Weakly Aligned Multi-part Bag-of-Poses for Action Recognition from Depth Cameras

Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti,
Alberto Del Bimbo, and Pietro Pala

University of Firenze, Firenze, Italy
lorenzo.seidenari@unifi.it
http://www.micc.unifi.it

Abstract. In this work, we propose an efficient and effective method to recognize human actions based on the estimated 3D positions of skeletal joints in temporal sequences of depth maps. First, the body skeleton is decomposed in a set of kinematic chains, and the position of each joint is expressed in a locally defined reference system, which makes the coordinates invariant to body translations and rotations. A multi-part bag-of-poses approach is then defined, which permits the separate alignment of body parts through a nearest-neighbor classification. Experiments conducted on the MSR Daily Activity dataset show promising results.

Keywords: depth camera, action recognition, nearest-neighbor classification.

1 Introduction

Recently, the use of RGB-D map sequences (sequences of synchronized and aligned RGB and depth images) is receiving an increasing attention in various applications, including recognition of human actions and gestures [2], human pose reconstruction and estimation [10,1,5], scene flow estimation [6], face super-resolution [3]. Approaches proposed for recognition of human actions and gestures are relevant in very different domains ranging from biomedicine (e.g., monitoring and analysis of patient movements for supervised rehabilitation), to video-surveillance and social behavior analysis (e.g., indicators of shame and embarrassment based on human gestures) [11]). These approaches can be grouped into three main categories: *skeleton based*, that estimate the positions of a set of joints in the human skeleton from the depth map, and then model the pose of the human body in subsequent frames of a sequence using the position and the relations between joints; *depth map based*, that extract volumetric and temporal features from the overall set of points of the depth maps in a sequence; and *hybrid* solutions, which combine information extracted from both the joints of the skeleton and the depth maps.

Skeleton based approaches have become popular thanks to the work of Shotton et al. [10], where a real-time method is defined to accurately predict 3D positions of 16 body joints in individual depth map without using any temporal information. Relying on the joints location provided by Kinect, in [13] an approach for human action recognition is proposed, which computes histograms of the locations of 12 3D joints as a compact representation of postures. The histograms computed from the action depth

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 446–455, 2013.

sequences are then projected using LDA and clustered into k posture visual words, which represent the prototypical poses of actions. The temporal evolutions of those visual words are modeled by discrete HMMs. Results were provided on the Microsoft Research (MSR) Action3D dataset [7]. In [14], human actions recognition is obtained by extracting pair-wise differences of joint positions in the current frame, between the current and the preceding frame, and between the current frame and the initial frame of the sequence (assumed as neutral). PCA is then used to reduce redundancy and noise in the feature and to obtain a compact *EigenJoints* representation for each frame. Finally, a naïve-Bayes nearest-neighbor classifier is used for multi-class action classification on the MSR Action3D dataset. With respect to the other classes of approaches, skeleton based solutions have the main merit in being simple and efficient, which is of paramount importance in real-time action recognition.

In this work, we propose a *skeleton based* solution for human action recognition from sequences of depth maps acquired with a Kinect camera. The key idea of our approach is to use joint positions to align multiple-parts of the human body using a bag-of-poses solution applied in a nearest-neighbor framework. We develop on the human body representation model proposed in [9], which is based on four kinematic chains. The coordinates of each joint in a chain are expressed in a local reference system, which is defined at the preceding joint. In this way, the coordinates are invariant to translation and rotation of the body, and each part of the body is modeled separately, to allow each part to be aligned independently. Hence, if the full body feature is noisy, the classifier can still obtain a strong score by aligning sub-parts of the body. Experimental results evidence competitive results in comparison to existing *skeleton based* solutions.

The rest of the paper is organized as follows: In Sect. 2, the proposed skeletal representation of the human body is described. This representation is then exploited in a multi-part nearest-neighbor classifier to perform action classification, as discussed in Sect. 3. Results obtained using the proposed framework on the MSR Daily Action 3D dataset are reported in Sect. 4. Finally, discussion and conclusions are drawn in Sect. 5.

2 Skeletal Representation

The proposed action recognition system relies on a skeletal based representation of the human body. This is provided by the Kinect platform that outputs a wireframe skeleton at a rate of 30 fps for each human body recognized in the acquired RGB-D datastream. Each skeleton part — forearm, upper arm, torso, head, etc. — is modelled as a rigid body. The position of the skeleton joints are provided as (x, y, z) coordinates in an absolute reference system that places the Kinect device at the origin with the positive z-axis extending in the direction in which the device is pointed, the positive y-axis extending upward, and the positive x-axis extending to the left. However, this absolute representation is highly inefficient and redundant since the coordinates of joints are mutually correlated. A much more convenient and generally adopted solution models the movements of the human body using kinematic chains, the root of the kinematic tree being the torso (base body) and the position of each joint being expressed relative to its parent joint. We adopt the same representation model proposed in [9] and assume the relative position of joints of the human torso — composed of the left and right

shoulders, the base of the neck and the left and right hips — does not change over time. Thus, the entire torso is modeled as a rigid part and the remaining joints are classified into *first* and *second* degree joints. The first degree joints are those adjacent to the torso: the *elbows* and the *knees*. The second degree joints are the children of the first degree joints in the four kinematic chains: the *wrists* and the *feet*.

The position of each first degree joint is expressed in a coordinate system which is derived from the *torso frame*. This is a 3D orthonormal basis $\{\boldsymbol{u}, \boldsymbol{r}, \boldsymbol{t}\}$ resulting from the PCA of the positions of the torso joints. The torso frame is translated so as to express the coordinates of the first degree joints (see Fig. 1). Coordinates $[u, r, t]_0$ of the left elbow joint are expressed in the torso frame coordinate system translated so as to center the origin at the left shoulder joint. Coordinates $[u, r, t]_1$ of the left knee joint are expressed in the torso frame coordinate system translated so as to center the origin at the left hip joint. Similarly, coordinates of the right elbow and knee are expressed in a torso frame coordinate system centered at the right shoulder and hip, respectively: $[u, r, t]_3$ and $[u, r, t]_2$. It should be noticed that this solution differs from the one proposed in [9] where two angular variables are used to represent the first degree joints in polar coordinates. Differently, we represent the coordinates of the first degree joints in Cartesian coordinates $[u, r, t]$, which makes the representation system immune to the well known "gimbal lock" problem.

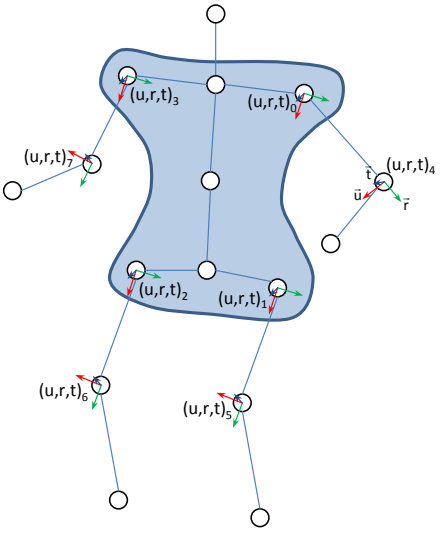

Fig. 1. The body skeleton and the first and second degree coordinate systems

The position of each second degree joint is expressed in a coordinate system that is derived from the coordinate system used to represent the position of its parent joint. Given a second degree joint, the $\{\boldsymbol{u}, \boldsymbol{r}, \boldsymbol{t}\}$ system with origin centered at the root of its kinematic chain is rotated and translated so as to center its origin at the parent first degree joint. The applied rotation is such that the direction of $\boldsymbol{r}$ matches the direction

of the link between the root of the kinematic chain and the parent first degree joint. In this way, four new coordinate systems $[u, r, t]_k$, $k = 4, \ldots, 7$ are created with origin at the left elbow, left knee, right knee and right elbow, respectively (see Fig. 1). Based on this representation system, a generic body pose is represented by a 24-dimensional feature vector $h = [u_0, r_0, t_0, \ldots, u_7, r_7, t_7]$ measuring the coordinates of the first and second degree joints in their coordinate systems. All of the vectors $v_j = [u_j, r_j, t_j]$ are L2-normalized in order to obtain robustness to the body size of different people and to the noise in estimating 3D joint position due to distance from the sensor.

3 Action Classification

State of the art methods for image classification are based on parametric classifiers, like SVM, Boosting, etc., which require an intensive learning/training stage. In contrast, non-parametric Nearest-Neighbor (NN) based classifiers have some favorable properties [4]: Naturally deal with a large number of classes; Avoid the overfitting problem; Do not require parameters learning.

Following this idea, in our approach a Naïve-Bayes Nearest-Neighbor (NBNN) classifier is applied for action recognition. For each frame in a sequence of depth maps, a feature vector is computed and used without quantization as frame descriptor, as detailed in Sect. 2. Considering M classes of actions to be recognized C_k, $k = 1, \ldots, M$, a number of labelled sequences per class is used as "training" set. Actually, this step does not include any learning/training of parameters, but the frame descriptors of these labelled sequences just serve as prototypes of a class.

According to this, given a depth frame f_i of a query sequence and its descriptor h_i, for each class C_k the training frame is searched which minimizes the distance:

$$d_i^{C_k} = \|h_i - NN^{C_k}(h_i)\|^2, \tag{1}$$

where $NN_{C_k}(h_i)$ is the NN-descriptor of h_i in the training frames of class C_k. Repeating this step for each frame $f_i, i = 1, \ldots, S$ of the query sequence, a set of M *class-reconstructed* sequences are derived, each comprising the NN-frames in the class C_k. Based on the distance between a query frame descriptor and its NN-frame descriptor, a *goodness* value is then associated to each of the *class-reconstructed* sequences:

$$G^{C_k} = \frac{1}{S}\sum_{i=1}^{S} g_i^{C_k} = \frac{1}{S}\sum_{i=1}^{S} \exp(d_i^{C_k}/\sigma^2). \tag{2}$$

3.1 Weak Temporal Alignment of Bag of Poses

The goodness value computed between two sequences does not account for their temporal ordering. Due to this, frames in the *class-reconstructed* sequences could have a meaningless temporal ordering when compared to the query sequence. So, in order to account for the temporal correlation between two sequences, we found beneficial to add an extra feature to the feature vector obtaining $h = \left[u_0\ r_0\ t_0, \ldots, u_7\ r_7\ t_7,\ \beta\frac{s}{S}\right]$, where s is the frame index and S is the sequence length in frames. The constant β ensures that

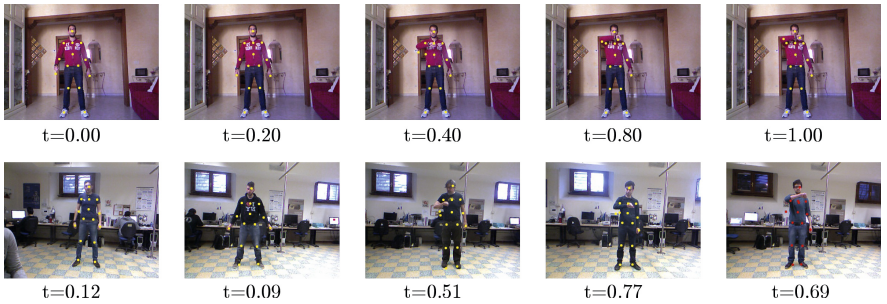

Fig. 2. Weak alignment between query (top) and (bottom) reconstructed sequence from the correct class with the normalized time-stamps

the weight of the temporal feature is not discarded because of the high dimensionality of the vector and it is selected by cross-validation. To encode short time temporal relationships, we also add to vector h temporal derivatives $[\mathrm{d}u_j\ \mathrm{d}r_j\ \mathrm{d}t_j]$. The final feature set is $h = \left[u_0\ r_0\ t_0,\ \mathrm{d}u_0\ \mathrm{d}r_0\ \mathrm{d}t_0, \ldots,\ \beta\frac{s}{S}\right]$. Adding the normalized timestampe $\beta\frac{s}{S}$ makes frame in the same relative position in a sequence closer; this perform a weak alignment of sequences.

For efficiency reasons, the frame descriptors of the training sequences of a class are stored in a KD-tree (a total of M trees are constructed). Using a KD-tree, the *class-reconstructed* sequence of a query with S frames is constructed with S searches, each search having a logarithmic cost in the number of frames in the tree. As it can be observed in Fig. 2, our approach performs an implicit *sequence-to-class* alignment procedure picking for each query frame the best exemplar without taking into account the sequence, but only the relative positioning. Dynamic Time Warping (DTW) instead, performs a *sequence-to-sequence* alignment; thus our method can leverage a lot more data since virtually any combination of frames from a class can be used to reconstruct the query sequence.

3.2 Multi-part Models

Following the approach proposed in [12] based on learning relevant depth and joint features for each action class, we improve our model by combining multiple local body descriptors computed hierarchically. Let δ_p be a binary vector representing a selector for part p, that picks a subset of the features such that $\delta_p \circ h = \left[u_p\ r_p\ t_p,\ \beta\frac{s}{S}\right]$; we simply zero all features except the one not belonging to the part and the normalized frame. To define a higher order feature it is sufficient to OR two selectors: $\delta_{LA} = \delta_{LE} \vee \delta_{LH}$, where LA, LE and LH indicate the left arm, elbow and hand, respectively. The legs, torso and lower body selector can be obtained as such. This procedure also applies to derivatives separately. For a multi-part model the NBNN classifier becomes:

$$C_{NBNN} = \arg\max_{C_k} \frac{1}{S} \sum_{p \in P} \sum_{i=1}^{S} \exp\left(\|\delta_p \circ h_i - NN^{C_k}(\delta_p \circ h_i)\|^2 / \sigma_p^2\right), \quad (3)$$

given a set of parts P. We estimate the σ_p value as:

$$\sigma_p = \frac{1}{S(S-1)/2} \sum_{i \in D} \sum_{j \in D} \|\delta_p \circ h_i - \delta_p \circ h_j\|, \forall i < j \in D, \quad (4)$$

with a sample of the training data D. The value σ_p is fixed for each part and does not depend on the category. The same approach is used to tune the σ in (3). Note that we are not learning the feature representation, but the key idea is to align separately meaningful body parts by seeking the best sequence able to independently align the sub-parts. As an example, if the full body feature is noisy, the classifier can still obtain a strong score from aligning the torso or the arms in actions such as *drinking* or *eating*.

4 Experimental Results

The proposed method has been evaluated on the Microsoft Research (MSR) Daily Activity 3D dataset [12]. Results scored by our approach on this benchmark were also compared against those reported by state of the art solutions on the same benchmark.

The Daily Activity 3D dataset was captured at MSR using a Kinect device [12]. There are 16 activities: *drink*, *eat*, *read book*, *call cellphone*, *write on a paper*, *use laptop*, *use vacuum cleaner*, *cheer up*, *sit still*, *toss paper*, *play game*, *lie down on sofa*, *walk, play guitar*, *stand up*, *sit down*. There are 10 subjects in the dataset. Each subject performs each activity twice, once in "standing" position, and once in "sitting on sofa" position. The total number of the activity samples is $16 \times 2 \times 10 = 320$. This dataset has been designed to cover humans daily activities in a living room. As a consequence, when the user stands close to the sofa or sits on the sofa, the 3D joint positions extracted by the skeleton tracker are very noisy. In addition, most of the activities involve humans-object interactions, thus making this dataset quite challenging.

Experiments have been conducted using a cross-actor training/testing setup. Specifically, we left out each actor from the training set and repeated an experiment for each of them (leave-one-actor-out). The confusion matrix obtained using the multi-part variant of our approach is reported on Fig. 3. It can be noted as the most critical actions to classify correspond to the cases where subjects interact with external objects, rather than to pose variations alone. Our algorithm occupy a very tiny time-slot ($<$10 ms) with respect to the user detection and tracking. For a single user, our system runs at 20 fps on standard hardware.

4.1 Comparative Evaluation

Results of a comparative analysis of the proposed approach to alternative solutions are reported on Table 1. The first investigation aims to evidence accuracies obtained by using the different variants of our solution. In particular, in the Table we indicate our base solution with "NBNN", its variants adding separately time and parts as, respectively, "NBNN+parts" and "NBNN+time," and the solution which accounts for time and parts together as "NBNN+parts+time". Results obtained on the DailyActivity3D datasets show that the "time" feature is as relevant as the part based modeling in improving the performance of the NBNN base approach; both cues combined together

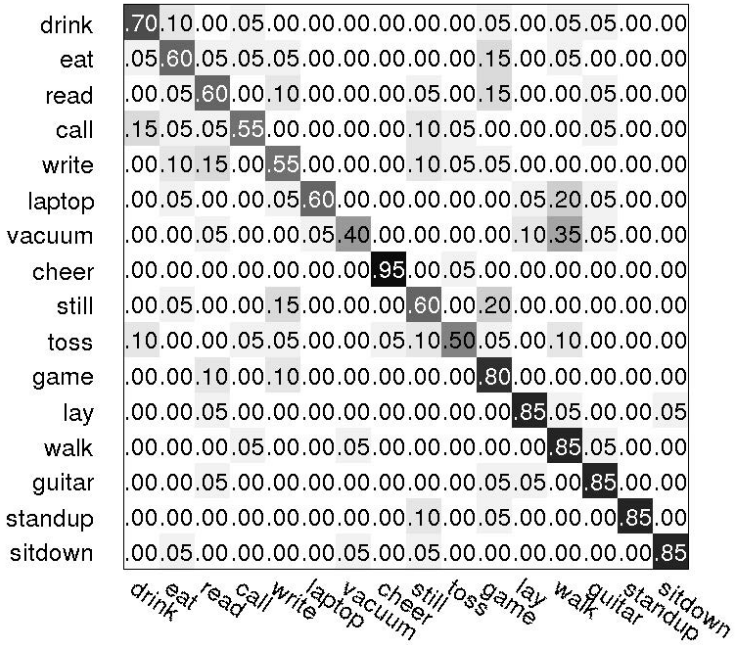

Fig. 3. MSR Daily Activity 3D dataset: Confusion matrix

yield state of the art results. On the MSR Daily Activity 3D dataset, we also compared our approach with the solutions obtained with [8] and [12] reported in [12]. The solution in [8] uses Dynamic Temporal Warping (DTW) to match the 3D joint positions to a template, and action recognition can be done through a NN-classification method. The method in [12], instead, uses the estimated 3D joint positions and a Local Occupancy Pattern as local feature for human body representation. Since our method only exploits the joints positions, for a fair comparison in the Table we report the results of [12] obtained only using the joints positions, as given by the authors. On the MSR Daily Activity, we can observe a diffused confusion in the upper left quadrant of the confusion matrix relative to {*drink*, *eat*, *call*, *eat*, *write*}. Also, since we are not employing features other than the joints representation, our approach has not very high accuracy on actions mainly defined by the presence of an object, like *vacuum*, *laptop*, *read* or *write*.

In Tab. 2 we also report the accuracy scored by individual body parts. In particular, "ubody" and "lbody" include, respectively, joints in the arms and joints in the legs, with "dubody" and "dlbody" indicating the differential features computes on the "ubody" and "lbody" parts. The rightmost columns report the accuracy resulting by combining together all the parts and their differential components (all parts), and that of the complete approach that also explicitly includes the temporal feature (all parts+time). From the Table, it can be observed as for some actions, like lay or walk, the lower part of the body provides more accuracy, whereas for other actions, like cheer, read and

Table 1. Recognition accuracy comparison. For the method in [12], results obtained using only the joints position are reported

Method	MSR Daily
NBNN + parts + time	**0.70**
NBNN + time	0.62
NBNN + parts	0.60
NBNN	0.53
Actionlets [12]	0.68
DTW [8]	0.54

guitar, is the upper part to provide significantly better results. Noticeably, in some cases (bold-underlined in the table) individual parts perform better than their combination.

Finally, in Fig. 4 we report the variation of the recognition accuracy as a function of the β parameter, which weights the mutual relevance of the spatial and temporal component in the feature vector (Sect. 3.1). Values are reported for both the cases in which the parts are used or not. The accuracy gain using the temporal cue can be appreciated.

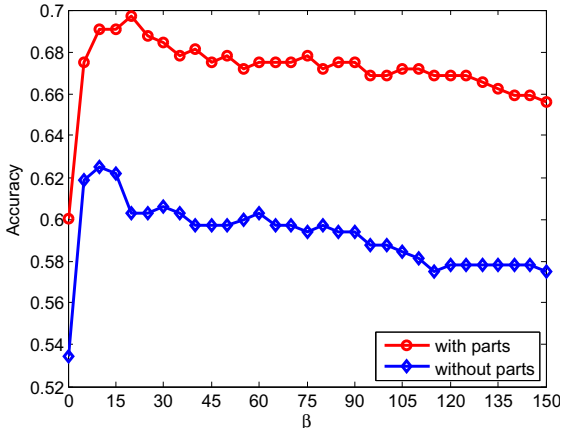

Fig. 4. Recognition accuracy as a function of β (for $\beta = 0$ no temporal information is used)

5 Conclusions

In this work, we proposed a method for human action recognition, which is based on weakly aligning the 3D coordinates of joints in multiple parts of the skeleton. First, four kinematic chains, each modeling a limb of the human body are defined, then the 3D coordinates of each joint in a chain are expressed in a locally defined reference system, which permits coordinates invariance with respect to rotations and translations. The coordinates of the joints, the temporal derivatives of the coordinates as well as a temporal

Table 2. Accuracy per class and body part (in bold the best part accuracy). In some cases (underlined values), the best part accuracy is better than the accuracy scored by all the parts+time

	ubody	lbody	dubody	dlbody	all parts	all parts + time
drink	**0.65**	0.20	0.60	0.30	0.65	0.70
eat	**0.55**	0.15	0.35	0.30	0.60	0.60
read	0.60	0.30	**0.70**	0.25	0.45	0.60
call	**0.40**	0.10	0.10	0.00	0.30	0.55
write	**0.60**	0.40	0.25	0.05	0.45	0.55
laptop	0.25	**0.40**	0.00	0.15	0.50	0.60
vacuum	0.40	**0.45**	0.30	0.40	0.50	0.40
cheer	0.95	0.35	**1.00**	0.40	0.95	0.95
still	0.40	0.10	**0.65**	0.45	0.55	0.60
toss	**0.45**	0.00	0.35	0.05	0.45	0.50
game	**0.70**	0.45	0.25	0.05	0.70	0.80
lay	0.80	**0.90**	0.80	0.80	0.70	0.85
walk	0.55	0.75	0.45	**0.90**	0.70	0.85
guitar	**0.85**	0.10	0.20	0.00	0.70	0.85
standup	0.60	**0.70**	0.50	0.65	0.70	0.90
sitdown	0.50	0.70	0.25	**0.75**	0.75	0.85
mean	0.58	0.38	0.42	0.34	0.60	0.70

feature are used as feature vector representing the human body in each frame. In order to make the approach robust to noise, a part based solution has been also deployed, which permits alignment of sub-sets of the joints. A sequence-to-class nearest-neighbor classifier has been used to score the similarity of a query action. Experiments carried out on a benchmark dataset support the applicability of the proposed solution, also showing competitive performance when compared to other skeletal- based solutions.

Acknowledgments. This research is carried out in the context of the RIS project that is funded with support from the *Programma Operativo Regionale* co-funded by FESR for the objective *Competitivitá regionale e occupazione* years 2007-2013.

References

1. Baak, A., Muller, M., Bharaj, G., Seidel, H.P., Theobalt, C.: A data-driven approach for real-time full body pose reconstruction from a depth camera. In: Proc. of Int. Conf. on Computer Vision, Barcelona, Spain, pp. 1092–1099 (November 2011)
2. Bagdanov, A.D., Del Bimbo, A., Seidenari, L., Usai, L.: Real-time hand status recognition from RGB-D imagery. In: Proc. of Int. Conf. on Pattern Recognition, Tsukuba, Japan, pp. 2456–2459 (November 2012)

3. Berretti, S., Del Bimbo, A., Pala, P.: Superfaces: A super-resolution model for 3D faces. In: Proc. of Work. on Non-Rigid Shape Analysis and Deformable Image Alignment, Florence, Italy, pp. 73–82 (October 2012)
4. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Anchorage, Alaska, USA, pp. 1–8 (June 2008)
5. Girshick, R., Shotton, J., Kohli, P., Criminisi, A., Fitzgibbon, A.: Efficient regression of general-activity human poses from depth images. In: Proc. of Int. Conf. on Computer Vision, Barcelona, Spain (November 2011)
6. Hadfield, S., Bowden, R.: Kinecting the dots: Particle based scene flow from depth sensors. In: Proc. of Int. Conf. on Computer Vision, Barcelona, Spain, pp. 2290–2295 (November 2011)
7. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: Work. on Human Communicative Behavior Analysis, San Francisco, California, pp. 9–14 (June 2010)
8. Muller, M., Röder, T.: Motion templates for automatic classification and retrieval of motion capture data. In: ACM SIGGRAPH/Eurographics Symp. on Computer Animation, Vienna, Austria, pp. 137–146 (September 2006)
9. Raptis, M., Kirovski, D., Hoppe, H.: Real-time classification of dance gestures from skeleton animation. In: ACM SIGGRAPH/Eurographics Symp. on Computer Animation, Vancouver, Canada, pp. 147–156 (August 2011)
10. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: IEEE Conf. on Computer Vision and Pattern Recognition, Colorado Springs, Colorado, pp. 1–8 (June 2011)
11. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vison Computing 27(12), 1743–1759 (2009)
12. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: IEEE Conf. on Computer Vision and Pattern Recognition, Providence, Rhode Island, pp. 1–8 (June 2012)
13. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: Work. on Human Activity Understanding from 3D Data, Providence, Rhode Island, pp. 20–27 (June 2012)
14. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: Work. on Human Activity Understanding from 3D Data, Providence, Rhode Island, pp. 14–19 (June 2012)