Space-Time Pose Representation for 3D Human Action Recognition

Maxime Devanne[1,2,3], Hazem Wannous[1], Stefano Berretti[3], Pietro Pala[3], Mohamed Daoudi[2], and Alberto Del Bimbo[3]

[1] University of Lille 1 - LIFL (UMR Lille1/CNRS 8022)
[2] Institut Mines-Telecom
[3] University of Firenze

Abstract. 3D human action recognition is an important current challenge at the heart of many research areas lying to the modeling of the spatio-temporal information. In this paper, we propose representing human actions using spatio-temporal motion trajectories. In the proposed approach, each trajectory consists of one motion channel corresponding to the evolution of the 3D position of all joint coordinates within frames of action sequence. Action recognition is achieved through a shape trajectory representation that is learnt by a K-NN classifier, which takes benefit from Riemannian geometry in an open curve shape space. Experiments on the MSR Action 3D and UTKinect human action datasets show that, in comparison to state-of-the-art methods, the proposed approach obtains promising results that show the potential of our approach.

Keywords: 3D human action, activity recognition, temporal modeling.

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, scene flow estimation, hand gesture recognition, face super-resolution. 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 this context, 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

A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, LNCS 8158, pp. 456–464, 2013.

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. [5], 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 [12] 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 [4].

In [13], human actions recognition is obtained by extracting three features for each joint which are based on pair-wise differences of joint positions: differences between joints in the current frame; between joints in the current frame and in 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.

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. Several approaches are used for action recognition like 3D silhouettes [4], *Comparative Coding Descriptor* [2], or *Histogram of Oriented Gradient* (HOG) on *Depth Motion Maps* (DMM) [14]. Other methods represent the action sequence as a 4D shape and extract *Spatio-Temporal Occupancy Pattern* features (STOP) [9], or *Random Occupancy Pattern* features (ROP) [10].

Hybrid solutions try to combine positive aspects of both skeleton and depth-map based methods. The approach in [11] proposes a *Local Occupancy Pattern* (LOP) around each 3D joint as local feature for human body representation.

Relying on the observation that most human gestures can be recognized using only the shape of the skeleton of the human body, most of the human action approaches focus on the positions of 3D joints as features for recognition. The most important advantage of these features is that they are easy to extract with new depth cameras. Except that, the choice of good features to model the shape

of the human body is not the only issue in human action recognition. Even if accurate 3D joints positions are available, action recognition task is still difficult due to significant spatial and temporal variations in an action for different, or even the same, actor. Feature space representation and similarity metric are also important factors for recognition effectiveness.

1.2 Proposed Approach

In this paper, we explore the joint positions as gesture representation and we model the dynamics of the full skeleton as a trajectory using shape analysis on Riemannian manifolds for human actions recognition. Our proposal in this work is motivated by: (1) The fact that many features in computer vision applications lie on curved space due to the geometric nature of the problems; (2) The shape and dynamic cues are very important for modeling human activity and their effectiveness have been demonstrated in several works in the state-of-the-art [8,1]; (3) Using such manifold offers a wide variety of statistical and modeling tools for gesture and action recognition.

The rest of the paper is organized as follows: Sect. 2 describes our approach including the spatio-temporal representation of action, the elastic metric used to compare action sequences and the recognition method used for classification; Sect. 3 discusses about the experimental results; Sect. 4 concludes the paper also prospecting future research directions.

2 Spatio-temporal Representation

In this work, 3D human actions are represented by spatio-temporal motion trajectories of pose vectors in an Euclidian space. Trajectories are represented as curves in the Riemannian manifold of open curve shape space in order to model the dynamics of temporal variations of pose as the action progresses. The shape of each trajectory is viewed as a point on the shape space of open curves and, hence, the similarity between two trajectories is qualified by an elastic distance between their corresponding points in shape space. Finally, a classification process is performed on shape space manifold. This approach is schematized in Fig. 1.

2.1 Space of Trajectories

Using the Kinect, we can easily obtain in real-time the 3D location of body parts, called joints. In each frame, the 3D positions of 20 joints are available. As there are 20 joints and each has 3 coordinates, the whole body pose at each frame can be represented by a vector in a 60-dimensional space (*pose space*). An instance of action will be regarded as a *trajectory of poses* or an open curve in the Euclidian space. Each trajectory consists of one motion channel corresponding to the evolution of all 3D joint coordinates. This is summarized on the left of

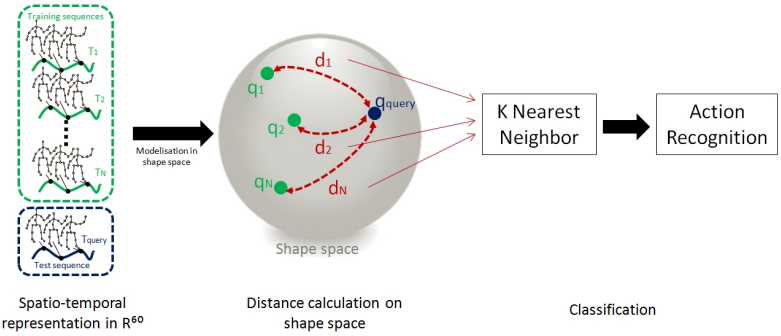

Fig. 1. Overview of our approach

Fig. 1, where each action is represented by a curve (for visualization, curves are shown in 2D, but they actually lie in a 60-dimensional space).

For each action sequence, we have a corresponding trajectory in a space of 60 dimensions. As the 3D position of each joint is represented by 3 different dimensions among the 60 dimensions, we are interested in the evolving shapes of these trajectories (curves) during actions. This motivated us to analyze the shape of the trajectories in order to compare action sequences. For this purpose, a measure representing the distance between the shape of two curves is needed. Since the actions are not realized at the same speed, and do not start and finish at the same time, the distance should be invariant to the temporal elasticity.

2.2 Trajectory Projection in Shape Space

In order to analyze human action trajectories independently to the elasticity (speed, time), we employ an elastic metric within a Riemannian shape space. Since a manifold is considered as a topological space which is locally similar to an Euclidean space, it can be seen as a continuous surface lying in a higher dimensional Euclidean space [3].

We can therefore represent the trajectory by $\beta : I \to \mathbb{R}^{60}$, for an interval $I = [0,1]$. To analyze the shape of β, we shall represent it mathematically using a square-root representation. We define its square-root velocity function (SRVF) $q : I \to \mathbb{R}^{60}$, given by:

$$q(t) \doteq \frac{\dot{\beta}}{\sqrt{\|\dot{\beta}\|}} \tag{1}$$

where $q(t)$ is a special function introduced in [3] that captures the shape of β and is particularly convenient for shape analysis. Its effectiveness has been shown in [6] for human body extremal curves in $\mathbb{R}^3$ in order to compute poses similarities. Our goal is to extend its use to spatio-temporal trajectories in $\mathbb{R}^n$.

As shown in [3], the $\mathbb{L}^2$ norm represents the elastic metric to compare the shape of two curves, under the SRVF representation. We define the set of curves:

$$\mathcal{C} = \{q : I \to \mathbb{R}^{60} | \|q\| = 1\} \subset \mathbb{L}^2(I, \mathbb{R}^{60}). \tag{2}$$

With the $\mathbb{L}^2$ norm on its tangent space, $\mathcal{C}$ becomes a Riemannian manifold and the distance between two elements of this manifold, q_1 and q_2, is given by:

$$d_c(q_1, q_2) \doteq cos^{-1}(\langle q_1, q_2 \rangle) \tag{3}$$

This distance measures the geodesic length between two trajectories represented in the manifold $\mathcal{C}$.

In our case, we need to compare the shape of the trajectories independently of the elasticity. So, we need to be invariant to the re-parametrization of the curves. We define the parametrization group Γ which is the set of all orientation-preserving diffeomorphisms of I to itself. The elements $\gamma \in \Gamma$ are the re-parametrization functions. For a curve $\beta : I \to \mathbb{R}^{60}$, $\gamma \circ \beta$ is a re-parametrization of β. As shown in [7], the SRVF of $\gamma \circ \beta$ is given by $\sqrt{\dot{\gamma}(t)}(q \circ \gamma)(t)$. We define the equivalent class containing q as:

$$[q] = \{\sqrt{\dot{\gamma}(t)}(q \circ \gamma)(t) | \gamma \in \Gamma\}. \tag{4}$$

The set of such equivalence classes is called the shape space of elastic curves, noted $\mathcal{S}$. In practise, dynamic programming is performed for optimal re-parametrization.

The shortest geodesic path between $[q_1]$ and $[q_2]$ in the shape space of open curves $\mathcal{S}$ is given by:

$$\alpha(\tau) = \frac{1}{\sin(\theta)} \left(\sin((1-\tau)\theta)q_1 + \sin(\theta\tau)q_2^*\right) \ , \tag{5}$$

where $\theta = d_s([q_1], [q_2]) = d_c(q_1, q_2^*)$.
In the above equations, q_2^* is the optimal element associated with the optimal re-parametrization γ^* of the second curve q_2. This defined distance allows comparing the trajectories shape regardless to elastic deformation.

2.3 Recognition Algorithm

Let $\{(X_i, y_i)\}$, $i = 1, \ldots, N$, be the training set with respect to class labels, where $X_i \in \mathbb{M}$, $y_i \in \{1, \ldots, N_c\}$, where N_c is the number of classes and $\mathbb{M}$ is a Riemannian manifold. We want to find a function $\mathrm{F(X)} : \mathrm{M} \longmapsto 1, \ldots, N_c$ for clustering data lying in different submanifolds of a Riemannian space, based on the training set of labeled items of the data. To this end, we propose a K-Nearest-Neighbor classifier on the Riemannian manifold, learned by the trajectories modeled on the open curve shape space. Such learning method exploits geometric properties of the open curve shape space, particularly its Riemannian metric. This indeed relies only on the computation of the (geodesic) distances to the nearest neighbors of each data point of training set.

The action recognition problem is reduced to classification in Riemannian space. More precisely, given a set of training trajectory samples $X_i : i = 1, \ldots, N$, they are represented by the underlying points $q_i : i = 1, \ldots, N$, which map trajectories on the shape space manifold (see the mapping between trajectories

and the shape space sphere in the middle of Fig. 1). Then, for any trajectory query sample X_q, a point representation q_q is obtained by mapping on the shape space manifold. Finally, a geodesic-based classifier is performed to find the K-closed trajectories of the query samples and to label X_q using the elastic metric computed via tangent spaces as given in Eq. (3).

3 Experimental Results

The proposed approach has been evaluated on two different datasets: MSR Action 3D and UTKinect. For each dataset, we compare our approach with state of the art methods which have been evaluated on these datasets.

3.1 MSR Action 3D Dataset

The MSR Action 3D dataset is a public dataset [4] on which many methods have been evaluated. This dataset includes 20 actions performed by 10 persons facing the camera. Each action is performed 2 or 3 times. In total, 567 sequences are available. For each sequence, the dataset provides depth information, color information and skeleton information. In our case, we only use the skeleton data. As reported in [11], 10 actions are not used in the experiments because the skeletons are either missing or too erroneous. For our experiments, we use 557 sequences.

In order to fairly compare our method with the state of the art, we follow the same experimental protocol as the works evaluated on MSR Action 3D. The sequences are split into three different subsets.

For each subset, we performed three different tests: Test One, Test Two, and Cross Subject Test. In Test One, 1/3 of the subset is used as training and the rest as testing. In Test Two, 2/3 of the subset is used as training and the rest as testing. In Cross Subject Test, one half of the subjects is used as training and the second half is used as test. The Cross Subject Test is more challenging because the subjects used as training are different from those used as testing. It is therefore more representative of a real case. In all our experiments, the data was randomly split into training and test sets. The random split was repeated 10 times and the average classification accuracy is reported here. Table 1 shows a comparison with the most significant state of the art methods on MSR Action 3D. Each comparison between a training action and a test action takes 45ms. The computation time to recognize a test action is 45ms multiplied by the number of training sequences.

We obtain an average accuracy of 93.1 for the Test One, 95.3 for the Test Two, and 92.8 for the Cross Subject Test. As shown in Tab. 1, we obtain competitive accuracies in the Test One and the Test Two, compared to the methods of the state of the art. In Cross Subject Test, we outperform existing methods.

First, we can see that for each test, we obtain better results with the Action Subset 3. Indeed, the actions in this subset are very different while most of the actions in subset 1 and subset 2 are quite similar. For example, we found actions

Table 1. MSR Action 3D: We compare our method with HO3DJ [12], EigenJoints [13], STOP [9], HOG [14], and Actionlet [11]. The method obtaining the best result in each experiment is evidenced in bold.

	HO3DJ	EigenJoints	STOP	HOG	**Our Method**
AS1 One	**98.5**	94.7	98.2	97.3	90.3
AS2 One	**96.7**	95.4	94.8	92.2	91.0
AS3 One	93.5	97.3	97.4	**98.0**	**98.0**
AS1 Two	98.6	97.3	**99.1**	98.7	93.4
AS2 Two	97.9	**98.7**	97.0	94.7	93.9
AS3 Two	94.9	97.3	**98.7**	**98.7**	98.6
AS1 CrSub	88.0	74.5	84.7	96.2	**90.1**
AS2 CrSub	85.5	76.1	81.3	84.1	**90.6**
AS3 CrSub	63.5	96.4	84.8	94.6	**97.6**

using hands or feet in subset 3 like *high throw* and *forward kick*. In subset 1 and 2, most of the actions are using only the hands, and especially only the left hand like *hammer*, *draw circle*, *forward punch*, or *draw X*.

We can also notice that we obtain similar accuracies for each subset regardless the test performed. Even if the subject who is performing the action is not present in the training set, we obtain good accuracies. Indeed, thanks to our spatio-temporal representation, an action performing by two different subjects are represented by similar trajectories in term of the shape. In real case, the subject performs an action for the first time in front of the recognition system. The Cross Subject Test is therefore the most representative test of a real case.

Finally, we observed that the accuracies are very low for some actions like *hammer* and *hand catch*, compared to the other actions. This can be explained by the fact that the way of performing these two actions varies a lot depending on the subjects. For example, some subjects repeat two or three times these actions while other subjects performs each action only once.

3.2 UTKinect Dataset

In order to confirm the effectiveness of our approach, we also evaluate the proposed method on a second dataset: UTKinect [12]. In this dataset, 10 subjects perform 10 different actions two times, for a total of 200 sequences. The actions include: *walk*, *sit-down*, *stand-up*, *pick-up*, *carry*, *throw*, *push*, *pull*, *wave* and *clap-hand*. The dataset provides color information, depth information, and skeleton information. This dataset presents three main challenges: First, the action sequences are registered from different views; Second, there is human-object interaction for some actions; Third, another difficulty is added by the presence of occlusions, caused by human-object interaction or by the absence of some body parts in the field of view.

To be comparable to the work in [12], we follow the same experimental protocol. We use the Leave One sequence Out Cross Validation method (LOOCV).

For each iteration, one sequence is used as test and all others sequences are used as training. The operation is repeated such that each sequence is used once as testing. We obtain an accuracy corresponding to the mean value of the accuracies obtained in each iteration. We also compute a mean accuracy obtained for each action separately (see Tab. 2).

Table 2. UTKinect dataset: We compare our method with HO3DJ [12]

Action	Walk	Sit	Stand	Pickup	Carry	Throw	Push	Pull	Wave	Clap	**Overall**
HO3DJ	**96.5**	91.5	93.5	97.5	**97.5**	59.0	81.5	92.5	**100**	**100**	90.9
Our	90.0	**100**	**100**	**100**	68.4	**95**	**90**	**100**	**100**	80.0	**91.5**

As we can see in the Tab. 2, we obtain an accuracy similar to the work in [12]. We remark that most of the wrongly classified sequences are due to actions that include human-object interaction. As our skeleton based approach is not able to detect objects, we expect these sequences to be the main source of error for our method. To investigate this point, we manually removed sequences with human-object interaction (*pick-up*, *carry*, *throw*) and repeated the classification experiments on this reduced dataset. We can see in Tab. 3 that removing actions with human-object interaction substantially improve the accuracy.

Table 3. Reduced version of UTKinect: Comparison of our method with HO3DJ [12]

Action	Walk	Sit	Stand	Push	Pull	Wave	Clap	**Overall**
HO3DJ	96.5	91.5	93.5	81.5	92.5	**100**	**100**	90.9
Our 100	**100**	**100**	**95**	**100**	**100**	**100**	80.0	**96.4**

4 Conclusions

We have proposed an effective human action recognition method by using a spatio-temporal motion trajectory representation. We take as input the 3D position of each joint of the skeleton in each frame of the sequence and use them to compute a corresponding trajectory. To compare the shape of the trajectories, we compute a distance between the projected trajectories in a shape space. Finally, we use a K-Nearest-Neighbor method to classify the action sequences which takes benefits from Riemannian geometry in open curve shape space. The experimental results on MSR Action 3D and UTKinect demonstrate that our approach outperforms the existing state-of-the-art in some cases. As future work, we plan to investigate other descriptors based on both depth and skeleton information to manage the problem of human-object interaction. We also plan to analyse and deal with specific cases where our method gives lower accuracies, like in sequences where actions are performed more than once. Finally, we would like to explore different applicative contexts and other available datasets.

References

1. Abdelkader, M.F., Abd-Almageed, W., Srivastava, A., Chellappa, R.: Silhouette-based gesture and action recognition via modeling trajectories on riemannian shape manifolds. Computer Vision and Image Understanding 115(3), 439–455 (2011)
2. Cheng, Z., Qin, L., Ye, Y., Huang, Q., Tian, Q.: Human daily action analysis with multi-view and color-depth data. In: Proc. of Work. on Consumer Depth Cameras for Computer Vision, Florence, Italy, pp. 52–61 (October 2012)
3. Joshi, S.H., Klassen, E., Srivastava, A., Jermyn, I.: A novel representation for riemannian analysis of elastic curves in rn. In: Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern. Recognit., July 16 (2007)
4. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: Proc. of Work. on Human Communicative Behavior Analysis, San Francisco, California, USA, pp. 9–14 (June 2010)
5. 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: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Colorado Springs, Colorado, USA, pp. 1–8 (June 2011)
6. Slama, R., Wannous, H., Daoudi, M.: Extremal human curves: a new human body shape and pose descriptor. In: 10th IEEE International Conference on Automatic Face and Gesture Recognition, Shanghai (2013)
7. Srivastava, A., Klassen, E., Joshi, S.H., Jermyn, I.: Shape analysis of elastic curves in euclidean spaces. IEEE Trans. Pattern Anal. Mach. Intell. 33, 1415–1428 (2011)
8. Turaga, P., Veeraraghavan, A., Srivastava, A., Chellappa, R.: Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(11), 2273–2286 (2011)
9. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.F.: Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. In: 17th Iberoamerican Congress on Pattern Recognition, Buenos Airies (2012)
10. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3D action recognition with random occupancy patterns. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 872–885. Springer, Heidelberg (2012)
11. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Providence, Rhode Island, USA, pp. 1–8 (June 2012)
12. Xia, L., Chen, C.C., Aggarwal, J.K.: View invariant human action recognition using histograms of 3D joints. In: Proc. of Work. on Human Activity Understanding from 3D Data, Providence, Rhode Island, USA, pp. 20–27 (June 2012)
13. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: Proc. of Work. on Human Activity Understanding from 3D Data, Providence, Rhode Island, USA, pp. 14–19 (June 2012)
14. Yang, X., Zhang, C., Tian, Y.: Recognizing actions using depth motion maps-based histograms of oriented gradients. In: Proc. of ACM Int. Conf. on Multimedia, Nara, Japan, pp. 1057–1060 (October 2012)