

Global Flow and Temporal-shape Descriptors for Human Action Recognition from 3D Reconstruction Data

Georgios Th. Papadopoulos and Petros Daras

Information Technologies Institute
Centre for Research and Technology Hellas
GR 57001 Thessaloniki, Thessaloniki, Greece

Abstract. In this paper, global-level view-invariant descriptors for human action recognition using 3D reconstruction data are proposed. 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 and the presence of (self-) occlusions. Initially, a set of calibrated Kinect sensors are employed for producing a 3D reconstruction of the performing subjects. Subsequently, a 3D flow field is estimated for every captured frame. For performing action recognition, a novel global 3D flow descriptor is introduced, which achieves to efficiently encode the global motion characteristics in a compact way, while also incorporating spatial distribution related information. Additionally, a new global temporal-shape descriptor that extends the notion of 3D shape descriptions for action recognition, by including temporal information, is also proposed. The latter descriptor efficiently addresses the inherent problems of temporal alignment and compact representation, while also being robust in the presence of noise. Experimental results using public datasets demonstrate the efficiency of the proposed approach.

Keywords: Action recognition, 3D reconstruction, 3D flow, 3D shape

1 Introduction

Efficiently and accurately recognizing human actions has emerged as one of the most challenging and active areas of research in the computer vision field over the past decades [1, 11, 20]. This is mainly due to the very wide set of possible applications 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. For achieving robust recognition performance, the typical requirements for rotation, translation and scale invariance need to be incorporated. Additional significant challenges need also to be sufficiently 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.

Two of the most significant challenges in 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, global-level view-invariant descriptors for human action recognition using 3D reconstruction data are proposed. 3D reconstruction techniques are employed in this work for addressing two of the most significant challenges in 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). In the first step, a 3D reconstruction of the performing subjects is generated using a set of calibrated Kinect sensors. Subsequently, a 3D flow field is estimated for every captured frame. A novel global 3D flow descriptor is proposed for performing action recognition. Among the advantages of this descriptor is that it efficiently encodes the global motion characteristics in a compact way, while also incorporating spatial distribution related information. Additionally, a new global temporal-shape descriptor that extends the notion of 3D shape descriptions for action recognition, by including temporal information, is also introduced. The latter descriptor efficiently addresses the challenging 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). Experimental results as well as comparative evaluation using datasets from the Huawei/3DLife 3D human reconstruction and action recognition Grand Challenge demonstrate the efficiency of the proposed approach.

The remainder of the paper is organized as follows: Previous work is reviewed in Section 2. Section 3 describes the 3D information processing. The descriptor extraction procedure is detailed in Section 4. Section 5 outlines the adopted action recognition scheme. Experimental results are presented in Section 6 and conclusions are drawn in Section 7.

2 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 over-

coming 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 is readily available by applying widely-used open-source software (e.g. OpenNI API¹, Kinect SDK², etc.). In [24], a depth similarity feature is proposed for describing the local 3D cuboid around a point of interest with an adaptable supporting size. Cheng et al. [4] introduce a descriptor of depth information, which depicts the structural relations of spatio-temporal points within action volumes, making use of the distance information in the depth data. Additionally, Wang et al. [21] introduce the so-called semi-local random occupancy pattern (ROP) features, which employ a sampling scheme that explores a large sampling space. In [22], an actionlet ensemble model is learnt to represent each action and to capture the intra-class variance. Moreover, Xia et al. [25] utilize histograms of 3D joint locations (HOJ3D) as a compact representation of human postures. The spherical angles between selected joints, along with the respective angular velocities, are calculated in [15].

2.1 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, 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. 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 [12, 17]. 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. [9] 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 [12], 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 [26], along with other descriptions (spatio-temporal interesting points, depth data, body posture). Gori et al. [8] 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 [16], a local-level 3D flow descriptor is introduced, which among others incorporates spatial and surface information in the flow representation and efficiently handles the problem of defining 3D orientation at every local

¹ <http://structure.io/openni>

² <http://www.microsoft.com/en-us/kinectforwindows/>

neighborhood. Furthermore, Fanello et al. [6] 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, most of them 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 information, view-invariance, introduction of a compact global representation, etc.

2.2 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. [23] 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 [7], 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. [5] 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. [10] present time-filtered and shape-flow descriptors for assessing the similarity of 3D video sequences of people with unknown temporal correspondence. In [2], 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. [27] 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 [18].

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.

3 3D information processing

The 3D information processing step is initiated by the application of a 3D reconstruction algorithm, which makes use of a set of calibrated Kinect sensors. Output of this algorithm is a uniform 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 each voxel corresponds to a cuboid region in the real 3D space with edge length equal to 10mm and t denotes the currently examined frame. It is considered that $v_t(x_g, y_g, z_g) = 1$ if $v_t(x_g, y_g, z_g)$ belongs to the subject's surface and $v_t(x_g, y_g, z_g) = 0$ otherwise.

For 3D flow estimation, an approach similar to the one described in [9] is followed, where pixel correspondences (obtained by the application of 2D flow estimation algorithms) are converted to voxel correspondences. Output of this procedure is the computation of a flow field $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ for every voxel grid VG_t .

4 Descriptor extraction

4.1 Global flow descriptor

For extracting a discriminative global 3D flow descriptor, the following challenges need to be addressed: a) the difficulty in introducing a consistent orientation definition for different action instances, in order to produce comparable low-level descriptions, and b) the incorporation of spatial distribution information in a compact way, while maintaining 3D rotation invariance.

The fundamental problem of orientation definition is addressed in this work 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. Subsequently, 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. 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 con-centric ring-shaped areas are defined, according to the following expressions:

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

where $\kappa \in [1, K]$, $\lambda \in [1, A]$, $\gamma = \zeta/A$ and $\delta = (y_{max}^c - y_{min}^c)/K$. For describing the flow information in every $B_{\kappa,\lambda}$ 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 $B_{\kappa,\lambda}$ 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:

$$\begin{aligned} \psi &= \tan^{-1}\left(\frac{z_g - z_{cg}}{x_g - x_{cg}}\right) - \tan^{-1}\left(\frac{\bar{\mathbf{F}}_{z,t}^{3D}(x_g, y_g, z_g)}{\bar{\mathbf{F}}_{x,t}^{3D}(x_g, y_g, z_g)}\right) \\ o &= \cos^{-1}\left(\frac{\langle(0, 1, 0), \bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)\rangle}{\|\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)\|}\right) \end{aligned} \quad (2)$$

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 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: $A = 4$, $K = 4$, $b_\psi = 6$ and $b_o = 3$. An example of ring-shaped $B_{\kappa,\lambda}$ areas formation for a ‘push away’ action instance is given in Fig. 1.

4.2 Global shape descriptor

As described in Section 2.2, 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,

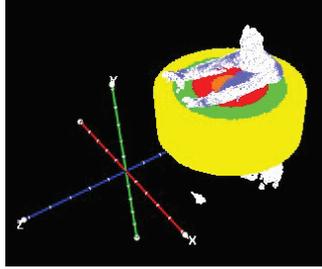


Fig. 1. Example of ring-shaped areas $B_{\kappa, \lambda}$ formation for $\kappa = 3$ and $\lambda \in [1, 4]$ for a ‘push-away’ action instance.

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, as it will be demonstrated in the experimental evaluation. Indicative examples of VG_t^{co} estimation for different human actions are depicted in Fig. 2.

For producing a compact temporal-shape descriptor, the descriptor vector sequence \mathbf{q}_t is initially adjusted to a predefined length H forming sequence $\bar{\mathbf{q}}_h$, 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 $\bar{\mathbf{q}}_{s,h}$ that are formed by considering the s -th element of $\bar{\mathbf{q}}_h$ each time. For frequency domain analysis, the Discrete Cosine Transform (DCT) is applied to $\bar{\mathbf{q}}_{s,h}$, as follows:

$$fc_s(\beta) = \sum_{h=1}^H \bar{\mathbf{q}}_{s,h} \cos \frac{\pi}{H} [(h-1) + \frac{1}{2}(\beta-1)] \quad (3)$$

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

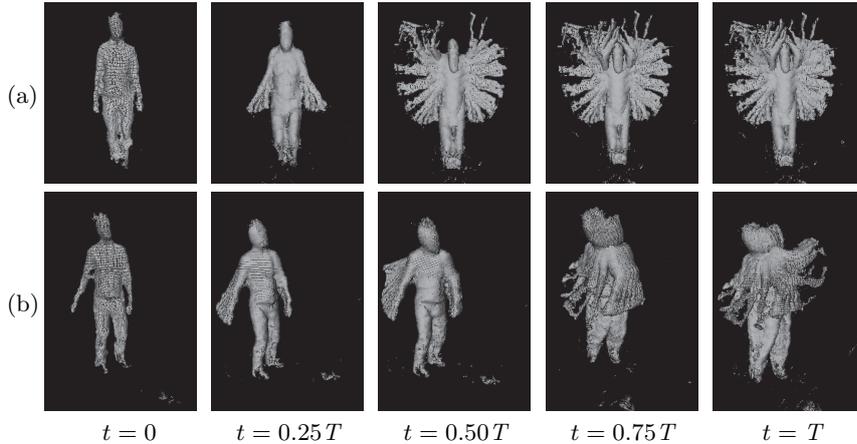


Fig. 2. Indicative examples of composite voxel grid VG_t^{co} estimation for actions: (a) jumping-jacks and (b) tennis-forehand. T denotes the overall duration of the action.

not lead to increased performance compared with the one received 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 $\bar{\mathbf{q}}_{s,h}$ 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 $\bar{\mathbf{q}}_{s,h}$ sequences during the descriptor extraction procedure led to inferior recognition performance, mainly due to overfitting occurrences.

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 [14] (3D distance histogram) was utilized; this was experimentally shown to lead to better overall action recognition performance than other common shape descriptors. In [10], description and comparative evaluation of different static 3D shape descriptors for action recognition are given.

5 Action recognition

After extracting the global 3D flow/shape descriptors for every examined human action (as detailed in Section 4), each descriptor is L1-normalized for incorporating invariance with respect to the execution speed of different instances of the same action from the same or different subjects. Action recognition is then realized using multi-class Support Vector Machines (SVMs).

6 Experimental results

In this section, experimental results from the application of the proposed approach to the Huawei/3DLife³ datasets for 3D human reconstruction and action recognition, which were used in the ACM Multimedia 2013 ‘Multimedia Grand Challenge’ and are among the most comprehensive and broad ones in the literature, are presented. In particular, the first (dataset D_1) and the second (dataset D_2) sessions of the first dataset are used, which provide RGB-plus-depth video streams from five and two Kinect sensors, respectively. For dataset D_2 , the data stream from only the frontal Kinect was utilized. D_1 and D_2 include captures of 17 and 14 human subjects, respectively, and each action is performed at least 5 times by every individual. Out of the available 22 supported actions, the following set of 17 dynamic ones were considered for the experimental evaluation: $E = \{e_g, g \in [1, G]\} \equiv \{\text{Hand waving, Knocking the door, Clapping, Throwing, Punching, Push away, Jumping jacks, Lunges, Squats, Punching and kicking, Weight lifting, Golf drive, Golf chip, Golf putt, Tennis forehand, Tennis backhand, Walking on the treadmill}\}$. 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. Performance evaluation was realized following the ‘leave-one-out’ methodology, where in every iteration one subject was used for performance measurement and the remaining ones were used for training.

In Fig. 3, 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 4.2) 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 esti-

³ <http://mmv.eecs.qmul.ac.uk/mmgc2013/>

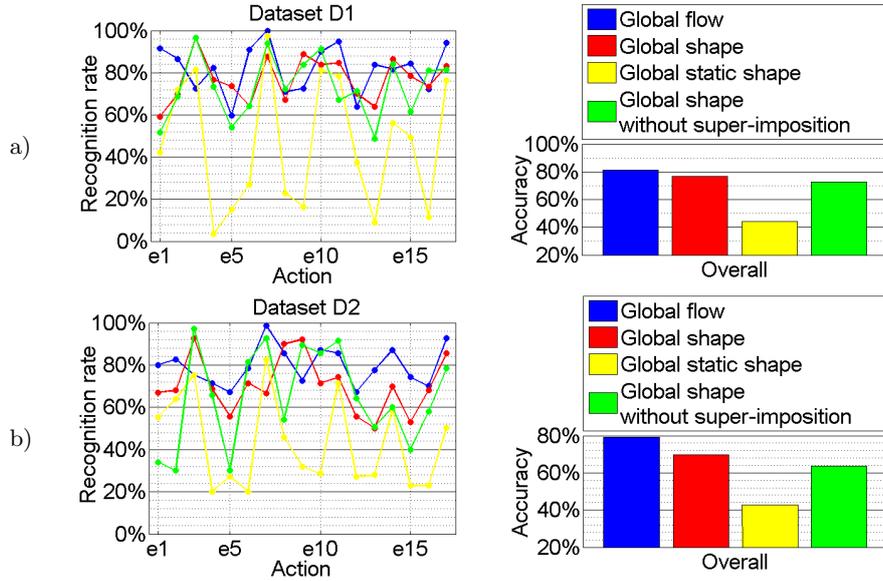


Fig. 3. Action recognition results for a) D_1 and b) D_2 datasets.

inating a 3D shape descriptor of the generated volume (like in [7] [23]). 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. From the results presented in Fig. 3, 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.

6.1 Parameter selection

In order to apply and evaluate the performance of the proposed 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 in the followings experimental results are given only for D_1 , while similar behavior of the proposed descriptors has been observed in D_2 . In particular, the descriptor behavior for different values of the following parameters, along with justification where particular values were selected, is investigated:

Table 1. Global flow descriptor parameter selection

Parameters	Accuracy
$K=3, \Lambda=3, b_\psi=6, b_o=3$	75.44%
$K=4, \Lambda=4, b_\psi=6, b_o=3$	81.27%
$K=5, \Lambda=5, b_\psi=6, b_o=3$	80.88%
$K=4, \Lambda=4, b_\psi=6, b_o=3$	81.27%
$K=4, \Lambda=4, b_\psi=4, b_o=3$	79.97%
$K=4, \Lambda=4, b_\psi=6, b_o=3$	81.27%
$K=4, \Lambda=4, b_\psi=6, b_o=6$	77.22%

- 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 4.1), respectively. Additionally, b_ψ and b_o define the number of bins of the histograms calculated with respect to angles ψ and o (Section 4.1), respectively. In Table 1, 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 H : This adjusts the length of the shape descriptor vector sequence $\bar{\mathbf{q}}_h$ (Section 4.2). 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 4.2). The performance obtained by the application of the proposed temporal-shape descriptor for different values of P is given for both datasets in Table 2. From the presented results, it can be seen that the best performance is achieved when only relatively few frequency coefficients are used; 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.

6.2 Comparative evaluation

Comparative evaluation results of the proposed descriptors (and their combination) with similar literature approaches are reported in this section. In particular,

Table 2. Temporal-shape descriptor parameter selection

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

in Fig. 4, quantitative results in terms of the estimated recognition rates and overall accuracy are given for the following cases: i) The HOF3D descriptor with ‘vertical rotation’ [9], which is a local histogram-based 3D flow descriptor that does not incorporate spatial information. ii) The local 3D flow descriptor of [16], which is again a local histogram-based 3D flow descriptor; however, it incorporates spatial and surface information in the flow representation. iii) The proposed global flow descriptor. iv) The LC-LSF local 3D shape descriptor of [13], which employs a set of local statistical features for describing non-rigid 3D models. v) The global 3D temporal-shape descriptor of [10], where 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. vi) The proposed global shape descriptor. vii) The overall proposed approach, which combines the proposed global flow and shape descriptors, by means of simple concatenation in a single feature vector. viii) The skeleton-tracking-based methods of [15] [19] [3], which estimate human posture representations at every frame, making use of the detected human joints (for the methods of [19] and [3], only the reported overall classification accuracy in D_2 is provided in Fig. 4).

From the presented results, it can be seen that the proposed global flow descriptor performs significantly better than the local flow ones of [9] and [16]. This is mainly due to the inefficiency of the local descriptors in fully addressing the problem of defining a consistent orientation, during the analysis in local 3D neighborhoods for descriptor extraction. The aforementioned difference in performance is more pronounced and clear in D_1 , i.e. the most challenging dataset due to the relatively increased presence of noise in the provided depth maps, which is in turn mainly due to the increased interference among the higher number of Kinect sensors used in D_1 . Similar observations can be made for the case of the proposed global shape descriptor, which performs equally or significantly better than the local 3D shape descriptor of [13] for the same reasons described above. It worths noticing that the proposed global shape descriptor outperforms the local flow ones by a large margin, despite the fact that 3D flow is a more discriminative information source. Additionally, the proposed temporal-shape descriptor is also shown to outperform the temporal-shape method of [10]. 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. Moreover, the overall proposed approach (which consists of simple concatenation of the

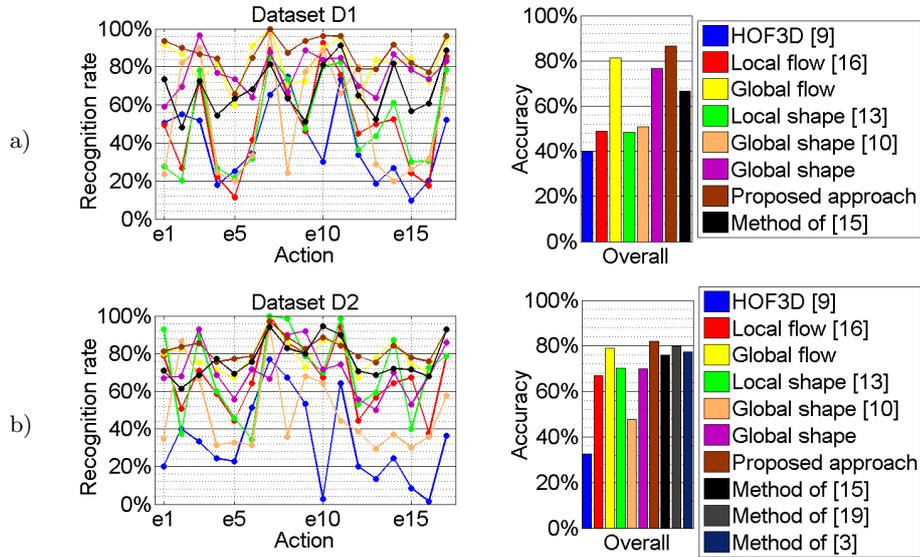


Fig. 4. Comparative evaluation results for a) D_1 and b) D_2 datasets.

proposed global flow and shape descriptors in a single feature vector) achieves increased performance, compared with the cases of using each individual descriptor alone. The latter demonstrates the complementarity of the introduced descriptors and dictates that a robust human action recognition framework should incorporate multiple information sources for accomplishing increased recognition performance. Furthermore, it is shown that the proposed approach outperforms the skeleton-tracking-based methods of [15] [19] [3]. This difference in performance demonstrates the superiority of the proposed method in capturing the action dynamics more efficiently, mainly due to the combination of multiple and complementary information sources (i.e. flow/shape information). On the other hand, the methods of [15] [19] [3] suffer from the limitations of the employed skeleton tracker, which is reported not to perform well in case of fast movements and in the presence of noise in the captured depth maps. Another interesting observation is that the performance difference between the proposed approach and the method of [15] is higher in D_1 (i.e. the dataset with increased presence of noise) than in D_2 ; this again highlights the robustness in the presence of noise of the proposed method, while the method of [15] is significantly affected by the limitations of the employed skeleton tracker (as described above). It needs to be highlighted that an additional advantageous characteristic of the proposed approach is that it does not make the assumption of human(s) being present in the scene, i.e. it does not use domain-specific knowledge and it can be applied with any other type of object being captured.

7 Conclusions

In this work, the problem of human action recognition using 3D reconstruction data was examined and novel global 3D flow/shape descriptors were introduced. Exploitation of 3D reconstruction techniques facilitates towards addressing two of the most challenging issues in human action recognition, namely view-variance and the presence of (self-) occlusions. This choice is further endorsed by the recent introduction of low-cost, portable, high-quality and accurate motion capturing devices. The proposed global 3D flow descriptor efficiently encodes the global motion characteristics in a compact way. It was observed that this descriptor led to the best action recognition results. Additionally, the proposed global temporal-shape descriptor efficiently addresses the inherent problems of temporal alignment and compact representation. The proposed descriptors were experimentally shown to outperform similar methods of the literature, using publicly available datasets for the evaluation. Moreover, the comparative evaluation of the overall proposed approach (concatenation of the introduced global 3D flow/shape descriptors) was shown to outperform action recognition methods that rely on human skeleton-tracking methodologies.

Acknowledgment

The work presented in this paper was supported by the European Commission under contract H2020-700367 DANTE.

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*, 23(11):1993–2008, 2013.
2. C. Budd, P. Huang, M. Kludiny, and A. Hilton. Global non-rigid alignment of surface sequences. *International Journal of Computer Vision*, 102(1-3):256–270, 2013.
3. 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*, 18(2):141–154, Feb 2016.
4. Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian. Human daily action analysis with multi-view and color-depth data. In *Computer Vision–ECCV 2012. Workshops and Demonstrations*, pages 52–61. Springer, 2012.
5. 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*, pages 726–733. IEEE, 2003.
6. 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*, 14(1):2617–2640, 2013.
7. L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(12):2247–2253, 2007.

8. 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*, pages 126–131. IEEE, 2013.
9. 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*, 6(5):553–565, 2012.
10. P. Huang, A. Hilton, and J. Starck. Shape similarity for 3d video sequences of people. *International Journal of Computer Vision*, 89(2-3):362–381, 2010.
11. X. Ji and H. Liu. Advances in view-invariant human motion analysis: a review. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 40(1):13–24, 2010.
12. M. Munaro, G. Ballin, S. Michieletto, and E. Menegatti. 3d flow estimation for human action recognition from colored point clouds. *Biologically Inspired Cognitive Architectures*, 5:42–51, 2013.
13. 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*, pages 593–598. IEEE, 2012.
14. R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin. Shape distributions. *ACM Transactions on Graphics (TOG)*, 21(4):807–832, 2002.
15. 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*, pages 473–483, 2014.
16. G. T. Papadopoulos and P. Daras. Local descriptions for human action recognition from 3d reconstruction data. In *IEEE International Conference on Image Processing (ICIP '14)*, pages 2814–2818, Nov. 2014.
17. M. Sizintsev and R. P. Wildes. Spatiotemporal stereo and scene flow via stequel matching. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(6):1206–1219, 2012.
18. R. Slama, H. Wannous, and M. Daoudi. 3d human motion analysis framework for shape similarity and retrieval. *Image and Vision Computing*, 32(2):131–154, 2014.
19. L. Sun and K. Aizawa. Action recognition using invariant features under unexampled viewing conditions. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 389–392. ACM, 2013.
20. P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, 2008.
21. J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3d action recognition with random occupancy patterns. In *Computer Vision–ECCV 2012*, pages 872–885. Springer, 2012.
22. 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*, pages 1290–1297. IEEE, 2012.
23. D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, 104(2):249–257, 2006.
24. 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*, pages 2834–2841. IEEE, 2013.

25. 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*, pages 20–27. IEEE, 2012.
26. L. Xia, I. Gori, J. Aggarwal, and M. Ryoo. Robot-centric activity recognition from first-person rgb-d videos. 2015.
27. T. Yamasaki and K. Aizawa. Motion segmentation and retrieval for 3d video based on modified shape distribution. *EURASIP Journal on Applied Signal Processing*, 2007(1):211–211, 2007.