# A Grassmannian Framework for Face Recognition of 3D Dynamic Sequences with Challenging Conditions

Taleb Alashkar, Boulbaba Ben Amor, Mohamed Daoudi, Stefano Berretti

# A Grassmannian Framework for Face Recognition of 3D Dynamic Sequences with Challenging Conditions

Taleb Alashkar[1], Boulbaba Ben Amor[1], Mohamed Daoudi[1], and Stefano Berretti[2]

[1]Télécom Lille/LIFL (UMR CNRS/Lille1 8022), Lille, France
[2]University of Florence, Florence, Italy

**Abstract.** Modern face recognition approaches target successful person identification in challenging scenarios, where uncooperative subjects are captured under unconstrained imaging conditions. With the introduction of a new generation of 3D acquisition devices capable of dynamic acquisitions, this trend is now emerging also in 3D based approaches. Motivated by these considerations, in this paper we propose an original and effective framework to address face recognition from 3D temporal sequences acquired in adverse conditions, including internal and external occlusions, pose and expression variations, and talking. Due to the novelty of the proposed scenario, a new database has been collected using a single-view structured light scanner with a large field of view, which allows free movement of the acquired subjects. The 3D temporal sequences are divided into fragments each modeled as a linear subspace in order to embody the shape and the motion of the facial surfaces. In virtue of the Riemannian geometry of the space of real $k$-dimensional linear subspaces, called Grassmann manifold, a new formulation of the matching between 3D temporal sequences has been developed. An unsupervised clustering over the Grassmann manifold is also introduced for efficient recognition. The proposed approach achieves promising results, without requiring any prior training or manual intervention.

**Keywords:** Face recognition, 3D dynamic face sequences, Grassmann manifold

## 1 Introduction

Early biometric solutions using the face for recognizing persons' identity were based on the face appearance in 2D still images acquired in controlled ambient, with ideal illumination conditions and with cooperative subjects. However, these over constrained solutions are of limited utility in real contexts, such as law enforcement, surveillance systems and access control, where occlusions, pose variation, illumination changes and facial expressions are present. Limitations of methods based on 2D still images have stimulated the investigation of new solutions, which also exploit the temporal dimension of 2D-videos acquired with

2D cameras. In fact, it is a shared conviction that motion information can improve the recognition rate, especially under uncontrolled viewing conditions [1, 5, 10]. However, evaluations on unconstrained face recognition (FR) from 2D still images and videos, such as the Multiple Biometric Grand Challenge [16] showed that FR under pose variations is still a distant goal [1]. This boosted a large corpus of ongoing work focusing on FR in the "wild" [19, 21, 23].

Recently, the availability of 3D acquisition systems opened the way to 3D face recognition solutions. Since these approaches use the 3D geometry of the face, they have the advantage of being robust against illumination and pose variations [2, 3]. However, most of the existing solutions are tested on datasets collected under well-controlled settings using static acquisition systems [17], though some methods have recently appeared that account for pose variation, facial expressions and occlusions [7, 15]. Most recent advancements of 3D technologies, like structured-light and time-of-flight scanners, made 3D dynamic acquisition systems available in the market at lower cost. These devices have still optical capabilities that are far from those exhibited by 2D cameras and often differ in terms of operating distance and resolution (for example, Kinect-like devices operate up to some meter, but with low resolution). Despite of these limitations, they make possible real-time capturing of a continuous flow of 3D scans, thus opening the way to solutions capable of performing face and facial expression recognition from dynamic sequences of 3D face scans. Apart for its technical practicability, adding the temporal dimension to 3D acquisitions is motivated by the observation that the face is a deformable 3D surface changing over time, so that using the temporal component can be essential to improve recognition, especially under adverse acquisition conditions. A clear example of this is given by spoofing attacks that can be difficult to detect in 2D still images or even in 2D videos, but result much more evident when the 3D temporal component is considered.

Works addressing FR from temporal sequences of 3D scans are still a few, with some of them restricted to RGB-D Kinect-like sensors [12, 14]. For example, Min et al. [14] proposed a real-time 3D face recognition system using multiple RGB-D instances. The approach does not exploit temporal correlation; however, it shows that exploiting majority voting between multiple instances provides better recognition rate than using static scans. Similarly, working on RGB-D acquisitions, Li et al. [12] proposed an algorithm for face recognition under varying poses, expressions, illumination and disguise. To the best of our knowledge, the only approach addressing FR from dynamic sequences of 3D face scans is that proposed by Sun et al. [20], where a 3D dynamic spatio-temporal approach is derived by computing a local descriptor based on the curvature values at vertices of 3D faces. Spatial and temporal Hidden Markov Models 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 that 2D videos and 3D static scans. However, the applicability of this work remains limited, since it requires 3D high resolution

scans in the sequences. In addition, the method requires scans in frontal pose, without pose variation or occlusions.

In this work, a new FR approach from temporal sequences of 3D scans acquired in adverse conditions, including internal and external occlusions, large and free pose variations, facial expressions and talking is proposed. To the best of our knowledge, this is the first work proposing this new paradigm to overcome the 2D video and 3D static based limitations. A subspace-based modeling approach is introduced, where the spatial-temporal data are modeled as a finite-dimensional linear subspace. Thus, each linear subspace is considered as an element on a Grassmann manifold. This formulation has some interesting aspects: *(i)* Comparing two subspaces is cheaper than comparing two 3D dynamic fragments; *(ii)* It is more robust to noise and missing data, which are common in realistic scenarios. In addition, this approach uses a holistic descriptor based on shape normals, without requiring any manual/automatic landmarking. The facial motion is also modeled and exploited in the recognition process.

According to the proposed representation, each subject in the gallery is represented by several 3D subsequences as instances, thus resulting in a large gallery set. Therefore, to optimize the efficiency of recognition, an unsupervised clustering approach over the Grassmannian of the gallery instances is applied. Due to the absence of databases collecting 3D dynamic sequences for FR under adverse conditions, we constructed a new database, which includes scans exhibiting free pose variations, facial expressions, talking, internal and external occlusions. In so doing, our dataset differs from the few existing 3D dynamic face databases (also called 4D datasets) [6, 13, 24], which are collected for facial expressions and/or action units recognition under highly conditioned settings and using high-resolution 3D acquisition.

In summary, the main contributions of this work are:

- A new FR scenario, where 3D dynamic sequences of the face are compared in order to permit FR under occlusions, pose variations and expressions;
- A new representation of 3D dynamic face sequences, which exploits relevant geometry tools on Grassmannian manifold, and unsupervised clustering of 3D temporal sequences;
- A new 3D dynamic face database, which includes well-known FR challenges in realistic scenarios.

The rest of the paper is organized as follows: Our FR approach between 3D dynamic sequences is presented in Sect. 2; In Sect. 3, the gallery clustering strategy for optimizing the efficiency of recognition is described; Experiments on the BU-4DFE database and the new 3D dynamic face dataset we collected are reported in Sect. 4; Discussion and conclusions are given in Sect. 5.

## 2 Face recognition from 3D temporal sequences

In the proposed scenario, we consider 3D scans of the face that are acquired continuously through a 3D camera, thus constituting a temporal 3D sequence

with dynamic variations of the geometry of the face. Using these data, the proposed approach is designed to exploit the spatio-temporal information available in 3D dynamic sequences of the face. To achieve this goal, a subspace modeling framework is applied. The basic idea of this solution is to extract a set of 3D temporal subsequences (fragments) from each 3D full temporal sequence, each constituted by a predefined number of 3D frames, and model each fragment $f$ as a linear subspace $\mathcal{P}_f$, which can be represented as an element on a Grassmann manifold. According to this, given a 3D temporal sequence $G$ in the gallery constituted by the concatenation of $N$ 3D temporal fragments $g_i$ indexed by $i$, so that $G = \{g_{i,(i=1,\ldots,N)}\}$, and a probe 3D temporal fragment $f$ with $m$ successive frames $f_{probe} = [f_1, \ldots, f_m]$, the process of comparing a probe fragment with a gallery sequence can be formulated as follows:

$$g^* = \arg\min_i \; d(\mathcal{P}_{f_{probe}}, \mathcal{P}_{g_i}) \;, \tag{1}$$

where $d(.,.)$ denotes the geodesic distance between two linear subspaces, and $g^*$ is the 3D temporal fragment of the gallery closer to the probe fragment according to the used distance. The complete recognition process is then obtained by extending this analysis to all the gallery sequences.

In order to apply the above representation and matching strategy, several steps are required for the scan preprocessing and subspace modeling, as illustrated in Fig. 1. After the acquisition, the face region of each frame in a 3D temporal sequence is cropped. Due to pose variations and the scanner technology, the number of vertices representing the surface of the face mesh varies in the same session and from one session to another. For the subspace modeling approach, it is important to have the same number of vertices representing the face in each frame of a sequence. To this end, a down-sampling is applied to each frame, so as to produce a constant number of $n$ vertices per frame. Then, the normal at each vertex is estimated based on the neighborhood vertices included in a sphere of radius $R$ around the vertex [18]. The set of estimated normals at the vertices of each frame capture the shape of the face, and is used as a spatial holistic descriptor of the face surface.

However, 3D frames constituting the 3D temporal sequences do not show a correspondence between their respective vertices, which is indeed necessary to develop the proposed linear subspace representation. In order to establish a rough and fast correspondence between frames, a normal shooting technique [4] is used between each two successive frames. As a result of this process, each 3D temporal fragment can be modeled as a matrix $\mathcal{S}$ of size $n \times \omega$, where $n$ is the number of vertices, and $\omega$ is the number of frames in the 3D temporal fragment. Each column of $\mathcal{S}$ is given by the $z$ component of the estimated normals at each vertex of one frame, so that each row embodies the motion information originated from the variability over time of the $z$ component of the normal of one vertex of the face surface. The main reason for using only the $z$ component, rather than $x$ and/or $y$ of the estimated normal is that $z$ provides a discriminative signature between faces of different subjects, whereas the other two components are more similar in inter-class cases, thus leading to less discrimination in the

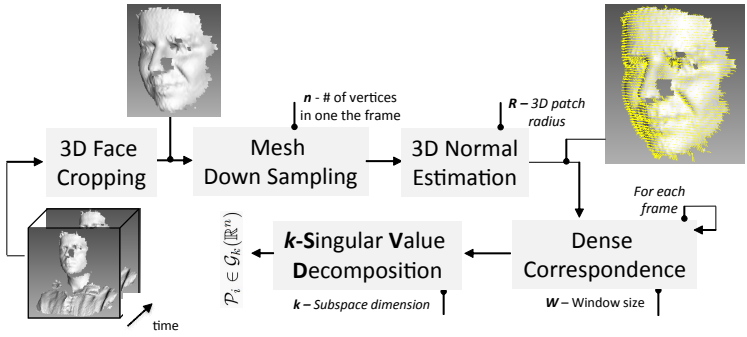


**Fig. 1.** Overview of the proposed approach

feature vector. Finally, a $k$-Singular Value Decomposition of the obtained matrix is performed $\mathcal{S} = U\Sigma V^t$. The eigenvectors matrix $U$ is an orthonormal basis of the subspace $\mathcal{P} = \text{span}(U)$, which is an element on the Grassman manifold $\mathcal{G}_k(\mathbb{R}^n)$. As a result of this pipeline, each 3D temporal fragment is viewed as an element of the Grassmannian manifold, and the original problem of comparing temporal sequences of 3D face scans is turned into a distance measurement between the elements over the Grassmannian manifold corresponding to the 3D temporal fragments.

### 2.1 Matching of 3D temporal fragments on the Grassmann manifold

Let $\mathcal{G}_k(\mathbb{R}^n)$ be the Grassmann manifold of a set of $k$-dimensional linear subspaces of $\mathbb{R}^n$, and $\mathcal{X}$, $\mathcal{Y}$ denote a pair of subspaces on $\mathcal{G}_k(\mathbb{R}^n)$. Formally, 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), as it is depicted in Fig. 2.

Golub and Loan [9] introduced an intuitive and computationally efficient way of defining the distance between two linear subspaces using the principal angles. In fact, there is a set of principal angles $\Theta = [\theta_1, \ldots, \theta_k]$ $(0 \leq \theta_1, \ldots, \theta_k \leq \pi/2)$, between the subspaces $\mathcal{X}$ and $\mathcal{Y}$ of size $n \times k$, recursively defined as follows:

$$\theta_k = \cos^{-1}\left(\max_{u_k \in \mathcal{X}} \max_{v_k \in \mathcal{Y}} \langle u_k^t, v_k \rangle\right) , \tag{2}$$

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: (1) $\langle u_k^t, u_k \rangle = \langle v_k^t, v_k \rangle = 1$, being $\langle ., . \rangle$ the inner product in $\mathbb{R}^n$; and (2) $\langle u_k^t, u_i \rangle = \langle v_k^t, v_i \rangle = 0$ $(i = 1, \ldots, k-1)$. In other words, the first principal angle $\theta_1$ is the smallest angle between all pairs of unit basis vectors in the two subspaces. The rest of the principal angles are defined in a similar manner.

Based on the definition of the principal angles, the geodesic distance between $\mathcal{X}$ and $\mathcal{Y}$ can be defined as [8]: $d^2(\mathcal{X}, \mathcal{Y}) = \sum_i \theta_i^2$. This distance is used to measure the similarity between two 3D temporal fragments, permitting to smooth

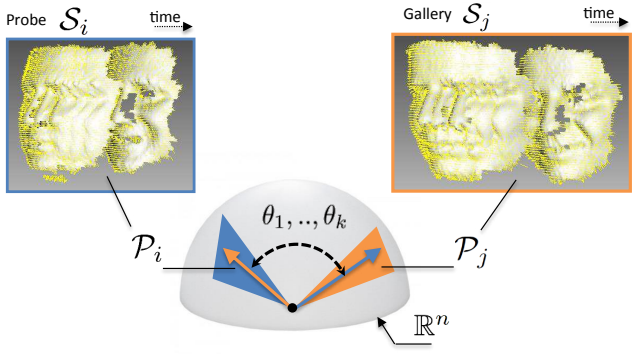


**Fig. 2.** Principal angles $\Theta = [\theta_1, .., \theta_k]$ computed between two linear subspaces $\mathcal{P}_i$ and $\mathcal{P}_j$ of the Grassmannian manifold $\mathcal{G}_k(\mathbb{R}^n)$

the effect of noisy data, at the same time showing robustness with respect to acquisition variations.

However, the combination of free pose variations of the subjects, and the use of a single-view 3D scanner for acquisition can result in many frames with missing parts of the face due to self occlusions. As a consequence, it is not possible to find correspondence and track vertices throughout successive frames of the whole 3D video. The proposed solution for this problem is to consider a sliding window of size $\omega$ containing an affordable pose variation. According to this, each subsequence of size $\omega$, called 3D temporal fragment, represents approximately one pose of the moving face. In this way, each subject in the gallery is represented by multiple instances. Besides, this step helps to solve the problem of pose variation, in that it keeps the motion information coming from the variability of the face surface embodied in the linear subspace of each instance. The same procedure is applied to the probe sequence, where each $\omega$ successive frames are modeled as one probe to be recognized. According to this, the matching between probe and gallery is modeled as a multiple instance matching, with the final recognition decision based on majority voting. Figure 3 summarizes this process, showing how each subject in the gallery can be represented by several instances. In addition, it is also shown how using majority voting to accumulate the recognition decision coming from several successive instances can improve the accuracy of the final recognition decision. In particular, the recognition rate increases over time since more instances from the probe session give more chance to find similar poses in the gallery session of the subject.

## 3 Gallery clustering for efficient recognition

As it is mentioned in Sect. 2, to solve the problem of pose variations the sequence of each subject in the gallery is divided into multiple instances over time. The same procedure is applied to probe sequences. So, each 3D temporal fragment

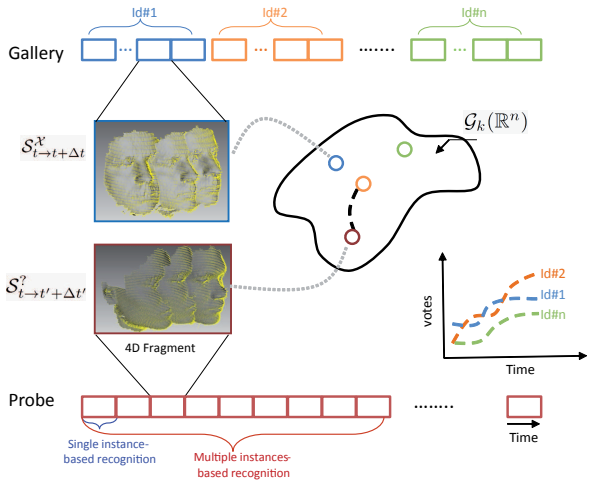


**Fig. 3.** Face recognition based on the match of multiple 3D temporal fragments derived from gallery and probe sequences: The sequences in the gallery (top of the figure) and the probe sequence (bottom of the figure) are divided into multiple 3D temporal fragments; each 3D temporal fragment is then regarded as a point on the Grassmannian manifold (mapping on $\mathcal{G}_k(\mathbb{R}^n)$, in the middle of the figure); the geodesic distance on the manifold is computed between every pair of fragments; the final recognition decision exploits majority voting over the time of successive instances (2D plots on the right)

of a probe is compared with all the 3D temporal fragments in the gallery. This exhaustive search can be avoided by clustering gallery instances according to the main pose of the 3D frames. After applying this unsupervised clustering, each cluster uses the *Karcher* mean [11] of the final elements included in the cluster as representative element. In this way, each probe sequence is compared just with the clusters' representative in order to recognize to which cluster it belongs to. Then, the comparison is extended to the instances that belong to this cluster in the gallery. This method significantly reduces the recognition time by avoiding the comparison of the probe instance with all the gallery instances.

The proposed clustering is performed using the K-means algorithm on the Grassmann manifold [22], as reported in Algorithm 1. This directly descends from the modeling of each 3D temporal fragment as a linear subspace (that is, an element on the Grassmann manifold with a well defined geodesic distance between any two elements). Let us consider a set of points on the Grassmann manifold $\mathcal{P} = \{\mathcal{P}_i\}_{i=1}^n$. They should be clustered in $k$ clusters $\mathcal{C} = \{\mathcal{C}_i\}_{i=1}^k$. Each cluster has a mean on the Grassmann manifold $(\mu_1, \mu_2, \ldots, \mu_k)$. Each mean point should satisfy this condition: the sum of geodesic distances between the class mean and its elements is minimized. To solve this problem, an expectation maximization (EM) method is used. First, $k$ points from $\mathcal{P}$ are initialized at random as cluster centers $(\mu_1^0, \mu_2^0, \ldots, \mu_k^0)$. Each point in $\mathcal{P}$ is then assigned to the nearest center in the E-step. Then, in the M-step, the cluster centers

**Algorithm 1** – Intrinsic K-means clustering on $\mathcal{G}_k(\mathbb{R}^n)$

**Require:** $\mathcal{P}_{i(1\leq i\leq m)} \in \mathcal{G}_k(\mathbb{R}^n)$, $N$: Number of iterations,
  Initialize cluster centers $(\mu_1^0, \ldots, \mu_k^0)$ at random
  $j \leftarrow 0$
  **while** $(j < N)$ **do**
    Assign each $\mathcal{P}_i$ to the nearest cluster $\mathcal{C}_t$ by computing $d^2(\mathcal{P}_i, \mu_t) \leftarrow |exp_{\mu_t}^{-1}(\mathcal{P}_i)|^2$, being $\mu_t$ the center of $\mathcal{C}_t$
    Recompute cluster centers $(\mu_1^j, \ldots, \mu_k^j)$ using Algorithm 2
    $j \leftarrow j + 1$
  **end while**
**Ensure:** $\mathfrak{C} = \mathcal{C}_{t(1\leq t\leq K)}$ the obtained clusters

are recomputed using the Karcher mean algorithm described in Algorithm 2 as detailed in [22]. Applying this clustering procedure to the gallery instances results in aggregations, which are produced according to the instance poses. As a consequence, instances from the same subject or from different subjects that have similar poses fall in the same cluster.

**Algorithm 2** – Computation of Karcher Mean on $\mathcal{G}_k(\mathbb{R}^n)$

**Require:** $\mathcal{P}_{i(0\leq i\leq n)} \in \mathcal{G}_k(\mathbb{R}^n), \epsilon > 0$
  Initialize $\mu_0 \leftarrow \mathcal{P}_0$, $i \leftarrow 0$,
  **repeat**
    Compute $\nu_i \leftarrow exp_{\mu_i}^{-1}(\mathcal{P}_j)$ for $j = 0, \ldots, n$
    Compute the average tangent vector $\bar{\nu} \leftarrow \frac{1}{n}\sum \nu_i$
    Move $\mu_i$ according to $\mu_{i+1} \leftarrow exp_{\mu_i}(\epsilon\bar{\mu})$
    $i \leftarrow i + 1$
  **until** $(||\bar{\nu}|| \neq \epsilon)$
**Ensure:** $\mu$: Karcher Mean of $\{\mathcal{P}_i\}$

## 4 Experimental results

The proposed approach for face recognition from 3D dynamic sequences has been evaluated on the BU-4DFE dataset, in order to compare our solution with results reported by state of the art solutions, and on a new 3D dynamic face database that we present in Sect. 4.2.

### 4.1 Evaluation on BU-4DFE database

Binghamton University 4D Facial Expression (BU-4DFE) database [24] is a 3D dynamic facial expression public database containing 101 subjects. For each subject there are 6 different facial expression sessions. Each session lasts about 4 seconds containing approximately 100 3D scans (frames). The number of vertices

in each 3D frame is between 35,000 to 40,000. All the frames show a frontal pose without any kind of occlusion. The same evaluation protocol used in [20] for expression dependent experiment is followed to validate our framework on the same database and to compare the performances, even though it is not the best settings for face recognition challenges. The six sessions of each subject are divided into two halves, each comprising 50 frames: the first half is used for training and the second for testing. For each subject there are 6 instances in the gallery, each of them belonging to one of the different basic expression session (i.e., angry, disgust, fear, happy, sad and surprise), and the same as probe. For this experiment, 60 subjects are selected as in [20]. Each instance in the gallery and the probe are modeled as a 10-dimensional linear subspace. The number of vertices in each scan is downsampled to $n = 10,000$. Comparison is performed over the Grassmann manifold by finding the smallest geodesic distance between the probe instance and gallery instances. The achieved recognition rate using single-based method is 92%. In [20], Sun et al. achieved 97.5%. However, 22 facial landmarks are manually annotated for vertex flow tracking, while the proposed approach uses an automatic tracking method. In [20], a training stage is also applied on the gallery data before recognition. Table 1 reports a comparison of the two approaches by considering efficiency and effectiveness aspects.

**Table 1.** Performance analysis and comparison

| | **Sun et al. [20]** | **This work** |
|---|---|---|
| One frame processing | 15 sec | 3 sec |
| One probe recognition | 5 sec | 3 sec |
| CPU used | 3.2 GHz | 2.66 GHz |
| FR rate | 97.5% | 92% |

### 4.2 Experimental results on a new 3D dynamic face dataset

Few 3D dynamic face databases, such as the BU-4DFE [24], D3DFACS [6], Hi4D-ADSIP [13] have recently appeared for the purpose of facial expressions and/or action units recognition. However, other FR challenges, like pose variation and occlusion are not considered. As an additional contribution of this work, we constructed a 3D dynamic face database, which presents the following features: (1) It includes most of the FR challenges that occur in realistic scenarios, like pose variation, facial expressions, talking, internal and external occlusions, which are not considered in current 3D dynamic databases; (2) The collected scans have low resolution, which is more convenient for real-world applications (for example, the number of vertices in each scan is about $4,000$, which is 10 times less than BU-4DFE); (3) The field of view of the used 3D scanner is wide enough to

permit non-cooperative free movement of the subject; (4) A single-view structured light system is used for 3D dynamic sequence acquisition, which permits real time capturing. Such system is more convenient in real world applications than multiple view systems used for current dynamic databases, which need long offline registration stages and highly conditioned acquisition environments.

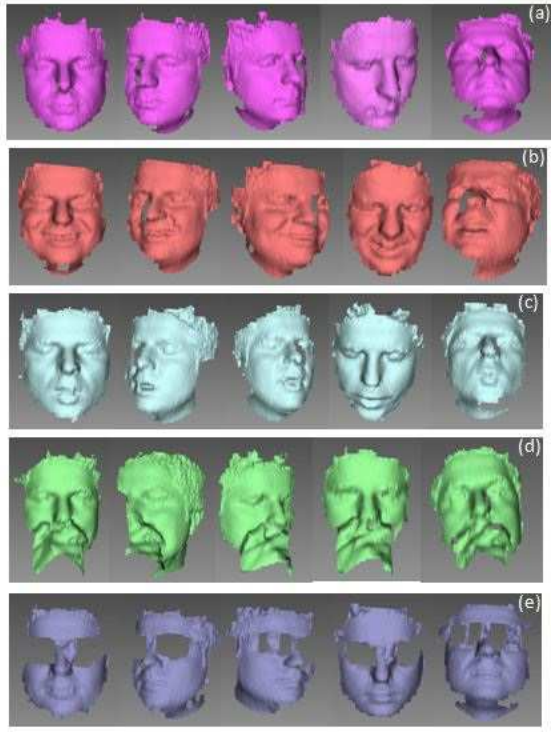


**Fig. 4.** Example 3D dynamic sequences from our face database acquired under unconstrained conditions: (a) neutral; (b) expressive; (c) talking; (d) internal occlusions (hand, hair); (e) and external occlusions (glasses, scarf)

In the proposed database, for each subject we have: *(i)* A full 3D static model with texture acquired using the *Artec* MHT 3D scanner, without any kind of occlusion or expression in the daylight with closed eyes; *(ii)* Six 3D dynamic sessions recorded using the *Artec* L 3D scanner. Each session lasts over 20 seconds, with 15fps as a temporal resolution (i.e., 300 frames in each session). These sessions represent five different unconstrained scenarios, namely: *neutral* (Ne), *facial expression* (Fe), *talking* (Tk), *external occlusion* (Eo) by scarf or sun glasses, and *internal occlusion* (Io) by hand or hair (examples are shown in Fig. 4). Two sessions are acquired for the neutral case, and one session for the other four scenarios. All the six 3D dynamic sessions are acquired under

uncontrolled pose variations around pitch and yaw axes, where subjects are free to move at normal speed. So far, 58 subjects have been collected, 23 females and 35 males. The average number of vertices is about 4,000 per frame (or mesh) for 3D dynamic videos, and around 50,000 for 3D models. The dataset is made freely available by request. Comparison with other existing 3D dynamic face datasets is given in Table 2.

**Table 2.** Comparison between existing 3D dynamic face datasets

| Dataset | #subjects | Temporal resolution | Spatial resolution | Illumination conditions | Pose changes |
|---|---|---|---|---|---|
| Posed BU-4DFE [24] | 101 | 25 | 35,000 | controlled | No |
| Spontaneous BU-4DFE [25] | 41 | 25 | 40,000 | controlled | Limited |
| D3DFACS [6] | 10 | 60 | 30,000 | controlled | No |
| Hi4D-ADSIP [13] | 80 | 60 | 20,000 | controlled | No |
| **Our Database** | 58 | 15 | 4,000 | un-controlled | Free |

**Evaluation on the proposed 3D dynamic face dataset** In a first experiment, we considered a subset of 13 subjects, with one of the two neutral sessions used as gallery and four sessions (i.e., neutral, facial expression, talking, and external occlusion) as probes. The goal behind doing these experiments on a small set of our database is to show how the performance of our framework vary according to different settings. This allows us to select the best setting to run on the whole subjects as it is reported in the next experiment. In so doing, we down sampled the 3D scans to $n = 3,500$ vertices, with a radius of $R = 15mm$ for the neighborhood sphere in the 3D normal estimation. The effect of varying the window size $\omega$ used to derive the 3D temporal fragment is explored by repeating the experiment for $\omega = \{5, 10, 15, 20\}$. This value is also used to change the number of eigenvectors which are considered after applying $k$-SVD (i.e., the obtained basis of a subspace). Each session contains about 300 frames, and the number of instances in the gallery differs according to $\omega$ (e.g., for $\omega = 15$ each subject has 20 instances, with 260 total instances in the gallery). In the testing, the four different testing sessions (i.e., Ne, Fe, Tk, Eo) for each subject are divided into multiple instances too. Each instance in the probe sessions is considered as a separate probe and is compared against the 260 instances of a neutral session in the gallery. The Nearest-Neighbor (NN) classifier given in Eq. (1) is applied to find the identity of the probe instances.

Recognition rates are reported in Fig. 5(a), for the four scenarios and the different window size. The best recognition rate is obtained for $\omega = 15$, confirm-

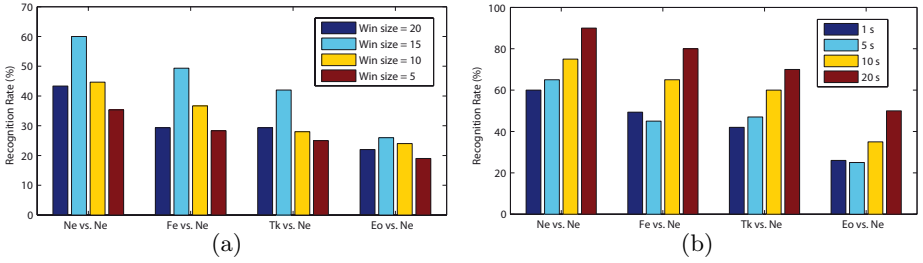


**Fig. 5.** (a) Single-instance based FR results as a function of the window size $\omega$; (b) Multiple-instances based recognition rates

ing the intuition that embodying motion information, coming from the temporal variability on the face surface, provides additional discriminative features for face recognition. For $\omega = 20$ and greater values, this approach scores lower recognition rate due to large pose variations, which make no more possible to track all the vertices from the first to the last frame of a 3D temporal fragment. It also results that the most challenging problem is the occlusion by glasses, where the facial shape is corrupted and presents missing data in the eyes region, as illustrated in Fig. 4(e).

In the previous experiment, only one instance out of the probe session is used for recognition. Due to noise and low resolution of the 3D videos this can be not effective. Using multiple instances of the probe session it is expected to improve the performance. To verify this intuition, the effect on the recognition rate deriving by the application of majority voting is investigated. In this case, the recognition rate is evaluated by combining decisions of multiple successive instances (5, 10 and 20, in addition to the single instance) of the probe session and considering them as one probe. Then, the majority voting is applied to determine which subject in the gallery obtains more votes from the successive probe instances. Figure 5(b) shows the results, where the whole session over 20 seconds and 20 instances are considered ($\omega = 15$ frames per instance are used). It clearly emerges that using more instances as one probe for voting provides better recognition rate in all the investigated scenarios.

**Clustering based face recognition** The matching approach proposed above is based on an exhaustive comparison of the 3D temporal fragments constituting a probe against all the gallery 3D temporal fragments, which results in a time consuming recognition process. To optimize the recognition time, the unsupervised clustering method described in Sect. 3 is applied. The 260 instances in the gallery are clustered into five clusters representing the main poses of the face. Experimentally, this number of classes gives better recognition rate than others. The results of clustering based 3D dynamic FR are reported for both single instance and multiple-instances based methods. In Fig. 6, exhaustive vs. clustering based recognition rates are presented for all the five scenarios. For the solution using a 3D single-instance (i.e., in the figure it corresponds to the required time

of 1$s$, since this includes the 15 3D frames of a fragment), the clustering-based approach gives lower recognition rate than exhaustive search for all the cases, since the gallery neutral sessions are acquired under unconstrained random pose variations. Thus, it is not necessary to find the pose of each probe instance in the gallery session of this subject. Nevertheless, after applying majority voting on multiple-instances based results, the recognition rates start to converge to exhaustive search rates when using 5 and 10 instances for voting. When the whole session, i.e., 20 instances is used as one probe, the clustering based recognition rate overcomes the exhaustive search method for the *Ne* scenario and they are comparable for the other three scenarios (*Fe*, *Tk* and *Eo*). The number of comparisons needed in the *Ne* scenario with clustering is 4 times less than in the exhaustive search method.

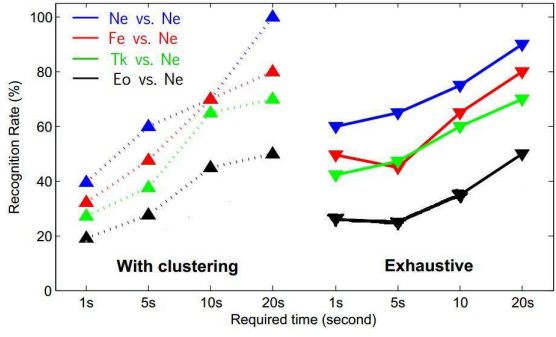


**Fig. 6.** Clustering-based vs. exhaustive recognition results for each test scenario, when the time required for recognition (i.e., number of 3D temporal fragments in majority voting) is varied

Results of these pilot experiments, have been used to set the best parameters for the approach (i.e., window size $\omega$=15, that is 20 instances are used for each subject, with majority voting applied using all the instances). Using this setting, an experiment on all the 58 subjects of the dataset has been conducted. The recognition rate for the four scenarios (*Ne*, *Fe*, *Tk*, and *Eo*) resulted equal to 72%, 62%, 65%, and 36%, respectively. Compared to the results reported in Fig. 5 for the train sample of the dataset, and for the same number of instances used in the majority voting (i.e., 20), it can be observed just a small decrease in the performance for the *Tk* case. A more marked decrease is observed instead in the cases of neutral (*Ne*), facial expressions (*Fe*), and external occlusion (*Eo*).

## 5 Conclusions

In this work, a geometric framework based on Grassmann manifold representation for face recognition is proposed, which exploits the advantages of 3D

dynamic faces. This approach allows us to compare two 3D face videos and to compute statistics (e.g., mean, clustering of a set of 3D face videos). Applying our approach on BU-4DFE [24] database, we have obtained a recognition rate of 92%. The proposed approach does not require to manually annotate landmarks of the face for vertex tracking, and can naturally handle several challenges, like large pose variations, facial expressions, talking and external occlusions. In order to address face recognition in such challenging conditions, a 3D dynamic face recognition database has been also constructed and made publicly available. Single and multiple-instances based recognition results are reported on this new dataset, showing that a majority voting strategy improves the performance in all the scenarios.

## References

1. Barr, J., Bowyer, K., Flynn, P., Biswas, S.: Face recognition from video: a review. International Journal of Pattern Recognition and Artificial Intelligence 26(5) (2012)
2. Berretti, S., Del Bimbo, A., Pala, P.: 3D face recognition using iso-geodesic stripes. IEEE Transaction on Pattern Analysis and Machine Intelligence 32(12), 2162–2177 (2010)
3. Bowyer, K., Chang, K., Flynn, P.: A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition. Computer Vision and Image Understanding 101(1), 1–15 (2006)
4. Chen, Y., Medioni, G.: Object modeling by registration of multiple range images. In: Robotics and Automation. vol. 3, pp. 2724–2729 (1991)
5. Chen, Y.C., Patel, V., Phillips, P., Chellappa, R.: Dictionary-based face recognition from video. In: European Conf. on Computer Vision. pp. 766–779 (2012)
6. Cosker, D., Krumhuber, E., Hilton, A.: A facs valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In: Int. Conf. on Computer Vision. pp. 2296–2303 (2011)
7. Drira, H., Ben Amor, B., Srivastava, A., Daoudi, M., Slama, R.: 3D face recognition under expressions, occlusions, and pose variations. IEEE Transaction on Pattern Analysis and Machine Intelligence 35(9), 2270–2283 (2013)
8. Edelman, A., Arias, T., Smith, S.: The geometry of algorithms with orthogonality constraints. Siam J. Matrix Anal. Appl. 20(2), 303–353 (1998)
9. Golub, G., Van Loan, C.: Matrix computations (3rd edition). Johns Hopkins University Press, Baltimore, MD, USA (1996)
10. Hadid, A., Pietikainen, M.: Manifold learning for video-to-video face recognition. In: Biometric ID Management and Multimodal Communication, vol. 5707, pp. 9–16. Springer (2009)
11. Karcher, H.: Riemannian center of mass and mollifier smoothing. Communications on Pure and Applied Mathematics 30, 509–541 (1977)
12. Li, B., Mian, A., Liu, W., Krishna, A.: Using kinect for face recognition under varying poses, expressions, illumination and disguise. In: Applications of Computer Vision (WACV), 2013 IEEE Workshop on. pp. 186–192 (Jan 2013)
13. Matuszewski, B., Quan, W., Shark, L.k., McLoughlin, A., Lightbody, C., Emsley, H., Watkins, C.: Hi4d-adsip 3D dynamic facial articulation database. Image and Vision Computing 30(10) (2012)

14. Min, R., Choi, J., Medioni, G., Dugelay, J.L.: Real-time 3D face identification from a depth camera. In: Int. Conf. on Pattern Recognition. pp. 1739–1742 (2012)
15. Passalis, G., Perakis, P., Theoharis, T., Kakadiaris, I.: Using facial symmetry to handle pose variations in real-world 3D face recognition. IEEE Transaction on Pattern Analysis and Machine Intelligence 33(10), 1938–1951 (2011)
16. Phillips, P., Flynn, P., Beveridge, J., Scruggs, W., O'Toole, A., Bolme, D., Bowyer, K., Draper, B., Givens, G., Lui, Y., Sahibzada, H., Scallan, J., Weimer, S.: Overview of the multiple biometrics grand challenge. In: IAPR/IEEE Int. Conf. on Biometrics. pp. 705–714 (2009)
17. Phillips, P., Flynn, P., Scruggs, W., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: Int. Conf. on Computer Vision and Pattern Recognition. pp. 947–954 (2005)
18. Rusu, R.: Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments. Ph.D. thesis, Computer Science department, Technische Universitaet Muenchen, Germany (2009)
19. Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: British Machine Vision Conference (2013)
20. Sun, Y., Chen, X., Rosato, M., Yin, L.: Tracking vertex flow and model adaptation for three dimensional spatiotemporal face analysis. IEEE Transaction on Systems, Man, and Cybernetics, Part A 40(3), 461–474 (2010)
21. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Int. Conf. on Computer Vision and Pattern Recognition (2014)
22. Turaga, P., Veeraraghavan, A., Srivastava, A., Chellappa, R.: Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. IEEE Transaction on Pattern Analysis and Machine Intelligence 33(11), 2273–2286 (2011)
23. Wang, H., Kang, B., Kim, D.: Pfw: A face database in the wild for studying face identification and verification in uncontrolled environment. In: IAPR Asian Conf. on Pattern Recognition (ACPR). pp. 356–360 (2013)
24. Yin, L., Chen, X., Sun, Y., Worm, T., Reale, M.: A high-resolution 3D dynamic facial expression database. In: Face and Gesture Recognition. pp. 1–6 (2008)
25. Zhang, X., Yin, L., Cohn, J., Canavan, S., Reale, M., Horowitz, A., Liu, P.: A high-resolution spontaneous 3d dynamic facial expression database. In: 10th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG'10). pp. 1–6 (April 2013)