Human Action Recognition Using 3D Reconstruction Data

Georgios Th. Papadopoulos, *Member, IEEE,* and Petros Daras, *Senior Member, IEEE*

***Abstract*—In this paper, the problem of human action recognition using 3D reconstruction data is deeply investigated. 3D reconstruction techniques are employed for addressing two of the most challenging issues related to human action recognition in the general case, namely view-variance (i.e. when the same action is observed from different viewpoints) and the presence of (self-) occlusions (i.e. when for a given point of view a body-part of an individual conceals an other body-part of the same or an other subject). The main contributions of this work are summarized as follows: i) Detailed examination of the use of 3D reconstruction data for performing human action recognition. The latter includes: a) the introduction of appropriate local/global, flow/shape descriptors, and b) extensive experiments in challenging publicly available datasets and exhaustive comparisons with state-of-art approaches. ii) A new local-level 3D flow descriptor, which incorporates spatial and surface information in the flow representation and efficiently handles the problem of defining 3D orientation at every local neighborhood. iii) A new global-level 3D flow descriptor that efficiently encodes the global motion characteristics in a compact way. iv) A novel global temporal-shape descriptor that extends the notion of 3D shape descriptions for action recognition, by incorporating the temporal dimension. The proposed descriptor efficiently addresses the inherent problems of temporal alignment and compact representation, while also being robust in the presence of noise (compared with similar tracking-based methods of the literature). Overall, this work significantly improves the state-of-art performance and introduces new research directions in the field of 3D action recognition, following the recent development and wide-spread use of portable, affordable, high-quality and accurate motion capturing devices (e.g. Microsoft Kinect).**

***Index Terms*—Action recognition, 3D reconstruction, 3D flow, 3D shape.**

I. Introduction

AUTOMATIC and accurate recognition of the observed human actions has emerged as one of the most challenging and active areas of research in the broader computer vision community over the past decades [1]. This is mainly due to the very wide set of possible application fields with great commercialization potentials that can benefit from the resulting accomplishments, such as surveillance, security, human computer interaction, smart houses, helping the elderly/disabled, gaming, e-learning, to name a few. Methods in this research area incorporate the typical requirements for rotation, translation and scale invariance for achieving robust recognition performance. However, additional significant challenges, regardless of the particular application field, need also to be efficiently addressed, like the differences in the appearance of the subjects, the human silhouette features, the execution of the same actions, etc. Despite the fact that human action recognition constitutes the central point of focus for multiple research groups/projects and that numerous approaches have already been proposed, significant obstacles towards fully addressing the problem in the general case still remain.

Action recognition approaches can roughly be divided into the following three categories [2], irrespectively of the data that they receive as input (i.e. single-camera videos, multi-view video sequences, depth maps, 3D reconstruction data, motion capture data, etc.): a) spatio-temporal shape- [3], [4], b) tracking- [5]–[9] and c) Space-Time Interest Point (STIP)-based [10]–[13]. Spatio-temporal shape-based approaches rely on the estimation of global-level representations for performing recognition, using e.g. the outer boundary of an action; however, they are prone to the detrimental effects caused by self-occlusions of the performing subjects. The efficiency of tracking-based approaches, which are based on the tracking of particular features or specific human body parts in subsequent frames (including optical-flow-based methods), depends heavily on the robustness of the employed tracker that is often prone to mistakes in the presence of noise. STIP-based methods perform analysis at the local-level. They are robust to noise, while they are shown to satisfactorily handle self-occlusion occurrences; however, they typically exhibit increased computational complexity for reaching satisfactory recognition performance.

Two of the most challenging issues related to human action recognition in the general case (i.e. in unconstrained environments) that current state-of-art algorithms face are view-variance and the presence of (self-) occlusions. In order to simultaneously handle both challenges in a satisfactory way, 3D reconstruction information is used in this work. This choice is further dictated by the recent technological breakthrough, which has resulted in the introduction of portable, affordable, high-quality and accurate motion capturing devices to the market; these devices have already gained tremendous acceptance in several research and daily-life application fields.

In this paper, the problem of human action recognition using 3D reconstruction information [14] is investigated. In particular, the main contributions of this work are the following:

- A **thorough examination of the use of 3D reconstruction data for realizing human action recognition**. This involves: a) the introduction of appropriate local/global, flow/shape descriptors, and b) extensive

Georgios Th. Papadopoulos and Petros Daras are with the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece.

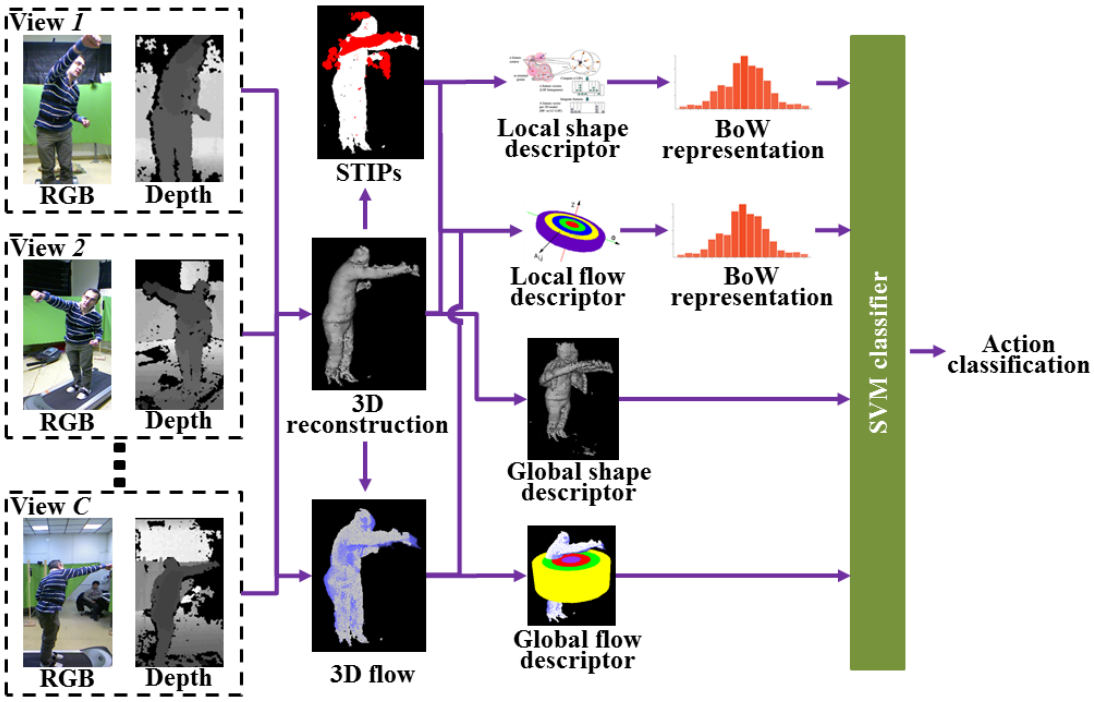


Fig. 1. Proposed 3D human action recognition framework

experiments in challenging publicly available datasets and exhaustive comparisons with state-of-art methods. This comprehensive study sheds light on several aspects of the problem at hand and proposes solutions to the encountered challenges.

- A **new local-level 3D flow descriptor**, which incorporates spatial and surface information in the flow representation (i.e. moving away from current naive histogram-based approaches) and efficiently handles the problem of defining a consistent 3D orientation at every local neighborhood (which in turn ensures the extraction of directly comparable low-level descriptions at different locations).
- A **new global-level 3D flow descriptor**, which efficiently encodes the global motion characteristics in a compact way (i.e. surpassing the need for complex, detailed and susceptible to noise global flow representations).
- A **novel global temporal-shape descriptor** that extends the notion of 3D shape descriptions for action recognition, by incorporating temporal information. The proposed descriptor efficiently addresses the inherent problems of temporal alignment and compact representation, while also being robust in the presence of noise (as opposed to similar tracking-based methods of the literature).

It must be noted that the descriptors utilized in this work cover all three main action recognition methodologies that were outlined in the beginning of this section; hence, meaningful conclusions regarding the advantages/disadvantages of every methodology can be reached by the examination of the conducted experimental evaluation. A graphical representation of the proposed 3D human action recognition framework is illustrated in Fig. 1, while its constituent parts are detailed in the subsequent technical sections of the manuscript.

The remainder of the paper is organized as follows: Previous work is reviewed in Section II. Section III describes the 3D information processing methodology. The descriptor extraction procedure is detailed in Section IV. Section V outlines the adopted action recognition scheme. Experimental results are presented in Section VI and conclusions are drawn in Section VII.

II. Previous work on 3D action recognition

The recent introduction of accurate motion capturing devices, with the Microsoft Kinect being the most popular one, has given great boost in human action recognition tasks and has decisively contributed in shifting the research focus towards the analysis in 3D space. This is mainly due to the wealth of information present in the captured stream, where the estimated 3D depth maps facilitate in overcoming typical barriers (e.g. scale estimation, presence of occlusions, etc.) of traditional visual analysis on the 2D plane and hence significantly extending the recognition capabilities. The great majority of the methods that belong to this category typically exploit human skeleton-tracking or surface (normal vectors) information, which are readily available by applying widely-used open-source software (e.g. OpenNI SDK[1], Kinect SDK[2], etc.). In [15], a depth similarity feature is proposed for describing the local 3D cuboid around a point of interest with an adaptable supporting size. Additionally, Zhang *et al.* [16] introduce 4D color-depth (CoDe4D) local spatio-temporal features that incorporate both intensity and depth information acquired from RGB-D cameras. In [17], an actionlet ensemble model is learnt to represent each action and to capture the intra-class variance. Xia *et al.* [18] utilize histograms of 3D

[1]http://structure.io/openni
[2]http://www.microsoft.com/en-us/kinectforwindows/

joint locations (HOJ3D) as a compact representation of human postures. In [19], an action graph is employed to model explicitly the dynamics of the actions and a bag of 3D points to characterize a set of salient postures that correspond to the nodes in the action graph. Moreover, Papadopoulos *et al.* [20] calculate the spherical angles between selected joints, along with the respective angular velocities. In [21], different slicing views of the spatiotemporal volume are investigated. Multiple configuration features (combination of joints) and movement features (position, orientation and height of the body) are extracted from the recovered 3D human joint and pose parameter sequences in [5]. Furthermore, Sun *et al.* [22] extract view-invariant features from skeleton joints. In [23], the distribution of the relative locations of the neighbors for a reference point in the human body point cloud is described in a compact way.

A. Flow descriptors

Although numerous approaches to 3D action recognition have already been proposed, they mainly focus on exploiting human skeleton-tracking or surface (normal vectors) information. Therefore, more elaborate information sources, like 3D flow [24] (which is the counterpart of 2D optical flow in the 3D space) have not received the same attention yet. The latter is mainly due to the increased computational complexity that inherently 3D flow estimation involves, since its processing includes an additional disparity estimation problem. In particular, 3D flow estimation algorithms, exhibiting satisfactory flow field calculations, have been presented [25], [26], requiring few seconds for processing per frame. However, methods that emphasize on reducing the required computational complexity, by adopting several optimization techniques (hardware, algorithmic, GPU implementation), have achieved processing rates up to 20Hz [27], [28]. Consequently, these recent advances have paved the way for introducing action recognition methods that make use of 3D flow information.

Regarding methods that utilize 3D flow information for recognizing human actions, Holte *et al.* [10] introduce a local 3D motion descriptor; specifically, an optical flow histogram (HOF3D) is estimated, taking into account the 4D spatio-temporal neighborhood of a point-of-interest. In [28], a 3D grid-based flow descriptor is presented, in the context of a real-time human action recognition system. Additionally, histograms of 3D optical flow are also used in [29], along with other descriptions (spatio-temporal interest points, depth data, body posture). Gori *et al.* [30] build a frame-level 3D Histogram of Flow (3D-HOF), as part of an incremental method for 3D arm-hand behaviour modelling and recognition. In [31], a 3D flow descriptor is derived by performing a multiple-window partition of a silhouette's bounding box and subsequently concatenating the average flow values of each formed window. Furthermore, Fanello *et al.* [32] present an approach to simultaneous on-line video segmentation and recognition of actions, using histograms of 3D flow.

Although some works have recently been proposed for action recognition using 3D flow information, they mainly rely on relatively simple local/global histogram- or grid-based representations. Therefore, significant challenges in 3D flow processing/representation still remain partially addressed or even unexplored, like incorporation of spatial and surface information, view-invariance, etc. Additionally, for the particular case of local-level flow representations, a satisfactory solution to the problem of defining a consistent 3D orientation at different locations (e.g. at different points-of-interest) has not been introduced yet.

B. Shape descriptors

Concerning the exploitation of 3D shape information for action recognition purposes, the overpowering majority of the literature methods refers to the temporal extension of the corresponding 2D spatial analysis (i.e. analysis in the $xy+t$ 3D space), which is typically initiated by e.g. concatenating the binary segmentation masks or outer contours of the examined object in subsequent frames. Consequently, analysis in the 'actual' xyz 3D space (or equivalently analysis in the $xyz+t$ 4D space, if the time dimension is taken into account) is currently avoided. In particular, Weinland *et al.* [3] introduce the so called Motion History Volumes (MHV), as a free-viewpoint representation for human actions, and use Fourier transforms in cylindrical coordinates around the vertical axis for efficiently performing alignment and comparison. In [4], human actions are regarded as three-dimensional shapes induced by the silhouettes in the space-time volume and properties of the solution to the Poisson equation are utilized to extract features, such as local space-time saliency, action dynamics, shape structure and orientation. Additionally, Efros *et al.* [33] present a motion descriptor based on optical flow measurements in a spatio-temporal volume for each stabilized human figure and an associated similarity measure.

Towards the goal of performing shape analysis for action recognition in the above-mentioned $xyz+t$ 4D space, Huang *et al.* [34] present time-filtered and shape-flow descriptors for assessing the similarity of 3D video sequences of people with unknown temporal correspondence. In [35], an approach to non-sequential alignment of unstructured mesh sequences that is based on a shape similarity tree is detailed, which allows alignment across multiple sequences of different motions, reduces drift in sequential alignment and is robust to rapid non-rigid motions. Additionally, Yamasaki *et al.* [36] present a similar motion search and retrieval system for 3D video based on a modified shape distribution algorithm. The problem of 3D shape representation, which is formulated using Extremal Human Curve (EHC) descriptors extracted from the body surface, and shape similarity in human video sequences is the focus of the work in [37].

Despite the fact that some works on temporal-shape descriptions have already been proposed, their main limitation is that they include in their analysis the problem of the temporal alignment of action sequences (typically using common techniques, like e.g. dynamic programming, Dynamic Time Warping, etc.). The latter often has devastating effects in the presence of noise or leads to cumulative errors in case of misalignment occurrences. To this end, a methodology that would alleviate from the burden of the inherent problem of

temporal alignment when performing temporal-shape analysis, while maintaining a compact action representation, would be beneficial.

III. 3D INFORMATION PROCESSING

A. Reconstruction

In order to efficiently address two of the most important problems inherent in human action recognition, namely view-variance and the presence of (self-)occlusions, 3D reconstruction techniques are employed in this work. In particular, the 3D reconstruction algorithm of [14], which makes use of a set of calibrated Kinect sensors, is utilized for generating a 3D point-cloud of the performing subjects, where the estimated points correspond to locations on the surface of the human silhouette. After the point-cloud is generated, it undergoes a 'voxelization' procedure for computing a corresponding voxel grid $VG_t = \{v_t(x_g, y_g, z_g) : x_g \in [1, X_g], y_g \in [1, Y_g], z_g \in [1, Z_g]\}$, where t denotes the currently examined frame. In the current implementation, a uniform voxel grid is utilized. Additionally, it is considered that $v_t(x_g, y_g, z_g) = 1$ (i.e. $v_t(x_g, y_g, z_g)$ belongs to the subject's surface) if $v_t(x_g, y_g, z_g)$ includes at least one point in the corresponding real 3D space and $v_t(x_g, y_g, z_g) = 0$ otherwise.

B. Flow estimation

For 3D flow estimation, a gradual approach is proposed in this work that relies on the application of a 2D flow estimation algorithm to every Kinect RGB stream and the subsequent fusion of the acquired results. In particular, a 2D optical flow estimation algorithm is initially applied to every captured RGB frame of the c-th ($c \in [1, C]$) employed Kinect and the resulting 2D optical flow field is denoted $\mathbf{f}_{c,t}^{2D}(x_{rgb}, y_{rgb})$, where (x_{rgb}, y_{rgb}) are coordinates on the 2D RGB plane and the algorithm receives as input the frames at times t and $t-1$. The optical flow algorithm of [38] was selected using the implementation provided by [39], since it was experimentally shown to produce satisfactory results [39]. In parallel, a $3D$ point-cloud $W_{c,t}^{3D}(x_l, y_l, z_l)$ is estimated from the corresponding depth map $D_{c,t}^{2D}(x_d, y_d)$, where (x_l, y_l, z_l) and (x_d, y_d) denote coordinates in the real 3D space and on the 2D depth map plane corresponding to the c-th Kinect, respectively. Subsequently, a 3D flow field $\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)$ is estimated by converting the pixel correspondences in $\mathbf{f}_{c,t}^{2D}(x_{rgb}, y_{rgb})$ to point correspondences; the latter is realized by considering the point-clouds $W_{c,t}^{3D}(x_l, y_l, z_l)$ and $W_{c,t-1}^{3D}(x_l, y_l, z_l)$. It must be noted that $\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)$ flow vectors that involve points that correspond to 'holes' (i.e. missing depth estimations from the Kinect), background or different human body parts are discarded. Points in $W_{c,t}^{3D}(x_l, y_l, z_l)$ are considered to belong to the background if their depth value z_l exceeds threshold T_b, while two points are assumed to correspond to different body parts if their depth difference is greater than threshold T_l (T_l=25mm in this work). $\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)$ may contain significant amounts of noise, due to the frequent failure of the Kinect sensor to provide accurate depth estimations. For tackling this noise, a reliability value is associated with every $\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)$ vector. More specifically, the reliability value $r_{c,t}^{3D}(x_l, y_l, z_l)$ of point (x_l, y_l, z_l) is approximated by the reliability value $r_{c,t}^{2D}(x_d, y_d)$ of its corresponding point (x_d, y_d) in $D_{c,t}^{2D}(x_d, y_d)$, which is calculated as follows:

$$r_{c,t}^{2D}(x_d, y_d) = \frac{\sum_{x'_d=x_d-Q}^{x_d+Q} \sum_{y'_d=y_d-Q}^{y_d+Q} b(x'_d, y'_d)}{(2Q+1)^2} \quad (1)$$

where $r_{c,t}^{2D}(x_d, y_d) \in [0, 1]$. $b(x'_d, y'_d) = 0$ if point (x'_d, y'_d) corresponds to background/hole or a different body part than the reference point (x_d, y_d) and $b(x'_d, y'_d) = 1$ otherwise. $r_{c,t}^{2D}(x_d, y_d) = 0$ if (x_d, y_d) belongs to the background or a hole in $D_{c,t}^{2D}(x_d, y_d)$. $Q = 20$ based on experimentation.

For estimating the 3D flow field $\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)$, the mappings from the (x_{rgb}, y_{rgb}), (x_d, y_d) spaces to each other as well as to the (x_l, y_l, z_l) one are required. If the (x_{rgb}, y_{rgb}) and (x_d, y_d) spaces are not strictly aligned during capturing (i.e. $(x_{rgb}, y_{rgb}) \neq (x_d, y_d)$), a typical calibration model [40] requires the setting of the following parameters: a) depth camera intrinsics and extrinsics, b) RGB camera intrinsics, c) distortion model of both cameras, and d) a Rotation-Translation (RT) transform from (x_d, y_d) to (x_{rgb}, y_{rgb}). Concerning the datasets utilized in this work (Section VI-A), for the Huawei/3DLife the aforementioned parameters are publicly available, while for the NTU RGB+D and UTKinect the best fit parameters from the Kinect sensors in the authors' laboratory were used.

Having estimated the flow fields $\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)$, the next step is to compute a 3D flow field $\mathbf{F}_t^{3D}(x_g, y_g, z_g)$ in VG_t. For achieving this, every Kinect c is initially examined separately. In particular, for every voxel $v_t(x_g, y_g, z_g)$ a flow vector $\mathbf{F}_{c,t}^{3D}(x_g, y_g, z_g)$ is estimated according to the following expression:

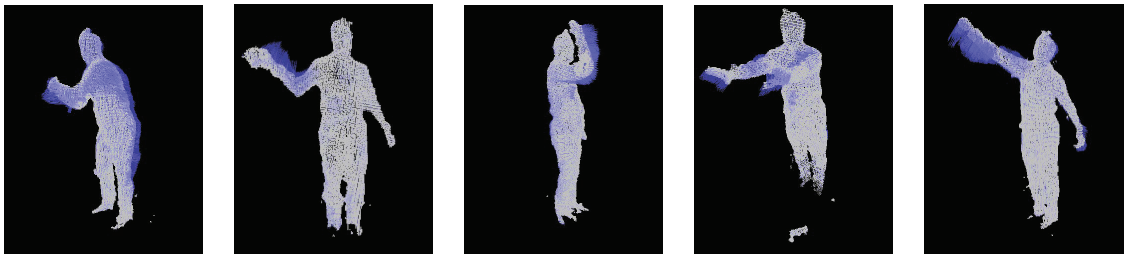
$$\mathbf{F}_{c,t}^{3D}(x_g, y_g, z_g) = \frac{\sum_{\mathbf{\Delta}} r_{c,t}^{3D}(x_l, y_l, z_l) \Psi[\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)]}{M} \quad (2)$$

where $\mathbf{\Delta}$ comprises the points in $W_{c,t}^{3D}(x_l, y_l, z_l)$ that correspond to voxel $v_t(x_g, y_g, z_g)$ and for which flow vectors $\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)$ have been calculated, M is the total number of points in $\mathbf{\Delta}$ and $\Psi[.]$ denotes the extrinsic calibration-based transformation from the $W_{c,t}^{3D}(x_l, y_l, z_l)$ to the (x_g, y_g, z_g) space. It must be noted that a depth difference threshold T_g (similar to the T_l described above) is used for controlling the assignment of points in $W_{c,t}^{3D}(x_l, y_l, z_l)$ to voxels $v_t(x_g, y_g, z_g)$ in VG_t (T_g = 25mm in this work). For combining $\mathbf{F}_{c,t}^{3D}(x_g, y_g, z_g)$ vectors estimated from different Kinect sensors, the following reliability value is estimated for each voxel $v_t(x_g, y_g, z_g)$ that is visible from every Kinect c:

$$a_{c,t}^{3D}(x_g, y_g, z_g) = \langle \mathbf{m}_c(x_g, y_g, z_g), \mathbf{n}_t(x_g, y_g, z_g) \rangle \quad (3)$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product of two vectors, $\mathbf{m}_c(x_g, y_g, z_g)$ is the unit vector that connects voxel $v_t(x_g, y_g, z_g)$ with the center of the c-th Kinect, $\mathbf{n}_t(x_g, y_g, z_g)$ is the unit normal vector to the 3D reconstructed surface at voxel $v_t(x_g, y_g, z_g)$ and $a_{c,t}^{3D}(x_g, y_g, z_g) \in [0, 1]$. Subsequently, $\mathbf{F}_t^{3D}(x_g, y_g, z_g)$ is computed, as follows:

$$\mathbf{F}_t^{3D}(x_g, y_g, z_g) = \frac{\sum_{\mathbf{U}} a_{c,t}^{3D}(x_g, y_g, z_g) \mathbf{F}_{c,t}^{3D}(x_g, y_g, z_g)}{L} \quad (4)$$



(a) (b) (c) (d) (e)

Fig. 2. Indicative 3D flow field $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ estimation examples for actions: (a) golf-chip, (b) hand-waving, (c) knocking the door, (d) push-away and (e) throwing.

where $\mathbf{U}$ comprises the Kinect sensors from which $v_t(x_g, y_g, z_g)$ is visible and L ($L \leq C$) their number. From the above definition, it can be seen that 3D flow estimations originating from Kinect sensors, whose infrared illumination strikes perpendicularly to the surface to be reconstructed, are favored. For further noise removal, $\mathbf{F}_t^{3D}(x_g, y_g, z_g)$ is low-passed using a simple 11x11x11 mean filter; hence, resulting to flow field $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$. Indicative examples of $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ estimations for different actions are shown in Fig. 2.

IV. Descriptor extraction

A. Local descriptions

In order to analyse human motion at local level, Spatio-Temporal Interest Points (STIPs) need to be detected first. In this work, an extension of the 3D (xy- coordinates plus time) detector of [41] to its counterpart in 4D (xyz-coordinates plus time) has been developed. In particular, the voxel grid VG_t is processed by a set of separable linear filters, according to the following equations:

$$\begin{aligned} R(x_g, y_g, z_g, t) = \\ \{v_t(x_g, y_g, z_g) * k(x_g, y_g, z_g; \sigma) * h_{ev}(t; \tau, \omega)\}^2 \\ +\{v_t(x_g, y_g, z_g) * k(x_g, y_g, z_g; \sigma) * h_{od}(t; \tau, \omega)\}^2 \end{aligned} \quad (5)$$

where $R(x_g, y_g, z_g, t)$ is the response function, $*$ denotes the convolution operator, $k(x_g, y_g, z_g; \sigma)$ is a Gaussian smoothing kernel applied only to the spatial dimensions, $\omega = 4/\tau$ and $h_{ev}(t; \tau, \omega) = -\cos(2\pi t\omega)e^{-t^2/\tau^2}$, $h_{od}(t; \tau, \omega) = -\sin(2\pi t\omega)e^{-t^2/\tau^2}$ is a quadrature pair [42] of 1D Gabor filters applied temporally. From the above definition, it can be seen that the response function $R(x_g, y_g, z_g, t)$ is controlled by parameters σ and τ, which roughly correspond to the spatial and temporal scale of the detector, respectively. Thresholding the estimated values of $R(x_g, y_g, z_g, t)$ generates the detected STIPs. In the current implementation, $\sigma = 2.0$ and $\tau = 0.9$ were set based on experimentation.

1) Local flow descriptor: For extracting discriminative local-level 3D flow descriptors, the following challenges need to be addressed: a) the difficulty in introducing a consistent orientation definition at every STIP location for producing comparable low-level descriptions among different STIPs, and b) the incorporation of spatial distribution and surface information in a compact way, while maintaining 3D rotation invariance.

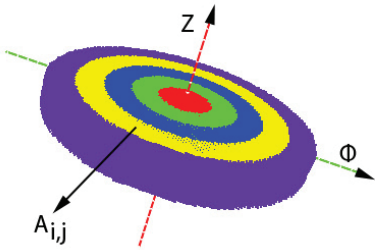


Fig. 3. Example of ring-shaped areas $A_{i,j}$ formation for $i = 0$ and $J = 5$ in the defined cylindrical coordinate system.

Under the proposed approach, a novel local-level 3D flow descriptor is introduced for efficiently addressing the aforementioned issues. Initially, the normal vector $\mathbf{n}_t^{stip}(x_g, y_g, z_g)$ at every STIP is used for defining a local cylindrical coordinate system (ϱ, ϕ, z), where the origin is placed at the STIP point $v_t^{stip}(x_g, y_g, z_g)$, the direction of the longitudinal axis Z coincides with vector $\mathbf{n}_t^{stip}(x_g, y_g, z_g)$ and the direction of the polar axis Φ (perpendicular to the longitudinal one) is selected randomly. Using this coordinate system, concentric ring-shaped areas are defined, according to the following expressions and depicted in Fig. 3:

$$A_{i,j} = \begin{cases} (j-1)\mu \leq \varrho \leq j\mu \\ \nu/2 + (i-1)\nu \leq z \leq \nu/2 + i\nu \end{cases}, \quad i > 0$$

$$A_{i,j} = \begin{cases} (j-1)\mu \leq \varrho \leq j\mu \\ -\nu/2 \leq z \leq \nu/2 \end{cases}, \quad i = 0 \quad (6)$$

$$A_{i,j} = \begin{cases} (j-1)\mu \leq \varrho \leq j\mu \\ -\nu/2 + i\nu \leq z \leq -\nu/2 + (i+1)\nu \end{cases}, \quad i < 0$$

where $i \in [-I, I]$ and $j \in [1, J]$ denote the indices of the defined areas $A_{i,j}$, $\mu = D_{cub}^s/J$, $\nu = D_{cub}^s/(2I+1)$ and D_{cub}^s is the spatial dimension of the spatio-temporal cuboid ($D_{cub}^s = 31$ is set experimentally), which is defined around its central point $v_t^{stip}(x_g, y_g, z_g)$ and constitutes the support area for the respective descriptor extraction procedure. From the expressions in (6), it can be seen that the direction of the polar axis, which is used for calculating angle ϕ, does not affect the formation of regions $A_{i,j}$ nor the estimation of the descriptor values, as it will be discussed in the sequel. In this work, $I = 2$ and $J = 5$ were set based on experimentation.

For describing the flow information in every $A_{i,j}$ region, a loose representation is required that will render the respective descriptor robust to differences in the appearance of the subjects and the presence of noise. To this end, a histogram-

based representation is adopted. In particular, for every $v_t(x_g, y_g, z_g) \in A_{i,j}$ for which a 3D flow $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ vector is estimated, the following angle is calculated:

$$\vartheta = \cos^{-1}\left(\frac{\langle \mathbf{n}_t(x_g, y_g, z_g), \bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)\rangle}{\|\mathbf{n}_t(x_g, y_g, z_g)\|\|\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)\|}\right) \quad (7)$$

where $\|.\|$ denotes the norm of a vector and $\vartheta \in [0, \pi]$. $\mathbf{n}_t(x_g, y_g, z_g)$ is used instead of $\mathbf{n}_t^{stip}(x_g, y_g, z_g)$ in (7) for implicitly encoding 3D surface information, i.e. for discriminating between an arm and a head that undergo a forward horizontal movement. Based on the calculated angles, a histogram is constructed for every region $A_{i,j}$, by uniformly dividing the interval $[0, \pi]$ into a set of p equal-length bins ($p = 8$ in this work). During the histogram estimation, $\|\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)\|$ is added to the appropriate bin value, when $v_t(x_g, y_g, z_g)$ is processed. By concatenating the histograms that have been computed for all regions $A_{i,j}$ in a single feature vector, the proposed local-level 3D flow descriptor for $v_t^{stip}(x_g, y_g, z_g)$ is formed. It must be noted that during the descriptor extraction procedure, the normal $\mathbf{n}_t(x_g, y_g, z_g)$ and the flow $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ vectors of all frames in the spatio-temporal cuboid defined for $v_t^{stip}(x_g, y_g, z_g)$ are considered; however, the cylindrical grid defined for frame t is used unaltered for all other frames as well. The temporal cuboid dimension D_{cub}^t, i.e. the total number of frames that it includes, is set equal to 3 in the current implementation. Additionally, for accounting for the difference in appearance and the execution of actions among different individuals (e.g. different velocity when the same action is performed by different individuals) the estimated 3D flow feature vector is L1 normalized.

2) Local shape descriptor: For reducing the effects of noise present in the 3D flow estimates and also for providing a more complete representation, 3D shape information is additionally extracted at every STIP position; however, only frame t is considered this time and not all frames in the STIP's cuboid. In the current implementation, the LC-LSF shape descriptor of [43], which employs a set of local statistical features for describing a 3D model, is used. The aforementioned descriptor was selected on the basis of its relatively low computational complexity and its increased efficiency in non-rigid 3D model retrieval.

B. Global descriptions

1) Global flow descriptor: For estimating a global 3D flow description, an approach similar to the one described in Section IV-A1 for local-level flow analysis is followed. In particular, the fundamental problem of orientation definition outlined in Section IV-A1 is addressed here by assuming a vertical direction consideration. The latter selection is justified by the fact that the angle of the principal axis of the 3D human silhouette with the vertical direction typically does not exhibit significant deviations among different instances of a given action.

The descriptor extraction procedure is initiated by estimating a vertically aligned minimum bounding cylinder of all $v_t(x_g, y_g, z_g)$ for which a flow vector $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ is estimated for all frames t that comprise the examined action.

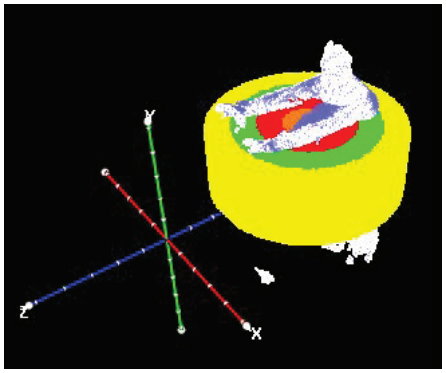


Fig. 4. Example of ring-shaped areas $B_{\kappa,\lambda}$ formation for $\kappa = 3$ and $\lambda \in [1, 4]$ for a 'push-away' action instance.

The center of the cylinder (i.e. the central point of its axis) is denoted $v_{cg}(x_{cg}, y_{cg}, z_{cg})$, while its radius is represented by ζ. Additionally, the upper and lower cylinder boundaries are denoted y_{max}^c and y_{min}^c, respectively. Then, a set of concentric ring-shaped areas are defined, similarly to (6):

$$B_{\kappa,\lambda} = \begin{cases} (\lambda-1)\gamma \le \xi \le \lambda\gamma \\ y_{min}^c + (\kappa-1)\delta \le y_g \le y_{min}^c + \kappa\delta \\ \xi = \sqrt{(x_g - x_{cg})^2 + (z_g - z_{cg})^2} \end{cases} \quad (8)$$

where $\kappa \in [1, K]$, $\lambda \in [1, \Lambda]$, $\gamma = \zeta/\Lambda$ and $\delta = (y_{max}^c - y_{min}^c)/K$. For every $B_{\kappa,\lambda}$ area, a 2D angle histogram is estimated, taking into account all flow vectors $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ during the whole duration of the examined action that correspond to voxels $v_t(x_g, y_g, z_g)$ that lie in that spatial area. More specifically, for each of the aforementioned $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$, the following two angles are calculated:

$$\psi = \tan^{-1}\left(\frac{z_g - z_{cg}}{x_g - x_{cg}}\right) - \tan^{-1}\left(\frac{\bar{\mathbf{F}}_{z,t}^{3D}(x_g, y_g, z_g)}{\bar{\mathbf{F}}_{x,t}^{3D}(x_g, y_g, z_g)}\right)$$
$$o = \cos^{-1}\left(\frac{\langle (0,1,0), \bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)\rangle}{\|\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)\|}\right) \quad (9)$$

where $\bar{\mathbf{F}}_{x,t}^{3D}(x_g, y_g, z_g)$ and $\bar{\mathbf{F}}_{z,t}^{3D}(x_g, y_g, z_g)$ are the x- and z-component of the flow vector $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$, respectively. $\psi \in [-\pi, \pi]$ corresponds to the angle between the horizontal projection of $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ and the projection of the vector connecting the cylindrical center (x_{cg}, y_{cg}, z_{cg}) with the examined voxel position (x_g, y_g, z_g) on the horizontal xz plane, while $o \in [0, \pi]$ corresponds to the angle of $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ with the vertical axis. Then, the above-mentioned 2D histogram for area $B_{\kappa,\lambda}$ is computed by partitioning the value ranges of ψ and o into b_ψ and b_o equal-width non-overlapping bins, respectively. During the calculations, $\|\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)\|$ is aggregated to the appropriate histogram bin. The global flow descriptor is computed by concatenating the estimated angle histograms of all $B_{\kappa,\lambda}$ areas, while it is subsequently $L1$ normalized for rendering the descriptor robust to the difference in the speed with which every action is executed. From the definitions of the ring-shaped areas $B_{\kappa,\lambda}$ and angle ψ, it can be justified that the proposed global-level 3D flow descriptor satisfies the requirement for rotation invariance, while it also incorporates spatial distribution-related information in the flow representation. In this work, the following parameter values were selected after experimentation: $\Lambda = 4$, $K = 4$, $b_\psi = 6$

and $b_o = 3$. An example of ring-shaped $B_{\kappa,\lambda}$ areas formation for a 'push away' action instance is given in Fig. 4.

2) Global shape descriptor: As described in Section II-B, current temporal-shape techniques include in their analysis the problem of the temporal alignment of the action sequences, which has devastating effects in the presence of noise or leads to cumulative errors in case of misalignment occurrences. To this end, a temporal-shape descriptor that encodes the dominant shape variations and avoids the need for exact action sequence alignment, while maintaining a compact shape representation, is proposed in this section.

The biggest challenge in using the temporal dimension for realizing 3D shape-based action recognition is the temporal alignment of different action executions, which is often misleading and causes devastating aggregated errors. Additionally, this alignment is more likely to lead to mismatches if high-dimensional vector representations need to be used, which is the case of 3D shape-based analysis. For overcoming these obstacles, a frequency domain analysis is followed in this work for identifying and modeling the dominant shape characteristics and their variation in time. In this way, the temporal sequence of the action constituent postures is captured, although this is not a strict temporal alignment of the respective action frames. In particular, for every frame t that belongs to the examined action segment an individual global 3D shape descriptor $\mathbf{q}_t$ is extracted. More specifically, for every frame t a composite voxel grid VG_t^{co} is computed, by superimposing all VG_t from the beginning of the action segment until frame t and estimating their outer surface. $\mathbf{q}_t$ is then computed by estimating a 3D shape descriptor for VG_t^{co}. Using VG_t^{co}, instead of VG_t, for descriptor extraction was experimentally shown to lead to better temporal action dynamics encoding. Indicative examples of VG_t^{co} estimation for different human actions are depicted in Fig. 5.

For producing a compact temporal-shape representation, the descriptor vector sequence $\mathbf{q}_t$ is initially adjusted to a predefined length H forming sequence $\overline{\mathbf{q}}_h$, using linear interpolation; the latter accounts for action sequences that typically consist of a different number of frames. $H = 20$ based on experimentation. Subsequently, 1D frequency domain analysis is applied to each of the value sequences $\overline{\mathbf{q}}_{s,h}$ that are formed by considering the s-th ($s \in [1, S]$) element of $\overline{\mathbf{q}}_h$ each time. For frequency domain analysis, the Discrete Cosine Transform (DCT) is applied to $\overline{\mathbf{q}}_{s,h}$, as follows:

$$fc_s(\beta) = \sum_{h=1}^{H} \overline{\mathbf{q}}_{s,h} \cos \frac{\pi}{H}[(h-1) + \frac{1}{2}(\beta - 1)] \qquad (10)$$

where $fc_s(\beta)$ are the estimated DCT coefficients and $\beta \in [1, H]$. The reason for using the DCT transform is twofold: a) its simple form requires relatively reduced calculations, and b) it is a frequency domain transform that receives as input a real sequence and its output is also a real set of values. Other common frequency analysis methods (e.g. Fourier transform) were also evaluated; however, they did not lead to increased performance compared with the one achieved when using DCT. Out of the H $fc_s(\beta)$ coefficients, only the first P are considered, since the remaining ones were experimentally shown to correspond mainly to noise or did not add to the discriminative power of the formed descriptor. The P selected coefficients for each $\overline{\mathbf{q}}_{s,h}$ are concatenated in a single vector that constitutes the proposed global 3D temporal-shape descriptor. It must be noted that modeling the correlations between different $\overline{\mathbf{q}}_{s,h}$ sequences during the descriptor extraction procedure (e.g. by directly applying 2D DCT transform on sequence $\overline{\mathbf{q}}_h$) led to inferior recognition performance, mainly due to overfitting occurrences. To this end, 1D DCT analysis, applied to each individual $\overline{\mathbf{q}}_{s,h}$ (as detailed in (10)), is adopted.

Although the proposed 3D temporal-shape descriptor extraction methodology is independent of the particular 3D static shape descriptor to be used, in this work the 'shape distribution' descriptor [50] (3D distance histogram) was utilized; this was experimentally shown to lead to better overall action recognition performance than other common shape descriptors. In [34], description and comparative evaluation of different static 3D shape descriptors for action recognition are given.

V. Action recognition

After extracting a set of local/global 3D flow/shape descriptors for every examined human action (as detailed in Section IV), an individual feature vector is estimated for every case. In particular, for the cases of the local 3D flow and shape information (where a set of STIPs is estimated for every action segment and subsequently local-level 3D flow/shape descriptions are extracted at every STIP location), the 'Bag-of-Words' (BoW) methodology [51] is followed for constructing a single vector description, where every action is represented by a L1-normalized histogram of 500 words. Additionally, the extracted global-level 3D flow/shape descriptors are also L1-normalized. In all cases, action recognition is realized using multi-class Support Vector Machines (SVMs).

VI. Experimental results

A. Datasets

In order to robustly, fairly and objectively evaluate the performance of the introduced human action recognition methods, as well as comparatively evaluating them against the respective literature approaches, challenging publicly-available datasets are required. In Table I, the most common and widely adopted datasets that have recently been introduced for the task of 3D human action recognition are reported. For every dataset, the most important characteristics are given, namely the number of action types, the total number of action instances, the number of individuals, the type of data and a short description of the capturing settings. The main criterion for selecting the datasets to be used in the experimental evaluation was the total number of action instances. In this respect, the two Huawei/3DLife datasets [49], which were used in the ACM Multimedia 2013 'Multimedia Grand Challenge', were utilized for designing and evaluating the proposed descriptors. In particular, in 'Dataset 1, session 1' (denoted D_1), 17 individuals were involved and $C = 5$ synchronized Kinect (placed at different viewpoints covering the whole body) were used for motion capturing. On the other hand, in 'Dataset 1,

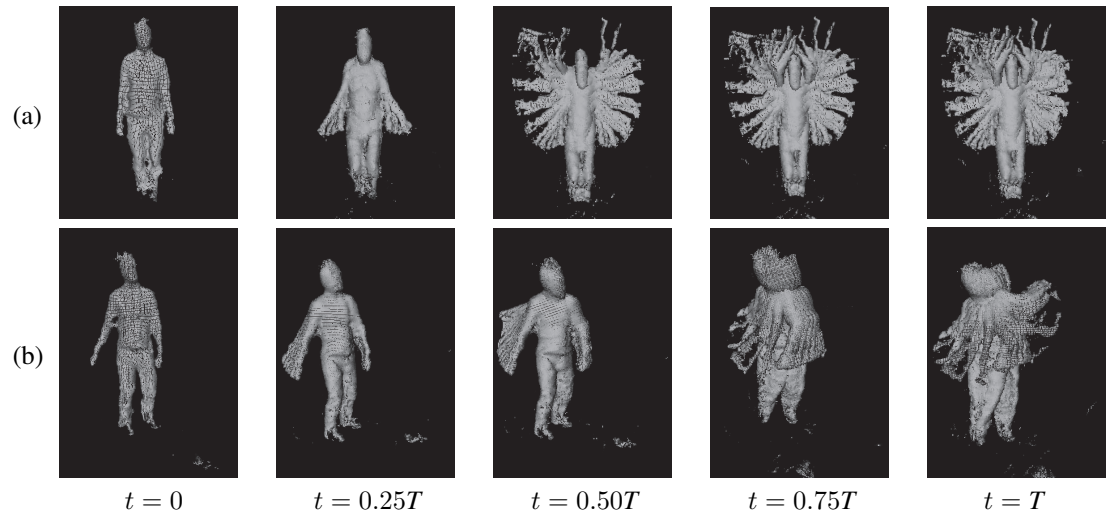


Fig. 5. Indicative examples of composite voxel grid VG_t^{co} estimation for actions: (a) jumping-jacks and (b) tennis-forehand. *T* denotes the overall duration of the action.

TABLE I
DATASETS FOR 3D HUMAN ACTION RECOGNITION

Dataset	No of action types	No of action instances	No of individuals	Type of data	Description
UTKinect [18]	10	200	10	RGB, depth	Single Kinect
Multiview Action3D [44]	10	1475	10	RGB, depth, human skeleton	3 synchronized Kinect
MSR Action3D [19]	20	567	10	Depth	Single depth sensor (using infra-red light)
MSR DailyActivity 3D [17]	16	320	10	RGB, depth, human skeleton	Single Kinect
Berkeley MHAD [45]	11	660	12	RGB, Mocap, depth accelerometer, audio	Multi-view multi-modal capturings
HuDaAct [46]	12	1189	30	RGB, depth	Single Kinect,capturing of human activities (30–150 sec)
CAD-60 [47]	12	∼350	4	RGB, depth, human skeleton	Single Kinect
i3DPost [48]	12	104	8	RGB, 3D meshes	Multi-view capturings
Huawei/3DLife [49] Dataset 1, session 1	22	∼1870	17	RGB, depth, audio, WIMU	5 synchronized Kinect multi-modal capturings
Huawei/3DLife [49] Dataset 1, session 2	22	∼1540	14	RGB, depth, audio, WIMU	2 synchronized Kinect multi-modal capturings
NTU RGB+D [7]	60	56880*	40	RGB, depth, human skeleton, infrared	3 synchronized Kinect multi-modal capturings

* total number of single-view action instances

session 2' (denoted D_2) actions of 14 human subjects were captured using $C = 2$ Kinect sensors. Out of the available 22 supported actions, the following set of 17 dynamic ones were considered for the experimental evaluation in this work: $E = \{e_g,\ g \in [1, G]\} \equiv$\{Hand waving, Knocking the door, Clapping, Throwing, Punching, Push away, Jumping jacks, Lunges, Squats, Punching and kicking, Weight lifting, Golf drive, Golf chip, Golf putt, Tennis forehand, Tennis backhand, Walking on the treadmill\}. The remaining 5 discarded actions (namely 'Arms folded', 'T-Pose', 'Hands on the hips', 'T-Pose with bent arms' and 'Forward arms raise') correspond to static ones that can be easily detected using a simple representation. At this point, the following facts need to be highlighted about these datasets: a) In D_2, the data stream from only the frontal Kinect was utilized, and b) In D_1, 5 Kinect were used for human motion capturing; however, the interference between multiple Kinect degrades the quality of the captured depth maps. This results into the introduction of noise (compared with D_2), which for example makes standard depth-based skeleton-tracking algorithms (e.g. the standard OpenNI skeleton-tracking module) less accurate. Additionally, the 'NTU RGB+D' [7], which is the broadest in the literature and is denoted D_3, and the 'UTKinect' [18] (denoted D_4) datasets were also used, mainly for comparatively evaluating the proposed approach. For the latter datasets only overall action recognition results are provided. In the sequel, for D_3 performance is measured using the 'cross-subject' and 'cross-view' evaluation criteria explicitly defined in [7]. For all other datasets, performance evaluation is realized following the 'leave-one-out' methodology, where in every iteration one subject is used for performance measurement and the remaining ones are used for training.

B. Local descriptors evaluation

In this section, experimental results, as well as comparative evaluation, from the application of the proposed local-level descriptors (presented in Section IV-A) are presented. In Fig. 6, quantitative action recognition results are presented in the form of the calculated recognition rates (i.e. the percentage of the action instances that were correctly identified), when local flow and shape information is used. Additionally, the value of the overall classification accuracy, i.e. the percentage of all

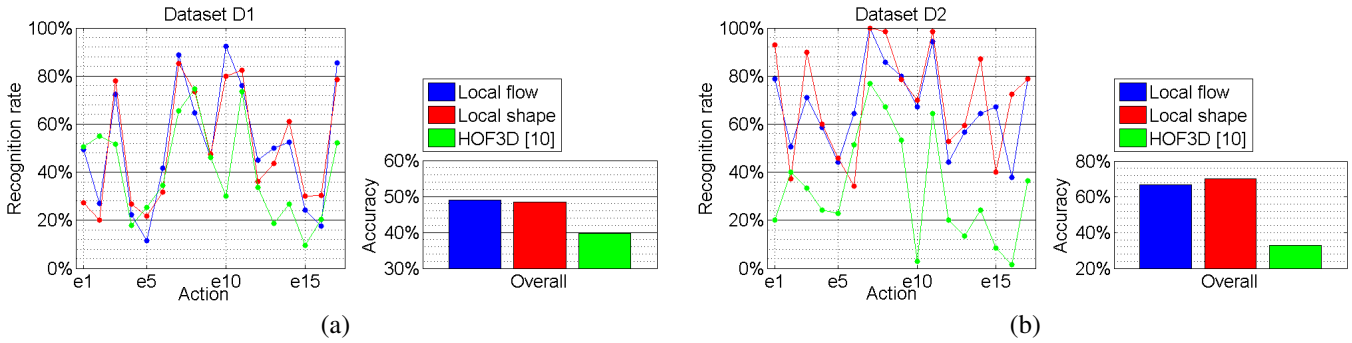


Fig. 6. Action recognition results using local-level descriptors for a) D_1 and b) D_2 datasets.

action instances that were correctly classified, is also given for every case. From the presented results, it can be seen that the proposed 3D flow descriptor leads to satisfactory action recognition performance (overall accuracy equal to 49.00% and 66.81% in D_1 and D_2, respectively). Examining the results in more detail, it is observed that there are actions that exhibit high recognition rates in both datasets (e.g. 'Jumping jacks', 'Punching and kicking' and 'Weight lifting'), since they present characteristic motion patterns among all subjects. However, there are also actions for which the recognition performance is not that increased (e.g. 'Punching', 'Throwing' and 'Tennis backhand'). This is mainly due to these actions presenting very similar motion patterns over a period of time during their execution with other ones (e.g. 'Throwing', 'Punching and kicking' and 'Tennis forehand', respectively). On the other hand, it can be seen that the 3D flow descriptor leads to slightly increased in D_1 and comparable performance in D_2, compared with the utilized 3D shape descriptor. 3D flow leads to this inferior performance in D_2 mainly due to the relatively lower quality of $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$, which in D_2 is estimated using a single Kinect.

The proposed 3D flow descriptor is comparatively evaluated with a similar approach of the literature, namely the HOF3D descriptor with 'vertical rotation' presented in [10]. HOF3D is also a local-level histogram-based descriptor. However, the local-level coordinate system is defined using the vertical axis and the horizontal component of the 3D flow vector at the examined STIP. Subsequently, a 3D flow histogram is constructed by uniformly dividing the corresponding 3D sphere into a set of orientation bins. It must be highlighted that the above-mentioned HOF3D descriptor is representative of a set of literature methods that employ local-level 3D flow histogram representations for performing action recognition, including the work of [29] (i.e. flow descriptions that do not take into account spatial distribution or surface information). From the results presented in Fig. 6, it is obvious that the proposed flow descriptor leads to increased performance compared with HOF3D in both datasets. This is due to the following advantageous characteristics that the proposed flow descriptor presents and which are described in details in Section IV-A1: a) introduction of a consistent orientation definition at every STIP location, based on the normal vector $\mathbf{n}_t^{stip}(x_g, y_g, z_g)$; on the contrary, HOF3D considers the vertical axis for orientation definition at all STIP positions. b) inclusion of spatial distribution related information, by estimating an individual flow histogram for every defined concentric ring-shaped area $A_{i,j}$; HOF3D does not incorporate information regarding the spatial distribution of the flow vectors in the neighborhood of each STIP location. c) inclusion of surface information, by constructing the histogram of the values of angle ϑ (i.e. the angle between the normal $\mathbf{n}_t(x_g, y_g, z_g)$ and the flow $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ vectors at every voxel $v_t(x_g, y_g, z_g)$ position, as defined in (7)) in each $A_{i,j}$ area; HOF3D does not incorporate surface information. The aforementioned advantageous characteristics result in the proposed local flow descriptor to exhibit approximately 9.22% and 34.13% increased overall performance, compared with HOF3D, in D_1 and D_2, respectively. This large performance difference in D_2 is in principle caused by the decreased quality of flow field $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ that is estimated using a single Kinect, while $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ in D_1 is computed by fusing information coming from 5 Kinect. In other words, the proposed local flow descriptor is shown to be more robust in the presence of noise and lower quality of the available flow field than HOF3D. Nevertheless, the performance of the proposed local flow descriptor is also affected, since in D_1 it slightly outperforms the proposed local shape descriptor, while in D_2 local shape information is advantageous.

1) Parameter selection: In order to apply and evaluate the performance of the proposed local-level descriptors, particular values inevitably need to be selected for the defined parameters. In this section, quantitative evaluation results are given for the most crucial parameters, aiming at shading light on the behavior of the respective descriptors. It must be noted that experimental results are given only for D_1, while similar behavior of the proposed local-level descriptors has been observed in D_2. For D_3 and D_4, the parameters selected for D_1 were used. In particular, the descriptor behavior for different values of the following parameters, along with justification where particular values were selected, has been investigated:

- Parameters σ and τ: These roughly correspond to the spatial and temporal scale of the employed STIP detector (Section IV-A). In this work, $\sigma = 2.0$ and $\tau = 0.9$ were set. These constitute values that are typically used in similar STIP detectors of the literature (analysis in the xy+t 3D space) [41] [52]. Additionally, the respective threshold value, which is used for generating the detected STIPs, was selected so as to lead to the estimation of at least 200 STIPs for any supported action type (e.g. even for the 'Clapping' action instances that typically

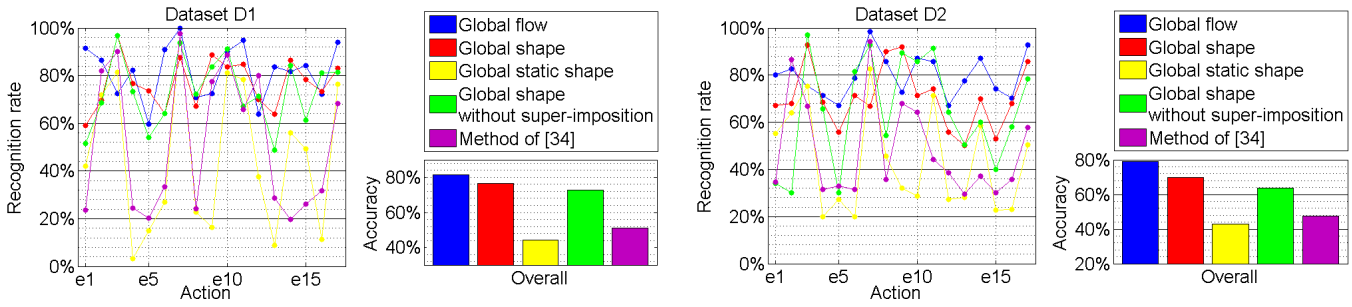


Fig. 7. Action recognition results using global-level descriptors for a) D_1 and b) D_2 datasets.

correspond to short temporal segments with no extensive motion observed). This was experimentally shown to be sufficient to lead to good recognition results. Using lower threshold values (i.e. generating more STIPs) did not lead to performance improvement, although it significantly increased the computational cost.

- Parameters I, J and p: These correspond to the number of segments along the longitudinal axis, number of segments along the polar axis and number of histogram bins during the local flow descriptor extraction procedure (Section IV-A1), respectively. In particular, the performance obtained by the application of the introduced local flow descriptor is given for different sets of values for the aforementioned parameters in Table II. It must be noted that the pair ($I = 0$, $J = 1$) corresponds to the case where no spatial information is incorporated in the extracted descriptor, i.e. a single flow histogram is estimated for the whole spatio-temporal cuboid defined for a given STIP $v_t^{stip}(x_g, y_g, z_g)$. From the first group of presented results, it can be observed that the pair ($I = 2$, $J = 5$) leads to the best overall performance. Additionally, based on the second group of results, it can be seen that using more histogram bins ($p = 8$) leads to slightly increased recognition performance.
- Parameter D^s_{cub}: This defines the spatial dimension of the spatio-temporal cuboid used for local flow/shape descriptor extraction (Section IV-A1). In Table III, the recognition performance accomplished using the proposed local flow and shape descriptor is reported. From the presented results, it can be seen that low values of D^s_{cub} (i.e. $D^s_{cub} = 21$) lead to decreased recognition performance, mainly due to the reduced size of the cuboid not being sufficient for efficiently capturing the local action dynamics. On the other hand, values greater than 31 lead practically to marginal variations in performance; however, the computational cost of the descriptor extraction procedure increases exponentially.
- BoW dimension: This defines the number of 'words' used in the BoW representation (Section V). In Table IV, the action recognition performance for different values of the BoW dimension for the proposed local-level descriptors is given. It can be observed that using a small number of 'words' leads to decreased performance, while using a number of approximately 500 'words' leads to the best recognition performance for both descriptors.
- Parameter D^t_{cub}: This defines the temporal dimension (i.e. number of frames) of the spatio-temporal cuboid used for local flow/shape descriptor extraction (Section IV-A1). In this work, $D^t_{cub} = 3$ was set. Greater values were also evaluated (i.e. $D^t_{cub} = 5, 7$); however, these resulted into negligible variations in the overall recognition performance, while significantly increasing the computational complexity at the same time.

TABLE II
LOCAL FLOW DESCRIPTOR PARAMETER SELECTION

Parameters	Accuracy
I=0, J=1, p=8	36.31%
I=0, J=3, p=8	37.23%
I=1, J=1, p=8	43.14%
I=1, J=3, p=8	44.91%
I=2, J=1, p=8	45.38%
I=2, J=5, p=8	**49.00%**
I=4, J=9, p=8	47.41%
I=2, J=5, p=4	48.11%
I=2, J=5, p=8	**49.00%**

TABLE III
SPATIAL CUBOID DIMENSION SELECTION

	D^s_{cub}			
Descriptor	21	31	41	51
Local flow	42.11%	**49.00%**	48.68%	48.71%
Local shape	41.56%	**48.40%**	47.98%	48.08%

TABLE IV
BOW DIMENSION SELECTION

	BoW dimension			
Descriptor	100	250	500	1000
Local flow	42.98%	47.52%	**49.00%**	47.25%
Local shape	42.48%	46.93%	**48.40%**	46.44%

C. Global descriptors evaluation

In this section, experimental results and comparative evaluation from the application of the proposed global-level descriptors (described in Section IV-B) are presented. In Fig. 7, quantitative results in terms of the estimated recognition rates and overall accuracy are given for the proposed global flow and shape descriptors. From the presented results, it can be seen that both descriptors achieve high recognition rates in both datasets; namely, the flow (shape) descriptor exhibits recognition rates equal to 81.27% and 78.99% (76.53% and 69.83%) in D_1 and D_2, respectively. From these results, it can be observed that the global flow descriptor outperforms the respective shape one in both utilized datasets; this is mainly due to the more detailed and discriminative information

contained in the estimated 3D flow fields. Due to the latter factor, the flow descriptor is advantageous for actions that incorporate more fine-grained body/body-part movements (e.g. 'Hand waving', 'Knocking the door', 'Punching and kicking' and 'Weight lifting'). On the other hand, the cases that the shape descriptor is better involve body movements with more extensive and distinctive whole body postures (e.g. actions 'Clapping' and 'Squats').

In order to investigate the behavior of the proposed global temporal-shape descriptor, comparison with the following benchmarks is performed: a) global static shape descriptor: A static shape descriptor (the 'shape distribution' descriptor described in Section IV-B2) is extracted for the composite voxel grid VG_t^{co} for $t = T$, i.e. when all constituent voxel grids VG_t of an action are superimposed. This can be considered as the counterpart of the respective volumetric descriptions for the 2D analysis case, i.e. methods that estimate a 3D volumetric shape of the examined action from the 2D video sequence and subsequently estimating a 3D shape descriptor of the generated volume (like in [4] [3]). b) variant of the proposed temporal-shape descriptor, where voxel grids VG_t are used instead of the composite ones VG_t^{co} during the descriptor extraction procedure. c) method of [34]: A self-similarity matrix is computed for every action (by means of static shape descriptor extraction for every frame) and subsequently a temporal shape descriptor is estimated by applying a time filter to the calculated matrix. From the results presented in Fig. 7, it can be seen that the proposed temporal-shape descriptor significantly outperforms the static one in both datasets. This fact highlights the significant added value of incorporating temporal information in the global 3D representation. Additionally, it can be observed that the use of the composite voxel grids VG_t^{co} is advantageous compared with when using the voxel grids VG_t. The latter implies that superimposing information from multiple frames during the descriptor extraction procedure can lead to more discriminative shape representations. Moreover, the proposed temporal-shape descriptor is also shown to outperform the temporal-shape method of [34]. This denotes the increased efficiency of the frequency domain analysis on top of the per-frame extracted shape descriptors in capturing and modeling the human action dynamics, compared with the case of estimating the self-similarity matrix of the same descriptors and applying time filtering techniques. It must be noted that a comparative evaluation of the proposed global 3D flow method has not been included, due to the following facts: a) the great majority of the global 3D flow descriptors of the literature employ simple global-level histogram representations (i.e. without considering the spatial distribution of the flow vectors), like in [30] [32]; hence, they lead to significantly lower action recognition performance compared with the proposed global 3D flow descriptor (similar observations with the comparison of the proposed local-level 3D flow descriptor and the HOF3D one in Section VI-B). b) the methods of [28] and [31] include spatial information in the estimated global 3D flow representation. However, these methods were also not included in the conducted comparative evaluation. This is due to the method of [28] being view-dependant, since it employs a static 3D space grid division that is defined according to the single Kinect sensor that is assumed to be present; hence, the comparison of the proposed global 3D flow descriptor with the method of [31] would not be fair. The same fact (i.e. view-variance) holds for the method of [31].

TABLE V
GLOBAL FLOW DESCRIPTOR PARAMETER SELECTION

Parameters	Accuracy
K=3, Λ=3, b_ψ=6, b_o=3	75.44%
K=4, Λ=4, b_ψ=6, b_o=3	**81.27**%
K=5, Λ=5, b_ψ=6, b_o=3	80.88%
K=4, Λ=4, b_ψ=6, b_o=3	**81.27**%
K=4, Λ=4, b_ψ=4, b_o=3	79.97%
K=4, Λ=4, b_ψ=6, b_o=3	**81.27**%
K=4, Λ=4, b_ψ=6, b_o=6	77.22%

TABLE VI
TEMPORAL-SHAPE DESCRIPTOR PARAMETER SELECTION

	Parameter P			
Dataset	5	10	15	20
D_1	**76.53**%	71.68%	68.11%	66.64%
D_2	**69.83**%	66.12%	61.64%	57.82%

1) Parameter selection: Similarly to the case of local descriptor behavior evaluation for different parameter value selection (Section VI-B1), the behavior of the proposed global flow and shape descriptors is detailed in this section. It must be noted that in the followings experimental results are given only for D_1, while similar behavior of the proposed global-level descriptors has been observed in D_2. For D_3 and D_4, the parameters selected for D_1 were used again. In particular, the performance accomplished by the application of the introduced global descriptors is investigated with respect to the values of the following key parameters:

- Parameters K, Λ, b_ψ, b_o: K and Λ control the partitioning of the longitudinal and the polar axis, when defining the ring-shaped areas $B_{\kappa,\lambda}$ (Section IV-B1), respectively. Additionally, b_ψ and b_o define the number of bins of the histograms calculated with respect to angles ψ and o (Section IV-B1), respectively. In Table V, action recognition results from the application of the proposed global flow descriptor for different sets of values of the aforementioned parameters are given. From the first group of experimental results, it can be seen that the ring-shape partitioning using $K = 4$ and $\Lambda = 4$ leads to the best overall performance. Additionally, the second group of experiments shows that using more bins in the histogram representation with respect to angle ψ, which corresponds to the angle between the horizontal projection of $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ and the projection of the vector connecting the cylindrical center (x_{cg}, y_{cg}, z_{cg}) with the examined voxel position (x_g, y_g, z_g) on the horizontal xz plane, is advantageous. On the other hand, using a decreased number of bins in the histogram representation with respect to angle o, which corresponds to the angle of $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ with the vertical axis, leads to increased performance (third group of experiments).
- Parameter S: This corresponds to the dimensionality of the global static shape descriptor that is used for estimating the proposed global temporal-shape one (Sec-

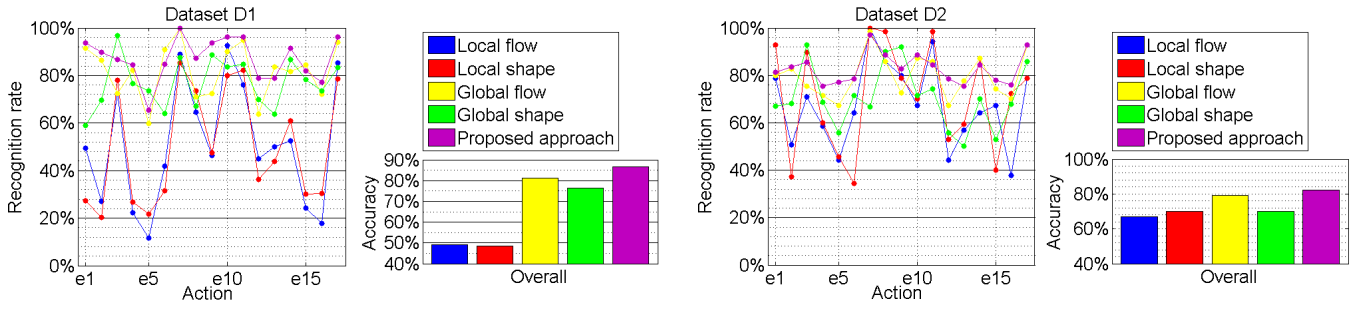


Fig. 8. Overall action recognition results for a) D_1 and b) D_2 datasets.

tion IV-B2). For the selected static 'shape distribution' descriptor [50], a 3D distance histogram of $S = 20$ bins exhibited maximum recognition performance (indicated as 'global static shape' descriptor in Fig. 7); greater values of S led to a gradual decrease in the overall action recognition rate.

- Parameter H: This adjusts the length of the shape descriptor vector sequence $\overline{\mathbf{q}}_h$ (Section IV-B2). In the current implementation, H was set equal to 20, which is close to the average action segment duration in frames in the employed datasets.
- Parameter P: This defines the number of selected DCT coefficients to be used in the produced global shape representation (Section IV-B2). The performance obtained by the application of the proposed temporal-shape descriptor for different values of P is given for both datasets in Table VI. From the presented results, it can be seen that the best performance is achieved when only relatively few frequency coefficients are used (i.e. $P = 5$); these are shown to be adequate for accomplishing a good balance between capturing sufficient temporal information and maintaining the dimensionality of the overall descriptor low. To this end, the dimensionality of the proposed global temporal-shape descriptor is this work is equal to $S \cdot P = 20 \cdot 5 = 100$.

D. Overall approach evaluation

Experimental results from the application of all proposed descriptors and their combination, as well as comparison with literature approaches, are reported in this section. In Fig. 8 and Tables VII-X, quantitative results in terms of the estimated recognition rates and overall accuracy are given for the introduced local/global flow/shape descriptors (for D_3 and D_4 only overall performance is reported). From the presented results, it can be seen that generally the global descriptors lead to higher recognition rates than the local ones. This observation implies that global-level representations are more discriminative and provide better modelling of the action dynamics than the local ones. Global descriptors are advantageous for a wide set of actions, including local hand movements (e.g. 'Knocking the door', 'Throwing and punching') as well as more extensive whole-body movements (like 'Push away' and 'Golf drive') in D_1 and D_2. Overall, the proposed global flow descriptor leads to the best recognition performance. However, the overall proposed approach (which consists of simple concatenation of all computed descriptors in a single feature vector) accomplishes to achieve increased performance, compared with all cases of using each individual descriptor alone. The latter demonstrates the complementarity of the proposed descriptors.

In Tables VII-X, the proposed approach is comparatively evaluated with numerous methods of the literature and it is shown to exhibit state-of-art performance in most cases (i.e. 3 out of 4 employed datasets). Generally, literature approaches can be divided into the following main categories, with respect to the type of information that they utilize for realizing 3D action recognition: surface, flow and skeleton-tracking ones.

From the presented results, it is shown that surface methods (either using only pure depth information, like the proposed temporal-shape descriptor, or also taking into account surface normal vectors, like the methods of [53] and [54]) exhibit satisfactory results across different datasets. Approaches of this category mainly aim at capturing the global posture of the performing subjects and through different methodologies (e.g. frequency domain analysis in the proposed temporal-shape descriptor, estimation of 'extended' normal vectors in the 4D space in [53] [54], etc.) to encode also temporal information about the action dynamics. Common characteristic of all methods is that they take into account all available points belonging to the human silhouette and assign them equal importance. Concerning the particular methods of [53] and [54], they focus on exploiting only the orientation of the calculated normal vectors; hence, they model local surface characteristics (including information about the observed motion), while however neglecting the magnitude of motion. Additionally, since both aforementioned methods do not incorporate any particular consideration for tackling the problem of view variance, they present a significant performance decrease for the 'cross-view' evaluation scenario (i.e. when training and test is realized under different viewing perspectives of the same actions) in D_3; such drawback is not present for the proposed shape descriptors.

Flow methods, like the proposed global flow descriptor, take into account both surface and RGB information, and provide a detailed representation of the action dynamics. They are shown to exhibit increased recognition performance, provided that a sufficient number of flow vectors is available for exploiting the full expressiveness capabilities of the corresponding descriptors. These methods perform at the signal level (flow field) and for their application require a pre-processing step of flow calculation. The proposed global flow descriptor exhibits recognition performance competitive to the state-of-art in most

TABLE VII
COMPARATIVE EVALUATION IN DATASET D1

Method	Accuracy
Local shape	48.40%
Local flow	49.00%
Global shape	76.53%
Global flow	81.27%
Proposed approach	**86.78**%
Method of [20]	66.62%

TABLE VIII
COMPARATIVE EVALUATION IN DATASET D2

Method	Accuracy
Local shape	70.09%
Local flow	66.81%
Global shape	69.83%
Global flow	78.99%
Proposed approach	**82.26**%
Method of [20]	76.03%
Method of [22]	79.78%
Method of [55]	77.43%

datasets, except from D_3 (as it will be discussed in the sequel). In particular, the proposed global flow descriptor outperforms all surface methods (proposed shape descriptors, HON4D [54], 'super normal vector' [53]), mainly due to exploiting more accurate/fine-grained information (i.e. both the direction and the magnitude of motion at every point of the human silhouette) and focusing only on points where motion is observed (also assigning varying importance to every point, based on the magnitude of the corresponding flow vector). Global flow performance is inferior compared to the state-of-art in D_3, due to the particular characteristics of this specific dataset, where all performing human subjects are positioned at a relatively increased distance from the capturing medium. As a consequence, the changes in the subjects' silhouette surface as well as the exhibited human motion is more difficult to be efficiently captured and modeled. In order to make this difference in the characteristics between the datasets more clear, it is observed that for most actions in D_3 only a set of approximately ~ 5K motion vectors in total are estimated for every frame, while in D_1 (where the global flow descriptor shows the best performance compared with the state-of-art) this number approaches the value of ~ 20K on average.

Skeleton-tracking approaches, which make extensive use of domain knowledge and rely on the prior application of a human joint detector, typically employ straightforward representations of the human posture. Their main drawback is that their efficiency largely relies on the robustness of the employed skeleton tracker. Most literature approaches for 3D human action recognition belong to this category and have also been evaluated in the employed datasets, including the works of [20] in D_1, [20] [22] [55] in D_2, [56] [9] [57] [8] [58] [7] in D_3 and [18] [59] [60] in D_4. From the presented results, it can be seen that skeleton-tracking approaches achieve high recognition rates in all datasets. Especially in D_3, skeleton-tracking methods are shown to be advantageous, compared with other unimodal approaches. This suggests that skeleton-tracking methods are advantageous when the silhouettes of the performing subjects are captured in lower resolution (e.g. when the performing subject is positioned in a relatively

TABLE IX
COMPARATIVE EVALUATION* IN DATASET D3

Method	Cross-subject Accuracy	Cross-view Accuracy
Local shape	32.49%	35.11%
Local flow	34.33%	37.42%
Global shape	46.24%	50.39%
Global flow	48.09%	52.44%
Proposed approach	58.48%	66.59%
HOG^2 [56]	32.24%	22.27%
Super Normal Vector [53]	31.82%	13.61%
HON4D [54]	30.56%	7.26%
Lie Group [9]	50.08%	52.76%
Skeletal Quads [57]	38.62%	41.36%
FTP Dynamic Skeletons [8]	60.23%	65.22%
HBRNN-L [58]	59.07%	63.97%
2 Layer RNN [7]	56.29%	64.09%
2 Layer LSTM [7]	60.69%	67.29%
2 Layer P-LSTM [7]	**62.93**%	**70.27**%

* results of literature methods are provided as reported in [7]

TABLE X
COMPARATIVE EVALUATION IN DATASET D4

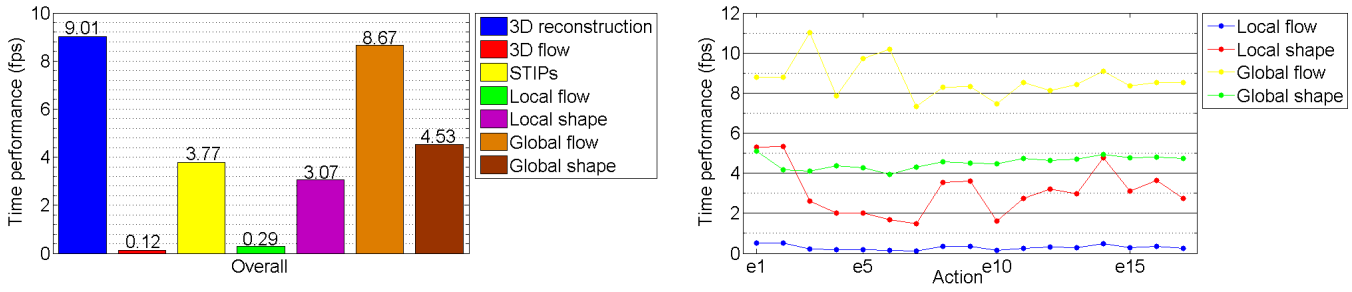
Method	Accuracy
Local shape	76.00%
Local flow	73.50%
Global shape	81.00%
Global flow	89.00%
Proposed approach	**93.00**%
Histogram of 3D Joints [18]	90.92%
Grassmann Manifold [59]	88.50%
Riemannian Manifold [60]	91.50%

greater distance from the capturing sensor). However, when a sufficiently large number of motion vectors are available (e.g. in D_1), skeleton-tracking methods are outperformed by flow ones, as discussed above.

From the results presented in Tables VII-X, it can be seen that the overall proposed approach, which combines local/global flow/shape descriptors, outperforms most literature methods and exhibits state-of-art performance in three out of the four employed datasets. Concerning comparison with the recent data-driven approaches in the computer vision field, the so called 'Deep Learning (DL)' approaches, results of the methods described in [58] and [7] are reported in Table IX. From the presented results, it can be seen that the overall proposed approach exhibits competitive recognition rates. It needs to be reminded, though, that skeleton-tracking approaches, including the DL methods described in [58] and [7], are favored in D_3, as already discussed.

E. Time performance

Having examined the action recognition performance of the proposed descriptors in the previous sub-sections, the time performance of the individual parts of the developed framework are investigated in details here. It must be highlighted that during the development of the proposed framework no particular attention was given on time performance-related issues, i.e. emphasis was primarily put on investigating and comparatively evaluating the recognition performance achieved by exploiting different information types (namely local/global 3D flow/shape information). In other words, significant time



(a) (b)

Fig. 9. Time performance of a) individual parts of the proposed framework and b) individual descriptors per action type.

performance improvements may be obtained with negligible or no variation in the corresponding recognition behavior, as it will become clear from the following discussion.

In Fig. 9a), the average time performance of the key information processing components of the proposed framework (depicted in Fig. 1) is given in terms of the calculated frames per second (fps) processing rate that has been measured in D_1. From the presented results, it can be seen that the adopted 3D reconstruction algorithm, which generates meshes comprising approximately $\sim$ 280K triangles, exhibits the highest processing rate. This is mainly due to the GPU-based implementation of the algorithm of [14] that is used in this work, as opposed to all remaining framework parts that are implemented in CPU. On the other hand, the 3D flow field calculation constitutes the most time consuming step, largely due to the selected 2D optical flow algorithm of [38] (Section III-B) that is applied on frames of dimension 640x480 pixels. However, significantly more time efficient optical flow algorithms can be used without significant decrease in the quality of the computed motion fields [61]. Implementations resulting in high 3D flow processing rates have also been reported (Section II-A). Additionally, the observed STIPs detection rate is measured for 3D grids comprising approximately 6.7 million voxels in total. Regarding the time efficiency of the proposed descriptors, the local flow one presents the lowest fps rate, which is mainly due to the increased computations that are required at every individual STIP position. Additionally, the extraction time of the proposed local and global shape descriptors for action recognition is mainly controlled by the computational complexity of the corresponding static 3D shape descriptors that are involved in their computation (Sections IV-A2 and IV-B2); the methods of [43] and [50] that have been used, respectively, were reported to present a good compromise between recognition performance and time complexity. Notably, the global flow descriptor (i.e. the best performing one among the introduced descriptors, as indicated in Section VI-D) also exhibits the best time performance. The latter has emerged as a consequence of its straight-forward computation (Section IV-B1).

Given the fact that the computational complexity of all preprocessing steps (namely 3D reconstruction, 3D flow estimation, STIPs detection) is not affected by the particular type of the observed human action, the average time performance of only the proposed descriptors per action type are reported in Fig. 9b). From the presented results, it can be seen that for the case of the local flow descriptor the very low processing rates are maintained for all action types. Additionally, for the global shape descriptor, the processing rate presents small variations among the different actions, mainly due to the fact that for the employed static 3D shape descriptor of [50] a constant number of points was used for its computation for each individual frame. On the other hand, the global flow descriptor exhibits the best performance for actions that do not involve extensive body movements (i.e. estimation of a relatively reduced total number of motion vectors), like actions 'Clapping' and 'Push away'. Moreover, the processing rate of the local shape descriptor (which is in principle determined by the total number of detected STIPs) exhibits its highest values for actions with not so rapid and intense changes in the shape of the performing subjects, e.g. actions 'Hand waving', 'Knocking the door' and 'Golf putt'. It needs to be mentioned that all reported processing rates were measured using a PC with Intel i7 processor at 3.5 GHz and a total of 16 GB RAM, while the capturing rate in D_1 is equal to 30 fps (i.e. the standard Kinect I frame rate).

VII. Conclusions

In this work, the problem of human action recognition using 3D reconstruction data was examined in detail and novel local/global 3D flow/shape descriptors were introduced. Additionally, comparative evaluation with multiple literature methods in different public datasets was provided, where the proposed approach was shown to be advantageous in most cases. Based on the conducted study, the following key conclusions can be drawn:

i) All proposed descriptors are experimentally shown to have a complementary nature and to contribute towards an increased overall recognition performance.

ii) Global descriptors provide more discriminant action representations than local ones, with the performance difference to be greater when higher quality voxel grids and 3D flow fields are available; the proposed global 3D flow descriptor achieves the best overall results in all datasets.

iii) Concerning time performance issues, global descriptors are advantageous, with global 3D flow, i.e. the best performing descriptor in terms of recognition accuracy, to exhibit the highest processing rate.

iv) Combining points (ii) and (iii), if the maximum overall recognition performance is not the sole criterion for building a 3D action recognition system, using only global descriptors is the most efficient solution, since they combine increased

recognition performance with high processing rates. In case of significant lack of computational resources, only the use of global 3D flow is suggested.

v) Concerning the per action type recognition performance, 3D flow is more efficient for modeling actions that incorporate more fine-grained body/body-part movements, while 3D shape is advantageous for actions with more extensive and distinctive whole body postures.

vi) Regarding the per action type time performance, actions that do not involve extensive body movements lead to higher processing rates the global flow and local shape descriptors, while the local flow and global shape ones exhibit relatively small variations in performance.

vii) With respect to the type of information that is used for realizing 3D action recognition, it is shown that surface methods exhibit satisfactory results across different experimental settings. Additionally, flow methods provide a detailed representation of the action dynamics and exhibit increased performance, provided that a sufficient number of flow vectors is available. On the other hand, skeleton-tracking approaches are shown to be advantageous when the silhouettes of the performing subjects are captured in lower resolution.

viii) Based on point (vii), it can be claimed that flow methods are more appropriate when fine-grained motion analysis is required and high-quality capturing is guaranteed (e.g. in security, sports or rehabilitation applications). On the other hand, in case of loose capturing settings, skeleton-tracking approaches are advantageous.

ix) A truly robust system in the general case should efficiently and adaptively combine all aforementioned information types (surface, flow, skeleton-tracking).

x) Concerning future research directions, the increased potentials of the 3D flow information stream could reasonably be combined with the recent advances in the data-driven 'Deep Learning' community. In particular, Convolutional Neural Networks (CNNs) could be employed for estimating discriminant features along the spatial dimensions of the flow field; current 'Deep Learning' methods only utilize Recurrent Neural Networks (RNNs) on top of skeleton-tracking data [7] [58], in order to model correlations in the temporal dimension. Additionally, a composite CNN-RNN architecture could potentially handle the challenge of multi-modal information fusion.

ACKNOWLEDGMENT

The work presented in this paper was supported by the European Commission under contract FP7-607480 LASIE.

REFERENCES

[1] P. V. K. Borges, N. Conci, and A. Cavallaro, "Video-based human behavior understanding: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 11, pp. 1993–2008, 2013.

[2] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[3] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[4] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 12, pp. 2247–2253, 2007.

[5] J. Gu, X. Ding, S. Wang, and Y. Wu, "Action and gait recognition from recovered 3-d human joints," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans. on*, vol. 40, no. 4, pp. 1021–1033, 2010.

[6] G. T. Papadopoulos, A. Briassouli, V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "Statistical motion information extraction and representation for semantic video analysis," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 19, no. 10, pp. 1513–1528, Oct. 2009.

[7] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2016.

[8] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5344–5352.

[9] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 588–595.

[10] M. B. Holte, B. Chakraborty, J. Gonzalez, and T. B. Moeslund, "A local 3-d motion descriptor for multi-view human action recognition from 4-d spatio-temporal interest points," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 5, pp. 553–565, 2012.

[11] S. Cho and H. Byun, "A space-time graph optimization approach based on maximum cliques for action detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 661–672, April 2016.

[12] T. V. Nguyen, Z. Song, and S. Yan, "Stap: Spatial-temporal attention-aware pooling for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 1, pp. 77–86, Jan 2015.

[13] X. Wu, D. Xu, L. Duan, J. Luo, and Y. Jia, "Action recognition using multilevel features and latent structural svm," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 8, pp. 1422–1431, Aug 2013.

[14] D. S. Alexiadis, D. Zarpalas, and P. Daras, "Real-time, full 3-d reconstruction of moving foreground objects from multiple consumer depth cameras," *Multimedia, IEEE Transactions on*, vol. 15, no. 2, pp. 339–358, 2013.

[15] L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2834–2841.

[16] H. Zhang and L. E. Parker, "Code4d: Color-depth local spatio-temporal features for human activity recognition from rgb-d videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 3, pp. 541–555, March 2016.

[17] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 1290–1297.

[18] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[19] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–14.

[20] G. T. Papadopoulos, A. Axenopoulos, and P. Daras, "Real-time skeleton-tracking-based human action recognition using kinect data," in *MultiMedia Modeling, Int. Conf. on*, 2014, pp. 473–483.

[21] Y. Shan, Z. Zhang, P. Yang, and K. Huang, "Adaptive slice representation for human action classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 10, pp. 1624–1636, Oct 2015.

[22] L. Sun and K. Aizawa, "Action recognition using invariant features under unexampled viewing conditions," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 389–392.

[23] Y. Song, J. Tang, F. Liu, and S. Yan, "Body surface context: A new robust feature for action recognition from depth videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 952–964, June 2014.

[24] T. Basha, Y. Moses, and N. Kiryati, "Multi-view scene flow estimation: A view centered variational approach," *International journal of computer vision*, vol. 101, no. 1, pp. 6–21, 2013.

[25] J. Cech, J. Sanchez-Riera, and R. Horaud, "Scene flow estimation by growing correspondence seeds," in *Computer Vision and Pattern*

Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011, pp. 3129–3136.

[26] D. Alexiadis, N. Mitianoudis, and T. Stathaki, "Multidimensional steerable filters and 3d flow estimation," in *Image Processing (ICIP), IEEE International Conference*, 2014.

[27] M. Sizintsev and R. P. Wildes, "Spatiotemporal stereo and scene flow via stequel matching," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 6, pp. 1206–1219, 2012.

[28] M. Munaro, G. Ballin, S. Michieletto, and E. Menegatti, "3d flow estimation for human action recognition from colored point clouds," *Biologically Inspired Cognitive Architectures*, vol. 5, pp. 42–51, 2013.

[29] L. Xia, I. Gori, J. Aggarwal, and M. Ryoo, "Robot-centric activity recognition from first-person rgb-d videos."

[30] I. Gori, S. R. Fanello, F. Odone, and G. Metta, "A compositional approach for 3d arm-hand action recognition," in *Robot Vision (WORV), 2013 IEEE Workshop on*. IEEE, 2013, pp. 126–131.

[31] K. Subramanian, S. Sundaram, and R. V. Babu, "3-d optical flow based human action recognition with meta-cognitive neuro-fuzzy inference system."

[32] S. R. Fanello, I. Gori, G. Metta, and F. Odone, "Keep it simple and sparse: Real-time action recognition," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2617–2640, 2013.

[33] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 726–733.

[34] P. Huang, A. Hilton, and J. Starck, "Shape similarity for 3d video sequences of people," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 362–381, 2010.

[35] C. Budd, P. Huang, M. Klaudiny, and A. Hilton, "Global non-rigid alignment of surface sequences," *International Journal of Computer Vision*, vol. 102, no. 1-3, pp. 256–270, 2013.

[36] T. Yamasaki and K. Aizawa, "Motion segmentation and retrieval for 3d video based on modified shape distribution," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 211–211, 2007.

[37] R. Slama, H. Wannous, and M. Daoudi, "3d human motion analysis framework for shape similarity and retrieval," *Image and Vision Computing*, vol. 32, no. 2, pp. 131–154, 2014.

[38] M. Proesmans, L. Van Gool, E. Pauwels, and A. Oosterlinck, "Determination of optical flow and its discontinuities using non-linear diffusion," in *Computer VisionECCV'94*. Springer, 1994, pp. 294–304.

[39] M. Mammarella, G. Campa, M. L. Fravolini, and M. R. Napolitano, "Comparing optical flow algorithms using 6-dof motion of real-world rigid objects," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 42, no. 6, pp. 1752–1762, 2012.

[40] D. Herrera, J. Kannala, and J. Heikkilä, "Joint depth and color camera calibration with distortion correction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2058–2064, 2012.

[41] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, IEEE Int. Workshop on*. IEEE, 2005, pp. 65–72.

[42] H. Knutsson and G. H. Granlund, *Signal processing for computer vision*. Springer, 1994.

[43] Y. Ohkita, Y. Ohishi, T. Furuya, and R. Ohbuchi, "Non-rigid 3d model retrieval using set of local statistical features," in *Multimedia and Expo Workshops (ICMEW), 2012 IEEE International Conference on*. IEEE, 2012, pp. 593–598.

[44] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning, and recognition," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 2649–2656.

[45] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*. IEEE, 2013, pp. 53–60.

[46] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *Consumer Depth Cameras for Computer Vision*. Springer, 2013, pp. 193–208.

[47] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human activity detection from rgbd images." *plan, activity, and intent recognition*, vol. 64, 2011.

[48] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database," in *Visual Media Production, 2009. CVMP'09. Conference for*. IEEE, 2009, pp. 159–168.

[49] Huawei/3DLife ACM Multimedia Grand Challenge for 2013: http://mmv.eecs.qmul.ac.uk/mmgc2013/.

[50] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions," *ACM Transactions on Graphics (TOG)*, vol. 21, no. 4, pp. 807–832, 2002.

[51] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, 2004, p. 22.

[52] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1948–1955.

[53] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 804–811.

[54] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 716–723.

[55] X. Cai, W. Zhou, L. Wu, J. Luo, and H. Li, "Effective active skeleton representation for low latency human action recognition," *IEEE Transactions on Multimedia*, vol. 18, no. 2, pp. 141–154, Feb 2016.

[56] E. Ohn-Bar and M. Trivedi, "Joint angles similarities and hog2 for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 465–470.

[57] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal quads: Human action recognition using joint quadruples," in *International Conference on Pattern Recognition*, 2014, pp. 4513–4518.

[58] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1110–1118.

[59] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, "Accurate 3d action recognition using learning on the grassmann manifold," *Pattern Recognition*, vol. 48, no. 2, pp. 556–567, 2015.

[60] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-d human action recognition by shape analysis of motion trajectories on riemannian manifold," *IEEE transactions on cybernetics*, vol. 45, no. 7, pp. 1340–1352, 2015.

[61] B. Galvin, B. McCane, K. Novins, D. Mason, S. Mills *et al.*, "Recovering motion fields: An evaluation of eight optical flow algorithms." in *BMVC*, vol. 98, 1998, pp. 195–204.

Georgios Th. Papadopoulos (S'08, M'11) was born in Thessaloniki, Greece, in 1982. He received the Diploma and Ph.D. degrees in Electrical and Computer Engineering from the Aristotle University of Thessaloniki (AUTH), Thessaloniki, Greece, in 2005 and 2011, respectively. He is currently a Postdoctoral Researcher at the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece. He has published 8 international journal articles and is a coauthor of 23 international conference proceedings. His research interests include computer vision, pattern recognition, semantic multimedia analysis, image and video processing, deep learning, context-based analysis and machine learning techniques.

Petros Daras is a Principal Researcher at the Information Technologies Institute (ITI) of the Centre for Research and Technology Hellas (CERTH). His main research interests include multimedia processing, multimedia & multimodal search engines, 3D reconstruction from multiple sensors, dynamic mesh coding, medical image processing and bioinformatics. He has co-authored more than 40 papers in refereed journals, 29 book chapters and more than 100 papers in international conferences. He has served as a regular reviewer for a number of international journals and conferences.