LOCAL DESCRIPTIONS FOR HUMAN ACTION RECOGNITION FROM 3D RECONSTRUCTION DATA

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

Information Technologies Institute, Centre for Research and Technology Hellas, Greece

ABSTRACT

In this paper, a view-invariant approach to human action recognition using 3D reconstruction data is proposed. 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, the 'Bag-of-Words' methodology is followed, where Spatio-Temporal Interest Points (STIPs) are detected in the 4D space (xyz-coordinates plus time). A novel 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 STIP location. Additionally, typical 3D shape descriptors of the literature are used for producing a more complete representation. 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.

Index Terms— Action recognition, view-invariance, 3D reconstruction, Kinect, 3D flow

1. INTRODUCTION

Over the past decades, human action recognition has received particular attention and has emerged as one of the most active topics in the computer vision research community [1, 2, 3]. This is mainly due to the very wide set of potential application fields that can benefit from the resulting accomplishments, such as surveillance, security, human computer interaction, smart houses, helping the elderly/disabled, to name a few. For achieving robust recognition results the typical requirements for rotation, translation and scale invariance need to be incorporated. Additional challenges that need to be efficiently addressed constitute the differences in the appearance of the subjects, the human silhouette features, the execution of the same actions, etc. Although multiple research groups focus on this topic and numerous approaches have already been presented, significant obstacles towards fully addressing the problem in the general case are still present.

Action recognition approaches can be roughly divided into the following three categories [4], irrespectively of the data that they receive as input (i.e. single-camera videos, multi-view video sequences, depth maps, 3D reconstruction data, etc.): spatio-temporal shape- [5, 6, 7], tracking- [8, 9, 10, 11, 12] and Space-Time Interest Point (STIP)-based [13, 14, 15]. Spatio-temporal shape 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. On the other hand, STIP-based methods perform analysis at the local-level. Although they typically exhibit increased computational complexity for reaching satisfactory recognition performance, they are robust to noise and they are shown to satisfactorily handle self-occlusion occurrences.

In this paper, a view-invariant approach to human action recognition using 3D reconstruction data is presented. A 3D reconstruction of the performing subjects is initially generated using a set of calibrated Kinect sensors, addressing in this way the inherent problems of view-variance and (self-)occlusions. Then, a 3D flow field is estimated for every captured frame, by appropriately combining the output of a 2D optical flow algorithm that is applied to the RGB stream of every employed Kinect. For realizing action recognition, the 'Bag-of-Words' methodology is adopted, where STIPs are identified using a 4D (xyz-coordinates plus time) detector. A novel local-level 3D flow descriptor is introduced for describing the 3D motion information at every STIP location. Among the advantages of the proposed descriptor is that it incorporates spatial and surface information in the flow representation and efficiently handles the problem of defining 3D orientation at every STIP position. Additionally, common 3D shape descriptors of the literature are used for producing a more complete representation. 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 paper is organized as follows: Section 2 describes the 3D information processing. The proposed local-level descriptor extraction procedure is detailed in Section 3. Section 4 outlines the adopted action recognition scheme. Experimental results are presented in Section 5 and conclusions are drawn in Section 6.

2. 3D INFORMATION PROCESSING

2.1. 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 volumetric 3D reconstruction algorithm of [16], which makes use of a set of calibrated Kinect sensors, is utilized for generating a 3D point-cloud of the performing subjects. 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 implementa-

The work presented in this paper was supported by the European Commission under contract FP7-601170 RePlay.

tion a uniform voxel grid is utilized, where each voxel corresponds to a cuboid region in the real 3D space with edge length equal to 10mm. 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.

2.2. Flow Estimation

The potential of exploiting 3D flow information for human action recognition, like e.g. in [13, 17], has not been extensively investigated, mainly due to the multiple challenges and the increased computational complexity that need to be tackled for achieving good flow estimation results.

In this work, a gradual approach is proposed for 3D flow estimation. 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 [18] was selected using the implementation provided by [19], since it was experimentally shown to produce satisfactory results [19]. 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 the mappings from the (x_{rgb}, y_{rgb}) and (x_d, y_d) spaces to the (x_l, y_l, z_l) one were estimated following a perspective projection modeling. Additionally, $\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)$ 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). For tackling the noise caused by the Kinect, 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:

$$r_{c,t}^{2D}(x_d, y_d) = \frac{\sum_{x'_d=x_d-Q}^{x_d+Q} \sum_{y'_d=y_d-Q}^{y_d+Q} b(x'_d, y'_d)}{(2Q+1)^2} \in [0, 1] , \quad (1)$$

where $b(x'_d, y'_d) = 0$ 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 $D_{c,t}^{2D}(x_d, y_d)$.

For computing a 3D flow field $\mathbf{F}_t^{3D}(x_g, y_g, z_g)$ in VG_t, 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:

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

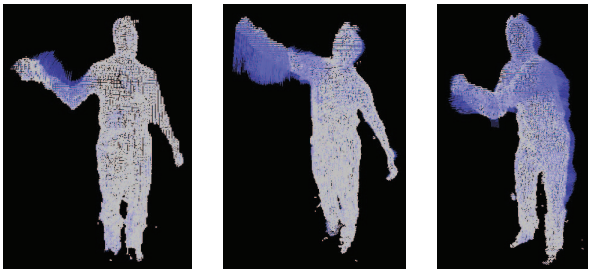
where $\mathbf{S}$ comprises the points in $W_{c,t}^{3D}(x_l, y_l, z_l)$ that correspond to voxel $v_t(x_g, y_g, z_g)$ and for which flow vectors $\mathbf{f}_{c,t}^{3D}(x_l, y_l, z_l)$ have been calculated, M is the number of points in $\mathbf{S}$ and $\Psi[.]$ denotes the extrinsic calibration-based transformation from the $W_{c,t}^{3D}(x_l, y_l, z_l)$ to the (x_g, y_g, z_g) space. 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 Kinects, the following reliability value is estimated for each voxel $v_t(x_g, y_g, z_g)$ that is visible from every Kinect c:

$$a_{c,t}^{3D}(x_g, y_g, z_g) = \langle \mathbf{m}_c(x_g, y_g, z_g), \mathbf{n}_t(x_g, y_g, z_g) \rangle \in [0, 1] , \quad (3)$$

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 and $\mathbf{n}_t(x_g, y_g, z_g)$ is the unit normal vector to the 3D reconstructed surface at voxel $v_t(x_g, y_g, z_g)$. Subsequently, $\mathbf{F}_t^{3D}(x_g, y_g, z_g)$ is computed, as follows:

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

where $\mathbf{U}$ comprises the Kinects from which $v_t(x_g, y_g, z_g)$ is visible and L their number. 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. 1.



(a) (b) (c)

Fig. 1. Indicative 3D flow field $\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)$ estimation examples for actions: (a) hand-waving, (b) throwing and (c) golf-chip.

3. DESCRIPTOR EXTRACTION

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 [20] 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 [21] 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.

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.

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. 2:

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

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

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

where $i \in [-I, I]$ and $j \in [1, J]$ are odd numbers denoting 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:

$$\vartheta = \arccos\left(\frac{\langle \mathbf{n}_t(x_g, y_g, z_g), \bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g) \rangle}{\|\mathbf{n}_t(x_g, y_g, z_g)\| \cdot \|\bar{\mathbf{F}}_t^{3D}(x_g, y_g, z_g)\|}\right) \in [0, \pi] , \quad (7)$$

where $\|.\|$ denotes the norm of a vector. $\mathbf{n}_t(x_g, y_g, z_g)$ 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.

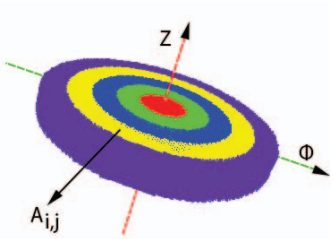


Fig. 2. Example of ring-shaped areas $A_{i,j}$ formation for $i = 0$ and $J = 5$ in the defined cylindrical coordinate system.

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 [22], 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.

4. ACTION RECOGNITION

After estimating a set of STIPs for every examined human action and subsequently extracting local-level 3D flow and shape descriptions at every STIP location (as described in Section 3), every action is represented with a single vector. For constructing the aforementioned vector, the 'Bag-of-Words' (BoW) methodology [23] is followed, where every action is represented by a L1-normalized histogram of 1000 words. Then, action recognition is realized using multi-class Support Vector Machines (SVMs).

5. EXPERIMENTAL RESULTS

In this section, experimental results from the application of the proposed approach to the Huawei/3DLife 3D human reconstruction and action recognition Grand Challenge datasets 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, which was used mainly for comparative evaluation purposes, 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 16 dynamic ones were considered for the experimental evaluation: $E = \{e_\lambda, \lambda \in [1, \Lambda]\}$ ≡{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, Ten-

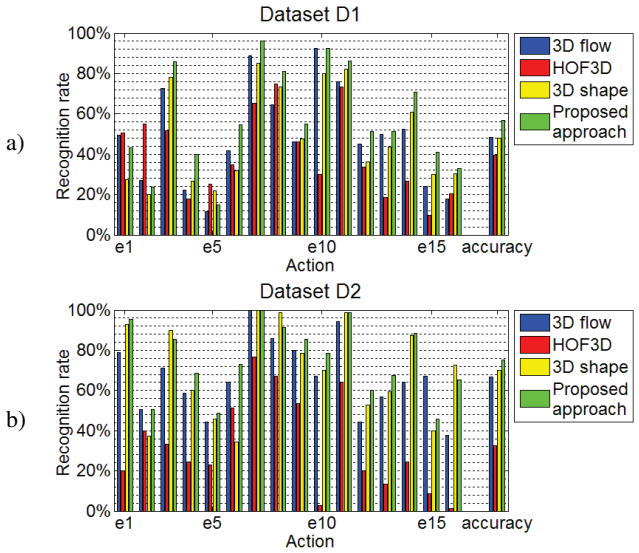


Fig. 3. Action recognition results for a) D_1 and b) D_2 datasets.

nis forehand, Tennis backhand}. Action 'Walking on the treadmill', although dynamic, was not included in the evaluation, due to the respective instances exhibiting significantly increased duration that would lead to a correspondingly increased computational time for processing them. 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 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 only flow, only shape and both flow and shape information is used. Additionally, the value of the overall classification accuracy, i.e. the percentage of all 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 48.50% and 66.67% in D_1 and D_2, respectively). Examining the results in details, 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). Additionally, 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. However, the combination of flow and shape information leads to improved recognition performance in both datasets (overall accuracy equal to 56.97% and 75.31% in D_1 and D_2, respectively), compared with the cases of using each of them alone; hence, demonstrating the complementary nature of the utilized descriptors.

The proposed 3D flow descriptor is comparatively evaluated with a similar approach of the literature, namely the HOF3D descriptor with 'vertical rotation' presented in [13]. 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; hence, ignoring both the spatial distribution of the 3D flow vectors and 3D surface information. From the results presented in Fig. 3, it can be seen that the proposed descriptor leads to significantly increased performance compared with HOF3D in both datasets. The latter verifies the capabilities of the proposed descriptor in efficiently defining an appropriate local-coordinate system and also encoding spatial distribution/surface-related information, as detailed in Section 3. It must be noted that the global 3D flow descriptor of [17] (mentioned in Section 2.2) was not included in the conducted comparative evaluation. This is due to the descriptor of [17] 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 with the view-invariant HOF3D and the proposed descriptor would not be fair.

5.1. Discussion

Dataset D_2 was included in the experiments mainly for comparing with the skeleton-tracking-based methods of [12] and [24] that have reported action recognition results for this dataset. More specifically, the authors' previous work [12] claimed accuracy equal to 76.03% using only the frontal Kinect (i.e. the same D_2 dataset in this work), while Sun and Aizawa [24] reported accuracy equal to 79.78% using both available Kinects. In other words, the proposed approach (with accuracy equal to 75.31% in D_2), achieved comparable performance with [12] and inferior compared with [24]. However, the following two important facts hold: a) The methods of [12] and [24] include in their evaluation the 'Walking on the treadmill' action, which is an easily recognizable action that increases the overall performance. Additionally, it is not clear if the work of [24] makes use of information from both available Kinects during training (i.e. being favored compared with the work of [12] and the proposed approach, due to using a training set of double size). b) both [12] and [24] methods are not in principle view-invariant (despite using depth information) and they extensively exploit domain specific knowledge. This is due to the human skeleton-tracking algorithm that they employ and which sets particular restrictions in the allowed poses of the captured subject to perform efficiently human calibration/adaptation/skeleton-tracking. On the contrary, the proposed approach is fully view-invariant, while it does not make the assumption of human(s) being present in the scene (i.e. it can be applied with any other type of object being captured). Moreover, the reported difference in performance is expected to be surpassed by extending the proposed 3D flow representation, in order to incorporate a global-level description of the 3D flow and its spatial distribution, using all detected STIPs.

6. CONCLUSIONS

In this paper, a view-invariant approach to human action recognition using 3D reconstruction data was presented and comparatively evaluated using two publicly available datasets. Future work includes the extension of the proposed 3D flow representation to include global-level distribution information and the investigation of incorporating 'discriminative' STIPs for local features extraction.

7. REFERENCES

[1] Xiaofei Ji and Honghai Liu, "Advances in view-invariant human motion analysis: a review," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Trans. on*, vol. 40, no. 1, pp. 13–24, 2010.

[2] Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, 2008.

[3] Paulo Vinicius Koerich Borges, Nicola Conci, and Andrea 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.

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

[5] Pavan Turaga, Ashok Veeraraghavan, and Rama Chellappa, "Statistical analysis on stiefel and grassmann manifolds with applications in computer vision," in *Computer Vision and Pattern Recognition, IEEE Conf. on*. IEEE, 2008, pp. 1–8.

[6] Daniel Weinland, Remi Ronfard, and Edmond Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[7] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 12, pp. 2247–2253, 2007.

[8] Junxia Gu, Xiaoqing Ding, Shengjin Wang, and Youshou 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.

[9] Bi Song, Ahmed T Kamal, Cristian Soto, Chong Ding, Jay A Farrell, and Amit K Roy-Chowdhury, "Tracking and activity recognition through consensus in distributed camera networks," *Image Processing, IEEE Transactions on*, vol. 19, no. 10, pp. 2564–2579, 2010.

[10] Imran N Junejo, Emilie Dexter, Ivan Laptev, and Patrick Pérez, "View-independent action recognition from temporal self-similarities," *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, vol. 33, no. 1, pp. 172–185, 2011.

[11] G. Th. 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.

[12] Georgios Th. Papadopoulos, Apostolos Axenopoulos, and Petros Daras, "Real-time skeleton-tracking-based human action recognition using kinect data," in *MultiMedia Modeling, Int. Conf. on*, 2014, pp. 473–483.

[13] Michael Boelstoft Holte, Bhaskar Chakraborty, Jordi Gonzalez, and Thomas 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.

[14] A. Haq, I. Gondal, and M. Murshed, "On temporal order invariance for view-invariant action recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 2, pp. 203–211, 2013.

[15] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra, "Effective codebooks for human action representation and classification in unconstrained videos," *Multimedia, IEEE Transactions on*, vol. 14, no. 4, pp. 1234–1245, 2012.

[16] Dimitrios S Alexiadis, Dimitrios Zarpalas, and Petros 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.

[17] Matteo Munaro, Gioia Ballin, Stefano Michieletto, and Emanuele Menegatti, "3d flow estimation for human action recognition from colored point clouds," *Biologically Inspired Cognitive Architectures*, 2013.

[18] Marc Proesmans, Luc Van Gool, Eric Pauwels, and André Oosterlinck, "Determination of optical flow and its discontinuities using non-linear diffusion," in *Computer VisionECCV'94*, pp. 294–304. Springer, 1994.

[19] Marco Mammarella, Giampiero Campa, Mario L Fravolini, and Marcello 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.

[20] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge 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.

[21] Hans Knutsson and Gösta H Granlund, *Signal processing for computer vision*, Springer, 1994.

[22] Yuki Ohkita, Yuya Ohishi, Takahiko Furuya, and Ryutarou 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.

[23] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, 2004, vol. 1, p. 22.

[24] Litian Sun and Kiyoharu 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.