Recognizing Actions from Depth Cameras as Weakly Aligned Multi-Part Bag-of-Poses

Lorenzo Seidenari,Vincenzo Varano, Stefano Berretti, Alberto Del Bimbo, Pietro Pala
University of Firenze
{lorenzo.seidenari, stefano.berretti, pietro.pala}@unifi.it, v.varano@ymail.com

Abstract

Recently released depth cameras provide effective estimation of 3D positions of skeletal joints in temporal sequences of depth maps. In this work, we propose an efficient yet effective method to recognize human actions based on the positions of joints. 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 Florence 3D Action dataset and the MSR Daily Activity dataset show promising results.

1. Introduction

Imaging technologies have recently shown a rapid advancement with the introduction of consumer depth cameras with real-time capabilities, like Microsoft Kinect or Asus Xtion PRO LIVE. These new acquisition devices have stimulated the development of various promising applications, including human pose reconstruction and estimation [1, 6, 12], scene flow estimation [7], hand gesture recognition [2], face super-resolution [3]. Encouraging results shown in these works have been made possible also thanks to the advantages that depth cameras have in comparison to conventional cameras, such as an easier foreground/background segmentation, and a lower sensitivity to lighting conditions.

In particular, an increasing attention has been directed to the task of recognizing human actions using depth map sequences. To this end, several approaches have been developed in the last few years that can be categorized as: *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. Following this categorization, existing methods for human action recognition with depth cameras are shortly reviewed below.

1.1. Related work

Skeleton based approaches have become popular thanks to the work of Shotton et al. [12], where a real-time method is defined to accurately predict 3D positions of body joints in individual depth map without using any temporal information. In that work, prediction accuracy results are reported for 16 joints, but the Kinect tracking system developed on top of this approach is capable to estimate 3D positions for 20 joints of the human skeleton. Relying on the joints location provided by Kinect, in [15] 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 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 Hidden Markov Models (HMMs). Results were provided on a proprietary dataset and on the public Microsoft Research (MSR) Action3D dataset [9]. In [16], human actions recognition is obtained by extracting three features for each joint which are based on pair-wise differences of joint positions, respectively: differences between joints in the current frame; between joints in the current frame and the preceding frame; and between joints in the current frame and in the initial frame of the sequence that is assumed to approximate the neutral posture. Since the number of these differences results in a high dimensional feature vector, PCA is 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.

978-0-7695-4990-3/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPRW.2013.77

Methods based on depth maps, do not rely on fitting a humanoid skeleton on the data, but use instead the entire set of points of depth map sequences to extract meaningful spatiotemporal descriptors. In [9], depth maps of a sequence are projected onto the three orthogonal Cartesian planes and a specified number of points at equal distance along the contours of the projections are sampled for each frame. Then, the 3D points that are nearest to the sampled 2D points are retrieved. Since the projections to the xz and zy plane can be very coarse due to the resolution of the depth map, interpolation may be required in order to construct these projections. In addition, each projection may have multiple unconnected regions, and in such a case contours of all regions are sampled. Finally, the sampled points are used as bag-of-points to characterize a set of salient postures that correspond to the nodes of an *action graph* used to model explicitly the dynamics of the actions. Experimental results reported on the MSR Action3D dataset have shown over 90% recognition accuracy by sampling only about 1% 3D points from the depth maps. In [13], a three-dimensional action sequence is treated as a 4D shape and random occupancy pattern (ROP) features are extracted. Since in the depth sequences many subvolumes do not contain useful information for classification, a weighted sampling approach is proposed based on the rejection sampling, which samples discriminative subvolumes with high probability. An Elastic-Net regularized classification model is then developed to further select the most discriminative features, and sparse coding is utilized to encode the features. A descriptor of depth information for action representation is also proposed in [5]. In particular, with this descriptor the structural information of spatiotemporal points within action volumes is captured using distance information in depth data. The approach in [17] projects depth maps onto three orthogonal planes and accumulates global activities through entire video sequences to generate Depth Motion Maps (DMM). In particular, DMMs are generated by projecting depth maps onto the three orthogonal Cartesian planes, computing a motion energy by thresholding the difference between two consecutive maps, and stacking the energies for each projection. Histograms of Oriented Gradients (HOG) are then computed from DMM as the representation of an action video and used as input to SVM classifiers. Recognition results on the MSR Action3D dataset are reported.

Hybrid solutions try to combine positive aspects of both skeleton and depth-map based methods. Relying on the depth data and the estimated 3D joint positions, the approach in [14] proposes a *Local Occupancy Pattern* (LOP) as local feature for human body representation. With this approach, each 3D joint is associated with a LOP which can be regarded as the "depth appearance" of the 3D joint. In addition, the temporal structure of an individual joint in an action is represented through a temporal pattern representation called *Fourier Temporal Pyramid*, which is quite insensitive to temporal sequence misalignment and noise. The concept of *actionlet* is then introduced to indicate a structure of the features originated by a particular conjunction of the features for a subset of the joints. Since the number of possible actionlets is intractable, a data mining solution is used to discover discriminative actionlets. Finally, an action is represented as an *Actionlet Ensemble*, that is a linear combination of the actionlets where the discriminative weights are learnt via a multiple kernel learning method.

1.2. Paper contribution and organization

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. In so doing, the approach results in a simple and efficient implementation that does not require any parameter learning. Experimental results also evidence competitive results of the approach in comparison to existing *skeleton based* solutions.

The contributions of this work are threefold. First, we propose an original representation of the human body, which is based on four kinematic chains. The coordinates of each joint in a kinematic chain are expressed in a local reference system which is defined at the preceding joint in the chain. This permits the coordinates to be expressed invariantly to translation and rotation of the body with respect to the camera reference system. In addition, Cartesian coordinates are used to avoid the "gimbal lock" problem. Second, we give a multi-part modeling of the body using the above joints representation, with the idea to align separately meaningful body parts. Our system seeks the best sequence able to independently align the sub-parts, so that if the full body feature may be noisy, the classifier will still obtain a strong score from aligning sub-parts of the body. Finally, we use nearest-neighbor alignment to perform action classification. On this latter point, we share some ideas with the approach in [16]. However, our solution is more general not requiring any re-training or assumption about the pose in the initial frame of a sequence. In fact, in [16] NN-classification is applied to the PCA projection of feature vectors that also include the difference between joint positions in the current frame and in the first frame of a sequence that is assumed to show a neutral pose of the body. As a consequence, PCA training should be recomputed when new classes or examples per class are added. In addition, the method strongly depends from the accurate sequence segmentation, which is necessary to guarantee the neutral pose hypothesis on the initial frame to hold. Both these limitations do not occur in our solution.

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 two benchmark datasets 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, for instance, a simple rotation of the upper arm results into a change of all the 3D coordinates of the wrist. A much more convenient and generally adopted solution relies on modeling 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 [11] and assume that 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 that are 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 that is derived from the *torso frame*. This is a 3D orthonormal basis $\{\vec{u}, \vec{r}, \vec{t}\}$ resulting from the Principal Component Analysis (PCA) of the positions of the torso joints. The first component $\vec{u}$ is aligned with the backbone and oriented downward. The second component $\vec{r}$ is aligned with the shoulders line and oriented from the right to the left shoulder. The third component $\vec{t}$ is the cross product of the first two components $\vec{t} = \vec{u} \times \vec{r}$.

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 [11] 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.

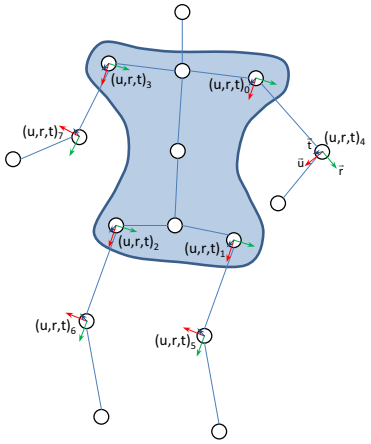

Figure 1. The body skeleton and the first and second degree coordinate systems.

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 $\{\vec{u}, \vec{r}, \vec{t}\}$ 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 $\vec{r}$ 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 different people body size and noise in the 3D estimation 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: Naturally deal with a large number of classes; Avoid the overfitting problem; Do not require parameters learning. However, the large performance gap between these two families of approaches rendered NN-based image classifiers useless. This position has been recently rebutted in [4], where it was observed that the effectiveness of NN for image classification has been largely underestimated due to the quantization of image descriptors and the computation of image-to-image distance. In particular, the experiments in [4] showed that frequent descriptors have low quantization error, but rare descriptors have high quantization error. Since discriminative descriptors tend to be rare, quantization can significantly degrade the discriminative power of descriptors. In addition, they observed that computing image-to-class distance, which depends on the distribution of the descriptor over the entire class, provides better generalization capability than image-to-image distance. Extension of these concepts to NN-based video classification for action recognition was also proposed in [16].

3.1. NBNN on bag of poses

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 a 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 than associated to each of the *class-reconstructed* sequences:

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

3.2. Weak temporal alignment of bag of poses

However, 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, the *Kendall rank correlation coefficient* (also known as *Kendall's τ coefficient*) is computed, which produces an index in the $[-1, 1]$ interval: τ equal to 0 means that the two sequences have independent ordering; values of τ equal to $+1$ or -1 indicate, respectively, that the two sequences have values that follow the same or opposite ordering. In our case, the *Kendall's τ coefficient* is computed between the S frames of a query sequence and the *class-reconstructed* sequence of each class C_k:

$$\tau^{C_k} = \frac{N_a^{C_k} - N_d^{C_k}}{\frac{1}{2}N(N-1)}, \tag{3}$$

where $N_a^{C_k}$ and $N_d^{C_k}$ represent, respectively, the number of observation pairs (i.e., frames) in the two sequences which are in agreement/disagreement. Finally, the two scores are combined together to obtain the overall classification score of a query sequence with respect to a class C_k. To this end the τ^{C_k} value is normalized in [0,1], that is: $T^{C_k} = (\tau + 1)/2$. The class C_k^* which maximizes the overall score is assumed as the label for the query sequence:

$$C_k^* = \arg\max_{C_k} (\alpha G^{C_k} + (1-\alpha) T^{C_k}). \tag{4}$$

In our preliminary experiments, we found that even for reasonably high α values (e.g., 0.8) the T^{C_k} scoring actually helps in disambiguating classes that appear similar if the order is not taken into account, like *sit down* and *stand up*, but may decrease the recognition accuracy for other classes. Instead, 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 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 $h = \left[u_0\ r_0\ t_0,\ \mathrm{d}u_0\ \mathrm{d}r_0\ \mathrm{d}t_0, \ldots,\ \beta\frac{s}{S}\right]$.

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

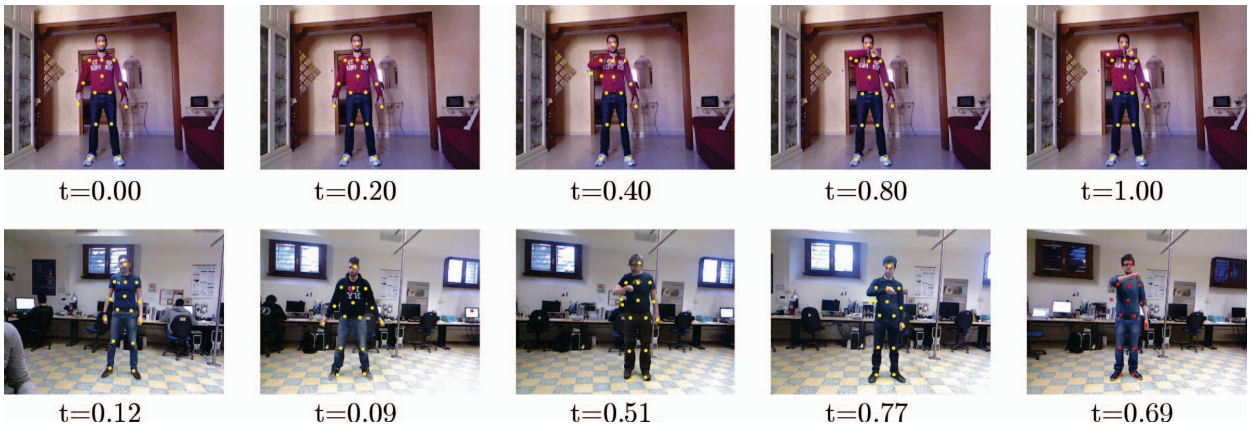

Figure 2. Weak alignment between query (top) and (bottom) reconstructed sequence from the correct class with the normalized time-stamps.

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.

3.3. Multi-part models

Following the approach proposed in [14] 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:

$$\arg\max_{C_k} = \frac{1}{S}\sum_{p\in P}\sum_{i=1}^{S}\exp\left(\|\delta_p \circ h_i - NN^{C_k}(\delta_p \circ h_i)\|^2/\sigma_p^2\right), \tag{5}$$

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

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

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 $sigma$ in Eq. 5. Note that we are not learning the feature representation, but the key idea is to align separately meaningful body parts. Our system seeks the best sequence able to independently align the subparts. As an example, if the full body feature may be noisy, the classifier will 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 two datasets: the Florence 3D Action Dataset [8]; and the Microsoft (MSR) Daily Activity 3D dataset [14]. Results scored by our approach on these datasets were also compared against those obtained by state of the art solutions.

4.1. Florence 3D Action dataset

The dataset collected at the University of Florence during 2012 [8], has been captured using a Kinect camera. It includes 9 activities: *wave*, *drink from a bottle*, *answer phone*, *clap*, *tight lace*, *sit down*, *stand up*, *read watch*, *bow*. During acquisition, 10 subjects were asked to perform the above actions for 2/3 times. This resulted in a total of 215 activity samples. As an example, frames in Fig. 2 are extracted from a *read watch* sequence used for test (upper line), and from *read watch* training sequences (lower line) of this dataset.

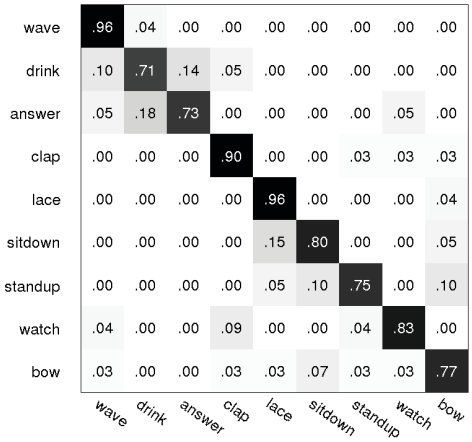

Figure 3. Florence 3D action dataset: Confusion matrix.

The confusion matrix for our approach on this dataset is reported in Fig. 3. The results are given for the multi-part variant of our approach that resulted superior to the basic solution as compared in Sect. 4.3. From the table it can be observed that, as expected, the larger classification errors are observed for actions that involve same joint groups, like for *drink from a bottle* and *answer phone* or for the *bow* which is confused with *sitdown*. It can be also observed as the temporal ordering accounted in the frame descriptors helps in reasonably distinguishing between actions that present same body postures but performed in different temporal ordering.

4.2. MSR Daily Activity 3D dataset

The Daily Activity 3D dataset was captured at Microsoft Research using a Kinect device [14]. 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.

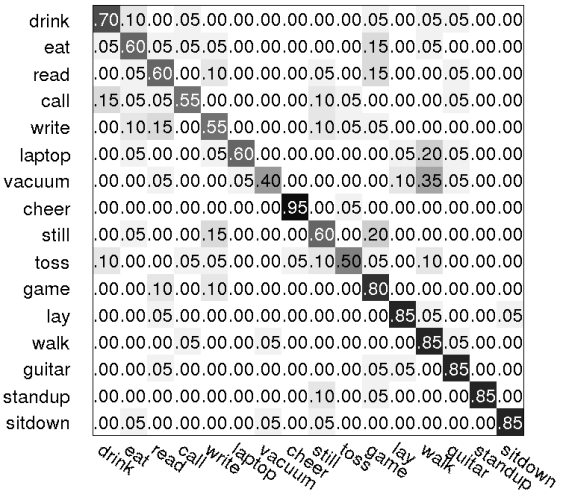

Figure 4. MSR Daily Activity dataset: Confusion matrix.

Experiments have been conducted, on both datasets, 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 in Fig. 4. For this dataset it can be noted as the more 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.

4.3. Comparative evaluation

The proposed approach has been also evaluated through a comparative analysis, which is reported in Tab. 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 both the Florence 3D and 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 yield state of the art results. It can be also noted that the improvement obtained with the "time" and "parts" features is lower for the Florence 3D Action dataset that has less classes and action samples are shorter on average. The Florence3D dataset is probably less difficult than MSR Daily because only a few actions are performed through external object interactions.

Method	Florence 3D	MSR Daily
NBNN + parts + time	**0.82**	**0.70**
NBNN + time	0.81	0.60
NBNN + parts	0.81	0.60
NBNN	0.78	0.53
Actionlets [14]	-	0.68
DTW [10]	-	0.54

Table 1. Recognition accuracy comparison for the Florence 3D and the MSR Daily Activity 3D datasets. For the method in [14], results obtained using only the joints position are reported.

On the MSR Daily Activity 3D dataset, we also compared our approach with the solutions reported in [10] and [14]. The solution in [10] uses Dynamic Temporal Warping (DTW) to match the 3D joint positions to a template, and action recognition can be done through a nearest-neighbor classification method. The method in [14], instead, uses the estimated 3D joint positions and a Local Occupancy Pattern (LOP) 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 [14] obtained only using the joints positions, as given by the authors. In both datasets we can observe that actions characterized by the same articular groups can be mistaken: {*tight lace*, *sit down*, *stand up*} and {*answer phone*, *drink*,

wave} on Florence 3D dataset; 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*.

5. Discussion and conclusions

In this work, we have proposed a method for human action recognition which is based on weakly aligning the 3D coordinates of joints in multiple parts of the skeleton. The approach first defines a kinematic representation of the human body which results into four chains, each modeling a limb. The 3D coordinates of each joint in a chain are then expressed in a locally defined reference system, which permits coordinates invariance with respect to rotations and translations. In the proposed basic approach, the coordinates of the joints are used as feature vector representing the human body in each frame. This basic solution is then extended with the use of temporal derivatives of the coordinates as well as with a temporal feature. In order to make the approach robust to noise, a part based solution has been also deployed with permits alignment of sub-sets of the joints. Both these extensions resulted beneficial in improving the performance of the approach. In all the cases, a sequence-to-class nearest-neighbor classifier has been used to score the similarity of a query action. Experiments carried out on two benchmark datasets support the applicability of the proposed solution. When compared to other skeletal-based solution our approach shows competitive performance.

Achieved results are still lower than those obtained by state of the art hybrid methods that exploit both joint and depth map information. We remark that the main aim of this work was to show the powerful of information that can be extracted from the 3D skeleton only, without requiring the additional processing of the entire depth maps of a sequence. The investigation of how to extend our solution also including such information is left as future work.

References

[1] A. Baak, M. Muller, G. Bharaj, H.-P. Seidel, and C. Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Proc. of Int. Conf. on Computer Vision*, pages 1092–1099, Barcelona, Spain, Nov. 2011.

[2] A. D. Bagdanov, A. Del Bimbo, L. Seidenari, and L. Usai. Real-time hand status recognition from RGB-D imagery. In *Proc. of Int. Conf. on Pattern Recognition*, pages 2456–2459, Tsukuba, Japan, Nov. 2012.

[3] S. Berretti, A. Del Bimbo, and P. Pala. Superfaces: A super-resolution model for 3D faces. In *Proc. of Work. on Non-Rigid Shape Analysis and Deformable Image Alignment*, pages 73–82, Florence, Italy, Oct. 2012.

[4] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classication. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage, Alaska, USA, June 2008.

[5] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian. Human daily action analysis with multi-view and color-depth data. In *Proc. of Work. on Consumer Depth Cameras for Computer Vision*, pages 52–61, Florence, Italy, Oct. 2012.

[6] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *Proc. of Int. Conf. on Computer Vision*, Barcelona, Spain, Nov. 2011.

[7] S. Hadfield and R. Bowden. Kinecting the dots: Particle based scene flow from depth sensors. In *Proc. of Int. Conf. on Computer Vision*, pages 2290–2295, Barcelona, Spain, Nov. 2011.

[8] http://www.micc.unifi.it/vim/datasets/3dactions/. Florence 3D action dataset, 2013.

[9] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *Proc. of Work. on Human Communicative Behavior Analysis*, pages 9–14, San Francisco, California, USA, June 2010.

[10] M. Muller and T. Röder. Motion templates for automatic classication and retrieval of motion capture data. In *Proc. of ACM SIGGRAPH/Eurographics Symp. on Computer Animation*, pages 137–146.

[11] M. Raptis, D. Kirovski, and H. Hoppe. Real-time classification of dance gestures from skeleton animation. In *Proc. of ACM SIGGRAPH/Eurographics Symp. on Computer Animation*, pages 147–156, Vancouver, Canada, Aug. 2011.

[12] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from single depth images. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Colorado Springs, Colorado, USA, June 2011.

[13] J. Wang, Z. Liu, J. Chorowski, Z. Chen, and Y. Wu. Robust 3D action recognition with random occupancy patterns. In *Proc. of European Conf. on Computer Vision*, pages 1–8, Florence, Italy, Oct. 2012.

[14] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Providence, Rhode Island, USA, June 2012.

[15] L. Xia, C.-C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *Proc. of Work. on Human Activity Understanding from 3D Data*, pages 20–27, Providence, Rhode Island, USA, June 2012.

[16] X. Yang and Y. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Proc. of Work. on Human Activity Understanding from 3D Data*, pages 14–19, Providence, Rhode Island, USA, June 2012.

[17] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proc. of ACM Int. Conf. on Multimedia*, pages 1057–1060, Nara, Japan, Oct. 2012.