Representation, Analysis and Recognition of 3D Humans: A Survey

Stefano Berretti, Mohamed Daoudi, Pavan Turaga, Anup Basu

Representation, Analysis and Recognition of 3D Humans: A Survey

STEFANO BERRETTI, University of Florence, Italy
MOHAMED DAOUDI, IMT Lille Douai, France
PAVAN TURAGA, Arizona State University, USA
ANUP BASU, University of Alberta, Canada

Computer Vision and Multimedia solutions are now offering an increasing number of applications ready for use by end users in everyday life. Many of these applications are centered of detection, representation, and analysis of face and body. Methods based on 2D images and videos are the most widespread, but there is a recent trend that successfully extends the study to 3D human data as acquired by a new generation of 3D acquisition devices. Based on these premises, in this survey, we provide an overview on the newly designed techniques that exploit 3D human data and also prospect the most promising current and future research directions. In particular, we first propose a taxonomy of the representation methods, distinguishing between spatial and temporal modeling of the data. Then, we focus on the analysis and recognition of 3D humans from 3D static and dynamic data, considering many applications for body and face.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Computing methodologies** → **Computer vision**; **Computer vision tasks**; **Computer vision problems**; *3D imaging*; *Shape representations*;

Additional Key Words and Phrases: 3D humans, 3D shape representation, 3D face and body representation, 3D face and body analysis and retrieval

ACM Reference Format:
Stefano Berretti, Mohamed Daoudi, Pavan Turaga, and Anup Basu. 2018. Representation, Analysis and Recognition of 3D Humans: A Survey. *ACM Trans. Multimedia Comput. Commun. Appl.* 0, 0, Article 1 (2018), 35 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

The idea of automatically discovering and representing the characteristics of the human body and face has been of interest in several areas of Computer Science for many years. For example, in Computer Graphics, modeling and animating human characters is central in games and movie industries; in Computer Vision, human action detection and identity recognition are at the base of many monitoring and surveillance applications; in Multimedia, understanding human body posture and gesture or facial expression is fundamental for developing advanced human-machine interfaces and innovative interaction modalities.

Authors' addresses: Stefano Berretti, University of Florence, Department of Information Engineering, Florence, 50139, Italy; Mohamed Daoudi, IMT Lille Douai, University of Lille, CNRS, UMR 9189 - CRIStAL - Centre de Recherche en Informatique Signal et Automatique de Lille, Lille, F-59650, France; Pavan Turaga, Arizona State University, Tempe, Arizona, USA; Anup Basu, University of Alberta, Edmonton, Alberta, Canada.

1551-6857/2018/0-ART1 $15.00
https://doi.org/0000001.0000001

For many years, 2D still images and videos have been used as the only sources to investigate methods for detecting, representing and analyzing human body and face [176]. Now, the interest for *non-Euclidean* data is growing. First attempts that tried to move from representations of 2D to 3D humans investigated how to automatically extract compact descriptors of the body and the face, mostly using synthetic models generated by dedicated software tools. The use of such 3D modeling tools required experienced operators, was time consuming, and resulted in limited realism. In addition, operating with synthetic data, hid most of the difficulties associated with the manipulation of the actual data acquired. Indeed, only recently the advent of new 3D acquisition technologies at affordable cost, including consumer depth cameras like Kinect, has made it possible to capture real human body and face in 3D. Such 3D scanners can be either static or dynamic (i.e., across time), with resulting scans obtained at high- or low-resolution. In the last few years, this has also allowed the production of large repositories of human samples, which has opened the way to substantial research advancements and new application domains.

Several surveys exist on 3D methods, but they are more focused on individual tasks, such as 3D face recognition [121, 152], 3D emotion and expression recognition [42, 139], 3D action recognition [95], and 3D retrieval [156]. Instead, our effort here is to provide a comprehensive and updated overview of what have been done in tasks that have the humans, body and face as the main focus of analysis or recognition. The usual framework for analysis and recognition of 3D body and face comprises the following steps. 3D data is typically noisy and irregularly formed so that some preprocessing is first applied. Then, a representation is built on lower-level descriptors that model the information embedded in the data. At large, the contra-position is between *hand-crafted* and *learned* features: these can capture the spatial information only or also account for the temporal dimension. Finally, such representations are the input for a classification stage that can rely on some classifier or be integrated into a (deep) learning framework.

In the following, we start by focusing on the representations (see Section 2) by distinguishing between methods that perform *spatial* or *temporal* modeling of human data (in Section 2.1 and Section 2.2, respectively). The former extract the representation by exploiting the data acquired as individual 3D scans, usually captured at high resolution with user cooperation. The latter also accounts for the temporal component in dynamic data (i.e., sequences of 3D scans acquired with 3D cameras). Differing from the static case, these scans can be acquired without user cooperation but, typically, at the cost of lower resolution. Spatial and temporal representations have been used in a variety of analysis and recognition tasks (see Section 3). In general, these applications are different for face and body so we will present them separately in Section 3.1 and Section 3.2, respectively.

2 REPRESENTATIONS OF 3D HUMANS

Representations of the 3D human body and face are usually built on low level descriptors that model static (spatial) and dynamic (spatio-temporal) data for extracting meaningful features. We summarize some of the most successful 3D descriptors, and outline how they have been used in the modeling of body and face. We organize in a taxonomy the representation methods proposed so far in the literature as illustrated in Figure 1. We further classify the methods based on the main characteristics of the 3D shape they capture. In particular, in the case of spatial modeling, we distinguish between representation methods in the following categories:

- **geometric**: account for the surface shape either in an *extrinsic*, or *intrinsic* way;
- **volumetric**: the volume delimited by the shape surface is accounted for by the representation;
- **topological**: topological variations of the shape are captured by the representation;
- **landmarks based**: the shape is represented by a set of landmarks (or fiducial points) with some local surface description attached to them.

An increasing importance is also assumed by learning solutions, where hand-crafted descriptors are substituted by *deep features* that are learned directly from the data. These deep features can be learned from point sets (and surfaces), landmarks and volumes.

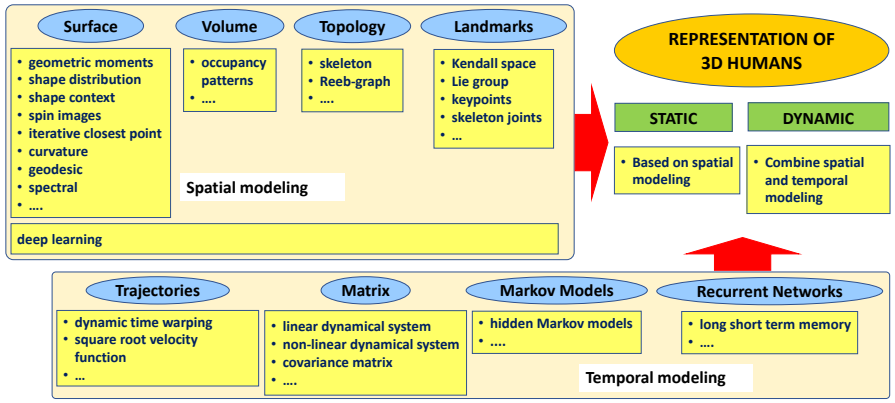


Fig. 1. Representations of 3D humans can be either *spatial* or *temporal*. The former, captures properties of the surface (or its point cloud), of the volume, or of some landmarks of the body and face; the latter, accounts for the dynamic evolution across the time of static descriptors.

In the case of dynamic acquisitions, independent of the final goal, methods have to address two tasks: *i)* representing the human subject (body or face) in a single frame, and *ii)* modeling the temporal evolution of the human subject across several frames. These tasks can be solved separately by addressing the the spatial (by frame) and temporal (across frames) representations in an independent way, or designing a combined solution. In the first case, methods in the same categories listed above for the static case can be applied to the frames individually. The frame-based individual representations can then be related together to form a *trajectory* across the temporal dimension. Combined solutions, instead, typically resort to *matrix* representations that incorporate both spatial and temporal dimensions. The existence of these two general approaches suggests to us the following categorization for dynamic methods:

- **trajectory based**: the idea of such solutions is to represent the temporal evolution of descriptors extracted from individual scans of a sequence as a trajectory in a multidimensional space and then analyze such trajectories;
- **matrix based**: a matrix representation is extracted from the overall sequence or its parts, like covariance or Hankel matrices. These matrices usually lie on special non-linear manifolds, whose geometry should be accounted for in computing distances and statistical operations;
- **Markov models**: there are cases where the dynamics can be modeled as a Markovian process and transitions between states can be used to model complex temporal events;
- **recurrent neural networks**: this category includes methods that learn dynamical patterns from the temporal data, rather than starting from predefined models. For example, this includes recent successful methods like the long-short term memory network.

2.1 Spatial Modeling

Initial solutions for representing 3D bodies and faces addressed static data and so they were required to model the spatial information only. Typically, these data were acquired with high-resolution scanners or synthetically generated by modeling softwares. Data can be in the form of point clouds, depth images (2.5D) or given as triangulated mesh manifolds.

2.1.1 Geometric (surface based) Representations. One intuitive idea for representing the human body and face is that of extracting information from the model's surface. The surface can be considered either as piecewise linear, if a mesh triangulation is used, or approximated through a continuous function. One possible way to group solutions for surface representation is that of separating them depending on whether they use *extrinsic* or *intrinsic* measurements. Extrinsic quantities are defined using surface normal vectors and/or surface embedding into a 3D space, while intrinsic quantities can be expressed exclusively in terms of distances computed on the surface.

Extrinsic Methods. The term "extrinsic" is used here in a way similar to that employed in differential geometry, i.e., we refer to methods that require the knowledge of the space embedding the surface ($\mathbb{R}^3$) for computing the shape representation.

For example, the work of Pozo et al. [126] introduced a fast exact algorithm and a series of faster approximate algorithms for the computation of 3D *geometric moments* from an unstructured surface mesh of triangles. Being based on the object surface, it reduces the computational complexity with respect to volumetric grid-based algorithms. In contrast, it can only be applied for the computation of geometric moments of homogeneous objects. This advantage and restriction is shared with other proposed algorithms based on the object boundary. In [116], Osada et al. proposed the idea of representing the signature of an object as a generalization of geometric histograms called *shape distribution* (SD). This is sampled from a *shape function* measuring global geometric properties of the object. For instance, one example SD, which is called *D2*, represents the distribution of Euclidean distances between pairs of randomly selected points on the surface of a 3D model. Intrinsic versions have been also proposed [12]. Thus, the challenges of this approach are to select discriminating shape functions, to develop efficient methods for sampling them, and to robustly compute the dissimilarity of probability distributions. Koertgen et al. [85] first extended the *Shape Context* (SC) descriptor, originally defined for 2D shapes, to the 3D domain. The representation of a 3D shape is a set of N histograms corresponding to N points sampled from the shape boundary. The surface of the shape is sampled with roughly uniform spacing adapting the SD method. Then, the set of vectors originating from one sample point to all other points in the shape is considered. These vectors express the appearance of the entire shape relative to the reference point. For a point P on the shape, a coarse histogram of the relative coordinates of the remaining N-1 points is computed. This histogram is defined as the SC of P. In general, histograms are based on a partitioning of the space in which the object reside, i.e., a complete and disjoint decomposition into cells, which correspond to the bins of the histograms. The SC at a point captures the distribution over relative positions of other shape points and thus summarizes global shape in a rich, local descriptor.

Johnson and Hebert [76] proposed a shape descriptor called *spin image* that is used to match surfaces represented as meshes. A spin image is created for an oriented point at a vertex in the surface mesh as follows. A 2D accumulator indexed by a and b is created. Next, the coordinates (a, b) are computed for a vertex in the surface mesh that is within the support of the spin image. The bin indexed by (a, b) in the accumulator is then incremented. This procedure is repeated for all vertices within the support of the spin image: the support distance corresponds to image width times the bin size and determines the amount of space swept out by a spin image; the support angle is the maximum angle between the direction of the oriented point basis of a spin image and the surface normal of points that are allowed to contribute to the spin image. The resulting accumulator can be thought of as an image; dark areas in the image correspond to bins that contain many projected points. Spin images have been defined as a general 3D shape descriptor, and found application in 3D facial normalization, and 3D facial landmarks localization.

In [23], Besl and McKay described a general-purpose method for the registration of 3D shapes, which is independent from the representation including free-form curves and surfaces. The method

handles the full six degrees of freedom and is based on the *Iterative Closest Point* (ICP) algorithm, which requires only a procedure to find the closest point on a geometric entity to a given point. The ICP algorithm always converges monotonically to the nearest local minimum of a mean-square distance metric, and the rate of convergence is rapid during the first few iterations. Therefore, given an adequate set of initial rotations and translations for a particular class of objects with a certain level of "shape complexity", one can globally minimize the mean-square distance metric over all six degrees of freedom by testing each initial registration. For example, Faltemier et al. [52] applied ICP to different parts of the face to perform recognition from 3D scans.

Intrinsic Methods. Kokkinos et al. [83] proposed the intrinsic shape context (ISC) descriptor for 3D shape surfaces as generalization of the 2D SC descriptor. To this end they generalized to surfaces the polar sampling of the image domain used in SC: they charted the surface by shooting geodesic outwards from the point being analyzed; "angle" was treated as equivalent to geodesic shooting direction, and radius as geodesic distance. The resulting charting method is intrinsic, and the resulting descriptor is a meta-descriptor that can be applied to any photometric or geometric property field defined on the shape. Another class of intrinsic similarity methods is based on the analysis of spectral properties of the Laplace-Beltrami operator. Reuter et al. [134] used Laplace-Beltrami spectra, referring to them as "shape DNA", for the characterization of surfaces. Since the Laplace-Beltrami operator is an intrinsic characteristic of the surface, it is insensitive to isometric deformations. Intrinsic similarity criteria based on spectral analysis are intimately related to methods used in manifold learning and data analysis. There are also methods that aim to guarantee invariance of the representation to isometric deformations. As the precursor of a recent trend in non-rigid shape analysis, we can consider the paper of Schwartz et al. [142], in which a method for the representation of the intrinsic geometry of the cortical surface of the brain using multidimensional scaling (MDS) was presented. MDS is a family of algorithms commonly used in dimensionality reduction and data analysis and graph representation. The idea of Schwartz et al. was extended by Elad and Kimmel [50], who proposed a non-rigid shape recognition method based on Euclidean embeddings. The key idea is to map the metric structure of the surfaces to a low-dimensional Euclidean space and compare the resulting images (called canonical forms) in this space. Grossman et al. [58] extended this method to objects represented as voxels. Applied to the problem of 3D face recognition, the canonical forms method proved to be very efficient in being able to recognize identity of people, while being insensitive to their facial expressions [31]. Ling and Jacobs used this method for recognition of articulated 2D shapes [97].

Other intrinsic solutions are based on surface *curvature* and *geodesics* computation, though extrinsic solutions have been also developed from them.

Curvature. Surface curvature provides a low level descriptor of how the surface deviates from a planar one, which can be used for deriving higher-level representations of the body or face. Though curvature is descriptive of the local shape of a surface, computing stable curvature values for real scans is a difficult task. One reason for this is the fact 3D scans are typically obtained as point clouds, while a mesh is then derived from the points trough a triangulation process. As a consequence, computed curvatures are sensitive to noisy points and different triangulations, thus making the final values not robust. The need for overcoming such difficulties inspired several methods for surface curvature computation [106, 157]. For example, Colombo et al. [40] were among the first to focus on the task of 3D face detection by analyzing the surface curvature. In their solution, signs of the mean and the Gaussian curvature were analyzed to obtain a concise description of the local behavior of the surface and detect the location of eyes and nose. The face was initially represented by a range image of points. For each point (x_i, y_i) on the grid, a bi-quadratic polynomial

approximation of the surface was considered. The coefficients of the polynomial were obtained by least squares fitting of the points in a neighborhood of (x_i, y_i). The derivatives of the surface in (x_i, y_i) were then estimated by the derivatives of the approximating polynomial.

To derive surface representations, statistics of the curvature values are usually computed, like local histograms. For example, Antini et al. [11] proposed curvature correlograms to capture the spatial distance of curvature values in 3D models retrieval. In their work, the mean curvature was computed according to the Taubin's approach [157]. Inspired by the success of Local Binary Patterns (LBP) in 2D face analysis, Werghi et al. [177] addressed the challenge of computing LBP on a mesh manifold. They proposed an original computational framework, called *mesh-LBP* that allows the extraction of LBP-like patterns directly from a triangular mesh manifold, without the need of any intermediate representation in the form of depth images. With this framework, it is possible to build on 2D-LBP analysis methods, extending them to mesh manifolds as well as to photometric information embedded into mesh models. Among the surface descriptors used for computing mesh-LBP, they considered different curvatures (mean, Gaussian) and obtained good results in the 3D face recognition task [178].

Geodesic. The idea of deriving representations from distance measures computed on the surface is attractive since these measures can preserve metric properties of the surface. In the usual case, where the surface is represented by a triangular mesh, the problem of computing the geodesic distance between two vertices of the mesh is solved mainly using two very well known methods: the *shortest path* algorithm, and the *fast marching* algorithm [143]. The former is easy to compute with the Dijkstra's algorithm, but approximates the true geodesic with the shortest path computed on the edges of the adjacency graph connecting the mesh vertices. The latter results in better approximations of the true geodesic by allowing paths that go across the triangular facets of the mesh, but at the cost of a higher computational complexity.

There are several examples in this class that have been used for representing body and face. One of the first work using geodesic distances is the *canonical* representation of the face proposed by Bronstein et al. [31]. In their work, face models were represented with the geometric moments up to the fifth order computed for the 3D face canonical surface. Canonical surfaces were obtained from face surfaces by warping according to a topology preserving transformation so that the Euclidean distance between two canonical surface points is equivalent to the geodesic distance between the corresponding points of the face surface. In the approach of Berretti et al. [20], the structural information of a face scan was captured through the 3D shape and relative arrangement of iso-geodesic stripes identified on the 3D surface. In this case, geodesic computation was used to define *iso-geodesic stripes* as the surface points whose normalized geodesic distance from the nose tip is comprised in a given range. This divided the face into a set of stripes whose mutual spatial relationships ni 3D were then modeled using weighted walktroughs.

2.1.2 Representations based on Volumetric Descriptors. Compared to the number of representations that use surface, volumetric solutions are just a few. For example, Rustamov [137] introduced a volume-based shape descriptor that is robust with respect to changes in pose and topology. To this end a modified SD was used in conjunction with the interior distances and barycentric potential based on barycentric coordinates. In this approach, SDs are aggregated throughout the entire volume contained within the shape, thus capturing information conveyed by the volumes of shapes. Volumetric descriptors have been developed also for the dynamic case, where volumes can evolve over time [167, 169]. For example, Vieira et al. [167] presented a method to compute Space-Time Occupancy Patterns from sequences of depth maps. In this representation, space and time axes are divided into multiple segments to define a 4D grid for each depth map sequence. Wang et al. [169] treated a 3D sequence as a 4D shape and proposed Random Occupancy Pattern (ROP) features,

which are extracted from randomly sampled 4D sub-volumes with different sizes and at different locations. This method treats a depth sequence as a 4D volume, and defines the value of a pixel in this volume $I(x, y, z, t)$ to be either 1 or 0, depending on whether or not there is a point in the 4D volume at this location. A 4D ROP is used to construct the features, whose value is defined to be the soft-thresholded sum of the pixels in a sub-volume.

2.1.3 Topological Representations. Methods in this class capture topological changes in the model surface or volume. Two relevant solutions are *skeleton* and *Reeb-graph*.

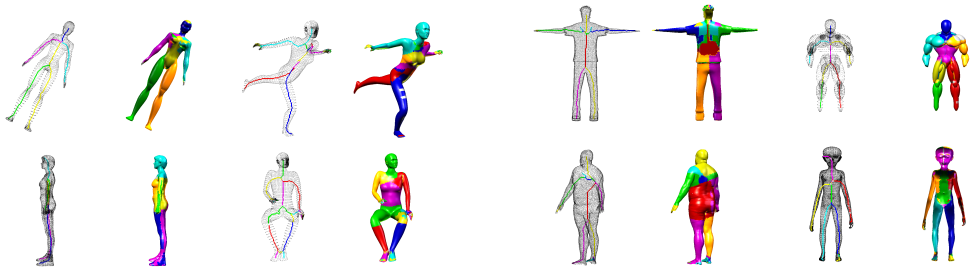


Fig. 2. Decomposition of human models by identification of distinct components in the skeleton [145].

Skeleton. One popular way of representing 3D objects is by using skeletons, or lines and curves inside the object. Skeletons can be used for a variety of applications including object recognition, matching shapes and animation of 3D models including humans. There are various techniques that can be used to extract skeletons, including shock graphs, potential fields and thinning [41]. Among these, an efficient fully parallel 3D thinning algorithm was developed by Wang et al. [173] and used to align 3D objects such as CT and MRI scans of the same patient at different points in time. Thinning algorithms can generate skeletons that are not properly centered inside an object. To address this problem, the Valence Normalized Spatial Median was proposed by Wang et al. [174] to make the skeletons more centered. This approach guarantees a unit width skeleton, which is essential for extracting meaningful segments from the skeleton. In the same work, skeletons were used for computing the 3D transformation between two poses of a shape and align objects. There are other approaches to human pose recognition, such as the Radon transform proposed by Singh et al. [147]. This method is also developed assuming an underlying skeleton representation of a human subject. In the work of Shi et al. [145], skeletonization was used for decomposing a 3D model into coherent parts. This is achieved by unit-width skeleton extraction, followed by using scale-space analysis to remove noise and make different segments smoother. Finally, the different branches of the skeleton were identified, assisting in the decomposition of a 3D model into visually meaningful parts. Figure 2 shows the results obtained by this algorithm. The resulting automatic decompositions were found to be very close to ground truth derived from anatomical specification.

Skeletons have been also used to generate animations. In this case, animation skeletons that best represent a 3D model should be generated. Usually, four steps need to be followed (see Hajari et al. [60]): *(a)* extracting a template skeleton that is provided in a Motion Capture (MoCap) file; *(b)* computing the curve skeleton (which may not have straight line segments) from the 3D model; *(c)* selecting important key points on the curve skeleton in order to generate an animation skeleton; and *(d)* optimizing the animation skeleton so as to minimize artifacts during animation.

Reeb-graph (RG). Effective methods for decomposing 3D articulated shapes were also obtained using RG. In this case, a sort of shape skeletonization is obtained by exploiting the capability of

the Reeb function computed on the shape surface to capture topological variations of the shape. Examples of this solution were given in [19, 158]. A RG is a schematic way to encode the behavior of a *Morse* function f on a surface. Nodes of the graph represent connected components of the level sets of f, contracted to points. Formally, given a surface S, the RG is the quotient space of f in $S \times \mathbb{R}$ formed by the equivalence relation $(v_1, f(v_1)) \sim (v_2, f(v_2))$, v_1 and v_2 being two points on S. According to this, given two points v_1 and v_2 on the model surface, the equivalence relation '$\sim$' holds if and only if both the following conditions are fulfilled:

(1) $f(v_1) = f(v_2)$,
(2) v_1 and v_2 are in the same connected component of $f^{-1}(f(v_1))$.

For each set of equivalent surface points, one node in the RG is defined. Arcs of the graph are used to represent adjacent sets of equivalent surface points.

2.1.4 Landmark based Representations. Landmarks refer to points of the body or the face that have a specific semantic meaning. For the face, some examples are the tip of the nose, the points at the inner and outer extrema of the eyes, or the corners of the mouth; for the body, landmarks can be posed at the joints of the main articulations of the skeleton (e.g., shoulder, elbow, wrist, knee, etc.). Landmarks can then be used as anchor points in several tasks: they can be used to normalize the pose of face scans, to derive measures of the face computing distances between them or to extract local descriptors that can be used to perform face or facial expression recognition, to analyze and recognize subject's movement from the temporal dynamics of the joints. Considering the specific case of body joints, Shotton et al. [146] were the first who proposed a method to quickly and accurately predict 3D positions of body joints from a single depth image, using no temporal information. To this end, they followed an object recognition approach, by designing an intermediate body parts representation that maps the pose estimation into a simpler per-pixel classification problem. A large and highly varied training dataset allowed the classifier to estimate body parts invariant to pose, body shape, clothing, etc.

Kendall Space. According to Kendall's definition: "Shape is all the geometrical information that remains when location, scale, and rotational effects are filtered out from an object." Starting from this, Kendall subsequently proposed landmark-based shape analysis [80], that was then advanced by several others. The basic idea is to sample the object at a number of points, called *landmarks*, and form polygonal shapes by connecting those points with straight lines. Of course, the number and locations of these points on the objects can drastically change the resulting polygonal shape. One can organize these landmarks in the form of a vector of coordinates and perform standard vector calculus. For simplicity, we consider the 2D case where $x \in \mathbb{R}^{n\times 2}$ represents n ordered points selected from the boundary of an object. It is often convenient to identify points in $\mathbb{R}^2$ with elements of $\mathbb{C}$, i.e., $x_i \equiv z_i = (x_{i,1} + jx_{i,2})$, where $j = \sqrt{-1}$. Thus, in this complex representation, a configuration of n points x is now $z \in \mathbb{C}^n$. An orthogonal section of that set can be formed by considering the joint action of the translation and scaling group on the set of such configurations. Each z is then represented by the corresponding element of the orthogonal section, often also called the *pre-shape space* $\mathcal{D} = \{z \in \mathbb{C}^n \mid \frac{1}{n}\sum_{i=1}^{n} z_i = 0,\ \|x\| = 1\}$. $\mathcal{D}$ is not a vector space because $a_1z_1 + a_2z_2$, for $a_1, a_2 \in \mathbb{R}$ and $z_1, z_2 \in \mathcal{D}$, is typically not in $\mathcal{D}$, due to the unit norm constraint. However, $\mathcal{D}$ is a unit sphere and one can utilize the geometry of a sphere to analyze points on it. Under the Euclidean metric, the shortest path between any two elements $z_1, z_2 \in \mathcal{D}$, also called a *geodesic* is given by the great circle. A natural application of Kendall's approach to shape analysis is given by human skeletons represented as sets of registered points (or landmarks). Ben Amor et al. [16] used this approach for analyzing shapes of human skeletons and their temporal evolutions. In their work, $x \in \mathbb{R}^{n\times 3}$ represents a skeleton or a configuration of n landmarks in $\mathbb{R}^3$.

Lie Group. The Special Euclidean group, denoted by $SE(3)$, is the set of all 4×4 matrices of the form (see [61] for a general introduction to Lie groups):

$$P(R, \vec{d}) = \begin{bmatrix} R & \vec{d} \\ 0 & 1 \end{bmatrix} , \tag{1}$$

where $\vec{d} \in \mathbb{R}^3$, and $R \in \mathbb{R}^{3\times3}$ is a rotation matrix. Members of $SE(3)$ act on points $z \in \mathbb{R}^3$ by rotating and translating them:

$$\begin{bmatrix} R & \vec{d} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z \\ 1 \end{bmatrix} = \begin{bmatrix} Rz + \vec{d} \\ 1 \end{bmatrix} . \tag{2}$$

Elements of this set interact by the usual matrix multiplication, and from a geometrical point of view, can be smoothly organized to form a curved 6D manifold, giving them the structure of a Lie group [61]. The 4×4 identity matrix I_4 is a member of $SE(3)$ and is referred to as the identity element of this group. Grounded on this idea, Vemulapalli et al. [163] proposed a skeletal representation that explicitly models the 3D geometric relationships between various body parts using rotations and translations in 3D space. Since 3D rigid body motions are members of $SE(3)$, the proposed skeletal representation lies in the *Lie* group $SE(3) \times \cdots \times SE(3)$, which is a curved manifold. Using this representation, human actions can be modeled as curves in the Lie group. For classification, the action curves are mapped from the Lie group to its Lie algebra, which is a vector space.

Representations based on Keypoints. Following the success of keypoints detectors and descriptors for 2D images, like SIFT and SURF, several keypoints detectors/descriptors have been defined for arbitrary 3D objects (see the survey of Tombari et al. [159] for an in-depth review and performance comparison). Here, we mention two solutions that have been applied to 3D faces.

In [189], Zaharescu et al. revisited local feature detectors/descriptors developed for 2D images and extended them to the more general framework of scalar fields defined on 2D manifolds. They provided methods and tools to detect and describe features on surfaces equipped with scalar functions, such as photometric information. They proposed a 3D feature detector (meshDOG) and a 3D feature descriptor (meshHOG) for uniformly triangulated meshes, invariant to changes in rotation, translation, and scale. The descriptor is able to capture the local geometric and/or photometric properties. Moreover, the method is defined generically for any scalar function, e.g., local curvature. MeshDOG is a generalization of the Difference Of Gaussian (DOG) operator, and it seeks the extrema of the Laplacian of a scale-space representation of any scalar function defined on a discrete manifold. MeshHOG is a generalization of the Histogram of Oriented Gradients (HOG) descriptor introduced for describing 2D images. The new descriptor is defined with respect to the measurements available at each of the discrete surface's vertices and it can work with photometric features as well as with geometric features, such as curvature, geodesic integral, etc.

Smeets et al. [151] proposed the meshSIFT algorithm and its use for 3D face recognition. This algorithm consists of four major components. First, salient points on the 3D facial surface are detected as mean curvature extrema in scale space. Second, orientations are assigned to each of these salient points. Third, the neighborhood of each salient point is described in a feature vector consisting of concatenated histograms of shape indices and slant angles. Fourth, the feature vectors of two 3D facial surfaces are reliably matched by comparing the angles in feature space. This results in an algorithm which is robust to expression variations, missing data and outliers.

2.1.5 Representations based on Deep Learning (DL). Since 2012 [86], the DL "revolution" has invented methods for image analysis and recognition. DL allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. It can discover intricate structures in large data sets by using the back-propagation

algorithm to indicate how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Deep Convolutional Neural Networks (CNNs) have brought about breakthroughs in processing images, videos, speech and audio, whereas Recurrent NN have shown light on sequential data such as text and speech [89]. However, these tools have been most successful on data with an underlying *Euclidean* or *grid-like* structure, and in cases where the invariance of these structures is built into networks used to model them. *Geometric* DL is a wide term for emerging techniques attempting to generalize (structured) deep neural models to non-Euclidean domains such as graphs, meshes and manifolds. Some preliminary results in this direction were presented by Bronstein et al. [32], where an overview of different examples of geometric DL problems and available solutions are presented, together with key difficulties, applications, and future research directions in this nascent field. In the following, we address some works that redefined CNN for volumes, surfaces, and point clouds.

In [179], Wu et al. proposed to represent a geometric 3D shape as a probability distribution of binary variables on a 3D voxel grid, using a Convolutional Deep Belief Network. This model, called 3D *ShapeNet*, learns the distribution of complex 3D shapes across different object categories and arbitrary poses from raw CAD data, and discovers hierarchical compositional part representations automatically. It naturally supports joint object recognition and shape completion from 2.5D depth maps, and it enables active object recognition through view planning. To train the 3D DL model, a large-scale 3D CAD model dataset (*ModelNet*) is constructed.

Surfaces serve as a natural parametrization to 3D shapes. Learning surfaces using CNNs is a challenging task. Current paradigms to tackle this challenge are to either adapt the convolutional filters to operate on surfaces, learn spectral descriptors defined by the Laplace-Beltrami operator, or to drop surfaces altogether in lieu of voxelized inputs. In the work of Sinha et al. [148], the 3D shape is converted into a "geometry image" (GI) so that standard CNNs can directly be used to learn 3D shapes. In particular, GIs are created using authalic parametrization on a spherical domain. This spherically parameterized shape is projected and cut to convert the original 3D shape into a flat and regular GI. Then, a way to implicitly learn the topology and structure of 3D shapes using GIs encoded with suitable features is proposed. Xie et al. [182] proposed a high-level shape feature learning scheme (*DeepShape*) to extract features that are insensitive to deformations via a discriminative deep auto-encoder. First, a multi-scale SD was developed for use as input to the auto-encoder. Then, by imposing the Fisher discrimination criterion on the neurons in the hidden layer, a discriminative deep auto-encoder was developed for shape feature learning. Finally, the neurons in the hidden layers from multiple discriminative auto-encoders were concatenated to form a shape descriptor. Fang et al. [54] developed novel techniques to extract a concise, but geometrically informative, shape descriptor and new methods of defining Eigen-shape and Fisher-shape descriptors to guide the training of a deep NN. The resulting deep shape descriptor tends to maximize the inter-class margin, while minimizing the intra-class variance.

Typical CNN architectures require highly regular input data formats, like those of image grids or 3D voxels, in order to perform weight sharing and other kernel optimizations. Since point clouds or meshes are not in a regular format, most researchers typically transform such data to regular 3D voxel grids or collections of images (e.g., views) before feeding them to a deep net architecture. This data representation transformation, however, renders the resulting data unnecessarily voluminous, while also introducing quantization artifacts that can obscure natural invariances of the data. To overcome such limitations, Qi et al. [129] designed a novel type of NN, named *PointNet*, that directly utilizes point clouds, which well respects the permutation invariance of points in the input.

2.1.6 ***Discussion: Issues and Perspectives.*** The innovation in the description and representation of 3D body and face is now undergoing an important change. On the one hand, there is a lot

of work done in the last two decades on hand-crafted descriptors. Several trends can be observed in this respect: *(i)* surface based representations are the most used, since point clouds and 3D volumes are more difficult to manage. However, accurate registration, and effective and efficient computations, e.g., curvature, geodesics, etc., on the mesh manifold support are not completely solved problems. Resorting to the 2.5D modality of depth images is also a very common solution; *(ii)* in most of the cases local representations have the capability to better adapt to the data. Thus, many representations have shifted from global to local (solutions based on 3D keypoints are examples of this trend); *(iii)* multi-modality has proved potential, but has not been fully exploited yet. Mostly photometric and geometric information are processed separately and later fused together, while a more integrated approach could results in more effective solutions.

On the other, it is emerging now the idea of transferring DL solutions to the 3D world. Some methods that adapt and extend DL solutions to the 3D (non-Euclidean) domain have been already proposed, and we can expect in the next few years a proliferation of new solutions specifically tailored for the representation and analysis of 3D data. In this respect, there is the need to overcome solutions that manage 3D data by projecting them to the depth (2.5D) images. In fact, the solution of using such images to fine tune networks that have been trained using image photometric information does not exploit the full power of 3D data. To obtain such advancement in learning from 3D data, there is an urgent need for larger 3D training datasets. In this respect, consumer depth cameras have the potential to acquire large amounts of 3D data as video streams, thus reducing the time of acquisition. Furthermore, these low-cost low-size devices can be deployed in a vast variety of real application contexts, that is often not possible with the complex equipment of high-resolution scanners. Finally, though several advancements have been done and some work provided interesting insights in the CNN learning mechanism [118], a more clear understanding of what actually is learned is required. Ultimately, this would provide a theoretical basis for designing effective NN from the scratch, without a tedious trial and error approach.

2.2 Temporal Modeling

In many applications, the relevant information that human representations aim to capture does not reside on static observations but, rather, is distributed across a temporal interval. For example, this is the case for recognizing actions from skeleton movement or detecting facial expressions. In such cases, human representations need to model the temporal dynamics of the underlying event.

2.2.1 Trajectory based Representations. A possible way of modeling temporal information is that of extracting appropriate features from each 3D frame (static data), and then compare sequences of those features. A convenient way to think of such features is as describing *trajectories* in certain feature-spaces, where the geometry of the feature-space needs to be considered while matching [62]. In the simplest case, one can think of the features as points in a high-dimensional vector space or, depending on additional constraints, one may think of them as sequences on certain non-Euclidean domains. This category of approaches has been used with success for representing human actions.

Dynamic Time Warping (DTW). One of the simplest ways to compare feature sequences, where temporal re-parametrization can be removed, is via DTW. DTW has its roots in speech analysis and its variants continue to be applied to human action recognition as well [162, 164]. Traditionally, DTW has enjoyed success owing to its simplicity and availability of various efficient algorithms for its computation. From a theoretical perspective, DTW has some limitations: *(i)* the resultant measure not being a true metric, but a divergence measure; *(ii)* without additional constraints, DTW can spuriously match very different sequences via temporal "pinching"; *(iii)* the DTW framework does not naturally enable statistical summaries to be drawn from sequence data, such as computation of average or measures of variance within sequences.

Shape Analysis. Other solutions explored more elaborate and effective ways to match the shape of trajectories. For example, in the work of Devanne et al. [45], the 3D position of each joint of the human skeleton extracted in each frame of a sequence is represented as a motion trajectory. Each motion trajectory is then expressed as a point in the open curve shape space. The *Square-Root Velocity Function* (SRVF) [153] has been applied to skeletal action recognition using joint location in $\mathbb{R}^3$. The action recognition problem is then formulated as the problem of computing the similarity between the shape of trajectories in a Riemannian manifold. The motion of a 3D joint is represented by a single trajectory. Formally, given a trajectory as a continuous parameterized function $\beta(t) \in \mathbb{R}^3$, $t \in [0, 1]$. β is first represented by its SRVF, q, according to:

$$q(t) = \frac{\dot{\beta}(t)}{\sqrt{\|\dot{\beta}(t)\|}}, \ t \in [0, 1] \ .$$

Then, with the $\mathbb{L}^2$-norm of the q functions scaled to 1 ($\|q\| = 1$), the space of such representation: $C = \{q : [0, 1] \rightarrow \mathbb{R}^3, \|q\| = 1\}$ becomes a Riemannian manifold and has a spherical structure in the Hilbert space $\mathbb{L}^2([0, 1], \mathbb{R}^3)$. Given two curves β_1 and β_2 represented by their SRVFs q_1 and q_2 on the manifold, the geodesic path connecting q_1, q_2 is given analytically by the minor arc of the great circle connecting them on C. It has been proved in [153] that under the $\mathbb{L}^2$-norm, the quantities $\|q_1 - q_2\|$ and $\|q_1 \circ \gamma - q_2 \circ \gamma\|$ are same, where the composition $(q \circ \gamma)$ denotes the function q with a new parameterization dictated by a non-linear function $\gamma : [0, 1] \rightarrow [0, 1]$. This important property allows curve registration by re-parameterization, and thus makes the curves registration easier. In fact, it allows one of the curves to be considered as reference, while searching for a $\gamma^* = \text{argmin}_{\gamma \in \Gamma}(\|q_1 - \sqrt{\dot{\gamma}} q_2 \circ \gamma\|)$, which optimally registers the two curves. This optimization is resolved by Dynamic Programming, as described in [153]. Some of these newer developments have been applied to activity recognition both from 2D and 3D data [10, 16, 162].

2.2.2 Matrix based Representations. In this class of representations, the idea is that of encoding spatio-temporal data into a tensorial form. This has the advantage of largely compressing the data and open the way to the vast set of mathematical tools on matrices.

Linear Dynamical Systems (LDSs). These are a powerful, yet compact, class of models for encoding high-dimensional spatio-temporal data. One of the first successful applications was in modeling dynamic textures in video. LDSs represent temporal sequences according to the following model:

$$\begin{cases} f(t) = Cz(t) + w(t), & w(t) \sim N(0, R) \\ z(t + 1) = Az(t) + v(t), & v(t) \sim N(0, Q) \end{cases} \tag{3}$$

where $f(t)$ is the feature at time t, $z \in \mathbb{R}^d$ is a hidden state vector, $A \in \mathbb{R}^{d \times d}$ is the transition matrix and $C \in \mathbb{R}^{p \times d}$ is the measurement matrix. w and v are noise components modeled as normal with mean equal to zero and covariance matrix $R \in \mathbb{R}^{p \times p}$ and $Q \in \mathbb{R}^{d \times d}$, respectively. The observability matrix encodes the space of all *expected* sequences that can be realized from arbitrary initial conditions of the system. A finite approximation of the observability matrix can be attained as: $[C^T, (CA)^T, (CA^2)^T, \ldots, (CA^{m-1})^T]$. The subspace spanned by the columns of this finite observability matrix (obtained by any orthogonalization procedure) corresponds to a point on a Grassmann manifold. The LDS then represents at each time-instant a point on the Grassmann manifold. Each video sequence is modeled as described above to become an element of the Grassmann manifold and the action learning and recognition problem is brought back to a classification problem on this manifold. A distance between two spatio-temporal sequences is defined as a distance between two subspaces on the Grassmann manifold. Statistical methods on the Grassmann manifold can be utilized to develop classifiers based on LDS parameters for a variety

of video applications [161]. For example, Slama et al. [150] addressed the problem of modeling and analyzing human motion by focusing on 3D body skeletons. An action is represented by a dynamical system, whose observability matrix is characterized as an element of a Grassmann manifold. Huang et al. [70] formulated LDS as an infinite Grassmann manifold. Other statistical approaches such as kernels on dynamical model parameters have also been proposed for improved recognition performance [37]. Extensions to bag-of-dynamical systems [132], and other related concepts like tensor-methods [102], and Hankel-matrix based methods [128] have also been investigated.

Non-linear Dynamical Systems (NLDSs). While LDSs are compact models for which efficient algorithms have been proposed for learning and inference, they do suffer from limited expressive power, due to assumptions of linearity, Markovian dynamics, and stationary noise processes. In order to address some of their limitations, methods from the NLDSs theory have also been extended to activity analysis. In the area of NL dynamics, one considers that the observations are outputs of a system whose parametric form is unknown, but a surrogate form (a topologically equivalent form) of its state-space can be recovered using the Takens' delay embedding theorem. This surrogate form is related to the original state-space via an unknown topology-preserving transformation. Thus, one extracts from this reconstructed *phase space* topologically invariant features such as Lyapunov exponents. There is a long history of applying NLD approaches for modeling human movement in the biomechanics community. This approach was applied for action recognition by Ali et al. [6] for video-based body-joint features. This approach was developed further by Venkataraman et al. [165] using a shape-theoretic framework for extracting discriminative features from the phase-space. Applications were shown in modalities as diverse as marker-based motion capture, RGBD sensors, and activity quality assessment for applications in stroke rehabilitation.

Covariance Matrix (CM). Let $f \in \mathbb{R}^d$ be a d-dimensional feature vector, and $D_{d\times n} = [f_1, \cdots, f_n]$ denote a set containing the d-dimensional feature descriptors of n images of an image set. The CM C of the set is defined as: $C = \frac{1}{(n-1)} \sum_{i=1}^{n} (f_i - \mu)(f_i - \mu)^T$, where μ is the sample mean. This representation allows compactly encoding a high-dimensional feature set. It is useful for cases where the temporal axis is not sufficiently densely sampled, or if the activity of interest does not need explicit encoding of dynamics. A CM can be viewed as an element of the Riemannian manifold Sym_d^+, i.e., the set of symmetric positive-definite (SPD) matrices [24]. This formulation makes it easier to compute a distance function between two videos, as well as simplifies the computation of statistics (e.g., mean). Covariance descriptors provide rich yet compact representations for several kinds of features, including image-features or even shape-landmarks. They also allow fusing various image cues, while attenuating the impact of noisy samples through their averaging process. While at the outset, this approach appears to ignore the temporal dynamics, there are some ways to include this information, for instance by concatenating time information to the feature itself. Several measures on Sym_d^+ have been proposed. The most widely used are the Affine Invariant Riemannian Metric [123], the log-Euclidean Riemannian Metric [13], and two other popular measures are Jensen-Bregman Log-det Divergence, and the KL-Divergence Metric [64, 191].

CM representation has limitations when applied to a small number of high-dimensional features, which is typical in video/activity analysis. Specifically, the situation where $n < d$ results in the covariance descriptor constructed to be rank-deficient. This is simply because the ambient dimensionality of the image descriptor is often greater than the number of available images in a set. Such CMs are positive semi-definite. Bonnabel and Sepulchre developed a metric on the space of fixed low-rank SPD matrices [29], thus extending the available tools for high-dimensional videos. Finally, given one of these metrics and a set of SPD matrices, the associated mean of this set exists.

2.2.3 Markov Models. This stochastic approach can be used to model randomly changing systems in the case the Markovian's property can be adopted. This is equivalent to assuming that future states depend only on the current state, not on the events that occurred before it. In the temporal analysis of humans, the most used model is the Hidden Markov Model (HMM) where, differing from the simpler Markov model, the state is not directly visible, but the output, dependent on the state, is visible. For example, HMMs have been used for facial expression recognition from dynamic 3D data [15, 154] or in activity recognition from depth sequences [74, 181].

2.2.4 Recurrent Neural Networks (RNN). Besides the aforementioned methods, RNNs based methods have shown great power in tackling time series tasks. In particular, the Long Short Term Memory (LSTM) networks have been proposed by Hochreiter and Schmidhuber [68] as an extension of RNNs, and have been refined by many researchers in following work. LSTMs were developed to deal with the exploding and vanishing gradient problem when training traditional RNNs. Relative insensitivity to gap length gives an advantage to LSTM over alternative RNNs, HMMs and other sequence learning methods so that they are now widely used since they work tremendously well on a large variety of problems. They have the capability of learning long-term dependencies so that several recent works have applied them, for example, to the case of human action recognition and re-identification from human depth data [63, 98, 144].

2.2.5 ***Discussion: Issues and Perspectives.*** Representations that model 3D human data dynamically varying over the time are relatively more recent than spatial ones. One reason for this is the difficulty in capturing and processing 3D time varying data. As a consequence, existing temporal modeling solutions are somewhat more acerb than spatial one. Among the trends that emerged so far in this respect, we can identify: *(i)* solutions that model the dynamics through trajectories are appealing since they exploit an intuitive idea, and considering the trajectory as a shape can rely on the vast literature of shape matching methods; *(ii)* the need for better methods to model the non-linearities that are intrinsic in temporal phenomena. The idea of exploiting the geometry of underlying manifold is a promising direction as demonstrated by the success of several recent works; *(iii)* a more tight integration between the spatial and temporal components appear as an aspect that has been often solved with naive solutions that just later fuse the two components or compress too much the dynamics, thus loosing non-linearities and temporal ordering; *(iv)* the use of RNNs, and LSTM in particular, is emerging as a new direction that can push the research in this field, especially in the case of long sequences, where other solutions have shown limitations and difficulty in adapting to the data.

3 ANALYSIS AND RECOGNITION OF 3D HUMANS

The representations discussed above can be applied to perform analysis and recognition of 3D humans. These can concern the human face and target identity recognition, action unit detection, expression and emotion recognition, face modeling and face morphing (Section 3.1); or they can have the main goal of understanding and recognizing human action, activity and behavior, gait analysis, emotion recognition from body posture and motion, identification and re-identification from body shape, retrieval of 3D human models, and animation (Section 3.2).

3.1 Face

Methods developed for face analysis from 2D images and videos performed either face or facial expression recognition as main tasks; but, in both the cases, results were negatively influenced by variations in pose, illumination and resolution. Such variations have a lower impact on 3D data, which has opened the way to 3D solutions that can either substitute or complement those based on 2D data or serve as intermediate tools to simulate the image formation by rendering operations.

3D face data can be captured either by high- or low-resolution scanners. Typically, devices in the first category are very costly and provide detailed acquisitions of the face, but are static and require users cooperation. By contrast, low-cost low-resolution cameras are dynamic in most of the cases, capturing depth data at high frame rate (30fps or more) and without the need for users cooperation. Though the performance in terms of resolution of dynamic scanners is rapidly increasing, the contra-position between high- and low-resolution is often also between static and dynamic acquisitions. As a consequence, high-resolution acquisitions are mostly possible only during the enrollment of the subjects in a face verification system or in creating a training set of facial scans. Instead, for on-line operations, dynamic scanners with lower-resolution outputs are more interesting for concrete use. Figure 3 proposes a schematic view of the main factors that play a role when methods for 3D faces are concerned: the data source that can be either static or dynamic; the intermediate tools applied to the data (denoising/reconstruction, face 3DMM); and the ultimate applications. These aspects will be expanded in the rest of this section. In particular, first, we revise two common solutions that try to compensate between different resolutions: the former is based on the idea of reconstructing higher-resolution face models from low-resolution depth frames (Section 3.1.1); the latter, relies on the 3D morphable model as a way to fill the gap between different 3D sources, and between 3D and 2D data (Section 3.1.2). Then, we will focus on the use of 3D data for face and facial expression recognition (Section 3.1.3 and 3.1.4, respectively).

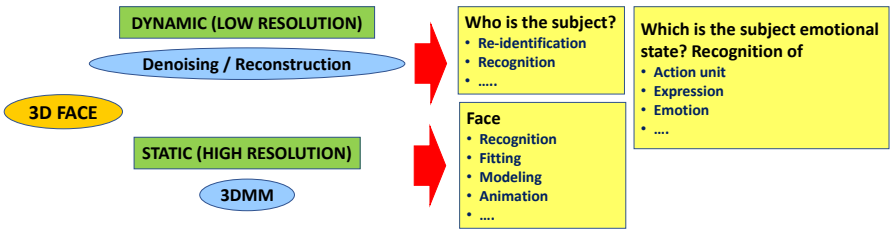


Fig. 3. Overview of 3D face analysis applications. For 3D faces acquired with dynamic (and typically low-resolution) scanners some denoising/reconstructing processing is required in most of the cases. Intermediate tools, such as 3D Morphable Models (3DMM) of the face can be derived instead from high-resolution data.

3.1.1 Denoising / Reconstruction. Consumer depth cameras like Kinect have enjoyed a large success in Computer Vision, Human Computer Interaction and Multimedia applications since they are cheap, easy to use, and provide RGB-D video data at high frame rate. However, such depth data have low resolution that can be not adequate in applications where fine details of captured objects are required, like for the face. This motivated the idea of constructing higher-resolution representations of imaged objects or scenes by using multiple lower-resolution observations, possibly altered by noise, blurring or geometric warping. Former solutions that addressed this problem extended ideas originally proposed for 2D still images, but subsequently this concept migrated to 3D generic data for recovering one high-resolution model from a set of low-resolution 3D acquisitions. However, the extension of 3D methods to depth videos is not straightforward due to the textureless nature of depth data, and to their high frequency content coupled with motion artifacts. Most of the solutions initially proposed to address this task combined low-resolution data captured by depth cameras with some ground-truth data. For example, Yang et al. [184] acquired data with a time-of-flight camera and used information from a high-resolution image of the same scene to up-sample and denoise the data. Time-of-flight data were processed also in the work of Schuon et al. [141] by using an energy minimization framework. Some works also focused on 3D

faces. Pan et al. [117] modeled the reconstruction process as a progressive resolution chain, whose features were computed as the solution to a Maximum a Posteriori estimation problem. Liang et al. [96] proposed an algorithm that takes a single face frame from a Kinect depth camera as input, and produces a high-resolution 3D mesh of the face as output.

The approaches above depend on a single 3D low-resolution scan, with the additional information used for reconstruction coming from multiple high-resolution scans that serve as reference. This completely disregards the temporal dimension available in depth sequences. To exploit such temporal information, some methods approached the problem of noise reduction in depth data by fusing the observations of multiple scans. Newcombe et al. [112] presented the Kinect Fusion system that operates on live depth data coming from a moving Kinect camera, and creates a high-quality 3D model of a scene. Later, Izadi et al. [73] added dynamic interaction to the system so that camera tracking was performed on a static background scene and a foreground object was tracked independently of camera tracking. However, this approach was targeted to generic objects rather than to faces. Newcombe et al. [111], further extended the approach to cope with non-rigid objects, such as faces, but results of non-rigid object denoising were demonstrated only in cases where the distance between the object and the camera was almost constant. Al Ismaeil et al. [2] proposed to enhance low-resolution dynamic depth videos containing non-rigidly moving objects using a dynamic multi-frame super-resolution algorithm. This was obtained by accounting for non-rigid displacements in 3D, in addition to 2D optical flow, and simultaneously correcting the depth measurement by Kalman filtering. This work was refined by Al Ismaeil et al. [1] into a so called up-sampling method for super-resolution (UP-SR) of dynamic depth scenes. Though able to handle nonrigid deformations, these latter methods were limited to lateral motions. The same authors in [3] improved on the UP-SR approach by proposing the recUP-SR algorithm. This method was designed to handle nonrigid deformations in 3D thanks to a per-pixel filtering that directly accounted for radial displacements in addition to lateral ones.

Just few works proposed specific solution for increasing the face resolution from depth sequences and proved applications. Hernandez et al. [66], obtained a 3D face model with an improved quality by using the frames captured from a user moving in front of a low resolution depth camera. The model was initialized with the first depth frame, and then each subsequent cloud of 3D points was registered to the reference one using a GPU implementation of the ICP algorithm. This approach was used by Choi et al. [39] to investigate whether a system that uses reconstructed 3D face models performs better than a system that uses the individual raw depth frames considered for the reconstruction. The approach proposed by Berretti et al. [22] relied on a constrained acquisition protocol, where subjects sit in front of the camera at a predefined distance, moving their head to the left/right side in order to expose different parts of the face to the sensor. This avoids scale and velocity problems obtaining a cumulated 3D point cloud by ICP registration of 3D frames of a sequence. An increased resolution of the cumulated point cloud was then obtained by using 2D-Box splines for up-sampling and approximation. Bondi et al. [28] generalized to the case where persons pass in front of the scanner, without assuming any particular cooperation. The 3D data were first registered combining rigid (ICP) and non-rigid (Coherent Point Drift [110]) registration, then filtered by combining a model of the expected distribution of the acquisition error with a variant of the *lowess* method to remove outliers and build the final face model.

3.1.2 Face Modeling. Having a set of high-resolution scans of the face spanning different characteristics (i.e., age, gender, ethnicity, etc.), it is possible to learn a generic 3D face model capable of generating new face instances with plausible shape and appearance. This can be done by capturing the face variability in a training set of 3D scans and constructing a statistical face model that includes an average component and a set of learned principal components of deformation. Such

a model can allow either to generate new face instances or deform and fit to 2D/3D target faces. Grounding on this idea, the 3D Morphable Face Model (3DMM) was proposed as an intermediate tool to enhance face analysis from 2D and 3D data.

Blanz and Vetter [25] first proposed to create a 3DMM from a set of exemplar 3D faces and showed its potential and versatility. They presented a complete solution to derive a 3DMM by transforming the shape and texture from a training set of 3D face scans into a vector space representation based on PCA. However, the training dataset had limited face variability, thus reducing the capability of the model to generalize to different ethnicities and non-neutral expressions. The 3DMM was further refined into the Basel Face Model by Paysan et al. [122]. This offered higher shape and texture accuracy thanks to a better scanning device, and a lower number of correspondence artifacts using an improved registration algorithm based on the non-rigid ICP [8]. The work by Booth et al. [30], introduced a pipeline for 3DMM construction. Initially, dense correspondence was estimated applying the non-rigid ICP to a template model. Then, the so called LSFM-3DMM was constructed using PCA to derive the deformation basis on a dataset of 9,663 scans with a wide variety of age, gender, and ethnicity. Though the LSFM-3DMM was built from the largest dataset compared to the current state-of-the-art, the face shapes still had neutral expression. Following a different approach, Patel and Smith [120] first used Procrustes analysis to establish correspondence between a set of 104 manually labeled landmarks of the face, and the mean coordinates of these landmarks were used as anchor points. A complete deformable model was then constructed by warping the landmarks of each sample to the anchor points and interpolating the regions between landmarks using showed that Thin-Plate Splines (TPS). Finally, consistent resampling was performed across all faces, but using the estimated surface between landmarks rather than the real one. Brunton et al. [33], instead, proposed a statistical model for 3D human faces with varying expressions. The approach decomposed the face using a wavelet transform, and learned many localized, decorrelated multilinear models on the resulting coefficients. In [100], Lüthi et al. presented a Gaussian Process Morphable Model (GPMM), which generalizes PCA-based Statistical Shape Models. GPMM was defined by a Gaussian process, which makes it inherently continuous.

The 3DMM has been used for face recognition and synthesis. In one of the first examples, Blanz and Vetter [26] used their 3DMM to simulate the process of image formation in 3D space, and estimated 3D shape and texture of faces from single images for face recognition. Later, Romdhani and Vetter [135] used the 3DMM for face recognition by enhancing the deformation algorithm with the inclusion of various image features. In all these cases, 3DMM was used mainly to compensate for the pose of the face, with some examples that also performed illumination normalization. Expressions were typically not considered. Ramanathan et al. [131] constructed a 3D Morphable Expression Model incorporating emotion-dependent face variations in terms of morphing parameters that were used for recognizing four emotions. Huber et al. [71] proposed a cascaded-regressor based face tracking and a 3DMM shape fitting for fully automatic real-time semi-dense 3D face reconstruction from monocular in-the-wild videos. A new approach for constructing a 3D Morphable Shape Model (called DL-3DMM) was proposed by Ferrari et al. [55]. They started from a set of 3D face scans with large variability in terms of ethnicity and expressions. Across these training scans, they computed a point-to-point dense alignment, which is accurate also in the presence of topological variations of the face. The DL-3DMM was constructed by learning a dictionary of basis components on the aligned scans. The model was then fitted to 2D target faces using an efficient regularized ridge-regression guided by 2D/3D facial landmark correspondences in order to generate pose-normalized face images. Li et al. [93] learned a facial model from thousands of accurately aligned 3D scans. The resulting FLAME model (Faces Learned with an Articulated Model and Expressions) is designed to work with existing graphics software and is easy to fit to the data. FLAME uses a linear shape space trained from 3,800 scans of human heads, and combines it with an articulated

jaw, neck, and eyeballs, pose-dependent corrective blendshapes, and additional global expression blendshapes. The pose and expression dependent articulations are learned from 4D face sequences in the D3DFACS dataset along with additional 4D sequences. Tuan Tran et al. [160] described a method for regressing discriminative 3DMM. They used a CNN to regress 3DMM shape and texture parameters directly from an input photo. They overcome the shortage of training data required for this purpose by offering a method for generating huge numbers of labeled examples. Coupled with a 3D-3D face matching pipeline, they also showed face recognition results on the LFW, YTF and IJB-A benchmarks using 3D face shapes as representations.

3.1.3 Face Recognition. Face recognition from 3D data has received great attention in recent years since 3D scans are less affected by illumination, pose and expression changes with respect to their 2D counterpart. Many approaches have been proposed in the literature, and going through all of them is out of the scope of this summary. We refer interested readers to the recent survey by Patil et al. [121] that focuses on features, databases, algorithms and challenges, and to the work by Soltanpour et al. [152] who presented a state-of-the-art review, with the main focus on the extraction and use of local features of the face. Instead, in the following, we will address some of the most effective solutions based on *hand-crafted* ("shallow") features, and on the recent and promising trend of using *learned* ("deep") features.

Hand-crafted Features. Among methods that represent the face by defining ad-hoc features that capture salient traits of the face, effective results have been reported by local methods that are capable of supporting partial face matching, as can occur in the case of expression variations or missing parts. Examples are the methods that detect *fiducial points* of the face (being them either anthropometric landmarks, points of a predefined grid, or sparse keypoints, see also Section 2.1.4), and compute local descriptors of surface patches centered at the fiducial points. One of the first approaches following this framework was proposed by Mian et al. [108], who designed a 3D keypoints detector and descriptor inspired by SIFT. This detector/descriptor was used to perform 3D face recognition through a multi-modal 2D+3D approach that also used SIFT to index 2D images of the face. In [151], Smeets et al. reformulated the framework of SIFT keypoints detector to operate on 3D face meshes by defining the mesh-SIFT detector and local descriptor. Effective local solutions based on fiducial points have been also reported in [92], where Li et al. used mesh-SIFT to detect feature points on 3D face scans. Then, the quasi-daisy local shape descriptor at each feature point was obtained using multiple order histograms of differential quantities extracted from the surface. Finally, these local descriptors were matched by computing their orientation angles. In [21], Berretti et al. used a similar paradigm by considering different varieties of histogram descriptors computed at mesh-DOG 3D keypoints [189]. The keypoints matching was also improved using the RANSAC algorithm. Drira et al. [47] proposed a method that represents facial surfaces by radial curves emanating from the nose tip and used elastic shape analysis of these curves to develop a Riemannian framework for analyzing shapes of full facial surfaces. To handle missing parts and partial occlusions due to glasses, hair, etc., they removed occlusions and attempted to restore the missing parts by exploiting statistical shape analysis of radial curves.

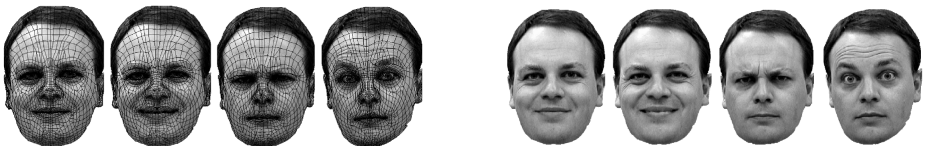
Multi-modal solutions are also attractive since they combine multiple processing paths (typically in 2D and 3D) into a coherent architecture. Mian et al. [107] assembled a fully automated system performing: pose correction, automatic region segmentation to account for local variations of the face geometry, quick filtering of distant faces using SIFT and 3D Spherical Face Representation, and matching of the remaining faces applying a modified ICP to a few regions of the face (eyes, forehead, and nose) that are less sensitive to face expressions. The similarity scores provided by the two matching engines were fused into a single similarity measure. Werghi et al. [178], presented a 3D face recognition approach which is both local and multi-modal. Using the mesh-LBP

framework [177], they computed shape and texture LBP directly on a mesh surface. To this end, they constructed a grid of the regions on the facial surface that can accommodate global and partial descriptions. Compared to its depth-image counterpart, the approach inherits the intrinsic advantages of mesh surface (e.g., preservation of the full geometry), does not require normalization, and can accommodate partial matching. In addition, it allows early-level fusion of texture and shape modalities. This is one of the few examples where the fusion of the descriptors is performed on the mesh, with the photometric information processed as a surface property of mesh triangles.

Deep Learning (DL). Recently, the performance of 2D face recognition systems was boosted significantly with the popularization of Deep CNN [119]. However, extension to the 3D domain is not straightforward. Compared to publicly available 2D face databases, 3D scans are hard to acquire, and the number of scans and subjects in public 3D face databases is limited. According to the survey in [121], the biggest 3D dataset is that collected at the University of Notre Dame in 2006 [52], which contains 13,450 scans over 888 individuals. It is small compared to publicly available labeled 2D faces, and may not be sufficient to train a DCNN from scratch. To overcome such difficulties, Kim et al. [81] proposed to leverage existing networks trained for 2D face recognition (VGG-Face [119]), and fine-tune them with a small amount of 3D scans in order to perform 3D to 3D surface matching. The work of Gilani et al. [193] presents three contributions in the direction of developing DL solutions for 3D face data: *(i)* a method for generating a large corpus of labeled 3D face data for training CNNs (the resulting dataset contains 3.1M scans of 100K identities); *(ii)* a large-scale test dataset that merges the most challenging existing public 3D face datasets (i.e., it contains 31,860 scans of 1,853 identities); *(iii)* The Deep 3D Face Recognition Network specifically designed for 3D face recognition and trained on 3.1M 3D faces.

Dynamic Data. Face recognition from sequences of 3D face scans has been proposed in few works [17, 91]. In this respect, a quite new challenge, which is now attracting an increasing interest is that of performing face recognition across 3D scans with different resolutions, and in particular for dynamic scans acquired with consumer depth cameras. Results on this emerging task are still preliminary. For example, Bondi et al. [28] tackled the problem by first constructing a higher resolution model of the face from a depth sequence to be used for the match (see also Section 3.1.1). Lee et al. [90] proposed to first perform a step for enhancing the quality of depth data, then used fine tuning of a trained deep neural network combining both the depth and color channels.

3.1.4 Facial Expression Recognition (FER). Automatically detecting human emotions can be useful in human-computer interaction, human facial behavior research, as well as in several other tasks. Since the work of Darwin that first studied the link between human emotions and facial expressions, most of the studies tried to derive emotional states from facial expressions. This has been done either associating the expression to the deformation of the entire face or categorizing expressions in terms of the configuration and strength of Action Units (AUs). These are atomic units of face deformation associated with the action of individual or group of muscles. In particular, automatic emotion analysis is mainly guided by the discrete categorization into six basic classes, namely, *anger*, *disgust*, *fear*, *happiness*, *sadness* and *surprise* as proposed by Ekman [48]. One of the earliest works in this area is associated with the Facial Action Coding System (FACS) by Ekman et al. [49]. FACS has evolved over the years with many of the details being refined. There are around 9 AUs for the upper face and 18 for the lower face; in addition there are several eye/head positions and movements defined along with other parameters. The idea behind FACS is to use a combination of these AUs to model various facial expressions, like sad, angry, surprise and so on. Figure 4 shows the end results of synthesizing expressions combining adaptive face modeling, expression representation and texture mapping [187].



(a) (b)

Fig. 4. (a) Face model adaptation; (b) Synthesized face expressions with texture mapping.

Automatic detection of facial expression has been tried using 2D RGB or thermal still images and 3D data (see the survey by Corneanu et al. [42] for a general overview). Methods working on 3D data have been summarized by Sandbach et al. [139] in their survey paper. In the following, we focus on methods that work on 3D data either static or dynamic.

Static Data. Most of the works can be categorized as *face model-driven* or *face model-free*.

In the first category, some prior knowledge on the human face (e.g., feature points, shape and texture variations or local geometry labels) is used. For example, a general 3D model of the face can be trained and subsequently deformed. This solution was followed by Ramanathan et al. [131], who minimized an energy function to establish a correspondence between face scans with expression and their neutral pair. A *Morphable Expression Model* (MEM) was constructed by applying PCA to different expressions so that new expressions can be projected into points in a low-dimensional space constructed by the eigen-expressions obtained by MEM. In the work of Mpiperis et al. [109], face and facial expression recognition were performed jointly by decoupling identity and expression components with a bilinear model. To this end, an elastically deformable model algorithm that established correspondence among a set of faces was proposed. Gong et al. [56], approximated the shape of an expressional 3D face as the sum of a basic facial shape component, representing the basic face structure and neutral-style shape, and an expressional shape component that contained shape changes caused by facial expressions. Ferrari et al. [55], constructed a 3DMM and used it as an intermediate mean to deform on 2D images, and then rendered pose normalized images for emotion and AU recognition. In some other approaches, prior knowledge of the face is in the form of a set of landmarks with known spatial layout. Then, landmarks can be used to extract features from the 3D scan and classify them into different expressions. Wang et al. [171], proposed a feature based facial expression descriptor, and the BU-3DFE database was used for the first time. The face was subdivided into seven regions using manually annotated landmarks, and primitive surface features were classified into basic categories using surface curvatures and their principal directions. In the work of Tang et al. [155], a set of candidate features composed of normalized *Euclidean* distances between the 83 facial landmarks of the BU-3DFE database were first extracted. Then, a feature selection method based on maximizing the average relative entropy of marginalized class-conditional feature distributions was used to retain just the most informative distances. A regularized multi-class *AdaBoost* algorithm provided the final classification. Venkatesh et al. [166] used a modified PCA to classify facial expressions using only the shape information at a finite set of fiducial points, which were extracted from the 3D neutral and expressive faces of the BU-3DFE database. Maalej et al. [104], proposed an approach based on the shape analysis of local facial patches. The patches were extracted around the facial landmarks of the BU-3DFE database, and the shape of each patch was described by a set of curves representing the surface points at the same Euclidean distance from the landmark. A Riemannian framework was then applied to compare the shape of curves undergoing to different facial expressions. Berretti et al. [18], first identified a set

of facial keypoints, then computed SIFT feature descriptors of depth images of the face around sample points defined starting from the facial keypoints, and selecting the subset of features with maximum relevance. A SVM classifier was trained for each facial expression to be recognized, subsequently combining them in a multi-class classifier.

In *model-free* methods, facial expression analysis does not depend on any prior shape model of the face. In 2D, a common approach is to extract a high-dimensional dense feature set from the images, then apply dimensionality reduction (e.g., by feature selection). These methods can be applied to 3D data as well, provided that they are first converted to some geometry map. For example, Savran and Sankur [140] proposed a feature extraction approach for 3D FER by incorporating non-rigid registration in face-model-free analysis. Facial information is adapted to the input faces via shape model-free dense registration, which provides a dynamic feature extraction mechanism. This approach eliminates the necessity for complex feature representations as required in the case of static feature extraction methods. Hariri et al. [65] uniformly selected feature points over the whole 3D surface as the center of patches of the face. Each point has a region of influence, which was characterized by the covariance of its geometric features of different type, dimension or scale.

Dynamic Data. Sun and Yin [154] were the first to propose FER from dynamic sequences of 3D scans. Their approach used a generic deformable 3D model whose changes were tracked both in space and time in order to extract a spatio-temporal description of the face. In the temporal analysis, a vertex flow tracking technique was applied to adapt the 3D deformable model to each frame of a 3D face sequence. Correspondences between vertices across the 3D dynamic facial sequences provided a set of motion trajectories (vertex flow) of 3D face scans. Once spatio-temporal features were extracted, a 2D HMM was used for classification (i.e., a spatial HMM and a temporal HMM were used to model the spatial and temporal relationships between the extracted features). The approach proposed by Sandbach et al. [138] exploited the dynamics of 3D facial movements to analyze expressions. This was obtained by first capturing motion between frames using Free-Form Deformations and extracting motion features using a quad-tree decomposition of several motion fields. HMMs were used for temporal modeling of the full expression sequence, which is represented as the composition of *neutral*, *onset*, *apex*, *offset* temporal segments. In [88], Le et al. proposed a level curve based approach to capture the shape of 3D facial models. The level curves were parameterized using the arc-length function. These spatio-temporal features were then used to train a HMM. Fang et al. [53] proposed a fully automatic 4D FER approach with a particular emphasis on 4D data registration and dense correspondence between 3D meshes along the temporal line. The variant of the LBP descriptor which computes LBP on three orthogonal planes was used as face descriptor along the sequence. Reale et al. [133] proposed a 4D spatio-temporal feature named *Nebula* for facial expressions and movement analysis from a volume of 3D data. After fitting the volume data to a cubic polynomial, they proposed to build histograms for different facial regions using geometric features, as curvatures and polar angles. Ben Amor et al. [15] represented 3D faces by collections of radial curves and a Riemannian shape analysis was applied to effectively quantify the deformations induced by the facial expressions in a given subsequence of 3D frames. This was obtained from the dense scalar field, which denoted the shooting directions of the geodesic paths constructed between pairs of corresponding radial curves of two faces. 3D motion extraction with temporal HMM and mean deformation captured with random forest were then used for classification.

However, deploying the above methods in real world applications encounters serious challenges because the human affect states are more complex than the six-classes representation, and spontaneous facial expressions are different from deliberate ones. In particular, several common affects in our daily life communication, like *confused*, *thinking*, *sadness* and *depressed* are not covered by such categorization. To address the limitations of the categorical affect description, a continuous 2D

arousal-valence space has been proposed by Russell and Mehrabian [136]. In this space, the *valence* dimension measures how a human feels, from positive to negative; The *arousal* dimension measures whether humans are more or less likely to take an action under the emotional state, from active to passive [168]. In contrast to the categorical representation, the *arousal-valence* representation enables labeling of a wider range of emotions. More details on emotion representation in continuous space are given by Gunes and Björn [59]. As an example, Alashkar et al. [4] proposed a framework for on-line spontaneous emotion detection, such as happiness or physical pain, from depth videos. The approach mapped the video streams onto a Grassmann manifold (i.e., space of k-dimensional linear subspaces) to form time-parameterized trajectories.

3.1.5 ***Discussion: Issues and Perspectives.*** Acquiring 3D data at high resolution is time demanding and costly. As a consequence, datasets for 3D face recognition still are quite small (few thousands) when compared to 2D images or video datasets (millions of instances). This limits the potential of learning methods that obtained effective results on 2D images. Things can change moving to depth data acquired with cost effective cameras, but this requires methods for de-noising and resolution improvement. In addition, outdoor and at distance 3D acquisition is difficult or even not possible. Missing parts and differences in resolutions are also challenging for 3D data. About aspects that current solutions do not address properly, we can mention the low-resolution of data, as well as the fact that many methods cannot work in real-time, while this constraint can be relevant for interactive applications. The use of posed data of the sole face is also a limitation that is expected to overcome in the near future, with more effective solutions. There is convergence in the idea that multi-modal 2D and 3D solutions can provide best solutions. 3D data are also useful as an intermediate tool for pose estimation and recovery. Better acquisition devices (i.e., with less cost and increased resolution) are expected to be in the market in the near future.

For what concerns FER, one trend is that of overcoming the rigid division into the six universal expressions, considering spontaneous emotions instead. This is expected to involve jointly the face and the upper body, and multi-modal solutions that combine together geometric 3D data with photometric and thermal images. As for a most promising solution, surely deep learning could be successfully applied also in this context. To the best of our knowledge, currently there is no work on deep models for 3D FER. One reason for this, is the absence of abundant training data, which are mandatory for the performance of deep architectures to not degrade significantly. We expect this gap to be closed in the comig years also with the use of data from consumer depth cameras.

3.2 Body

Similar to what was illustrated for face, the use of dynamic or static data is the first aspect that orient methods for body analysis. Data resolution, instead, is less important for body than it is for face. Following this initial subdivision, Figure 5 proposes a categorization of the main body analysis applications. For example, dynamic data acquired with cost effective cameras can be used for understanding what the subjects are doing (action, behavior, activity), which are the subjects' intentions or emotions, or if they interact with other persons or objects. Patterns of movement can be also used for user training or rehabilitation, for identity recognition or re-identification. For these analyses, 3D low-resolution or even skeletal data alone can enable effective results. Static scans of the body are used less frequently. In fact, they can provide higher accuracy in the reconstruction of the 3D body, but they are costly and necessitate user cooperation since a rotating platform is typically used to acquire the whole body. Static acquisitions can be useful for 3D body similarity retrieval or person identity recognition based on body measures. Medical applications where the accurate knowledge of the body dimension is important are also of interest. Construction of a 3DMM of the body, for example to be used for virtual fitting of clothes, or of human characters for

animation and movies is also possible. Based, on the above categorization, in the following, we summarize the literature in this field also discussing current issues and future perspectives.

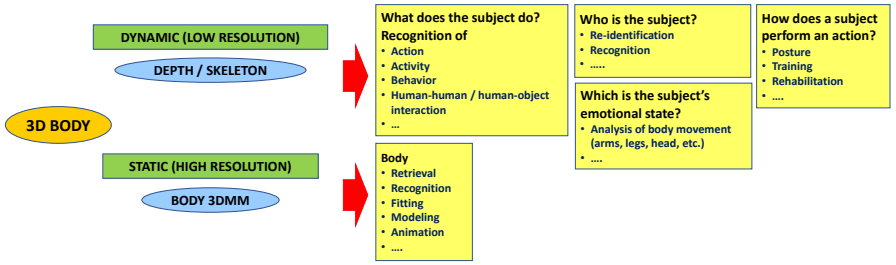


Fig. 5. Overview of 3D body analysis methods and applications. Data of the full body can be acquired with dynamic, and typically low-resolution, scanners. Skeleton and depth data enable a multitude of possible applications that aim answering different questions about the human body movement. Static, and typically high-resolution, acquisitions can be used in further processing, for example by constructing a 3DMM of the body, or directly applied in tasks that go from retrieval of similar 3D bodies, to modeling, and animation.

3.2.1 Static Data. 3D models of humans are commonly used in Computer Graphics and Vision, and so the ability to distinguish between body shapes is an important shape similarity problem. These models are typically acquired with static and high-resolution scanners or produced by modeling softwares. To evaluate the similarity between 3D body models, crucial are the methods used to describe their surface. Many of the descriptors addressed in Section 2 can be used, in combination with different similarity measures. A comparison of such methods is given in the work by Pickup et al. [124], where a total of 25 different shape retrieval methods are evaluated on common benchmark datasets of 3D human models.

Recent developments focused on elastic shape analysis motivated by the fact that it provides a comprehensive framework for simultaneous registration, deformation, and comparison of shapes. These methods achieve computational efficiency using certain square-root representations that transform invariant elastic metrics into Euclidean metrics, allowing for the application of standard algorithms and statistical tools. For analyzing shapes of embeddings, Jermyn et al [75] introduced Square-Root Normal Fields (SRNFs), which transform an elastic metric, with desirable invariant properties, into the metric. These SRNFs are essentially surface normals scaled by square-roots of infinitesimal area elements. Laga et al. [87] took a numerical approach, and derived an multiresolution algorithm, based on solving an optimization problem in the surface space, that estimates surfaces corresponding to given SRNFs. This solution was found to be effective even for complex shapes that undergo significant deformations including bending and stretching, e.g., human bodies.

3.2.2 Dynamic Data. Most of the dynamic data of the body available nowadays come from depth cameras, like Kinect. These cameras have the property of capturing both depth and RGB data of the body at high-frame rate (i.e., 30fps). In addition, the 3D position of a number of joints (20 for Kinect-1, 24 for Kinect-2) that coincide with the main articulations of the human body and other salient points can be estimated. These joints can be used to derive a skeletal representation of the body. Applications that analyze the dynamics of body movement can target different objectives: *i)* understanding what subjects are doing or intend to do. Depending on the duration of the temporal frames across which the analysis is performed and the ultimate goal, this analysis can result in *action*, *behavior* or *activity* recognition, in the prediction of subjects' *intention* as well as in the recognition of subject/subject or subject/object *interaction*. Other possible analysis include how a

certain movement is performed, which has applications in training and rehabilitation; *ii)* recognize or re-identify a subject from the whole body or by characteristic gait patterns only; *iii)* use the body to recognize the emotional state of a subject. This too can involve the whole body or its parts (e.g., upper body, arms or legs). A number of different methods can exploit dynamic data using representations as those introduced in Section 2.2. Before summarizing methods that operate on low-resolution depth data, we focus on 4D body modeling from high-resolution scans.

Body Modeling. Less attention in the Computer Vision research has focused on 4D data; that is 3D scans of moving nonrigid objects, captured over time. To be useful for vision research, such 4D scans need to be registered, or aligned, to a common topology. Consequently, extending mesh registration methods to 4D is important. This problem was addressed by Bogo et al. [27] who created a dataset of high-resolution 4D scans of human subjects in motion, captured at 60fps. They proposed a mesh registration method that uses both 3D geometry and texture information to register all scans in a sequence to a common reference topology. The approach exploits consistency in texture over both short and long time intervals and deals with temporal offsets between shape and texture capture. Pons-Moll et al. [125], learned a model of soft-tissue deformations from examples using a high-resolution 4D capture system and a method that accurately registers a template mesh to sequences of 3D scans. Using over 40,000 scans of ten subjects, we learn how soft-tissue motion causes mesh triangles to deform relative to a base 3D body model. Our Dyna model uses a low-dimensional linear subspace to approximate soft-tissue deformation and relates the subspace coefficients to the changing pose of the body. Dyna uses a second-order auto-regressive model that predicts soft-tissue deformations based on previous deformations, the velocity and acceleration of the body, and the angular velocities and accelerations of the limbs. Dyna also models how deformations vary with a person's body mass index, producing different deformations for people with different shapes. Dyna realistically represents the dynamics of soft tissue for previously unseen subjects and motions.

Action, Behavior, and Interaction. In recent years, recognition and understanding of human behavior by analyzing depth data has attracted increasing interest. While some methods focus on the analysis of human motion in order to recognize human *gestures* or *actions*, other approaches try to also model more complex behaviors (*activities*), which include object interaction. These solutions focus on the analysis of short sequences, where one single behavior is performed along the sequence. The approaches proposed so far can be grouped into three main categories, according to the way they use the depth channel: *skeleton*-based, *depth map*-based and *hybrid* approaches. *Multi-modal* methods that exploit both depth and photometric information are also possible.

Skeleton based approaches have become popular thanks to the work of Shotton et al. [146], where a real-time method to accurately predict the 3D positions of body joints in individual depth maps, without using any temporal information, was described. The basic idea of these methods is to model the pose of the human body in subsequent frames of a sequence using the position and the relationships between joints. In [181], Xia et al., proposed an action recognition approach based on the histograms of the position of 12 joints. The histograms were projected using LDA and clustered into k posture visual words, representing the prototypical poses of the actions. The temporal evolution of these visual words was modeled by discrete HMM. Yang and Tian [185], performed human action recognition by extracting three features for each joint, based on pair-wise differences of joint positions in the current frame, between the current frame and the previous frame, and between the current frame and the initial frame of the sequence. PCA was used to reduce redundancy and noise, and to obtain a compact *EigenJoints* representation of each frame. Finally, a naïve-Bayes nearest-neighbor classifier was used for multi-class action classification. Similar features were used by Luo et al. [103], but pair-wise differences were computed only in the current frame and with respect to only one reference joint (the hip joint). To better represent

these features, they proposed a dictionary learning method based on group sparsity and geometry constraints. The classification of sequences was performed using SVM. Zanfir et al. [190] proposed the Moving Pose feature, capturing for each frame the human pose information as well as the speed and acceleration of body joints within a short temporal window. A modified k-NN classifier was employed to perform action recognition. Hongzhao et al. [38] introduced a part-based feature vector to identify the most relevant body parts in each action sequence. Recent works addressed more complex challenges in on-line action recognition systems, where a trade-off between accuracy and latency becomes an important goal. For example, Ellis et al. [51] targeted this trade-off by adopting a Latency Aware Learning method for reducing latency when recognizing human actions. A logistic regression-based classifier was trained on 3D joint position sequences to search a single canonical posture for recognition. Other approaches used differential geometry to represent skeleton data. In [163], Vemulapalli and Chellappa represented each skeleton as one element on the Lie-group, and the sequence corresponds to a curve on this manifold. In [150], Slama et al. expressed the time series of skeletons as one point on a Grassmann manifold, where the classification is performed benefiting from Riemannian geometry of this manifold. Anirudh et al. [9], modeled actions as trajectories on a Riemannian manifold, and analysis of such trajectories using Transport Square-Root Velocity Function is employed for action recognition. The solution proposed by Devanne et al. [45], developed on the idea of representing the 3D coordinates of the joints of the skeleton and their change over time as a trajectory in a suitable *action space*. This solution is capable of capturing both the shape and the dynamics of the human body simultaneously. The action recognition problem is then formulated as the problem of computing the similarity between the shape of trajectories in a Riemannian manifold. A more comprehensive review on space-time representations of people based on 3D skeletal data can be found in the survey of Han et al. [62], while the survey of Lo Presti and La Cascia [99] focuses on 3D skeleton methods for action classification.

Depth map based approaches extract volumetric and temporal features directly from the overall set of points of the depth maps in the sequence. The approach by Li et al. [94] employed 3D human silhouettes to describe salient postures and used an action graph to model the dynamics of the actions. Yang et al. [186], modeled the action dynamics by using Depth Motion Maps, which highlight areas where some motion takes place. Other methods, such as Spatio-Temporal Occupancy Pattern [167], Random Occupancy Pattern [169] and Depth Cuboid Similarity Feature [180], proposed to work on the 4D space divided into spatio-temporal boxes to extract features representing the depth appearance in each box. Such features are extracted from Spatio-Temporal Interest Points. A similar method was proposed by Rahmani et al. [130], where keypoints were detected and the point cloud was described within a volume using the Histogram of Principal Components. Oreifej and Liu [115] proposed a method to quantize the 4D space using vertices of a polychoron and then modeled the distribution of the normal vectors for each cell. Althloothi et al. [7] represented 3D shape features based on spherical harmonics representation and 3D motion features using kinematic structure from skeleton. Both features were then merged using a multi-kernel learning method. A depth feature to describe shape geometry and motion, called Range-Sample, was proposed by Lu and Tang [101]. Depth information can also be used in combination with color images as in [113].

Analyzing human motion, however, may not be sufficient to understand more complex behaviors involving human interaction with the environment (i.e., what we call *activities*). Hybrid solutions are often proposed that use depth maps for modeling scene objects and body skeleton for modeling human motion. For example, Wang et al. [170] used Local Occupancy Patterns to represent the observed depth values in correspondence to skeleton joints. Ohn-Bar et al. [114], characterized actions using pairwise affinity measures between joint angle features and histogram of oriented gradients computed on depth maps. Other methods proposed to describe and model spatio-temporal interaction between humans and objects characterizing the activities, using Markov Random

Field [84]. A graphical model was also employed by Wei et al. [175] to hierarchically define activities as combination of sub-events including description of the human pose, the object and interaction between them. Yu and Liu [188] proposed to capture meaningful skeleton and depth features using a middle level representation called *orderlet*.

Additional challenges appear when several different behaviors are executed one after the other over a long sequence. In order to face these challenges, methods based on *online detection* have been proposed. Such methods can recognize behaviors before the end of their execution by analyzing short parts of the observed sequence; thus, multiple behaviors can be detected within a long sequence, which may not be the case for methods analyzing the entire sequence directly. Some of the works reviewed above have such *online* action recognition capability, as they compute their features within a short sliding window along the sequence [188]. For example, Huang et al. [69] proposed and applied the Sequential Max-Margin Event Detector algorithm on long sequences comprising many actions in order to perform on-line detection by successively discarding not corresponding action classes. Devanne et al. [44] proposed a framework for analyzing and understanding human behavior from depth videos. The solution first employed shape analysis of the human pose across time to decompose the full motion into short temporal segments representing elementary motions. Then, each segment was characterized by human motion and the depth appearance around hand joints so as to describe the change in pose of the body and the interaction with objects. Finally, the sequence of temporal segments was modeled through a Dynamic Naive Bayes classifier, which captured the dynamics of elementary motions characterizing human behavior.

Deep learning methods based on the CNN and RNN architectures have been also adopted for motion recognition using RGB-D data. This by itself includes a large corpus of methods that can be broadly categorized into four groups, depending on the modality adopted for recognition: RGB-based, depth-based, skeleton-based and multi-modal-based. We refer to the survey by Wang et al. [172] for a detailed overview of recent advances in RGB-D-based motion recognition using DL. Advantages and limitations of existing techniques are discussed. Particularly, the methods of encoding spatial-temporal-structural information inherent in video sequence are highlighted, while potential directions for future research are discussed.

Gait and Re-Identification. Person (re-)identification through gait analysis, as an interesting and non-intrusive biometric means, has been extensively studied in recent years. Several methods have been proposed for gait recognition by using depth or skeleton data captured by RGB-D sensors.

Several depth-based studies have applied anthropometric and soft biometrics concepts to the case of 3D human skeleton [5]. Some works investigated 3D point clouds for person identification [72, 192], by using hand-crafted features (e.g., arm length, torso width) or low-level RGB features (e.g., SURF, SIFT). These methods have largely ignored spatio-temporal information. One idea that has been used to overcome such limitation is that of embedding temporal information onto a 2D image by averaging the silhouette across all frames of a video. For example, this solution has been used in the gait energy image [105] and its variants [14]. The gait energy image has been subsequently extended to 3D by using depth sensors. For example, Sivapalan et al. [149] proposed the Gait Energy Volume feature averaging the voxel volumes, derived from depth images, over an entire gait cycle. An extended feature was proposed by Chattopadhyay et al. [34] by combining depth and RGB data to reduce noise and considering only a set of key poses during the gait cycle. Conversely, other methods proposed to investigate skeleton data to derive motion features. Yang et al. [183] focused on human recognition with relative distance-based gait features. Preis et al. [127] employed skeleton data to define thirteen biometric features including subject height, length of several limbs, length of steps and speed. Similar soft-biometric features were computed by Chattopadhyay et al. [35] in addition to kinematic features analyzing variation of each skeleton joint. The combination

of such features improved the gait recognition accuracy. In [36], the same authors combined front and back views of the gait to compute their features. Due to the poor accuracy of skeleton data of the back view, depth-based features were computed in addition to skeleton features from the front view. Devanne et al. [46] proposed a gait analysis method from depth sequences by analyzing separately each step so as to be robust to gait duration and incomplete cycles. The shape of the motion trajectory was analyzed as signature of the gait, and shape variations within a Riemannian manifold were considered to learn step models. Haque et al. [63] presented an attention-based model that reasons on human body shape and motion dynamics to identify individuals in the absence of RGB information (i.e., in the dark). The approach leverages unique 4D spatio-temporal signatures to address the identification problem across days. Formulated as a reinforcement learning task, the model is based on a combination of Convolutional and Recurrent NNs with the goal of identifying small, discriminative regions indicative of human identity.

Emotion. The availability of whole-body sensing technology makes it possible to investigate the role played by body expressions as a powerful affective communication channel. For a comprehensive coverage of the topic, we refer to the surveys by Kleinsmith and Bianchi-Berthouze [82] that reviewed the literature on affective body expression perception and recognition, and by Karg et al. [78] that summarized methods to recognize affective expressions from body movements, and the converse problem of generating movements for virtual agents or robots.

Kapur et al. [77] were among the first to address these aspects in 3D. Using a Vicon Motion Capture system, they collected gestural sequence data depicting sadness, joy, anger, and fear emotions. The 3D position of 14 markers, plus their velocity and acceleration were calculated, and the mean values of velocity and acceleration and the standard deviation values of position, velocity and acceleration across the sequence were considered as descriptors. Classification was finally performed comparing the output of five different classifiers. Gong et al. [57], addressed the problem of recognizing affect from non-stylized human body motion using 3D joints of the skeleton. Motion capture data were represented by a descriptor based on the shape of signal probability density function, and SVM was used for classification. Karg et al. [79] analyzed the human gait to reveal persons affective state, comparing inter-individual versus person dependent recognition. The dynamics of the body was captured by measuring features such as the stride length, cadence, velocity, minimum mean and maximum values of angles between body parts. Then, these features were reduced using PCA, kernel PCA, LDA and GDA techniques, while classification was performed with nearest-neighbor, Naive-Bayes and SVM. They observed that automatic recognition based on gait patterns tends to better recognize *active* than *passive* emotional states. Hicheur et al. [67] presented a study aimed at investigating how emotion affect the kinematic aspect of human walking. They showed both step-related behavioral changes (in terms of step length, speed, etc.) that are common to different emotions and emotion-specific body configuration changes (mainly at the level of the upper body posture) during emotional gaits. In the work of Daoudi et al. [43], a scenario was proposed for which emotional states are related to 3D dynamics of the whole body motion (see Fig. 6). To address the complexity of human body movement, the covariance descriptors of the sequence of the 3D skeleton joints have been used, and represented them in the non-linear Riemannian manifold of SPD matrices.

3.2.3 ***Discussion: Issues and Perspectives.*** Analysis of human body from 3D dynamic data is a quite recent research topic. Several issues limit current solutions. One problem is technological and comes from the device that are available nowadays. Kinect like cameras allowed a great advancement in the field, but such devices still are limited to indoor environments with an operation range of few meters. MoCap systems smooth some of the constraints of depth cameras, but are costly, require complex setting, and use body markers. To come to a deployment in a larger spectrum of real

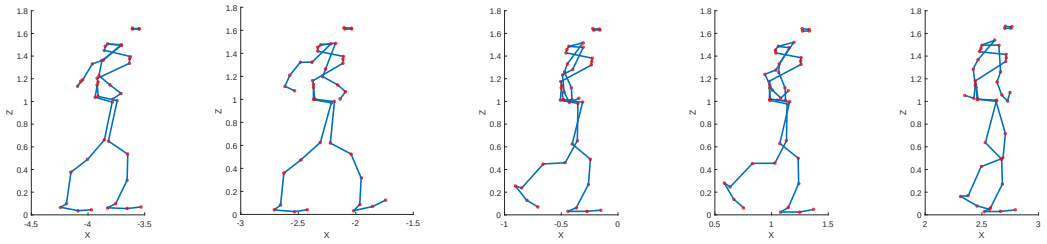


Fig. 6. Frames from a MoCap skeleton sequence, where an actor moves following a "U" shaped trajectory showing an *anger* emotion. In each frame, the skeleton is represented by 43 joints.

application contexts, new and more effective sensors are needed. In terms of applications, most of the current solutions focus on recognition of well defined actions in relatively small datasets. And, in most of the cases, this is done using just one media (depth, skeleton or RGB). In addition, the large majority of approaches consider off-line processing, requiring the entire sequence to analyze be available in order to elaborate it. This does not allow the deployment in online applications, where the dynamic stream must be analysed on-the-fly, while it is acquired.

In the representation and analysis of human body, there are many open questions that require further investigation and advancement. First, how to pass from action recognition to behavior and activity recognition is a problem that has not been addressed satisfactorily yet. This can require the capability of analysing online and across time the incoming stream of 3D frames, identifying the temporal intervals corresponding to the start and end of specific actions. A higher-level analysis could then try to infer behavioral understanding on the sequence of elementary actions. In performing such analysis, improvements can derive by the joint use and fusion of different data, such as skeleton, depth and RGB.

4 CONCLUSIONS

The interest in the automatic analysis of humans for a variety of applications that go from identity recognition to action detection and prediction has dramatically increased in the last few years. Though image and video data still dominate the scene, 3D data appear as one of the most attractive solutions to go beyond the limits imposed by the flat nature of 2D data, with the promise to open the way to the solution of problems that are difficult in 2D.

In this survey, we have presented a comprehensive view of the solutions for the representation, analysis and recognition of 3D humans. Though this area has attracted considerable interest only quite recently, vast literature already exists, which is continuously growing. This oriented us more in identifying classes of problems that have 3D humans at the center, providing methods, open issues, and prospecting new and interesting research directions, rather than going into depth with method details. In doing so, we renounced to be exhaustive in the number of methods reported, focusing more on the identification of mathematical and theoretical solutions that opened new directions and that have been adopted in different works. Finally, compared to surveys that address specific aspects in 3D data analysis, this work is unique in keeping its focus on humans data.

REFERENCES

[1] K. Al Ismaeil, D. Aouada, B. Mirbach, and B. Ottersten. 2016. Enhancement of dynamic depth scenes by upsampling for precise super-resolution (UP-SR). *Computer Vision and Image Understanding* 147 (2016), 38–49.

[2] K. Al Ismaeil, D. Aouada, T. Solignac, B. Mirbach, and B. Ottersten. 2015. Real-time non-rigid multi-frame depth video super-resolution. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. 8–16.

[3] K. Al Ismaeil, D. Aouada, T. Solignac, B. Mirbach, and B. Ottersten. 2017. Real-Time Enhancement of Dynamic Depth Videos with Non-Rigid Deformations. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 39, 10 (Oct 2017), 2045–2059.
[4] T. Alashkar, B. Ben Amor, M. Daoudi, and S. Berretti. 2018. Spontaneous Expression Detection from 3D Dynamic Sequences by Analyzing Trajectories on Grassmann Manifolds. *IEEE Trans. on Affective Computing* (to appear 2018).
[5] A. Albiol, A. Albiol, J. Oliver, and J. M. Mossi. 2012. Who is who at different cameras: people re-identification using depth cameras. *IET Computer Vision* 6, 5 (Sept 2012), 378–387.
[6] S. Ali, A. Basharat, and M. Shah. 2007. Chaotic Invariants for Human Action Recognition. In *IEEE Int. Conf. on Computer Vision*. 1–8.
[7] S. Althloothi, M. H. Mahoor, X. Zhang, and R. M. Voyles. 2014. Human Activity Recognition Using Multi-features and Multiple Kernel Learning. *Pattern Recognition* 47, 5 (May 2014), 1800–1812.
[8] B. Amberg, S. Romdhani, and T. Vetter. 2007. Optimal Step Nonrigid ICP Algorithms for Surface Registration. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 1–8.
[9] R. Anirudh, P. Turaga, J. Su, and A. Srivastava. 2015. Elastic functional coding of human actions: From vector-fields to latent variables. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 3147–3155.
[10] R. Anirudh, P. Turaga, J. Su, and A. Srivastava. 2017. Elastic Functional Coding of Riemannian Trajectories. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 39, 5 (May 2017), 922–936.
[11] G. Antini, S. Berretti, A. Del Bimbo, and P. Pala. 2005. 3D Mesh Partitioning for Retrieval by Parts Applications. In *IEEE Int. Conf. on Multimedia and Expo*. 1210–1213.
[12] D. Aouada, S. Feng, and H. Krim. 2007. Statistical Analysis of the Global Geodesic Function for 3D Object Classification. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 1. I–645–I–648.
[13] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. 2005. Fast and Simple Calculus on Tensors in the Log-Euclidean Framework. In *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention*. 115–122.
[14] K. Bashir, T. Xiang, and S. Gong. 2010. Gait recognition without subject cooperation. *Pattern Recognition Letters* 31, 13 (2010), 2052–2060.
[15] B. Ben Amor, H. Drira, S. Berretti, M. Daoudi, and A. Srivastava. 2014. 4-D Facial Expression Recognition by Learning Geometric Deformations. *IEEE Trans. on Cybernetics* 44, 12 (Dec 2014), 2443–2457.
[16] B. Ben Amor, J. Su, and A. Srivastava. 2016. Action Recognition Using Rate-Invariant Analysis of Skeletal Shape Trajectories. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 38, 1 (2016), 1–13.
[17] L. Benedikt, D. Cosker, P. L. Rosin, and D. Marshall. 2010. Assessing the Uniqueness and Permanence of Facial Actions for Use in Biometric Applications. *IEEE Trans. on Systems, Man, and Cybernetics - Part A: Systems and Humans* 40, 3 (May 2010), 449–460.
[18] S. Berretti, B. Ben Amor, M. Daoudi, and A. Del Bimbo. 2011. 3D facial expression recognition using SIFT descriptors of automatically detected keypoints. *The Visual Computer* 27, 11 (Jun 2011), 1021.
[19] S. Berretti, A. Del Bimbo, and P. Pala. 2009. 3D Mesh decomposition using Reeb graphs. *Image and Vision Computing* 27, 10 (2009), 1540 – 1554.
[20] S. Berretti, A. Del Bimbo, and P. Pala. 2010. 3D Face Recognition Using Isogeodesic Stripes. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32, 12 (Dec 2010), 2162–2177.
[21] S. Berretti, A. Del Bimbo, and P. Pala. 2013. Sparse Matching of Salient Facial Curves for Recognition of 3-D Faces With Missing Parts. *IEEE Trans. on Information Forensics and Security* 8, 2 (Feb 2013), 374–389.
[22] S. Berretti, P. Pala, and A. Del Bimbo. 2014. Face Recognition by Super-Resolved 3D Models From Consumer Depth Cameras. *IEEE Trans. on Information Forensics and Security* 9, 9 (Sept 2014), 1436–1449.
[23] P. J. Besl and N. D. McKay. 1992. A method for registration of 3-D shapes. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 14, 2 (Feb 1992), 239–256.
[24] R. Bhatia. 2007. *Positive Definite Matrices*. Princeton.
[25] V. Blanz and T. Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Annual Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*. 187–194.
[26] V. Blanz and T. Vetter. 2003. Face Recognition based on Fitting a 3D Morphable Model. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 25, 9 (Sept 2003), 1063–1074.
[27] F. Bogo, J. Romero, G. Pons-Moll, and M. J. Black. 2017. Dynamic FAUST: Registering Human Bodies in Motion. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 5573–5582.
[28] E. Bondi, P. Pala, S. Berretti, and A. Del Bimbo. 2016. Reconstructing High-Resolution Face Models from Kinect Depth Sequences. *IEEE Trans. on Information Forensics and Security* 11, 12 (2016), 2843–2853.
[29] S. Bonnabel and R. Sepulchre. 2010. Riemannian metric and geometric mean for positive semidefinite matrices of fixed rank. *SIAM J. Matrix Anal. Appl.* 31, 3 (2010), 1055–1070.
[30] J. Booth, A. Roussos, S. Zafeiriou, A. Ponniahand, and D. Dunaway. 2016. A 3D Morphable Model Learnt From 10,000 Faces. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 5543–5552.

[31] A. M. Bronstein, M. M. Bronstein, and R. Kimmel. 2005. Three dimensional face recognition. *Int. Journal of Computer Vision* 64, 1 (Aug. 2005), 5–30.
[32] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. 2017. Geometric Deep Learning: Going beyond Euclidean data. *IEEE Signal Processing Magazine* 34, 4 (July 2017), 18–42.
[33] A. Brunton, T. Bolkart, and S. Wuhrer. 2014. Multilinear Wavelets: A Statistical Shape Space for Human Faces. In *European Conf. on Computer Vision*. 297–312.
[34] P. Chattopadhyay, A. Roy, S. Sural, and J. Mukhopadhyay. 2014. Pose Depth Volume extraction from RGB-D streams for frontal gait recognition. *Journal of Visual Communication and Image Representation* 25, 1 (2014), 53–63.
[35] P. Chattopadhyay, S. Sural, and J. Mukherjee. 2014. Frontal Gait Recognition From Incomplete Sequences Using RGB-D Camera. *IEEE Trans. on Information Forensics and Security* 9, 11 (Nov 2014), 1843–1856.
[36] P. Chattopadhyay, S. Sural, and J. Mukherjee. 2015. Frontal gait recognition from occluded scenes. *Pattern Recognition Letters* 63, Supplement C (2015), 9–15.
[37] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. 2009. Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 1932–1939.
[38] H. Chen, G. Wang, J.-H. Xue, and L. He. 2016. A novel hierarchical framework for human action recognition. *Pattern Recognition* 55, Supplement C (2016), 148–159.
[39] J. Choi, A. Sharma, and G. Medioni. 2013. Comparing strategies for 3D face recognition from a 3D sensor. In *IEEE RO-MAN*. 19–24.
[40] A. Colombo, C. Cusano, and R. Schettini. 2006. 3D Face Detection using Curvature Analysis. *Pattern Recognition* 39, 3 (March 2006), 444–455.
[41] N. D. Cornea, D. Silver, and P. Min. 2007. Curve-skeleton properties, applications, and algorithms. *IEEE Trans. on Visualization and Computer Graphics* 13, 3 (2007), 530–548.
[42] C. A. Corneanu, M. O. Simon, J. F. Cohn, and S. Escalera Guerrero. 2016. Survey on RGB, 3D, Thermal, and Multimodal Approaches for Facial Expression Recognition: History, Trends, and Affect-Related Applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 38, 8 (January 2016), 1548–1568.
[43] M. Daoudi, S. Berretti, P. Pala, Y. Delevoye, and A. Del Bimbo. 2017. Emotion Recognition by Body Movement Representation on the Manifold of Symmetric Positive Definite Matrices. In *Int. Conf. on Image Analysis and Processing*. 550–560.
[44] M. Devanne, S. Berretti, P. Pala, H. Wannous, M. Daoudi, and A. Del Bimbo. 2017. Motion segment decomposition of RGB-D sequences for human behavior understanding. *Pattern Recognition* 61, Supplement C (2017), 222–233.
[45] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo. 2015. 3-D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold. *IEEE Trans. on Cybernetics* 45, 7 (July 2015), 1340–1352.
[46] M. Devanne, H. Wannous, M. Daoudi, S. Berretti, A. Del Bimbo, and P. Pala. 2016. Learning shape variations of motion trajectories for gait analysis. In *Int. Conf. on Pattern Recognition*. 895–900.
[47] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, and R. Slama. 2013. 3D Face Recognition under Expressions, Occlusions, and Pose Variations. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 35, 9 (2013), 2270–2283.
[48] P. Ekman. 1972. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*, Vol. 19. 207–283.
[49] P. Ekman and W. V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement.*
[50] A. Elad and R. Kimmel. 2003. On bending invariant signatures for surfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 25, 10 (Oct 2003), 1285–1295.
[51] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. La Viola Jr., and R. Sukthankar. 2013. Exploring the Trade-off Between Accuracy and Observational Latency in Action Recognition. *Int. Journal of Computer Vision* 101, 3 (2013), 420–436.
[52] T. C. Faltemier, K. W. Bowyer, and P. J. Flynn. 2007. Using a Multi-Instance Enrollment Representation to Improve 3D Face Recognition. In *IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems*. 1–6.
[53] T. Fang, X. Zhao, S.K. Shah, and I.A. Kakadiaris. 2011. 4D facial expression recognition. In *IEEE Int. Conf. on Computer Vision Workshop*. 1594–1601.
[54] Y. Fang, J. Xie, G. Dai, M. Wang, F. Zhu, T. Xu, and E. Wong. 2015. 3D deep shape descriptor. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 2319–2328.
[55] C. Ferrari, G. Lisanti, S. Berretti, and A. Del Bimbo. 2017. A Dictionary Learning-Based 3D Morphable Shape Model. *IEEE Trans. on Multimedia* 19, 12 (Dec 2017), 2666–2679.
[56] B. Gong, Y. Wang, J. Liu, and X. Tang. 2009. Automatic facial expression recognition on a single 3D face by exploring shape deformation. In *ACM Int. Conf. on Multimedia*. 569–572.
[57] L. Gong, T. Wang, C. Wang, F. Liu, F. Zhang, and X. Yu. 2010. Recognizing Affect from Non-stylized Body Motion Using Shape of Gaussian Descriptors. In *ACM Symposium on Applied Computing*. 1203–1206.

[58] R. Grossmann, N. Kiryati, and R. Kimmel. 2002. Computational surface flattening: a voxel-based approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24, 4 (Apr 2002), 433–441.
[59] H. Gunes and B. Schuller. 2013. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing* 31, 2 (2013), 120–136.
[60] N. Hajari, I. Cheng, and A. Basu. 2016. Robust human animation skeleton extraction using compatibility and correctness constraints. In *IEEE Int. Symposium on Multimedia*. 1–4.
[61] B. Hall. [n. d.]. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction.*
[62] F. Han, B. Reily, W. Hoff, and H. Zhang. 2017. Space-time representation of people based on 3D skeletal data: A review. *Computer Vision and Image Understanding* 158, Supplement C (2017), 85–105.
[63] A. Haque, A. Alahi, and L. Fei-Fei. 2016. Recurrent Attention Models for Depth-Based Person Identification. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 1229–1238.
[64] M. T. Harandi, R. I. Hartley, B. C. Lovell, and C. Sanderson. 2016. Sparse Coding on Symmetric Positive Definite Manifolds Using Bregman Divergences. *IEEE Trans. on Neural Networks and Learning Systems* 27, 6 (2016), 1294–1306.
[65] W. Hariri, H. Tabia, N. Farah, A. Benouareth, and D. Declercq. 2017. 3D facial expression recognition using kernel methods on Riemannian manifold. *Engineering Applications of Artificial Intelligence* 64, Supplement C (2017), 25–32.
[66] M. Hernandez, J. Choi, and G. Medioni. 2012. Laser scan quality 3-D face modeling using a low-cost depth camera. In *European Signal Processing Conf.* 1995–1999.
[67] H. Hicheur, H. Kadone, J. Grèzes, and A. Berthoz. 2013. The Combined Role of Motion-Related Cues and Upper Body Posture for the Expression of Emotions during Human Walking. In *Modeling, Simulation and Optimization of Bipedal Walking*. 71–85.
[68] S. Hochreiter and J. Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780.
[69] D. Huang, Y. Wang, S. Yao, and F. De la Torre. 2014. Sequential Max-Margin Event Detectors. In *European Conf. on Computer Vision*. 410–424.
[70] W. Huang, F. Sun, L. Cao, D. Zhao, H. Liu, and M. Harandi. 2016. Sparse Coding and Dictionary Learning With Linear Dynamical Systems. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 3938–3947.
[71] P. Huber, P. Kopp, W. Christmas, M. RÃďtsch, and J. Kittler. 2017. Real-Time 3D Face Fitting and Texture Fusion on In-the-Wild Videos. *IEEE Signal Processing Letters* 24, 4 (April 2017), 437–441.
[72] D. Ioannidis, D. Tzovaras, I. G. Damousis, S. Argyropoulos, and K. Moustakas. 2007. Gait Recognition Using Compact Feature Extraction Transforms and Depth Information. *IEEE Trans. on Information Forensics and Security* 2, 3 (Sept 2007), 623–630.
[73] S. Izadi, R. A. Newcombe, D. Kim, O. Hilliges, D. Molyneaux, S. Hodges, P. Kohli, J. Shotton, A. J. Davison, and A. Fitzgibbon. 2011. KinectFusion: Real-time Dynamic 3D Surface Reconstruction and Interaction. In *ACM SIGGRAPH*. 23:1–23:1.
[74] A. Jalal, Y.-H. Kim, Y.-J. Kim, S. Kamal, and D. Kim. 2017. Robust human activity recognition from depth video using spatio temporal multi-fused features. *Pattern Recognition* 61, Supplement C (2017), 295–308.
[75] I. H. Jermyn, S. Kurtek, E. Klassen, and A. Srivastava. 2012. Elastic Shape Matching of Parameterized Surfaces Using Square Root Normal Fields. In *European Conf. on Computer Vision*. 804–817.
[76] A. E. Johnson and M. Hebert. 1999. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 21, 5 (May 1999), 433–449.
[77] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen. 2005. Gesture-Based Affective Computing on Motion Capture Data. In *Int. Conf. on Affective Computing and Intelligent Interaction*. 1–7.
[78] A. Karg, A.-A. Samadani, R. Gorbet, K. Kühnlenz, J. Hoey, and D. Kulić. 2013. Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation. *IEEE Trans. on Affective Computing* 4, 4 (2013), 341–359.
[79] M. Karg, K. Kuhnlenz, and M. Buss. 2010. Recognition of Affect Based on Gait Patterns. *IEEE Trans. on Systems, Man, and Cybernetics, Part B* 40, 4 (2010), 1050–1061.
[80] D. G. Kendall. 1984. Shape manifolds, procrustean metrics and complex projective spaces. *Bulletin of the London Mathematical Society* 16 (1984), 81âĂŞ121.
[81] D. Kim, M. Hernandez, J. Choi, and G. Medioni. 2017. Deep 3D Face Identification. *CoRR* abs/1703.10714 (2017). arXiv:1703.10714 http://arxiv.org/abs/1703.10714
[82] A. Kleinsmith and N. Bianchi-Berthouze. 2013. Affective Body Expression Perception and Recognition: A Survey. *IEEE Trans. on Affective Computing* 4, 1 (Jan 2013), 15–33.
[83] I. Kokkinos, M. M. Bronstein, R. Litman, and A. M. Bronstein. 2012. Intrinsic shape context descriptors for deformable shapes. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 159–166.
[84] H. S. Koppula, R. Gupta, and A. Saxena. 2013. Learning human activities and object affordances from RGB-D videos. *Int. Journal of Robotics Research* 32, 8 (July 2013), 951–970.
[85] M. Körtgen, G.-J. Park, M. Novotni, and R. Klein. 2003. 3D Shape Matching with 3D Shape Contexts. In *Central European Seminar on Computer Graphics*.

[86] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Int. Conf. on Neural Information Processing Systems*, Vol. 1. 1097–1105.
[87] H. Laga, Q. Xie, I. H. Jermyn, and A. Srivastava. 2017. Numerical Inversion of SRNF Maps for Elastic Shape Analysis of Genus-Zero Surfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 39, 12 (Dec 2017), 2451–2464.
[88] V. Le, H. Tang, and T. S. Huang. 2011. Expression Recognition from 3D Dynamic Faces using Robust Spatio-temporal Shape Features. In *IEEE Conf. on Automatic Face and Gesture Recognition*. 414–421.
[89] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521 (May 2015), 436–444. Issue 7553.
[90] Y. Lee, J. Chen, C. W. Tseng, and S.-H. Lai. 2016. Accurate and robust face recognition from RGB-D images with a deep learning approach. In *British Machine Vision Conf.* 123.1–123.14.
[91] B. Y. L. Li, A. S. Mian, W. Liu, and A. Krishna. 2013. Using Kinect for face recognition under varying poses, expressions, illumination and disguise. In *IEEE Workshop on Applications of Computer Vision*. 186–192.
[92] H. Li, L. Chen, D. Huang, Y. Wang, and J. Morvan. 2015. Towards 3D Face Recognition in the Real: A Registration-Free Approach Using Fine-Grained Matching of 3D Keypoint Descriptors. *Int. Journal of Computer Vision* 113, 2 (June 2015), 128–142.
[93] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graphics* 36, 6 (2017), 194:1–194:17.
[94] W. Li, Z. Zhang, and Z. Liu. 2010. Action recognition based on a bag of 3D points. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. 9–14.
[95] B. Liang and L. Zheng. 2015. A Survey on Human Action Recognition Using Depth Sensors. In *Int. Conf. on Digital Image Computing: Techniques and Applications*. 1–8.
[96] S. Liang, I. Kemelmacher-Shlizerman, and L. G. Shapiro. 2014. 3D Face Hallucination from a Single Depth Frame. In *Int. Conf. on 3D Vision*, Vol. 1. 31–38.
[97] H. Ling and D. W. Jacobs. 2005. Using the inner-distance for classification of articulated shapes. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Vol. 2. 719–726.
[98] J. Liu, A. Shahroudy, D. Xu, A. Kot Chichung, and G. Wang. 2017. Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (to appear 2017).
[99] L. Lo Presti and M. La Cascia. 2016. 3D skeleton-based human action classification: A survey. *Pattern Recognition* 53, 5 (2016), 130–147.
[100] M. LÃijthi, T. Gerig, C. Jud, and T. Vetter. 2017. Gaussian Process Morphable Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2017), 1–1.
[101] C. Lu, J. Jia, and C.-K. Tang. 2014. Range-Sample Depth Feature for Action Recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 772–779.
[102] Y. M. Lui. 2012. Advances in matrix manifolds for computer vision. In *Image and Vision Computing*, Vol. 30. 380–388.
[103] J. Luo, W. Wang, and H. Qi. 2013. Group Sparsity and Geometry Constrained Dictionary Learning for Action Recognition from Depth Maps. In *IEEE Int. Conf. on Computer Vision*. 1809–1816.
[104] A. Maalej, B. Ben Amor, M. Daoudi, A. Srivastava, and S. Berretti. 2010. Local 3D Shape Analysis for Facial Expression Recognition. In *Int. Conf. on Pattern Recognition*. 4129–4132.
[105] J. Man and B. Bhanu. 2006. Individual recognition using gait energy image. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 28, 2 (Feb 2006), 316–322.
[106] M. Meyer, M. Desbrun, P. Schröder, and A. H. Barr. 2003. Discrete Differential-Geometry Operators for Triangulated 2-Manifolds. In *Visualization and Mathematic*. Springer, Berlin, Heidelberg, 35–57.
[107] A. S. Mian, M. Bennamoun, and R. Owens. 2007. An Efficient Multimodal 2D-3D Hybrid Approach to Automatic Face Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 29, 11 (Nov. 2007), 1927–1943.
[108] A. S. Mian, M. Bennamoun, and R. Owens. 2008. Keypoint Detection and Local Feature Matching for Textured 3D Face Recognition. *Int. Journal of Computer Vision* 79, 1 (Aug. 2008), 1–12.
[109] I. Mpiperis, S. Malassiotis, and M.G. Strintzis. 2008. Bilinear Models for 3-D Face and Facial Expression Recognition. *IEEE Trans. on Information Forensics and Security* 3, 3 (Sept. 2008), 498–511.
[110] A. Myronenko and X. Song. 2010. Point Set Registration: Coherent Point Drift. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32, 12 (Dec 2010), 2262–2275.
[111] R. A. Newcombe, D. Fox, and S. M. Seitz. 2015. DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 343–352.
[112] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. 2011. KinectFusion: Real-time dense surface mapping and tracking. In *IEEE Int. Symposium on Mixed and Augmented Reality*. 127–136.
[113] B. Ni, Y. Pei, P. Moulin, and S. Yan. 2013. Multi-level Depth and Image Fusion for Human Activity Detection. *IEEE Trans. on Cybernetics* 43, 5 (Oct 2013), 1383–1394.

[114] E. Ohn-Bar and M. M. Trivedi. 2013. Joint Angles Similarities and HOG^2 for Action Recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. 465–470.
[115] O. Oreifej and Z. Liu. 2013. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 716–723.
[116] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. 2002. Shape Distributions. In *ACM Trans. on Graphics*, Vol. 21. 807–832.
[117] G. Pan, S. Han, Z. Wu, and Y. Wang. 2006. Super-Resolution of 3D Face. In *European Conf. on Computer Vision*. 389–401.
[118] V. Papyan, Y. Romano, and M. Elad. 2017. Convolutional Neural Networks Analyzed via Convolutional Sparse Coding. *Journal of Machine Learning Research* 18, 83 (2017), 1–52.
[119] O. M. Parkhi, A. Vedaldi, and A. Zisserman. 2015. Deep Face Recognition. In *British Machine Vision Conf.*, Vol. 1. 1–12.
[120] A. Patel and W. A. P. Smith. 2009. 3D morphable face models revisited. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 1327–1334.
[121] H. Patil, A. Kothari, and K. Bhurchandi. 2015. 3-d face recognition: features, databases, algorithms and challenges. *Artificial Intelligence Review* 44, 3 (2015), 393–441.
[122] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. 2009. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*. 296–301.
[123] X. Pennec, P. Fillard, and N. Ayache. 2006. A Riemannian Framework for Tensor Computing. *Int. Journal of Computer Vision* 66, 1 (2006), 41–66.
[124] D. Pickup, X. Sun, P. L. Rosin, R. R. Martin, Z. Cheng, Z. Lian, M. Aono, A. Ben Hamza, A. Bronstein, M. Bronstein, S. Bu, U. Castellani, S. Cheng, V. Garro, A. Giachetti, A. Godil, L. Isaia, J. Han, H. Johan, L. Lai, B. Li, C. Li, H. Li, R. Litman, X. Liu, Z. Liu, Y. Lu, L. Sun, G. Tam, A. Tatsuma, and J. Ye. 2016. Shape Retrieval of Non-rigid 3D Human Models. *Int. Journal of Computer Vision* 120, 2 (2016), 169–193.
[125] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. 2015. Dyna: A Model of Dynamic Human Shape in Motion. *ACM Trans. on Graphics* 34, 4 (Aug. 2015), 120:1–120:14.
[126] J. M. Pozo, M. C. Villa-Uriol, and A. F. Frangi. 2011. Efficient 3D Geometric and Zernike Moments Computation from Unstructured Surface Meshes. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 33, 3 (March 2011), 471–484.
[127] J. Preis, M. Kessel, M. Werner, and C. Linnhoff-Popien. 2012. Gait Recognition with Kinect. In *Workshop on Kinect in Pervasive Computing*.
[128] L. Lo Presti, M. La Cascia, S. Sclaroff, and O. Camps. 2015. Hankelet-based dynamical systems modeling for 3D action recognition. *Image and Vision Computing* 44 (2015), 29–43.
[129] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. 2017. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Vol. 1. 652–660.
[130] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. 2014. HOPC: Histogram of Oriented Principal Components of 3D Pointclouds for Action Recognition. In *European Conf. on Computer Vision*. 742–757.
[131] S. Ramanathan, A. Kassim, Y. V. Venkatesh, and W. S. Wah. 2006. Human Facial Expression Recognition using a 3D Morphable Model. In *IEEE Int. Conf. on Image Processing*. 661–664.
[132] A. Ravichandran, R. Chaudhry, and R. Vidal. 2013. Categorizing Dynamic Textures Using a Bag of Dynamical Systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 35, 2 (2013), 342–353.
[133] M. Reale, X. Zhang, and L. Yin. 2013. Nebula feature: A space-time feature for posed and spontaneous 4D facial behavior analysis. In *IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition*. 1–8.
[134] M. Reuter, F.-E. Wolter, and N. Peinecke. 2006. Laplace-beltrami spectra as shape-DNA of surfaces and solids. *Computer-Aided Design* 38, 4 (April 2006), 342–366.
[135] S. Romdhani and T. Vetter. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *IEEE Conf. on Computer Vision and Pattern Recognition*, Vol. 2. 986–993.
[136] J. Russell and A. Mehrabian. 1977. Evidence for a Three-Factor Theory of Emotions. *Journal of Research in Personality* 11, 3 (Sep 1977), 273–294.
[137] R. M. Rustamov. 2010. Robust Volumetric Shape Descriptor. In *Eurographics Workshop on 3D Object Retrieval*. 1–5.
[138] G. Sandbach, S. Zafeiriou, M. Pantic, and D. Rueckert. 2011. A dynamic approach to the recognition of 3D facial expressions and their temporal models. In *IEEE Conf. on Automatic Face and Gesture Recognition*. 406–413.
[139] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. 2012. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing* 30, 10 (2012), 683–697.
[140] A. Savran and B. Sankur. 2017. Non-rigid registration based model-free 3D facial expression recognition. *Computer Vision and Image Understanding* 162, Supplement C (Sept. 2017), 146–165.
[141] S. Schuon, C. Theobalt, J. Davis, and S. Thrun. 2009. LidarBoost: Depth superresolution for ToF 3D shape scanning. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 343–350.

[142] E. L. Schwartz, A. Shaw, and E. Wolfson. 1989. A numerical solution to the generalized mapmaker's problem: flattening nonconvex polyhedral surfaces. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 11, 9 (Sep 1989), 1005–1008.

[143] J. A. Sethian. 1996. A Fast Marching Level Set Method for Monotonically Advancing Fronts. *Proceedings of the National Academy of Science* 93, 4 (1996), 1591–1595.

[144] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 1010–1019.

[145] L. Shi, I. Cheng, and A. Basu. 2011. Anatomy preserving 3D model decomposition based on robust skeleton-surface node correspondence. In *IEEE Int. Conf. on Multimedia and Expo*. 1–6.

[146] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. 2011. Real-time human pose recognition in parts from single depth images. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. 1–8.

[147] M. Singh, M. Mandal, and A. Basu. 2005. Pose recognition using the radon transform. In *IEEE Midwest Symposium on Circuits and Systems*, Vol. 2. 1091–1094.

[148] A. Sinha, J. Bai, and K. Ramani. 2016. Deep Learning 3D Shape Surfaces Using Geometry Images. In *European Conf. on Computer Vision*. 223–240.

[149] S. Sivapalan, D. Chen, S. Denman, S. Sridharan, and C. Fookes. 2011. Gait energy volumes and frontal gait recognition using depth images. In *Int. Joint Conf. on Biometrics*. 1–6.

[150] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava. 2015. Accurate 3D action recognition using learning on the Grassmann manifold. *Pattern Recognition* 48, 2 (feb 2015), 556–567.

[151] D. Smeets, J. Keustermans, D. Vandermeulen, and P. Suetens. 2013. meshSIFT: Local surface features for 3D face recognition under expression variations and partial data. *Computer Vision and Image Understanding* 117, 2 (Feb. 2013), 158–169.

[152] S. Soltanpour, B. Boufama, and Q. M. J. Wu. 2017. A survey of local feature methods for 3D face recognition. *Pattern Recognition* 72, Supplement C (2017), 391–406.

[153] A. Srivastava, E. Klassen, S. Joshi, and I. Jermyn. 2011. Shape Analysis of Elastic Curves in Euclidean Spaces. *IEEE Trans. on, Pattern Analysis and Machine Intelligence* 33, 7 (July 2011), 1415–1428.

[154] Y. Sun and L. Yin. 2008. Facial Expression Recognition Based on 3D Dynamic Range Model Sequences. In *European Conf. on Computer Vision*. 58–71.

[155] H. Tang and T. S. Huang. 2008. 3D Facial Expression Recognition Based on Automatically Selected Features. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*. 1–8.

[156] J.W.H. Tangelder and R.C. Veltkamp. 2008. A survey of content based 3D shape retrieval methods. In *Multimedia Tools and Applications*, Vol. 39. 441–471.

[157] G. Taubin. 1995. Estimating the Tensor of Curvature of a Surface from a Polyhedral Approximation. In *Int. Conf. on Computer Vision*. 902–907.

[158] J. Tierny, J. P. Vandeborre, and M. Daoudi. 2006. Invariant High Level Reeb Graphs of 3D Polygonal Meshes. In *Int. Symposium on 3D Data Processing, Visualization, and Transmission*. 105–112.

[159] F. Tombari, S. Salti, and L. Di Stefano. 2013. Performance Evaluation of 3D Keypoint Detectors. *Int. Journal of Computer Vision* 102, 1 (Mar 2013), 198–220.

[160] A. Tuan Tran, T. Hassner, I. Masi, and G. Medioni. 2017. Regressing Robust and Discriminative 3D Morphable Models With a Very Deep Neural Network. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 5163–5172.

[161] P. K. Turaga, A. Veeraraghavan, A. Srivastava, and R. Chellappa. 2011. Statistical Computations on Grassmann and Stiefel Manifolds for Image and Video-Based Recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 33, 11 (Nov. 2011), 2273–2286.

[162] A. Veeraraghavan, A. Srivastava, A. K. Roy-Chowdhury, and R. Chellappa. 2009. Rate-Invariant Recognition of Humans and Their Activities. *IEEE Trans. on Image Processing* 18, 6 (2009), 1326–1339.

[163] R. Vemulapalli, F. Arrate, and R. Chellappa. 2014. Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 588–595.

[164] R. Vemulapalli, F. Arrate, and R. Chellappa. 2016. R3DG features: Relative 3D geometry-based skeletal representations for human action recognition. *Computer Vision and Image Understanding* 152 (2016), 155–166.

[165] V. Venkataraman and P. Turaga. 2016. Shape Distributions of Nonlinear Dynamical Systems for Video-Based Inference. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 38, 12 (Dec 2016), 2531–2543.

[166] Y. V. Venkatesh, A. A. Kassim, and O. V. Ramana Murthy. 2009. A Novel Approach to classification of facial expressions from 3D-mesh datasets using modified PCA. *Pattern Recognition Letters* 30, 12 (Sept. 2009), 1128–1137.

[167] A. W. Vieira, E. R. Nascimento, G. L. Oliveira, Z. Liu, and M. F.M. Campos. 2012. STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences. In *Iberoamerican Congress on Pattern Recognition*. 252–259.

[168] A. Vinciarelli, M. Pantic, and H. Bourlard. 2009. Social Signal Processing: Survey of an Emerging Domain. *Image and Vision Computing* 27, 12 (2009), 1743–1759.
[169] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. 2012. Robust 3D Action Recognition with Random Occupancy Patterns. In *European Conf. on Computer Vision*. 1–8.
[170] J. Wang, Z. Liu, Y. Wu, and J. Yuan. 2012. Mining Actionlet Ensemble for Action Recognition with Depth Cameras. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 1290–1297.
[171] J. Wang, L. Yin, X. Wei, and Y. Sun. 2006. 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution. In *IEEE Int. Conf. on Computer Vision and Pattern Recognition*, Vol. 2. 1399–1406.
[172] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera. 2017. RGB-D-based Human Motion Recognition with Deep Learning: A Survey. *CoRR* abs/1711.08362 (2017).
[173] T. Wang and A. Basu. 2007. A note on a fully parallel 3D thinning algorithm and its applications. *Pattern Recognition Letters* 28, 4 (2007), 501–506.
[174] T. Wang and I. Cheng. 2008. Generation of Unit-Width Curve Skeletons Based on Valence Driven Spatial Median. In *Int. Symposium on Visual Computing*. 1051–1060.
[175] P Wei, Y. Zhao, N. Zheng, and S.-C. Zhu. 2013. Modeling 4D Human-Object Interactions for Event and Object Recognition. In *Int. Conf. on Computer Vision*. 3272–3279.
[176] D. Weinland, R. Ronfard, and E. Boyer. 2011. A Survey of Vision-based Methods for Action Representation, Segmentation and Recognition. *Computer Vision and Image Understanding* 115, 2 (Feb. 2011), 224–241.
[177] N. Werghi, S. Berretti, and A. Del Bimbo. 2015. The Mesh-LBP: A Framework for Extracting Local Binary Patterns From Discrete Manifolds. *IEEE Trans. on Image Processing* 24, 1 (Jan 2015), 220–235.
[178] N. Werghi, C. Tortorici, S. Berretti, and A. Del Bimbo. 2016. Boosting 3D LBP-Based Face Recognition by Fusing Shape and Texture Descriptors on the Mesh. *IEEE Trans. on Information Forensics and Security* 11, 5 (May 2016), 964–979.
[179] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 2015. 3D ShapeNets: A deep representation for volumetric shapes. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 1912–1920.
[180] L. Xia and J. K. Aggarwal. 2013. Spatio-Temporal Depth Cuboid Similarity Feature for Activity Recognition Using Depth Camera. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. 2834–2841.
[181] L. Xia, C.-C. Chen, and J. K. Aggarwal. 2012. View Invariant Human Action Recognition Using Histograms of 3D Joints. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. 20–27.
[182] J. Xie, Y. Fang, F. Zhu, and E. Wong. 2015. Deepshape: Deep learned shape descriptor for 3D shape matching and retrieval. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 1275–1283.
[183] K. Yang, Y. Dou, S. Lv, F. Zhang, and Q. Lv. 2016. Relative distance features for gait recognition with Kinect. *Journal of Visual Communication and Image Representation* 39, Supplement C (2016), 209–217.
[184] Q. Yang, R. Yang, J. Davis, and D. Nister. 2007. Spatial-Depth Super Resolution for Range Images. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 1–8.
[185] X. Yang and Y. Tian. 2012. EigenJoints-based Action Recognition Using Naive-Bayes-Nearest-Neighbor. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. 14–19.
[186] X. Yang, C. Zhang, and Y. Tian. 2012. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *ACM Int. Conf. on Multimedia*. 1057–1060.
[187] L. Yin and A. Basu. 1999. Integrating active face tracking with model based coding. *Pattern Recognition Letters* 20, 6 (1999), 651–657.
[188] G. Yu, Z. Liu, and J. Yuan. 2014. Discriminative Orderlet Mining For Real-time Recognition of Human-Object Interaction. In *Asian Conf. on Computer Vision*. 50–65.
[189] A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud. 2009. Surface feature detection and description with applications to mesh matching. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 373–380.
[190] M. Zanfir, M. Leordeanu, and C. Sminchisescu. 2013. The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection. In *IEEE Int. Conf. on Computer Vision*. 2752–2759.
[191] X. Zhang, Y. Wang, M. Gou, M. Sznaier, and O. Camps. 2016. Efficient Temporal Sequence Comparison and Classification Using Gram Matrix Embeddings on a Riemannian Manifold. In *IEEE Conf. on Computer Vision and Pattern Recognition*. 4498–4507.
[192] G. Zhao, G. Liu, H. Li, and M. Pietikainen. 2006. 3D gait recognition using multiple cameras. In *Int. Conf. on Automatic Face and Gesture Recognition*. 529–534.
[193] S. Zulqarnain Gilani and A. Mian. 2017. Learning from Millions of 3D Scans for Large-scale 3D Face Recognition. *ArXiv e-prints* (Nov. 2017). arXiv:1711.05942

Received December 2017; revised January 2017; accepted January 2017