the similarity between two subsequences is transformed to the problem of computing the distance between two points on the manifold. More specifically, let $\mathcal{X}$ and $\mathcal{Y}$ denote a

[1] The VTK library has been used: http://www.vtk.org.

Table 1
Subspace distances.

Subspace distance	Mathematical formulation
Min correlation	$d_{Min}(\mathcal{X},\mathcal{Y}) = \sin\ \theta_k$
Binet-Cauchy	$d_{BC}(\mathcal{X},\mathcal{Y}) = (1-\prod_i^k\ \cos^2\theta_i)^{1/2}$
geodesic	$d_{Geo} = \left(\sum_i^k \theta^2\right)^{1/2}$
Procrustes	$d_{Proc}(\mathcal{X},\mathcal{Y}) = 2\left(\sum_i^k\ \sin(\theta_i/2)\right)^{1/2}$
Max correlation	$d_{Max}(\mathcal{X},\mathcal{Y}) = \sin\ \theta_1$
Projection	$d^2_{proj}(\mathcal{X},\mathcal{Y}) = \sum_i^k\ \sin(\theta_i)^2$

pair of subspaces on $\mathcal{G}_k(\mathbb{R}^n)$: the Riemannian distance between $\mathcal{X}$ and $\mathcal{Y}$ is the length of the shortest path connecting the two points on the manifold (i.e., the geodesic distance). The problem of computing this distance can be solved using the notion of *Principle Angles* introduced by Golub and Loan [21] as an intuitive and computationally efficient way for defining the distance between two linear subspaces. In fact, there is a set of principal angles $\boldsymbol{\Theta} = [\theta_1,\ldots,\theta_k]$ $(0 \le \theta_1,\ldots,\theta_k \le \pi/2)$, between the subspaces $\mathcal{X}$ and $\mathcal{Y}$ of size $n \times k$, recursively defined as follows:

$$\theta_k = \cos^{-1}\left(\max_{u_k \in \mathcal{X}} \max_{v_k \in \mathcal{Y}} \langle u_k^T, v_k\rangle\right), \tag{1}$$

where u_k and v_k are the vectors of the basis spanning, respectively, the subspaces $\mathcal{X}$ and $\mathcal{Y}$, subject to the additional constraints: $\langle u_k^T, u_k\rangle = \langle v_k^T, v_k\rangle = 1$, being $\langle\cdot,\cdot\rangle$ the inner product in $\mathbb{R}^n$; and $\langle u_k^T, u_i\rangle = \langle v_k^T, v_i\rangle = 0$ $(\forall\ k,i:\ k \ne i)$.

Based on the notion of principle angles, several other distances and metrics on the Grassmann manifold have been proposed in the literature. Some of them are summarized in Table 1.

3.3. Karcher mean computation

As mentioned above, an important tool in shape (and its temporal evolution) analysis is given by the computation of statistical summaries. The idea here is that given a set of subspaces, which correspond to subsequences of 3D videos of the same person (or different persons), with the same expression (or different expressions), one would like to compute their statistical mean. For a set of given subspaces $\mathcal{P}_1,\ldots,\mathcal{P}_m \in \mathcal{G}_k(\mathbb{R}^n)$ (i.e., points on the underlying manifold), Karcher mean μ is defined as $\mu = \arg\min_{\mathcal{P}} \sum_{i=1}^m d_{Geo}(\mathcal{P},\mathcal{P}_i)^2$, is a point on the Grassmannian, which minimizes the mean squared error [22] with respect to the canonical metric d_{Geo} previously defined in Table 1.

4. Grassmann sparse representation

Recently, the sparse coding theory showed great success in several topics, like signal processing [23], image classification [24] and face recognition [15], where a given signal or image can be approximated effectively as a combination of few members (atoms) of a dictionary. The success of sparse coding motivated the extension of this approach to the space of linear subspaces [12], in order to represent a subspace as the combination of few subspaces of a dictionary. However, in so doing, the main issue is the inherent non-linearity of the Grassmann manifold, which implies using tools from differential geometry. Since these tools often require intensive computation, these solutions are less attractive for 2D and 3D video modeling and analysis.

The problem of *sparse coding* has been solved in Euclidean spaces in $\mathbb{R}^n$ by minimizing the following quantity, which includes a coding cost function with a penalty term related to the sparsity of the result:

$$l(x,\mathcal{D}) = \min_y \|x - \mathcal{D}y\|_2^2 + \lambda\|y\|_1, \tag{2}$$

where $x \in \mathbb{R}^n$ is the sample signal to be coded, $\mathcal{D}$ is a dictionary (a $n \times N$ matrix being N the number of training samples) with atoms $D_i \in \mathbb{R}^n$ in its columns, and λ the sparse regularization parameter. The vector $y \in \mathbb{R}^N$ is the new latent sparse representation of the original data, which contains many zeros. The problem of *dictionary learning* consists of minimizing the total coding cost for all the samples $\{x^t \in \mathbb{R}^n\}_{1 \le t \le N}$ of the training set, over all choices of codes and dictionaries as follows:

$$h(\mathcal{D}) = \min_{\{x^t\},\mathcal{D}} \frac{1}{N}\sum_{t=1}^{N} l(x^t,\mathcal{D}). \tag{3}$$

In order to combine advantages of subspace modeling mentioned in Section 3 with the powerful sparse coding representation, it is essential to handle the non-linearity of the Grassmann manifold. The first direction to tackle this problem is provided in the literature by the *extrinsic solution*, which relies on the basic idea of mapping the points of the Grassmann manifold into a fixed tangent space (i.e., a vector space) [17]. The main constraint of this method is the logarithm map function used for tangent space mapping, which does not have an explicit formula in the case of Grassmann manifold. This makes its estimation numerically not accurate, especially for the points far from the tangent space position. Also, it is time consuming, which makes the approach slow. To avoid these limitations, a second common *extrinsic method* consists of embedding the Grassmann manifold into a smooth sub-manifold of the space of symmetric matrices [25]. This embedding is performed by a projection mapping function [26,27]. For convenience, we recall here the main ideas and the derived algorithm.

Formally, let us have $\mathcal{P} = \{Span\{P_1\}, Span\{P_2\},\ldots,Span\{P_m\}\}$ as a set of points (i.e., subspaces), where $Span\{P_i\} \in \mathcal{G}_k(\mathbb{R}^n)$. We need to be able to represent each point as a linear combination of a few atoms $\{D_1, D_2,\ldots,D_j\}$ of a dictionary $\mathcal{D}$ using the sparse coding. For any $\mathcal{X} = Span(X) \in \mathcal{G}_k(\mathbb{R}^n)$ the mapping ϕ: $\mathcal{G}_k(\mathbb{R}^n) \to Sym(\mathbf{n})$, such that $\phi(\mathcal{X}) = XX^T = \hat{X}$ is computed. The mapping function ϕ is isometric, as it preserves the curve length between the Grassmann manifold and the manifold of symmetric matrices $Sym(\mathbf{n})$ [28]. A natural choice of metric on the manifold of symmetric matrices $Sym(\mathbf{n})$ is the Frobenius inner product, that is for any $Span(X)$, $Span(Y) \in \mathcal{G}_k(\mathbb{R}^n)$, $\delta_s(\hat{X},\hat{Y}) = \mathrm{Tr}(\hat{X},\hat{Y}) = \|X^TY\|_F^2$. With this embedding, Eq. (2) can be rewritten by considering the embedding $\hat{X}$ of a given query subspace $\mathcal{X}$:

$$l(\mathcal{X},\mathcal{D}) = \min_y \|\hat{X} - \hat{\mathcal{D}}y\|_F^2 + \lambda\|y\|_1, \tag{4}$$

where $\hat{\mathcal{D}}y$ denotes the dictionary with atom elements of $Sym(\mathbf{n})$, y the sparse representation, and λ is the regularization parameter that weighs the importance of the fitting of the model versus the magnitude of y. This convex optimization problem is solvable as a vectorized sparse coding problem, as depicted in Algorithm 1.

Algorithm 1. Sparse coding on $\mathcal{G}_k(\mathbb{R}^n)$.

Require A given dictionary $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^N \in \mathcal{G}_k(\mathbb{R}^n)$ where $\mathcal{D}_i = Span(D_i)$ of size N. Query subspace $\mathcal{X} \in \mathcal{G}_k(\mathbb{R}^n) = Span(X)$
for $i,j \leftarrow 1$ to N **do**
$\mathbb{K}(\mathcal{D})_{i,j} \leftarrow \|D_i^T D_j\|_F^2$
end for
$\mathbb{K}(\mathcal{D})_{N\times N} = U\Sigma U^T$
$A = \Sigma^{1/2}U^T$
for $i \leftarrow 1$ to N **do**
$\mathcal{K}(X,\mathcal{D})_i \leftarrow \|X^TD_i\|_F^2$

end for

$x^* \leftarrow \Sigma^{-1/2} U^T \mathcal{K}(X, D)$

ensure $y^* \leftarrow \arg\min_y \| x^* - Ay \|^2 + \lambda \| y \|_1$

In Algorithm 1, the training set of (labeled) subspaces is considered as the dictionary $\mathcal{D}$ of size N (i.e., the training set size); (i) a similarity matrix between dictionary elements $\mathbb{K}(\mathcal{D})$ is computed based on the Frobenius inner product; (ii) SVD is applied to $\mathbb{K}$ (i.e., $\mathbb{K} = U\Sigma V^T$) to compute the A matrix, which is the weighted singular vectors of $\mathbb{K}$; (iii) the similarity matrix $\mathcal{K}(X, \mathcal{D})$ between testing and training samples is computed on the induced space. The decomposition of Eq. (4) shows that the sparse coding problem can be formulated as:

$$l(\mathcal{X}, \mathcal{D}) = \min_y \| x^* - Ay \|^2 + \lambda \| y \|_1, \tag{5}$$

where $x^* = \Sigma^{-1/2} U^T \mathcal{K}(X, \mathcal{D})$.

We have used the implementation provided by Harandi et al. [18], and we refer the reader to their recently-published paper [29] for further mathematical details on their *extrinsic solution*. Using Algorithm 1, a new observation point in the symmetric matrices manifold after the embedding step, could be (sparsely) decomposed into a combination of atoms of the dictionary. From a classification perspective, it is now possible to use conventional classifiers, such as SVM (Support Vector Machine) or SRC (Sparse Representation Classification), since these features lie in an Euclidean space.

5. Identity recognition from 4D faces

To perform face recognition from the 3D facial shapes and their temporal evolution, the flow of curvature-maps is first divided into clips (subsequences) of size w. Then, each clip is modeled as an element on the Grassmann manifold via k-SVD orthogonalization. More formally, given a sequence of curvature-maps $\{m_0, \ldots, m_t\}$, a predefined size of a sliding window w, and a fixed subspace order k, the idea is to consider the maps under the temporal interval $[t-w+1, t]$, and to compute the corresponding subspace $\mathcal{P}_t$. This results in a collection of subspaces, elements of the Grassmann manifold, which represent the 3D video sequence (after curvature computation). The main goal of such representation is to capture the 3D shape of the face as well as its dynamics (spatio-temporal description) to perform face recognition.

5.1. Grassmann Nearest-Neighbor Classifier (GNNC)

In this approach, for each subject a Karcher mean subspace is computed out of the subspaces that belong to the training set of the subject (i.e., more subsequences are used as training for each individual). These means constitute the gallery subspaces used for recognition. According to this, given a probe subspace $\mathcal{X} = Span(X)$, it is compared against the gallery subspaces using one of the distances defined on the Grassmann manifold (see Table 1). Finally, the probe subspace is assigned to one class using the Grassmann Nearest-Neighbor Classifier.

5.2. Grassmann Sparse Representation Classifier (GSRC)

In this case, the classification is performed on the sparse representation computed according to Algorithm 1. In fact, given a test sample, its sparse representation is first computed using the dictionary on the training samples. Consequently, conventional classification methods, like SVM or Nearest-Neighbor can be applied. An alternative solution is to use the Sparse Representation Classifier (SRC) proposed in [14]. The main concept behind this classifier is to reproduce the testing query subspace from non-zero sparse codes that belong to one specific class only in the dictionary. Repeating this class-specific estimation, and computing the residual error between the estimation and the original query subspace gives a similarity indicator. The estimation from the correct class should give the minimum residual error.

In summary, face recognition is performed according to the following steps: (1) *Dictionary learning* on the Grassmann manifold – given a training subset of observations, a set of atoms (dictionary) is determined to describe the observations sparsely; (2) *Sparse representation* – given a dictionary and a probe on the underlying manifold, the probe is approximated using a sparse linear combination of atoms from the dictionary; (3) *GSR-based classification* – once the training and testing observations are expressed linearly using sparse representation, it is possible to perform the Grassmann Sparse Representation Classification.

6. Experimental results

To investigate the contribution of facial dynamics in identity recognition using 4D data, we conducted extensive experiments on the BU-4DFE dataset. This dataset has been collected at Binghamton University [30] and is currently used in several studies on 4D facial expression recognition. To our knowledge, only the work of Sun et al. [6] has reported identification performance on this dataset. The main characteristics of the BU-4DFE dataset are summarized in Table 2.

6.1. Experiments setting

Following the protocol proposed in [6], 60 subjects have been considered out of the BU-4DFE, and their sequences are partitioned into subsequences using a window size $w=6$ (with a shifting step of 3 frames). This results into 30 sub-sequences extracted out of every facial expression sequence of the 60 subjects (i.e., each sequence lasts approximately 90 frames). On these subsequences, experiments have been conducted following two settings:

- *Expression independent* (EI): One expression per subject is used for training, and this expression does not appear in the testing. All the remaining five expression sequences per subject are used for testing. Since 30 sub-sequences represent each expression sequence, for the 60 subjects a total of $30 \times 60 = 1800$ subsequences are used for training. Five expressions per subject are used for testing, i.e., for each subject we have $5 \times 30 = 150$ test subsequences, with a total for all the 60 subjects of $150 \times 60 = 9000$ subsequences.
- *Expression dependent* (ED): For each sequence, the first half (from neutral to nearby the apex of the expression) is used for training, while the remaining half (from the apex of the expression to neutral) is used for testing. As a consequence, the gallery and the probe samples convey similar dynamic behavior,

Table 2
BU-4DFE dataset main characteristics.

Number of subjects (male/female)	101 (43/58)
Age range	18–45
Number of 3D videos per subject	6
Expressions per subject	*Angry, Disgust, Fear, Happy, Sad, Surprise*
3D sequence duration	4 sec (25 fps) – about 90 frames per sequence
Vertices per 3D mesh	35,000 (Di4D capturing system)
Acquisition protocol	neutral–onset–apex–offset–neutral

Table 3
Recognition rates (RR%) for GNN-classification using different subspace distances.

Subspace distance	ED-RR (%)	EI-RR (%)
Min correlation	44.75	28.72
Binet-Cauchy	52.83	51.99
Geodesic	73.00	65.00
Procrustes	78.11	66.55
Max correlation	92.61	67.12
Projection	**93.69**	**68.88**

though with inverse temporal evolution. The number of training subsequences for every subject is $15 \times 6 = 90$, with a total for the 60 subjects of $90 \times 60 = 5400$ subsequences. The same number of subsequences is used for testing.

Using these settings, we report in the following experimental evaluation and comparative analysis of the proposed approaches using Grassmann Nearest-Neighbor Classification (GNNC) on the mean subspaces of each subject' class, and Grassmann Sparse-Representation (GSR) based classification computed on the sparse codes, in comparison to the current literature.

6.2. 4D face recognition using GNNC

In this experiment, a window of six frames $w=6$ and shifting step equals to 3 is used (the same as in [6]), with only the first two dominant components kept for representing the subspace ($k=2$). The GNN-classification method is based on a gallery of subspaces, one per subject, each computed as the mean of the training subsequences for the subject. With the setting above, in the EI scenario, one complete expression is used to compute the mean for each subject, i.e. 30 subsequences; in the ED scenario, the mean is computed on $15 \times 6 = 90$ subsequences with different expressions.

Using the GNN-classifier, a comparison is performed between the ED and EI experiments. Different distances are also considered, which involve the principal angles between subspaces (see Section 3.2). The average recognition rates are reported in Table 3.

Some observations can be derived from this GNN: (i) ED results outperform EI results for each distance measure. This is expected, since in the ED setting there are sequences of the same subject conveying the same expression both in the gallery and probe sets (though with inverse temporal evolution); (ii) the different RRs scored by the distances provide experimental evidence of the discriminative information distribution across the principle angles. In particular, the highest RR obtained with the *Projection* distance shows that all the singular vectors, and consequently the dynamic information of subsequences, helps in the recognition task by improving the result obtained using just one singular vector (i.e., *Max Correlation* distance). The lowest RR is scored by the *Min Correlation* distance, suggesting us that the subspaces on the manifold are sufficiently separated from each other, thus making them well suited for the identity recognition task.

Results reported in Table 3 have been obtained by comparing single instances (subspaces) in the video. Actually, since subsequences are part of a continuous video, it is possible to fuse the decisions of successive subsequence instances to perform recognition. This allows us to design an incremental recognition approach over time, where multiple instances are used instead of only one. This idea has been implemented using a majority voting fusion rule, at each time, using all available instances. The experimental results are reported in Fig. 3 to show the performance at increasing size of the data seen and analyzed along a sequence. From these plots, it is clear the performance increases with the fraction of the video seen. This trend is the same under ED and EI settings. The same conclusions drawn above are still valid, with a serious limitation in using a metric-based approach for comparing sub-sequences exhibiting different facial behavior. More elaborated approaches, which handle with these differences are suitable for performing 3D face recognition under EI settings.

6.3. 4D face recognition using GSRC

In these experiments, we use the proposed solution based on Sparse Coding on Grassmann manifold (GSR). A variant of the GDA Grassmann Discriminant Analysis algorithm [20], called GGDA (Graph-embedding GDA) [31] is also used as a baseline to evaluate the effectiveness of the GSR algorithm. In practice, the flow of curvature-maps, for the window of size w is first mapped to the Grassmann manifold using SVD. Then, the steps described in Section 5.2 are performed for training and testing. The value of the

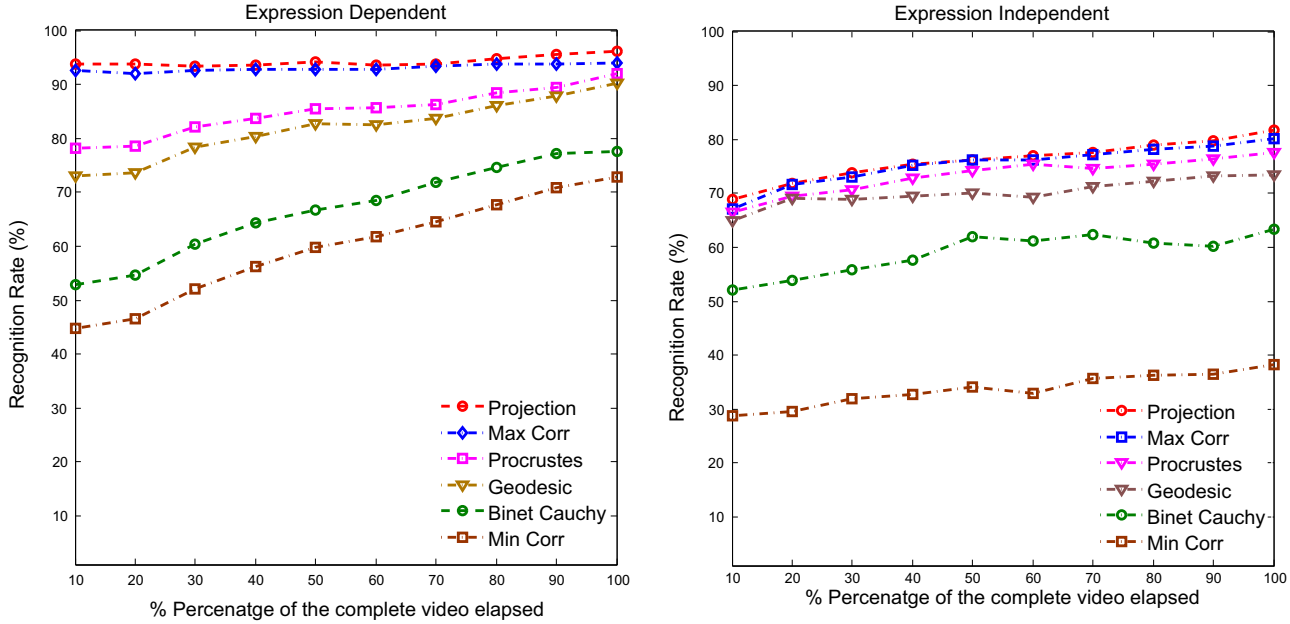

Fig. 3. Trade-off between accuracy and latency (fraction of the video seen) for different Grassmann metrics/distances in the ED and EI settings.

Table 4
EI experiment: effect of the subspace order k on the recognition rate for the GSR algorithm. Subsequences with window size $w=6$ have been used in all the cases.

Subspace order k	1	**2**	3	4	5	6
Y_k (%)	81	**90**	94	96	98	100
RR (%)	81.03	**84.13**	81.76	81.22	80.94	80.02

Table 5
EI experiment: effect of the window size w on the recognition accuracy for the GGDA and GSR algorithms. The subspace order k is set to keep 90% of the original data.

w, k	**Algorithm**	
	GGDA-RR (%)	**GSR-RR** (%)
6, 2	**64.24**	**84.13**
10, 4	61.15	79.89
15, 6	56.61	76.55
20, 9	50.50	76.59
25, 11	50.60	75.80

regularization parameter λ in Eq. (4) is selected empirically. Results under the ED and EI settings are reported. Results when varying the window size w, and the subspace order k are also reported and discussed.

6.3.1. Expression independent (EI) experiment

As a preliminary experiment, we investigated the effect of the subspace order k on the performance. To this end, we apply the GSR algorithm with a varying $k \in \{1,2,3,4,5,6\}$, while keeping a fixed window size $w=6$, and shifting step equals to 3. Thus, in this case, we have 30 training subspaces per subject, for a total of 1800 subspaces in the training set (dictionary).

The subspace order k (i.e., number of singular vectors considered) is also related to the information carried by the respective singular value λ_i, through the measure Y_k:

$$Y_k = \left(\sum_{i=1}^{k} \lambda_i\right) \Big/ \left(\sum_{i=1}^{w} \lambda_i\right), \tag{6}$$

where w is the window size or the maximum number of singular vectors (the length of the subsequence).

As shown in Table 4, the highest average recognition rate is 84.13%, obtained for $k=2$. This rate is 3% higher than the average recognition rate obtained for $k=1$ (using only the first dominant left-singular vector, which corresponds to the common data over the window).

This allows us to make two main conclusions: (i) the importance of the facial dynamics in improving the recognition performance. In fact, the optimal parameter $k=2$ implies that the mean and the first dominant deformations are important in the recognition process. They are given by the first and the second singular-vectors of the orthogonal matrix, respectively; (ii) the remaining left-singular vectors are less relevant in the recognition process, since they include the noise, which is present in the 4D acquisition. We note that $k=2$ allows capturing, on average, about 90% of the data available in the 4D sub-sequence. Based on these empirical observations, in our next experiments, we will consider 90% of the information for different size of the window w.

We are interested now on studying the effect of varying the size of the window w on the performance. In the following experiment, we considered $w \in \{6,10,15,20,25\}$. The subspace order k is defined as the number of left singular-vectors, which retains 90% of the original data. The corresponding recognition accuracies are reported in Table 5. It can be seen that the optimal window size is $w=6$ for both the GSR and the GGDA algorithms. One explanation for the decreasing accuracy at increasing size of w is the lack of temporal registration of the curvature-maps. In fact, a large difference between the frames across the window affects negatively the orthogonalization procedure, which assumes dense correspondence between the frames. Interestingly, the accuracy obtained using the GSR (84.13%) substantially improves the accuracy achieved using the GGDA (64.24%), and the GNN-classification (68.88%). This result also evidences the efficiency of sparse coding of subspaces in comparison to the discriminant analysis, which can be affected by the points' distribution over the Grassmannian manifold.

In Fig. 4, we show an example where the face is reconstructed by using only the first k-singular vectors out of the base, which contains 90% of the information as reported in Table 5. It is clear from this figure that the accuracy of reconstructing the face decreases by increasing the window size due to the lack of tracking.

To investigate the effect of the regularization parameter λ used in dictionary learning (see Eq. (4)), we report in Table 6 the recognition rate of the GSR method, using $w=2$ and $k=2$, for different values of λ.

Table 7 provides additional details by reporting the RR obtained separately for each test expression, by the GGDA and GSR algorithms, and the approach proposed in [6]. The average recognition rate achieved by GSR is 84%, which is about 10% lower than the accuracy reported in [6]. However, differently from the approach proposed by Sun et al., our solution does not require any manual or automatic landmarking of the face and it is computationally more efficient. In addition, the dense (vertex-level) registration of the 3D frames, which is computationally complex and time consuming, is not performed in our method. On an opposite side, this operation permits the approach presented in [6] to achieve comparable results throughout all the expressions. In our case instead, we observe the RR decreases of about 4% in the case of posed surprise expression, which includes topological variations of the face (i.e., mouth open). Another methodological difference between the two approaches is that Sun et al. designed and trained two separate HMMs called spatial and temporal; In our approach, only few coefficients of the sparse representation are sufficient to code the dynamics of a 3D facial sequence and can be used to perform GSR classification.

The recognition performance of our solution can be improved by using an increasing fraction of the video. This implies that more than one instance (subsequence) is used to recognize a subject. With this approach, the overall performance of GSR increases from 84.13% (using only one instance, which represents about 5% of the video) to 95.11% using the whole video (about 4 s). This is illustrated in Fig. 5, separately for each expression. This figure also confirms the difficulty in recognizing subjects with *Surprise* expression.

In the experiments above, recognition values have been obtained by averaging on six-folds, each of which uses one expression for (identity) training and the other five for testing. To investigate the importance of using larger training set and with different expressions, we have also performed experiments in the case the training includes five expressions (i.e., 9000 subsequences, 150 per subject), while the subsequences from the only remaining expression are used for testing (i.e., 1800 subsequences, 30 per subject). Results are reported in Table 8 in comparison with those obtained for training with one expression. Results (using GSR) show that increasing the number of samples and their dynamics (even though originated from different expressions) can significantly increase the recognition rate from 84.13% to 93.37%. We can also observe that identity recognition under *Surprise* expression is the most difficult, due to the large shape changes in

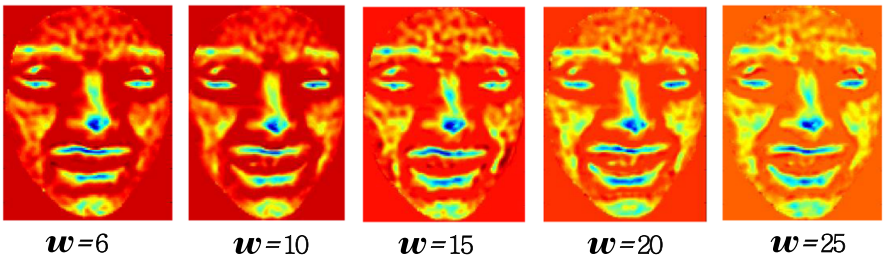

Fig. 4. Reconstructed faces from the first k-singular vectors for different window sizes.

Table 6
Recognition rate of the GSR algorithm for subspaces generated using different values of the regularization parameter λ.

λ	0.01	0.05	**0.1**	0.15	0.2	0.3
RR (%)	82.09	83.21	**84.13**	83.00	82.94	80.02

Table 7
EI experiment: recognition rate obtained using different training expression compared to the approach in [6].

Training expression	**Method**		
	Sun et al. [6] (%)	**GGDA** (%)	**GSR** (%)
Angry	94.12	61.26	85.20
Disgust	94.09	68.54	87.70
Fear	94.45	69.02	83.49
Happy	94.52	68.56	83.36
Sad	93.87	63.05	84.86
Surprise	95.02	56.07	80.49
Overall	94.37	64.42	84.13

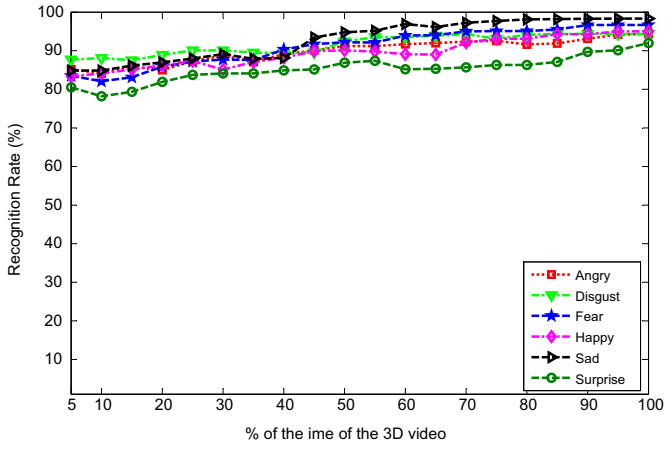

Fig. 5. EI experiment: trade-off between accuracy (RR%) and latency.

Table 8
Impact of the training set on the performance: training based on one expression vs. training based on five expressions.

Testing expression	**Training by one** (%)	**Training by five** (%)
Angry	83.27	94.50
Disgust	78.42	96.30
Fear	92.21	98.13
Happy	86.23	93.20
Sad	94.32	97.73
Surprise	69.75	80.40
Overall	**84.13**	**93.37**

Table 9
ED experiment: comparison between the recognition accuracy obtained for the methods proposed in this works, and for the 2D video, 3D static, and 3D dynamic (4D) approaches reported in [6].

Method	**RR** (%)
Gabor-wavelet on 2D videos (from [6])	85.09
LLE on static 3D (from [6])	82.34
PCA on static 3D (from [6])	80.78
LDA on static 3D (from [6])	91.37
ST-HMM on 4D (proposed in [6])	97.47
GNN on 4D	**93.69**
GGDA on 4D	**98.08**
GSR on 4D	**100**

Table 10
ED and EI results for 2D and 3D videos.

Method	**EI–RR** (%)	**ED –RR** (%)
2D video A-HMM [32] (from [6])	67.05	93.97
4D ST-HMM [6]	94.37	97.47
4D GSR	**84.13**	**100**

the mouth region, while identity recognition under *Sad* expression is the easiest across the time.

6.3.2. *Expression dependent (ED) experiments*

In this experiment, the window size is $w=6$, with shifting step equals to 3, and 30 sub-sequences are obtained from each facial expression sequence, half of which is used for training and half for testing. Thus, we have 90 training subspaces per subject, and a dictionary of 5400 subspaces. The GSR-based classifier is used in this experiment. Table 9 reports the results obtained using the GSR and the GGDA algorithms on 3D dynamic sequences (4D). In addition, for comparison purposes, we also reported in the table several results from [6], including Gabor wavelets on 2D videos, LLE, PCA and LDA on 3D static data, and the ST-HMM on 4D data.

It can be seen that both GGDA and GSR outperform previous approaches. In particular, their accuracy is close or equal to 100% under the ED-setting. Our explanation of the higher accuracy achieved by GGDA and GSR compared to existing methods is that the optimized SVD-based orthogonalization produces a matrix independent of the time-order of the 3D video clips. That is, comparing two video clips taken from the Onset-Apex and the Apex-Offset gives small distance as the temporal order of the curvature-maps is ignored. This demonstrates the ability of the Grassmann representation associated with the learning methods in 4D face recognition.

6.4. *Comparative study and discussions*

From the experimental results reported above, it emerges the proposed approach, which combines Grassmann representation with an extrinsic learning method achieved promising results in

Table 11
Processing time of the proposed pipeline compared to [6]. A 3.2 GHz CPU was used in [6], compared to the 2.6 GHz CPU used in our work.

Processing step	Processing time (s)	
	Sun et al. [6]	This work
One 3D frame processing	15	1
One probe recognition	5	0.73
Full video processing – 100 frames	1500	90

4D face recognition. This demonstrates the contribution of the facial dynamics in the recognition process. In Table 10, we summarize the obtained results under the ED and EI settings. To the best of our knowledge, only Sun et al. have investigated the problem of face recognition from 4D data. They have also studied the advantage of using the dynamic of shape (3D videos) compared to the dynamic of appearance (2D videos), as reported in the table.

It is clear from these results that the 3D video modality outperforms the 2D video modality. That is, the dynamics in 3D facial shapes has more discriminating power compared to the dynamics of 2D facial images. When the proposed approach is compared with [6], it is evident that the latter performs better in the EI case, where sequences with different expressions are compared. This indicates a good robustness to the expression differences of the method in [6]. This is mainly due to the dense temporal vertex-tracking approach required before training the HMMs. However, this comes at the cost of an increased computational complexity of the tracking, in addition to the required accurate manual/automatic landmarks detection in the first 3D frame of a sequence.

The computational aspect is evaluated in Table 11, which reports the processing time of the proposed pipeline compared to [6] (the original values reported in [6] are used here). From the table, it emerges the proposed approach is less demanding in processing time. In addition, it does not need any manual or automatic landmark detection of the face.

7. Conclusions and future perspectives

In this paper, we have proposed a comprehensive and effective 4D face recognition framework, which adopts a subspace-learning methodology. We have demonstrated that the shape dynamics (behavior) improves the recognition accuracy. This conclusion is valid even if the training samples (in the gallery) and the probes (to be recognized) present a different behavior. Leveraging the geometry of Grassmann manifolds, relevant geometric tools and advanced Machine Learning tools, i.e., dictionary learning and sparse coding on the underlying manifold, our approach is capable of managing face recognition from dynamic sequences of 3D scans in an effective and efficient way. The main advantages of this framework are: it is completely automatic and computationally less demanding compared to the current literature; it achieves promising face recognition accuracy; it outperforms previous approaches under the expression-dependent setting.

Conversely, the main limitation of the proposed approach is the lack of temporal dense correspondence across the curvature-maps of a sequence. Despite the limited size of the temporal window considered, this has a negative effect especially for expressions with large face variations. These aspects will be part of our next investigations and future research direction. Investigating the proposed framework for different face analysis tasks, such as facial expression recognition will be also considered as an important direction for future work.

Conflict of interest

None declared.

Acknowledgment

This research was supported by the Futur & Rupture program of the Institut Mines-Télécom, the MAGNUM project (BPI and Région Nord-Pas de Calais) and the PHC Utique 2016 program for the CMCU project number 34882WK.

References

[1] D. Kang, H. Han, A.K. Jain, S.-W. Lee, Nighttime face recognition at large standoff: cross-distance and cross-spectral matching, Pattern Recognit. 47 (12) (2014) 3750–3766.
[2] M. De la Torre, E. Granger, R. Sabourin, D.O. Gorodnichy, Adaptive skew-sensitive ensembles for face recognition in video surveillance, Pattern Recognit. 48 (11) (2015) 3385–3406.
[3] N. Batool, R. Chellappa, Fast detection of facial wrinkles based on Gabor features using image morphology and geometric constraints, Pattern Recognit. 48 (3) (2015) 642–658.
[4] A. Maalej, B. Ben Amor, M. Daoudi, A. Srivastava, S. Berretti, Shape analysis of local facial patches for 3D facial expression recognition, Pattern Recognit. 44 (8) (2011) 1581–1589.
[5] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, R. Slama, 3D face recognition under expressions, occlusions, and pose variations, IEEE Trans. Pattern Anal. Mach. Intell. 35 (9) (2013) 2270–2283.
[6] Y. Sun, X. Chen, M. Rosato, L. Yin, Tracking vertex flow and model adaptation for three dimensional spatiotemporal face analysis, IEEE Trans. Syst. Man Cybern. Part A 40 (3) (2010) 461–474.
[7] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, A. Srivastava, 4-D facial expression recognition by learning geometric deformations, IEEE Trans. Cybern. 44 (12) (2014) 2443–2457.
[8] S. Berretti, P. Pala, A. Del Bimbo, Face recognition by super-resolved 3D models from consumer depth cameras, IEEE Trans. Inf. Forensics Secur. 9 (9) (2014) 1436–1449.
[9] G.-S. Hsu, Y.-L. Liu, H.-C. Peng, P.-X. Wu, RGB-D-based face reconstruction and recognition, IEEE Trans. Inf. Forensics Secur. 9 (12) (2014) 2110–2118.
[10] R. Shigenaka, B. Raytchev, T. Tamaki, K. Kaneda, Face sequence recognition using grassmann distances and Grassmann kernels, in: IEEE International Joint Conference on Neural Networks (IJCNN), Brisbane, QLD, Australia, 2012, pp. 1–7.
[11] Y.M. Lui, J. Beveridge, B. Draper, M. Kirby, Image-set matching using a geodesic distance and cohort normalization, in: IEEE International Conference on Automatic Face Gesture Recognition (FG), Amsterdam, The Netherlands, 2008, pp. 1–6.
[12] P. Turaga, A. Veeraraghavan, A. Srivastava, R. Chellappa, Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition, IEEE Trans. Pattern Anal. Mach. Intell. 33 (11) (2011) 2273–2286.
[13] Z. Huang, R. Wang, S. Shan, X. Chen, Projection metric learning on Grassmann manifold with application to video based face recognition, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 140–149.
[14] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[15] M. Yang, D. Zhang, J. Yang, D. Zhang, Robust sparse coding for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 2011, pp. 625–632.
[16] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, IEEE Trans. Pattern Anal. Mach. Intell. 35 (11) (2013) 2765–2781.
[17] Y. Xie, J. Ho, B. Vemuri, On a nonlinear generalization of sparse coding and dictionary learning, in: International Conference of Machine Learning (ICML), Atlanta, GE, USA, 2013, pp. 1480–1488.
[18] M. Harandi, C. Sanderson, C. Shen, B. Lovell, Dictionary learning and sparse coding on grassmann manifolds: an extrinsic solution, in: IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 2013, pp. 3120–3127.
[19] M. Turk, A. Pentland, Face recognition using eigenfaces, in: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Maui, HI, USA, 1991, pp. 586–591.
[20] J. Hamm, D.D. Lee, Grassmann discriminant analysis: a unifying view on subspace-based learning, in: International Conference on Machine Learning (ICML), Helsinki, Finland, 2008, pp. 376–383.
[21] G. Golub, C. Van Loan, Matrix Computations, 3rd edition, Johns Hopkins University Press, Baltimore, MD, USA, 1996.
[22] H. Karcher, Riemannian center of mass and mollifier smoothing, Commun. Pure Appl. Math. 30 (1977) 509–541.

[23] Y. Xu, Z. Xiao, X. Tian, A simulation study on neural ensemble sparse coding, in: International Conference on Information Engineering and Computer Science (ICIECS), Wuhan, China, 2009, pp. 1–4.
[24] W. Zuo, D. Meng, L. Zhang, X. Feng, D. Zhang, A generalized iterated shrinkage algorithm for non-convex sparse coding, in: IEEE International Conference on Computer Vision (ICCV), Sydney, Australia, 2013, pp. 217–224.
[25] C. Yasuko, Statistics on special manifolds, in: Lecture Notes in Statistics, vol. 174, Springer, New York, 2003.
[26] A. Srivastava, A Bayesian approach to geometric subspace estimation, IEEE Trans. Signal Process. 48 (5) (2000) 1390–1400.
[27] R. Vemulapalli, J. Pillai, R. Chellappa, Kernel learning for extrinsic classification of manifold features, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 2013, pp. 1782–1789.
[28] J. Helmke, K. Huper, Newton's method on Grassmann manifold, Preprint, 2007, arXiv:0709.2205.
[29] M. Harandi, R. Hartley, C. Shen, B. Lovell, C. Sanderson, Extrinsic methods for coding and dictionary learning on Grassmann manifolds, Int. J. Comput. Vis. 114 (2) (2015) 113–136.
[30] L. Yin, X. Chen, Y. Sun, T. Worm, M. Reale, A high-resolution 3D dynamic facial expression database, in: IEEE Conference on Face and Gesture Recognition (FG), Amsterdam, The Netherlands, 2008, pp. 1–6.
[31] M.T. Harandi, C. Sanderson, S.A. Shirazi, B.C. Lovell, Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 2011, pp. 2705–2712.
[32] X. Liu, T. Chen, Video-based face recognition using adaptive hidden Markov models, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Madison, WS, USA, 2003, pp. 340–345.

Taleb Alashkar is a research assistant in Department of Electrical and Computer Engineering, Northeastern University, Boston, MA, USA. He received his Ph.D. degree in computer science from University of Lille 1 and Master degree in Computer Vision and Image Processing from University of Dijon in 2015 and 2012 respectively. Before that he finished Bachelor of Science degree in Computer Engineering from University of Aleppo, Syria. His research interests include Computer Vision, Machine Learning and Pattern Recognition.

Boulbaba Ben Amor is Associate Professor (with Habilitation) of computer science with the Institut Mines-Télécom/Télécom Lille and member of the CRIStAL Research Center (UMR CNRS 9189), since 2007. He received the Ph.D. degree from Ecole Centrale de Lyon (France) in 2006. During 2013–2014, he was a visiting research professor at Florida State University (USA). He served as Area Chair for the WACV'16 conference and Reviewer for several major conferences (ICCV, CVPR, EECV, ICPR, etc.) and Journals (T-PAMI, T-IP, T-IFS, T-Cybernetics, etc.) in computer vision. His research areas include 3D computer vision, 3D/4D shape analysis and pattern recognition.

Mohamed Daoudi is a Professor of Computer Science at Institut Mines-Télécom/Télécom Lille and the head of Image group at CRIStAL Laboratory (UMR CNRS 9189), France. He received his Ph.D. degree in Computer Engineering from the University of Lille 1 (France) in 1993 and Habilitation à Diriger des Recherches from the University of Littoral (France) in 2000. His research interests include pattern recognition, shape analysis, computer vision and 3D object processing. He has published over 150 research papers dealing with these subjects that have appeared in the most distinguished peer-reviewed journal and conference proceedings. He is the co-author of several books including 3D Face Modelling, Analysis and Recognition (Wiley 2013) and 3D Object Processing: Compression, Indexing and Watermarking (Wiley 2008). He has been Conference Chair of the Shape Modelling International Conference (2015) and several other national conferences and international workshops. He is Fellow of IAPR, Senior Member of IEEE and member of Association of Computing Machinery (ACM).

Stefano Berretti received the Ph.D. in Information and Telecommunications Engineering in 2001 from the University of Florence, Italy. Currently, he is an Associate Professor at the Department of Information Engineering and at the Media Integration and Communication Center of the University of Florence, Italy. His main research interests focus on 3D object retrieval and partitioning, face recognition and facial expression recognition from 3D and 4D data, 3D face super-resolution, human action recognition from 3D data. He has been visiting researcher at the Indian Institute of Technology (IIT), in Mumbai, India, and visiting professor at the Mines-Télécom/Télécom Lille, in Lille, France, and at the Khalifa University, Sharjah, UAE. Stefano Berretti is author of more than 120 papers appeared in conference proceedings and international journals in the area of pattern recognition, computer vision and multimedia. He is in the program committee of several international conferences and serves as a frequent reviewer of many international journals. He has been co-chair of the Fifth Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment (NORDIA 2012), held in conjunction with ECCV 2012. Since January 2016 he is Information Director of the ACM Transactions on Multimedia.