Real-time Skeleton-tracking-based Human Action Recognition Using Kinect Data

Georgios Th. Papadopoulos, Apostolos Axenopoulos and Petros Daras

Information Technologies Institute
Centre for Research & Technology - Hellas
Thessaloniki, Greece

Abstract. In this paper, a real-time tracking-based approach to human action recognition is proposed. The method receives as input depth map data streams from a single kinect sensor. Initially, a skeleton-tracking algorithm is applied. Then, a new action representation is introduced, which is based on the calculation of spherical angles between selected joints and the respective angular velocities. For invariance incorporation, a pose estimation step is applied and all features are extracted according to a continuously updated torso-centered coordinate system; this is different from the usual practice of using common normalization operators. Additionally, the approach includes a motion energy-based methodology for applying horizontal symmetry. Finally, action recognition is realized using Hidden Markov Models (HMMs). Experimental results using the Huawei/3DLife 3D human reconstruction and action recognition Grand Challenge dataset demonstrate the efficiency of the proposed approach.

Keywords: Action recognition, skeleton-tracking, action representation, depth map, kinect

1 Introduction

Action recognition constitutes a widely studied field and a very active topic in the computer vision research community [7]. This is due to the wide set of potential fields where the research outcomes can be commercially applied, such as surveillance, security, human computer interaction, smart houses, helping the elderly/disabled, to name a few. In order to develop a robust action recognition system, the respective algorithm needs to efficiently handle the differences in the appearance of the subjects, the human silhouette features, the execution of the same actions, etc. Additionally, the system should incorporate the typical rotation, translation and scale invariances. Despite the fact that multiple research groups focus on this topic and numerous approaches have already been presented, significant challenges towards fully addressing the problem in the general case are still present.

Action recognition approaches can be roughly divided into the following three categories [10], irrespectively of the data that they receive as input (i.e.

single-camera videos, multi-view video sequences, depth maps, 3D reconstruction data, etc.): Space-Time Interest Point (STIP)- [6][5][2], spatio-temporal shape- [13][14][3] and tracking-based [4][12][8][9]. STIP-based methods perform analysis at the local-level; however, they typically exhibit increased computational complexity for reaching satisfactory recognition performance. 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. On the other hand, the performance of tracking-based approaches, which rely on the tracking of particular features or specific human body parts in subsequent frames (including optical-flow-based methods) depends heavily on the efficiency of the employed tracker. Nevertheless, the advantage of the latter category of methods is that they can allow the real-time recognition of human actions.

In this paper, a real-time tracking-based approach to human action recognition is proposed. The method receives as input a sequence of depth maps captured from a single kinect sensor, in order to efficiently capture the human body movements in the 3D space. Subsequently, a skeleton-tracking algorithm is applied, which iteratively detects the position of 15 joints of the human body in every captured depth map. Then, a new action representation is introduced, which is based on the calculation of spherical angles between selected joints and the respective angular velocities, for satisfactorily handling the differences in the appearance and/or execution of the same actions among the individuals. For incorporating further invariance to appearance, scale, rotation and translation, a pose estimation step is applied prior to the feature extraction procedure and all features are calculated according to a continuously updated torso-centered coordinate system; this is different from the usual practice of using normalization operators during the analysis process [4]. Additionally, the approach incorporates a motion energy-based methodology for applying horizontal symmetry and hence efficiently handling left- and right-body part executions of the exact same action. Finally, action recognition is realized using Hidden Markov Models (HMMs). Experimental results using the Huawei/3DLife 3D human reconstruction and action recognition Grand Challenge dataset[1] demonstrate the efficiency of the proposed approach.

The paper is organized as follows: Section 2 outlines the employed skeleton-tracking algorithm. The proposed action recognition approach is described in Section 3. Experimental results are presented in Section 4 and conclusions are drawn in Section 5.

2 Skeleton-tracking

Prior to the application of the proposed action recognition approach, the depth maps captured by the kinect sensor are processed by a skeleton-tracking algorithm. The depth maps of the utilized dataset were acquired using the OpenNI

[1] http://mmv.eecs.qmul.ac.uk/mmgc2013/

API[2]. To this end, the OpenNI high-level skeleton-tracking module is also used for detecting the performing subject and tracking a set of joints of his/her body. More specifically, the OpenNI tracker detects the position of the following set of joints in the $3D$ space $G = \{g_i,\ i \in [1, I]\} \equiv \{Torso,\ Neck,\ Head,\ Left\ shoulder,\ Left\ elbow,\ Left\ wrist,\ Right\ shoulder,\ Right\ elbow,\ Right\ wrist,\ Left\ hip,\ Left\ knee,\ Left\ foot,\ Right\ hip,\ Right\ knee,\ Right\ foot\}$. The position of joint g_i is implied by vector $\mathbf{p}_i(t) = [x\ y\ z]^T$, where t denotes the frame for which the joint position is located and the origin of the orthogonal XYZ co-ordinate system is placed at the center of the kinect sensor. An indicative example of a captured depth map and the tracked joints is given in Fig. 1.

The OpenNI skeleton-tracking module requires user calibration in order to estimate several body characteristics of the subject. In recent versions of OpenNI, the 'auto-calibration' mode enables user calibration without requiring the subject to undergo any particular calibration pose. Since no calibration pose was captured for the employed dataset, the OpenNI's (v. 1.5.2.23) 'auto-calibration' mode is used in this work. The experimental evaluation showed that the employed skeleton-tracking algorithm is relatively robust for the utilized dataset. In particular, the position of the joints is usually detected accurately, although there are some cases where the tracking is not correct. Characteristic examples of the latter are the inaccurate detection of the joint positions when very sudden and intense movements occur (e.g. arm movements when performing actions like 'punching') or when self-occlusions are present (e.g. occlusion of the knees when extensive body movements are observed during actions like 'golf drive').

3 Action recognition

In this section, the proposed skeleton-tracking-based action recognition approach is detailed. The developed method satisfies the following two fundamental principles: a) the computational complexity needs to be relatively low, so that the real-time processing nature of the algorithm to be maintained, and b) the dimensionality of the estimated action representation needs also to be low, which is a requirement for efficient HMM-based analysis [11].

3.1 Pose estimation

The first step in the proposed analysis process constitutes a pose estimation procedure. This is performed for rendering the proposed approach invariant to differentiations in appearance, body silhouette and action execution among different subjects, apart from the typically required invariances to scale, translation and rotation. The proposed methodology is different from the commonly adopted normalization procedures (e.g. [4]), which in their effort to incorporate invariance characteristics they are inevitably led to some kind of information loss. In particular, the aim of this step is to estimate a continuously updated

[2] http://www.openni.org/

(a)

(b)

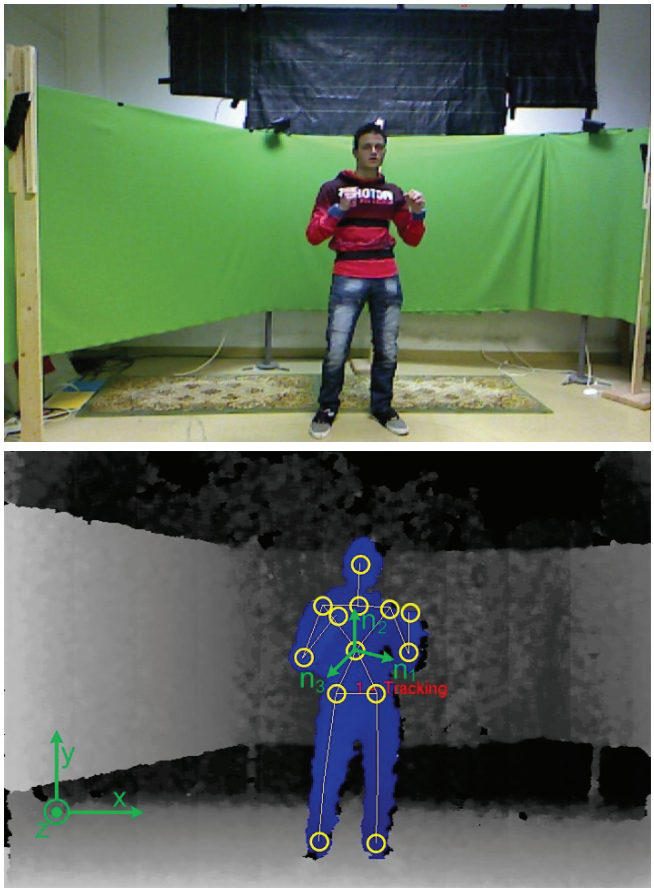

Fig. 1. Indicative pose estimation example: (a) examined frame and (b) captured depth map along with the tracked joints and the estimated orthogonal pose vectors.

orthogonal basis of vectors for every frame t that represents the subject's pose. The calculation of the latter is based on the fundamental consideration that the orientation of the subject's torso is the most characteristic quantity of the subject during the execution of any action and for that reason it could be used as reference. For pose estimation, the position of the following three joints is taken into account: *Left shoulder*, *Right shoulder* and *Right hip*. These comprise joints around the torso area, whose relative position remains almost unchanged during the execution of any action. The motivation behind the consideration of the three aforementioned joints, instead of directly estimating the position of the torso joint and the respective normal vector, is to reach a more accurate estimation of the subject's pose. It must be noted that the *Right hip* joint was preferred instead of the obvious *Torso* joint selection. This was performed so that the orthogonal basis of vectors to be estimated from joints with bigger in

between distances that will be more likely to lead to more accurate pose estimation. However, no significant deviation in action recognition performance was observed when the $Torso$ joint was used instead. In this work, the subject's pose comprises the following three orthogonal vectors $\{\mathbf{n}_1, \mathbf{n}_2, \mathbf{n}_3\}$ that are calculated as follows:

$$\mathbf{n}_1 = \frac{\mathbf{p}_7 - \mathbf{p}_4}{\|\mathbf{p}_7 - \mathbf{p}_4\|}, \ \mathbf{u} = \frac{\mathbf{p}_7 - \mathbf{p}_{13}}{\|\mathbf{p}_7 - \mathbf{p}_{13}\|}$$
$$\mathbf{n}_3 = \frac{\mathbf{n}_1 \times \mathbf{u}}{\|\mathbf{n}_1 \times \mathbf{u}\|}, \ \mathbf{n}_2 = \mathbf{n}_3 \times \mathbf{n}_1 \tag{1}$$

where subscript t is omitted in the above expressions for clarity, $\|.\|$ denotes the norm of a vector and $\times$ denotes the cross product of two vectors. An indicative example of the proposed pose estimation procedure is illustrated in Fig. 1.

3.2 Action representation

For realizing efficient action recognition, an appropriate representation is required that will satisfactorily handle the differences in appearance, human body type and execution of actions among the individuals. For that purpose, the angles of the joints' relative position are used in this work, which showed to be more discriminative than using e.g. directly the joints' normalized coordinates. In order to compute a compact description, the aforementioned angles are estimated in the spherical coordinate system and the radial distance is omitted, since it contains information that is not necessary for the recognition process. Additionally, building on the fundamental idea of the previous section, all angles are computed using the $Torso$ joint as reference, i.e. the origin of the spherical coordinate system is placed at the $Torso$ joint position. For computing the proposed action representation, only a subset of the supported joints is used. This is due to the fact that the trajectory of some joints mainly contains redundant or noisy information. To this end, only the joints that correspond to the upper and lower body limbs were considered after experimental evaluation, namely the joints $Left\ shoulder$, $Left\ elbow$, $Left\ wrist$, $Right\ shoulder$, $Right\ elbow$, $Right\ wrist$, $Left\ knee$, $Left\ foot$, $Right\ knee$ and $Right\ foot$. Incorporating information from the remaining joints led to inferior performance, partly also due to the higher dimensionality of the calculated feature vector that hindered efficient HMM-based analysis. For every selected joint g_i the following spherical angles are estimated:

$$\varphi_i = \arccos\left(\frac{\langle(\mathbf{p}_i - \mathbf{p}_0), \mathbf{n}_2\rangle}{\|\mathbf{p}_i - \mathbf{p}_0\| \cdot \|\mathbf{n}_2\|}\right) \in [0, \pi]$$
$$\theta_i = \arctan\left(\frac{\langle(\mathbf{p}_i - \mathbf{p}_0), \mathbf{n}_1\rangle}{\langle(\mathbf{p}_i - \mathbf{p}_0), \mathbf{n}_3\rangle}\right) \in [-\pi, \pi] \tag{2}$$

where subscript t is omitted for clarity, φ_i is the computed polar angle, θ_i is the calculated azimuth angle and $\langle\cdot,\cdot\rangle$ denotes the dot product of two vectors.

Complementarily to the spherical angles, it was experimentally shown that the respective angular velocities provide additional discriminative information. To this end, the polar and azimuth velocities are estimated for each of the selected joints g_i, using the same expressions described in (2). The difference is that instead of the position vector $\mathbf{p}_i$ the corresponding velocity vector $\mathbf{v}_i$ is used. The latter is approximated by the displacement vector between two successive frames, i.e. $\mathbf{v}_i(t) = \mathbf{p}_i(t) - \mathbf{p}_i(t-1)$.

The estimated spherical angles and angular velocities for frame t constitute the frame's observation vector. Collecting the computed observation vectors for all frames of a given action segment forms the respective action observation sequence h that will be used for performing HMM-based recognition, as will be described in the sequel.

3.3 Horizontal symmetry

A problem inherently present in action recognition tasks concerns the execution of the same action while undergoing either a right- or left-body part motion. In order to address this issue, a motion energy-based approach is followed in this work, which builds upon the idea of the already introduced subject's pose (Section 3.1). The proposed method goes beyond typical solutions (e.g. measuring the length or the variance of the trajectories of the left/right body joints) and is capable of identifying and applying symmetries not only concerning common upper limb movements (e.g. left/right-hand waving actions), but also more extensive whole-body actions (e.g. right/left-handed golf-drive). In particular, all joints defined in G are considered, except from the $Torso$, $Head$ and $Neck$ ones. For each of these joints, its motion energy, which is approximated by $\|\mathbf{v}_i(t)\| = \|\mathbf{p}_i(t) - \mathbf{p}_i(t-1)\|$, is estimated for every frame t. Then, the calculated motion energy value $\|\mathbf{v}_i(t)\|$ is assigned to the Left/Right Body Part (LBP/RBP) according to the following criterion:

$$if\ \langle(\mathbf{p}_i(t) - \mathbf{p}_0(t)), \mathbf{n}_1(t)\rangle = \begin{cases} > 0, \|\mathbf{v}_i(t)\| \rightarrow RBP \\ < 0, \|\mathbf{v}_i(t)\| \rightarrow LBP \end{cases} \tag{3}$$

where $\rightarrow$ denotes the assignment of energy value $\|\mathbf{v}_i(t)\|$ to LBP/RBP. By considering the overall motion energy that is assigned to LBP/RBP for the whole duration of the examined action, the most 'active' body part is estimated. Then, a simple reallocation of the extracted feature values and application of horizontal symmetry attributes is performed. This is realized as follows: If the most 'active' part is LBP, the representation described in Section 3.2 remains unchanged. In case that RBP is the most 'active' one, the feature values of horizontally symmetric joints (e.g. $Left\ shoulder$-$Right\ shoulder$) are exchanged in all observation vectors of the respective action observation sequence h, while the horizontal symmetry attribute is applied by inverting the sign of all estimated azimuth angles and azimuth angular velocities. In this way, efficient horizontal symmetry that covers both limb motions as well as more extensive body movements is imposed.

3.4 HMM-based recognition

HMMs are employed in this work for performing action recognition, due to their suitability for modeling pattern recognition problems that exhibit an inherent temporality [11]. In particular, a set of J HMMs is employed, where an individual HMM is introduced for every supported action a_j. Each HMM receives as input the action observation sequence h (described in Section 3.2) and at the evaluation stage returns a posterior probability $P(a_j|h)$, which represents the observation sequence's fitness to the particular model.

Regarding the HMM implementation details, fully connected first order HMMs, i.e. HMMs allowing all possible hidden state transitions, were utilized for performing the mapping of the low-level features to the high-level actions. For every hidden state the observations were modeled as a mixture of Gaussians (a single Gaussian was used for every state). The employed Gaussian mixture models (GMMs) were set to have full covariance matrices for exploiting all possible correlations between the elements of each observation. Additionally, the Baum−Welch (or Forward−Backward) algorithm was used for training, while the Viterbi algorithm was utilized during the evaluation. Furthermore, the number of hidden states of the HMMs was considered a free variable. The developed HMMs were implemented using the software libraries of [1].

4 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 dataset are presented. In particular, the second session of the first dataset is used, which provides RGB-plus-depth video streams from two kinect sensors. In this work, the data stream from the frontal kinect was used. The dataset includes captures of 14 human subjects, where 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 of the proposed approach: $A = \{a_j,\ j \in [1, J]\} \equiv \{Hand\ waving,\ Knocking\ the\ door,\ Clapping,\ Throwing,\ Punching,\ Push\ away,\ Jumping\ jacks,\ Lunges,\ Squats,\ Punching\ and\ kicking,\ Weight\ lifting,\ Golf\ drive,\ Golf\ chip,\ Golf\ putt,\ Tennis\ forehand,\ Tennis\ backhand,\ Walking\ on\ the\ treadmill\}$. The 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 action representation; hence, they were not included in the conducted experiments that aim at evaluating the performance of the proposed approach for detecting complex and time-varying human actions. 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; eventually, an average performance measure was computed taking into account all intermediate recognition results.

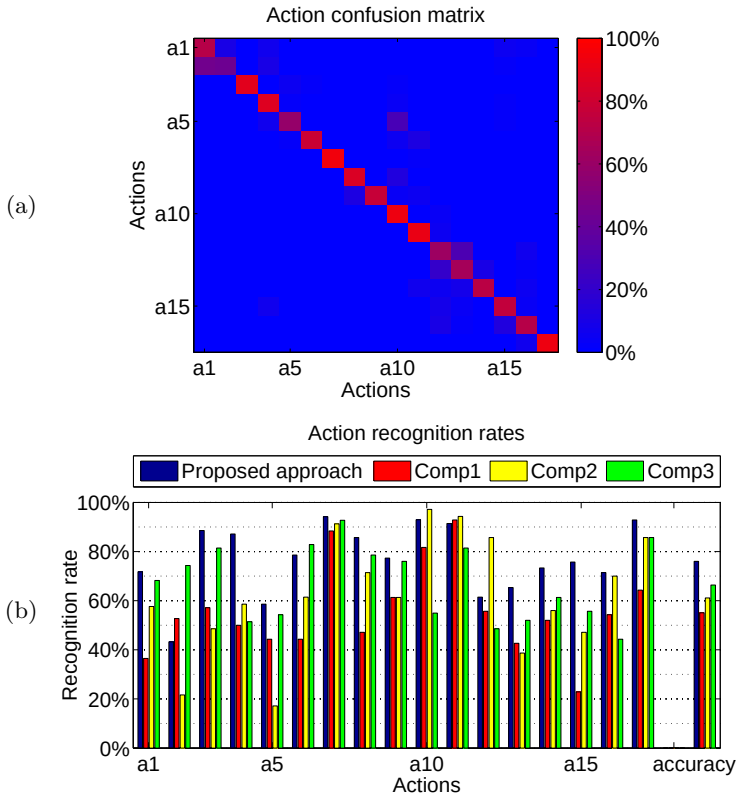

Fig. 2. Obtained action recognition results: (a) Estimated action confusion matrix and (b) calculated action recognition rates. Supported actions: a_1: Hand waving, a_2: Knocking the door, a_3: Clapping, a_4: Throwing, a_5: Punching, a_6: Push away, a_7: Jumping jacks, a_8: Lunges, a_9: Squats, a_{10}: Punching and kicking, a_{11}: Weight lifting, a_{12}: Golf drive, a_{13}: Golf chip, a_{14}: Golf putt, a_{15}: Tennis forehand, a_{16}: Tennis backhand, a_{17}: Walking on the treadmill.

In Fig. 2, quantitative action recognition results are presented in the form of the estimated confusion matrix (Fig. 2 (a)) and the calculated recognition rates (Fig. 2 (b)), i.e. the percentage of the action instances that were correctly identified. Additionally, the value of the overall classification accuracy, i.e. the percentage of all action instances that were correctly classified, is also given. For performance evaluation, it has been considered that $\arg\max_j(P(a_j|h))$ indicates the action a_j that is assigned to observation sequence h. From the presented results, it can be seen that the proposed approach achieves satisfactory action recognition performance (overall accuracy equal to 76.03%), which demonstrates the capability of the developed method to combine real-time processing with increased recognition rates. Examining the results in details, it can be seen that there are actions that exhibit high recognition rates (e.g. *Jumping jacks*, *Punching and*

Table 1. Time efficiency evaluation (per frame average processing times in msec)

Kinect sensor	Steps of the proposed approach		
Capturing period	Skeleton tracking	Feature extraction	HMM-based recognition
46.821	45.350	0.010	0.106

kicking and *Walking on the treadmill*), 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*, *Knocking the door* and *Golf drive*). This is mainly due to these actions presenting very similar motion patterns over a period of time during their execution with other ones (i.e. *Punching and kicking*, *Hand waving* and *Golf chip*, respectively). Moreover, the careful examination of the obtained results revealed that further performance improvement was mainly hindered due to the following two factors: a) the employed tracker sometimes provided inaccurate joint localizations (especially in cases of rapid movements or self-occlusions) and b) significant ambiguities in the execution of particular action pairs (e.g. some *Golf drive* instances presented significant similarities with corresponding *Golf chip* ones that even a human observer would be difficult to discriminate between them).

The proposed approach is also quantitatively compared with the following variants: a) use of normalized Euclidean distance of every selected joint from the *Torso* one (*Comp*1), b) use of normalized spherical angles of each joint from the *Torso* one (*Comp*2), and c) variant of the proposed approach, where the estimated spherical angles and angular velocities are linearly normalized with respect to the maximum/minimum values that they exhibit during the execution of a particular action (*Comp*3). Variants (a) and (b) follow the normalization approach described in [4], i.e. the human joint vector is position and orientation normalized without considering the relative position of the joints. From the presented results, it can be seen that the proposed method significantly outperforms both (a) and (b) variants. The latter demonstrates the usefulness of estimating the pose of the subject during the computation of the action representation, compared to performing a normalization step that does not take explicitly into account the relative position of the detected joints. The proposed method also outperforms variant (c), which suggests that performing a normalization of the angles/velocities within the duration of a particular action leads to decrease in performance.

The time efficiency of the proposed approach is evaluated in Table 1. In particular, the per frame average processing time, i.e. the time required for processing the data that correspond to a single frame, are given for every algorithmic step of the proposed method. More specifically, the following steps were considered: a) skeleton-tracking, described in Section 2, b) feature extraction, which includes the pose estimation, action representation computation and horizontal symmetry application procedures that are detailed in Sections 3.1, 3.2 and 3.3, respectively, and c) HMM-based recognition, outlined in Section 3.4. The dura-

tion of the aforementioned procedures is compared with the capturing period of the employed kinect sensor, i.e. the time interval between two subsequently captured depth maps. The average processing times given in Table 1 are obtained using a PC with Intel i7 processor at 2.67 GHz and a total of 6 GB RAM, while for their computation all video sequences of the employed dataset were taken into account. From the presented results, it can be seen that the employed skeleton-tracking algorithm constitutes the most time-consuming part of the proposed approach, corresponding to approximately 99.74% of the overall processing. The latter highlights the increased time efficiency (0.26% of the overall processing) of the proposed action recognition methodology (Section 3), characteristic that received particular attention during the design of the proposed method. Additionally, it can be seen that the overall duration of all algorithmic steps is shorter that the respective kinect's capturing period; in other words, all proposed algorithmic steps are completed for the currently examined depth map before the next one is captured by the kinect sensor. This observation verifies the real-time nature of the proposed approach. It must be highlighted that the aforementioned time performances were measured without applying any particular code or algorithmic optimizations to the proposed method.

5 Conclusions

In this paper, a real-time tracking-based approach to human action recognition was presented and evaluated using the Huawei/3DLife 3D human reconstruction and action recognition Grand Challenge dataset. Future work includes the investigation of STIP-based approaches for overcoming the inherent limitations of skeleton-tracking algorithms and the incorporation of multimodal information from additional sensors.

Acknowledgment

The work presented in this paper was supported by the European Commission under contract FP7-287723 REVERIE.

References

1. Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk
2. Ballan, L., Bertini, M., Del Bimbo, A., Seidenari, L., Serra, G.: Effective codebooks for human action representation and classification in unconstrained videos. Multimedia, IEEE Transactions on 14(4), 1234–1245 (2012)
3. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(12), 2247–2253 (2007)
4. Gu, J., Ding, X., Wang, S., Wu, Y.: Action and gait recognition from recovered 3-d human joints. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Trans. on 40(4), 1021–1033 (2010)

5. Haq, A., Gondal, I., Murshed, M.: On temporal order invariance for view-invariant action recognition. Circuits and Systems for Video Technology, IEEE Transactions on 23(2), 203–211 (2013)
6. Holte, M.B., Chakraborty, B., Gonzalez, J., Moeslund, T.B.: 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)
7. Ji, X., Liu, H.: Advances in view-invariant human motion analysis: a review. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Trans. on 40(1), 13–24 (2010)
8. Junejo, I.N., Dexter, E., Laptev, I., Pérez, P.: View-independent action recognition from temporal self-similarities. Pattern Analysis and Machine Intelligence, IEEE Trans. on 33(1), 172–185 (2011)
9. Papadopoulos, G.T., Briassouli, A., Mezaris, V., Kompatsiaris, I., Strintzis, M.G.: Statistical motion information extraction and representation for semantic video analysis. Circuits and Systems for Video Technology, IEEE Transactions on 19(10), 1513–1528 (Oct 2009)
10. Poppe, R.: A survey on vision-based human action recognition. Image and vision computing 28(6), 976–990 (2010)
11. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proc. of the IEEE 77(2), 257–286 (1989)
12. Song, B., Kamal, A.T., Soto, C., Ding, C., Farrell, J.A., Roy-Chowdhury, A.K.: Tracking and activity recognition through consensus in distributed camera networks. Image Processing, IEEE Transactions on 19(10), 2564–2579 (2010)
13. Turaga, P., Veeraraghavan, A., Chellappa, R.: Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In: Computer Vision and Pattern Recognition, IEEE Conf. on. pp. 1–8. IEEE (2008)
14. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding 104(2), 249–257 (2006)