# A 3D Dynamic Database for Unconstrained Face Recognition

Taleb ALASHKAR[1,*], Boulbaba Ben AMOR[1], Mohamed DAOUDI[1] , Stefano BERRETTI[2]
[1] Télécom Lille/LIFL (UMR CNRS/Lille1 8022), Villeneuve d'Ascq, France ;
[2] University of Florence, Florence, Italy

http://dx.doi.org/10.15221/14.357

## Abstract

In this paper, we present a new 3D dynamic face dataset dedicated to the development and test of algorithms which target face recognition under unconstrained conditions, from 3D videos. Several challenges which can occur in real world like scenarios are considered, such as continuous and freely-pose variation, expressive and talking faces, changes of the distance to the 3D camera, occlusions and multiple persons in the scene. In this database, a full 3D static model is collected for each subject, together with eight 3D video sequences. Each video lasts about 20 seconds, including challenging variations under continuous and freely pose variation. Single-view structured-light 3D scanners are used in the acquisition process. This dataset contains 58 subjects. To provide baseline recognition performance on this database, a 4D-to-4D subspace learning face recognition approach is introduced and experimented. To the best of our knowledge, this database is the first 3D dynamic database designed for the purpose of face recognition considering freely-moving 3D faces. As such, it provides to the research community a new benchmark that can stimulate investigation of face recognition algorithms under new and challenging conditions.

**Keywords:** 4D face, 3D dynamic face, face recognition.

## 1. Introduction

Face recognition (FR) is considered as one of the attractive biometrics for surveillance and access control systems due to its non-intrusive nature. The research filed has started with 2D still images 40 years ago. 2D-video face recognition systems appeared later to exploit the richer information included in videos. In fact, videos contain more viewpoints of the acquired subject [16], besides the dynamic information implied within successive frames [2]. In the last decade, due to the advancement of 3D imaging technologies, 3D face recognition systems appeared to overcome basic challenges in 2D still and dynamic FR systems, like pose variation, illumination and makeup [3, 6]. More recently, a few works have started to consider the fact that the face is a 3D dynamic surface, but mainly for facial expression [13] and action units analysis problems [11]. Up to now, to the best of our knowledge, there is no 3D dynamic face dataset designed specifically for the face recognition problem. All the available 3D dynamic databases are created to address the problem of facial expressions and action units recognition. There are several publicly available face datasets. In [1, 12], a general review of the available 3D databases for Face Recognition and Facial Expression Recognition (FER), respectively, can be found. Because the availability of public standard databases is essential for evaluating and comparing proposed face recognition approaches, we present in this paper a new 3D dynamic face recognition dataset, which includes the most common challenges in this field.

In this work, we present the first 4D FR database collected using single-view 3D scanners with temporal resolution around 15 frames per second. This database can make a real contribution in 4D FR research, especially for non-constrained scenarios. In this database, for each subject, we collected a full 3D static model and eight 3D videos, which are characterized by two neutral sessions, facial expression, talking, walking, internal occlusion by sunglasses, external occlusion by hand or hair, and multiple persons in the scene. All sessions are recorded under freely pose variation using 3D structured-light Artec scanner. A baseline algorithm is designed and evaluated as well, on the constructed dataset. It is based on the 4D-vs-4D face recognition paradigm. In the gallery, 3D videos of moving subjects are enrolled, and the probe consists of a 3D video too. In this approach, each 3D video is considered as a set of subsequences and modeled as a finite-dimensional linear subspace. The rest of this paper is organized as follows: In Sect. 2 the most common 3D dynamic databases are discussed; Sect. 3 describes our 3D dynamic database; A 4D-to-4D face recognition approach is presented in Sect. 4, which is used to provide a baseline performance evaluation of the database; the experimental results are reported in Sect. 5; conclusions and future work are discussed in Sect. 6.

* taleb.alashkar@telecom-lille.fr

## 2. Existing 3D Face Dynamic Datasets

In recent years, several facial 3D dynamic databases have been introduced to the community. All of them share two common points: (1) they are collected under highly conditioned environments using stereo systems; (2) none of them has been created for face recognition applications, but for facial expressions or action units issues. The main existing 3D dynamic databases and their main features are summarized below.

The **Bu-4DFE** dataset, by Yin et al. [14] is the first database consisting of 4D faces (sequences of 3D faces). The database includes 101 subjects and was created using the DI3D (Dimensional Imaging) dynamic face capturing system. It contains sequences of the six prototypical facial expressions with their temporal segments (neutral-onset-apex-offset-neutral) with each sequence lasting approximately 4 seconds. The temporal and spatial resolution is 25 fps and 35,000 vertices, respectively. The main limits of this database is that it contains posed facial expressions, and restricted acquisition environment (well-controlled illumination and frontal view of the subject's face), which make it far from real scenarios. Zhang et al. [15] created a High Resolution Spontaneous 3D Dynamic FE Database, the **Spontaneous BU-4DFE**. Also, for this dataset, the DI3D system was used for acquisition, but the expressions are not posed, instead they are spontaneously conveyed by the participants. Expressions include happiness or amusement, sadness, surprise, embarrassment, fear or nervous, physical pain, anger or upset and disgust. There are 41 participants in this database. For each subject, 3D and 2D videos lasting about one minute for each scenario are captured. Manually annotated action units (FACS AU) by a certified FACs coders, automatically tracked facial landmarks and head pose in 3D/2D videos are provided with the database as a metadata.

Cosker et al. [5] presented the first database that contains coded examples of dynamic 3D Action Units (AUs) in **D3DFACS**. There are 10 subjects in this dataset including 4 FACS experts asked to perform 38 AUs in various combinations. Totally, there are 519 AUs sessions at 60 fps temporal resolution. Each action unit consisting of 90 frames approximately. A FACS expert coded the peak of each sequence. The 3dMD Face Dynamic system was used for capturing this database [8]. It is more oriented for AUs recognition, captured under highly conditioned framework with posed facial expressions.

The **Hi4D-ADSIP** database, presented by Matuszewski et al. [9] is a 3D dynamic facial database containing facial articulation. The temporal resolution, 60 fps, and the spatial resolution, 2352 × 1728 pixels per frame are quiet high recorded using the Di3D. Totally, there are 80 subjects in this dataset with 3360 sequences. Subjects have various ages, genders and races. The seven basic facial expressions are included with seven facial articulations. The main reason to include these articulations is to support the clinical research on facial dysfunctions. Facial expression recognition algorithms are applied to validate the part of the database containing standard facial expressions. Two different algorithms in static and dynamic mode are applied to it. In addition, a psychophysical experiment that was used to formally evaluate the accuracy of recorded expressions is conducted. Table 1 presents a comparison between existing dynamic 3D face datasets and the dynamic part of our 3D-4D database.

*Table 1. Comparison between our database and other 4D Face databases in the community.*

| Database | # subjects | Temporal Resolution | Spatial Resolution | Illumination condition | Pose variation |
|---|---|---|---|---|---|
| Posed Bu4DFE [1] | 101 | 25 | 35,000 | Controlled | No |
| Spontaneous BU-4DFE [2] | 41 | 25 | 40,000 | Controlled | Limited |
| D3DFACS [3] | 10 | 60 | 30,000 | Controlled | No |
| Hi4DADSIP [4] | 80 | 60 | 20,000 | Controlled | No |
| Our 4D Database | 58 | 15 | 4,000 | Un-controlled | Free |

# 3. 4D/3D Dataset Description

A new 3D-4D database is created in order to allow development and performance evaluation of algorithms that address the problem of FR using 3D dynamic sequences data in non-cooperative scenarios. The key point of this work is that we propose a 4D database captured via a single view 3D scanner under unconstrained conditions. This database consists of two parts: first, the full 3D part, that contains a full 3D static face model of high resolution for each subject with texture information; second, the 4D part, that contains eight 3D videos with geometrical information and without texture for each subject. While the ***Artec MHT*** 3D scanner is used to acquire the 3D models, the ***Artec L*** 3D scanner, with larger field of view, is used to record 3D videos. The temporal resolution of the scanners is 15 frames per second. In Table 2, there is a detailed comparison between the two Artec scanners used. To the best of our knowledge, all current available 3D dynamic databases have been collected using view systems.

## 3.1. Full 3D Static Textured Model

To capture a full 3D static model for each subject in our database we need: 1) Operator: the person who is performing the scan; 2) the participant; 3) some utilities, like creeper seat, headband. The acquisition protocol is as follows. The subject should take off any kind of accessories, like glasses, hat, or anything may hide any part of the face. The subject puts on the headband if necessary to avoid hair occlusion of the front; The subject sits on the creeper seat, standing on his left-facing profile; The operator projects the MTH scanner from above viewpoint to take the upper part of the face. Distance between the scanner and the participant face is about 60 cm; After triggering the scanner, the subject is asked to rotate slowly around himself on the creeper seat anticlockwise $180^0$, then he/she should stop. It takes about 40 sec; The operator projects the MTH scanner from downside viewpoint in order to get the lower part of the face; The subject is now asked to rotate slowly clockwise on the creeper seat $180^0$ returning to the first position; The operator ends the scanning procedure. The acquisition procedure takes about 2 minutes. After that, obtained scans are post-processed via the Artec Studio Software to delete non-informative parts like small objects. Then, other post-processing steps are applied like registration, smoothing, holes filling, meshes simplification, mapping texture information to 3D full static model. Figure 1 illustrates an example of a resulted full 3D model.

## 3.2. 3D Dynamic Sequences

The 3D dynamic sequences have been captured using the ***Artec L*** single-view structured light 3D scanner, because its wide angle of view. The temporal resolution is $15\ fps$ and 3D resolution is up to $1.0\ mm$ without texture information. To collect a realistic database under non cooperative scenarios, seven different cases have been registered for each subject.

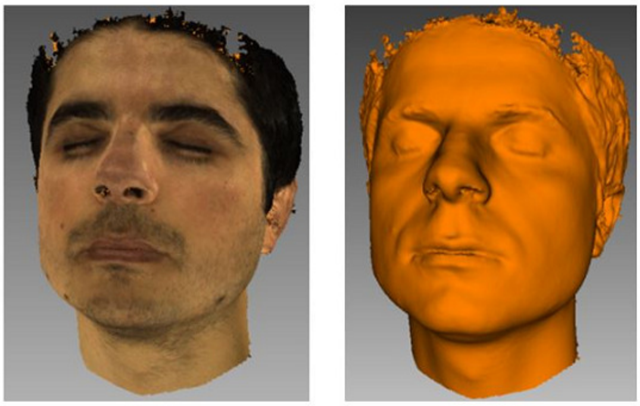

*Figure 1. A full 3D static model with and without texture*

Each one lasts for about 20 sec (i.e., 300 frames) under free pose variation. The acquired sequences are as follows:

- **Neutral (Ne) –** two 3D videos without any facial expression or any kind of occlusion under continuous free pose variation.
- **Facial Expression (Fe) –** the subject is asked to act more than one facial expression randomly with free pose variation.
- **Talking (Tk) –** the subject is asked to talk and move the head freely.
- **Internal Occlusion (hand or hair) (Io) –** part of the face of up to 40% is occluded via hand or by hair under pose variation.
- **External Occlusion (Eo) –** part of the face is occluded by sun glasses or a scarf which cover part of the face with pose variation.
- **Walking (Wk) –** the subject starts walking from a point far about 120 cm from the scanner and getting closer to the scanner up to 60 cm with free pose variation. The number of points is increasing in the frames of this scenario by approaching the scanner.
- **Multiple persons (MP) –** the subject is acquired with another person in the scene making a dialogue or acting facial expressions under pose variation.

The only preprocessing step is cropping the face region out of the whole scan to delete unwanted parts of the body. Figure 2 shows key frames from each 4D session for one subject. So far, 3D sessions for 58 subjects have been collected.

*Table 2. Comparison between Artec MHT and Artec L 3D Scanners.*

| | **Artec MHT 3D Scanner** | **Artec L 3D Scanner** |
|---|---|---|
| Ability to capture texture | Yes | No |
| 3D resolution | 0.5 mm | 1.0 mm |
| 3D point accuracy | 0.1 mm | 0.2 mm |
| 3D accuracy over 100 cm distance | 15% | 15% |
| Working distance | 0.4 to 1 m | 0.8 to 1.6 m |
| Angular field of view, HW | 30 × 21° | 41 × 32° |
| Video frame rate | 15 fps | 15 fps |

### 3.3. Data organization

This dataset is organized in such a way that a separate folder is created for each subject. They contain the full 3D static model with texture information and eight 3D dynamic videos for seven evaluation scenarios. Most of the participants in our database are students. The average age is about 23 years old from different ethnic groups. Up to now, 58 subjects have been collected, 23 females and 35 males. The average number of vertices is about 3,500 per frame (or mesh) for 3D dynamic videos, and around 50,000 for 3D models. The 3D frames are exported to PLY 3D file format. The dataset is made available by request.

## 4. 4D-vs-4D FR Baseline Algorithm

In this section, a 4D-to-4D face recognition baseline algorithm is proposed, which can operate under unconstrained conditions. This approach is designed to exploit the spatio-temporal information available in 3D dynamic sequences of the face. To this end, a subspace modeling method is applied. The basic idea of this solution is to extract a set of 4D fragments from each 4D sequence and model each fragment f as a linear subspace $P_f$ which is an element on a Grassmann manifold. According to this, suppose that there are $N$ **4D** fragments indexed by $i$ as gallery: $\mathrm{G} = \{\mathrm{g_i}; (\mathrm{i} = 1 \ldots \mathrm{N})\}$, and a probe input **4D** fragment with $m$ successive frames: $F_{input} = [f_1, \ldots, f_m]$. The recognition process can be formulated as follows:

$$g^* = \mathrm{arg_{min}}\ \mathrm{d}\left(P_{\mathrm{F_{input}}}, P_{\mathrm{gi}}\right) \qquad (1)$$

where d(.,.) denotes the geodesic distance between two linear subspaces, and $g^*$ is the gallery subject whose identity is recognized as corresponding to the probe fragment. The subspace modeling process is performed in several steps, as detailed below.

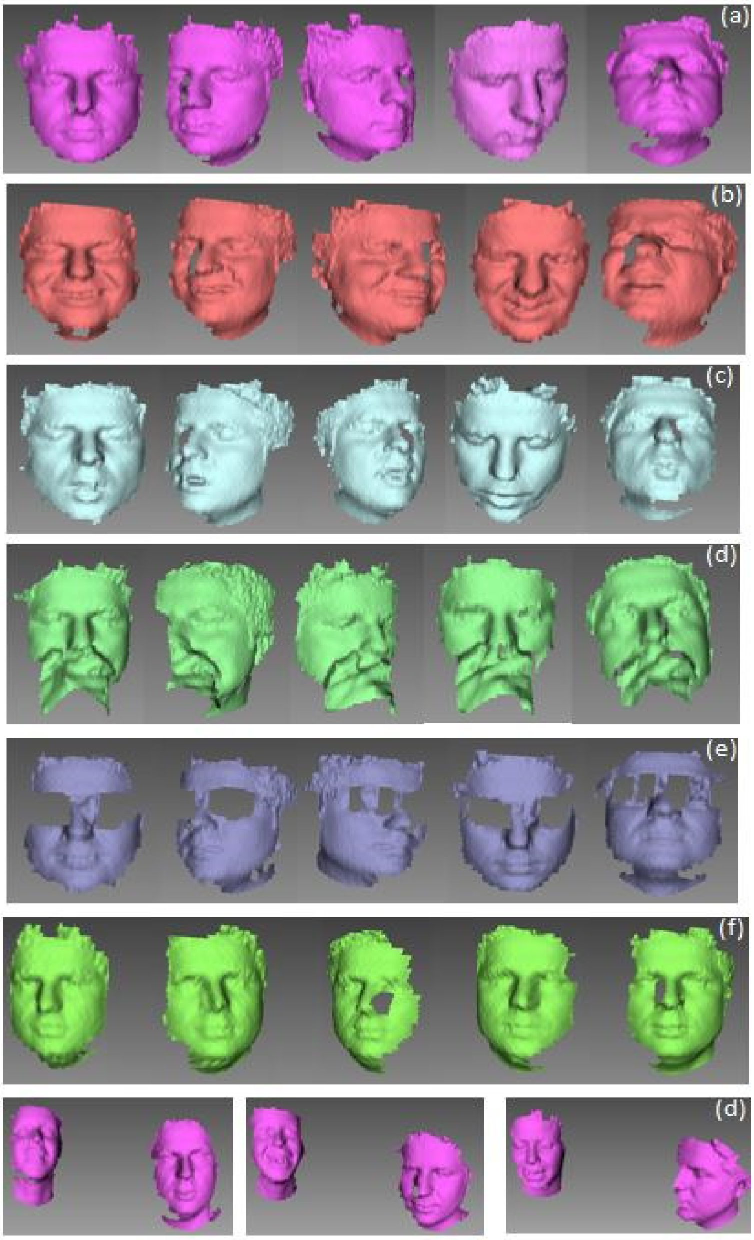

*Figure 2. The 4D sessions acquired under different conditions: (a) neutral; (b) expressive; (c) talking; (d) internal occlusion by hand; (e) external occlusion by sunglasses; (f) walking and (g) multiple persons.*

## 4.1. Overview of the 4D FR baseline

The outline of the proposed baseline algorithm is illustrated in Figure 3. After the 4D acquisition step, the face region of each frame in a 4D sequence is cropped. Due to pose variations and the properties of the 3D scanner used for acquisition, 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, having the same number of vertices representing the face in each frame of a sequence is important. To this end, a down-sampling function is applied to each frame to produce a constant number of vertices denoted by $n$. Then, at each vertex the normal is estimated directly from the point cloud. The surface normal estimation is based on neighborhood vertices in a sphere of radius $R$ around the vertex [10]. The set of estimated normals at the vertices of each frame captures the shape of the face, which is used as a spatial holistic descriptor of the face surface. However, frames constituting the 4D sequences do not have a correspondence between their respective vertices, which is indeed necessary to develop for 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.

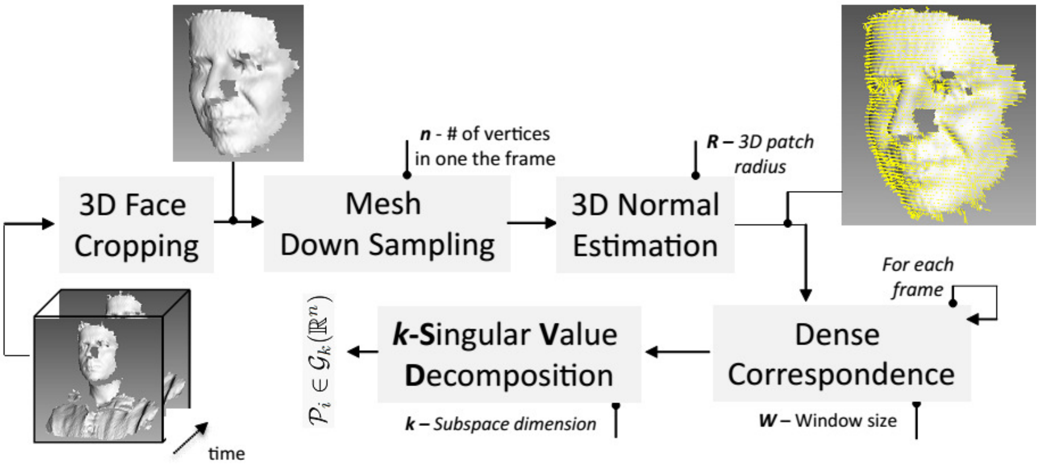

*Figure 3. Overview of the baseline 4D FR algorithm*

As a result of this process, each 4D fragment can be modeled as a matrix $S$ of size $n \times \omega$, where $n$ is the number of vertices, and $\omega$ is the number of frames in the 4D fragment. Each column of $S$ is a vectorization of the $z$ component of the estimated normals at each vertex of one frame, and each row embodies the motion information originated from the variability over time of one vertex of the face surface. Finally, a $k$-Singular Value Decomposition of the obtained matrix is performed $S = U\Sigma V^t$. The eigenvectors matrix is an orthonormal basis of the subspace $P = Span(U)$, which is an element on the Grassman manifold $G_k(R^n)$. As a result of this pipeline, each 4D fragment is viewed as an element of the Grassmannian, and the original problem of 4D-to-4D matching in turned into a distance measurement over the Grassmannian between the correspondent elements.

To measure the similarity between two 4D fragments, which every 4D fragment is represented by a linear subspace, Golub and Loan [8] introduced an intuitive and computationally efficient way of defining the distance between two linear subspaces $X, Y$ using the principal angles. This distance can be calculated numerically [7] as in the next equation:

$$d(X,Y) = \Sigma_i \theta_i^2 \qquad (2)$$

### 4.2. Multiple instances-based recognition

In the acquired 4D sequences, the combination of free pose variations of the subjects, and the use of a single-view 3D scanner for acquisition, results in many frames with missing parts of the face due to self-occlusions. As a consequence, it is hard to find correspondence and track vertices within successive frames. 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 4D fragment, represents approximately one pose of the moving face. As a result, each subject in the gallery and the probe is represented by multiple instances. Applying majority voting techniques on the sub-sequences of any probe session will improve the recognition rate by the time.

## 5. Experiment and Evaluation

To validate our database, we applied the proposed 4D-to-4D framework to recognize faces on the 58 subjects. The first neutral session is considered as a gallery, and four scenarios (*Ne, Fe, Tk, Eo*) are tested separately as a probe session in this experiment. The following parameters are empirically setup to these values – threshold of downsampling step: $n = 3500$ vertices per 3D frame; radius of considered neighborhood disk for 3D normal estimation: $R = 15\,mm$ the window size has been posed as $\omega = 15$, that is 20 instances are used for each subject. (these values are chosen after experimental analysis). The majority voting is applied over time using all the instances. The recognition rate for the four scenarios (*Ne, Fe, Tk, and Eo*) resulted equal to 72%; 62%; 65%; 36%, respectively. Figure 4 demonstrates how the recognition rate increases by time along the video for the four scenarios (*Ne, Fe, Tk, Eo*). The recognition rate is calculated and updated after each new sub-sequence arrives along the video considering all previous ones.

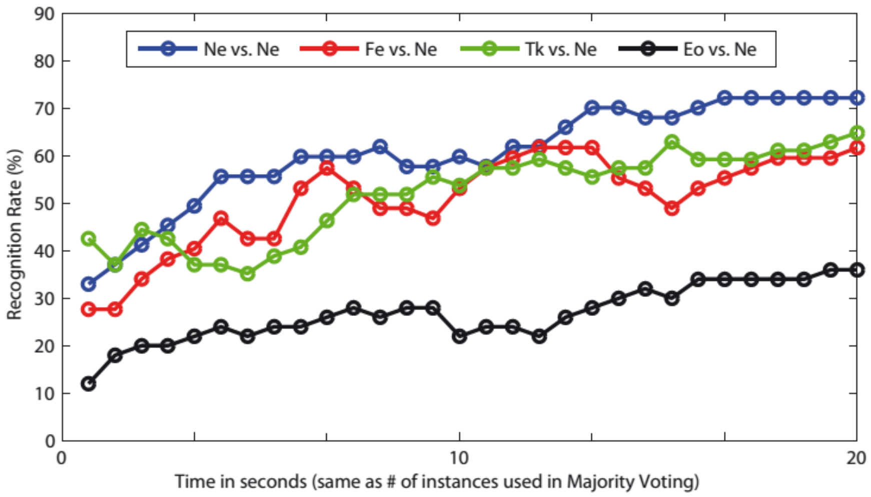

*Figure 4. Trade-off between accuracy and latency for the 4D face recognition baseline algorithm*

## 6. Discussion and Conclusions

In this paper, a 3D dynamic face recognition database is introduced. This database is the first one designed to include most face recognition challenges in realistic scenarios using 3D face sequences. Overall, $58$ subjects are collected with a 3D full static model and eight 3D videos each using structured-light single view scanners. Each video contains one or more variations, like large pose variation, facial expression, talking, internal and external occlusion, walking and multiple persons. The fact that a single view scanner is used for acquiring 3D videos makes it closer to realistic scenarios. The low spatial resolution, about $3500$ vertices per frame, and with $15\,fps$ as temporal resolution make it a more challenging database. This database will be made freely available to the 3D face recognition community, with the aim to provide new challenging scenarios for the development and test of a new generation of face recognition approaches capable to work in 3D real-world conditions. A baseline 4D-to-4D approach is presented to validate the performance of this challenging new database.

## References

[1] C. Anitha et al. "A survey on facial expression databases". Int. Journal of Engineering Science and Technology, 2(10):5158–5174, 2010.

[2] J. R. Barr et al. "Face recognition from video: a review". Int. Journal of Pattern Recognition and Artificial Intelligence, 26(5), 2012.
http://dx.doi.org/10.1142/S0218001412660024

[3] S. Berretti et al. "3D face recognition using isogeodesic stripes". IEEE Trans. on Pattern Analysis and Machine Intelligence, 32(12):2162–2177, 2010.
http://dx.doi.org/10.1109/TPAMI.2010.43

[4] Y. Chen and G. Medioni. "Object modeling by registration of multiple range images". In Robotics and Automation, volume 3, pages 2724–2729, 1991.
http://dx.doi.org/10.1109/ROBOT.1991.132043

[5] D. Cosker et al. "A facs valid 3D dynamic action unit database with applications to 3d dynamic morphable facial modeling". In IEEE Int. Conf. on Computer Vision (ICCV), pages 2296–2303, Nov 2011. http://dx.doi.org/10.1109/ICCV.2011.6126510

[6] H. Drira et al, "3D face recognition under expressions, occlusions, and pose variations". IEEE Trans. Pattern Analysis and Machine Intelligence, 35(9):2270–2283, 2013. http://dx.doi.org/10.1109/TPAMI.2013.48

[7] A. Edelman et al. "The geometry of algorithms with orthogonality constraints". SIAM Journal on Matrix Analysis and Applications, 20(2):303–353, 1998.

[8] G. H. Golub and C. F. Van Loan. "Matrix computations" (3rd edition). Johns Hopkins University Press, Baltimore, MD, USA,1996.

[9] B. J. Matuszewski et al. "Hi4dadsip 3-d dynamic facial articulation database". Image and Vision Computing, 30(10), 2012. http://dx.doi.org/10.1016/j.imavis.2012.02.002

[10] R. B. Rusu. "Semantic 3D Object Maps for Everyday Manipulation in Human Living Environments". PhD thesis, Computer Science department, Technische Universitaet Muenchen, Germany, 2009.
http://dx.doi.org/10.1007/s13218-010-0059-6

[11] G. Sandbach et al. "Local normal binary patterns for 3d facial action unit detection". In 19th IEEE Int. Conf. on Image Processing (ICIP), pages 1813–1816, Sept. 2012. http://dx.doi.org/10.1109/ICIP.2012.6467234

[12] G. Sandbach et al. "Static and dynamic 3d facial expression recognition: A comprehensive survey". Image and Vision Computing, 30(10):683–697, Oct.2012. http://dx.doi.org/10.1016/j.imavis.2012.06.005

[13] Y. Sun et al. "Tracking vertex flow and model adaptation for three-dimensional spatiotemporal face analysis". IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans, 40(3):461–474, May 2010.
http://dx.doi.org/10.1109/TSMCA.2010.2041659

[14] L. Yin et al. "A high-resolution 3d dynamic facial expression database". In 8th IEEE Int. Conf. on Automatic Face Gesture Recognition (FG'08), pages 1–6, Sept. 2008.
http://dx.doi.org/10.1109/AFGR.2008.4813324

[15] X. Zhang et al. " BP4D-Spontaneous: A high-resolution spontaneous 3D dynamic facial expression database". In Image and Computer Vision, July-2014.
http://dx.doi.org/10.1016/j.imavis.2014.06.002

[16] W. Zhao et al. "Face recognition: A literature survey". ACM Computing Surveys, 35(4):399–458, 2003.
http://dx.doi.org/10.1145/954339.954342