3D Point Cloud Descriptor for Posture Recognition

Margarita Khokhlova, Cyrille Migniot and Albert Dipanda
Le2i, FRE CNRS 2005, Univ. Bourgogne Franche-Comté, France

Keywords: 3D Descriptor, Point Cloud Structure, 3D Posture.

Abstract: This paper introduces a simple yet powerful algorithm for global human posture description based on 3D Point Cloud data. The proposed algorithm preserves spatial contextual information about a 3D object in a video sequence and can be used as an intermediate step in human-motion related Computer Vision applications such as action recognition, gait analysis, human-computer interaction. The proposed descriptor captures a point cloud structure by means of a modified 3D regular grid and a corresponding cells space occupancy information. The performance of our method was evaluated on the task of posture recognition and automatic action segmentation.

1 INTRODUCTION

3D pose estimation is a common task in Computer Vision applications. In the case of a rigid object, pose estimation seeks to capture the appearance of an object under certain viewing conditions. This task is challenging for natural images due to the ambiguity of an object representation in 2D, poor texture and varying view-points. With the introduction of consumer 3D sensors, this problem has been revisited by researchers developing a broad range of new descriptors. They may be both handcrafted (Hinterstoisser et al., 2012) or automatic (Wohlhart and Lepetit, 2015), and capture information from both global and local scales.

Non-rigid object pose estimation is inherently more complicated. A human body is an articulated object, and its motion can be build up from rigid and non-rigid motion parts. Articulated pose estimation seeks to estimate the configuration of a human body in a given image or video sequence. Recognition of body postures is an important step towards the fully automatic classification of human motion.

A canonical work on human posture estimation using RGBD camera data is by Shotton et al. (Shotton et al., 2013), which proposes a real-time algorithm which segments a human body from a corresponding depth map and locates skeleton joints. This algorithm shows good results and its variations are widely used today. However, it has certain limitations: in presence of severe occlusions and noise, the positions of the joints cannot be estimated correctly; it gives approximate joint positions and therefore coarse pose estimation and is not able to capture very subtle variations between postures. For this reason, joint-based posture estimation methods, although simple and powerful, will fail if the initial joints were estimated wrongly, which gives the way to low-level attributes based methods.

This paper proposes a simple yet effective descriptor for pose recognition based directly on point cloud data. The algorithm takes a holistic pose estimation approach, capturing the slightest posture changes using accumulated point cloud features. Our descriptor is based on the space occupancy for cells of a modified 3D regular grid, super-imposed on a point cloud. It is translation, scale, and rotation invariant.

Originally, we aim at a descriptor which can be used for a gait analysis. The proposed design should be able to reliably detect different postures in human gait, where the precision of skeleton data is not sufficient (the Kinect reliability is evaluated by (Cippitelli et al., 2015) for the side and front (Mentiplay et al., 2015) views). The second problem addressed is the symmetry of the gait which should be evaluated based on the point cloud data. However, resulting descriptor is very general and can be used as an intermediate step in a great number of computer vision applications such as action recognition, gait analysis, smart homes, assessing the quality of sports actions, human-computer interaction and others, where posture estimation is an essential intermediate step. This work presents the descriptor in the context of action recognition, and postures are estimated from frames of video sequences from MSR Action3D database.

Khokhlova, M., Migniot, C. and Dipanda, A.
3D Point Cloud Descriptor for Posture Recognition.
DOI: 10.5220/0006541801610168
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP*, pages 161-168
ISBN: 978-989-758-290-5

The paper is organized as follows. Section II overviews existing methods for human posture recognition. Section III introduces the descriptor and its parameters. Section IV describes the data used in experiments proposed in section V. Section VI summarizes the results, proposes possible applications and outlines the future work.

2 RELATED WORK

Most methods for human pose estimation are based on variations of a so called pictorial structures model, which represents human body configuration as a collection of connected rigid parts (Chen and Yuille, 2014)(Jhuang et al., 2013)(Agarwal and Triggs, 2006)(Pishchulin et al., 2013). To model an articulation, parts of the structure are parameterized by their spatial location and orientation.

Holistic approaches (Agarwal and Triggs, 2006)(Pishchulin et al., 2013)(Vieira et al., 2012) and middle-part (Yang and Ramanan, 2011) based methods form the other research direction in posture recognition. Holistic approaches aim to directly predict positions of body parts from image features without relying on an intermediate part-based representation. Part-based approaches first detect intermediate parts independently or with some constraints on body joints spatial relations.

Recently researchers significantly advanced posture recognition from natural images with the increasing popularity of machine learning based approaches (Tompson et al., 2014)(Chéron et al., 2015)(Chen and Ramanan, 2017). Cheron et al (Chéron et al., 2015) proposed a new Pose-based Convolutional Neural Network descriptor (P-CNN) for 2D action recognition. A pre-trained CNN learns the features corresponding to 5 pre-selected body parts based on quantized motion flow data for each frame. Chen and Ramanan (Chen and Ramanan, 2017) extend an estimated 2D model, using a neural network, to 3D using a simple Nearest Neighbor pose matching algorithm. A good review on recent advances in 3D articulated pose estimation is proposed by Sarafianos et al. (Sarafianos et al., 2016). Posture recognition is a part of action recognition, since actions can be modeled as a postures evaluation in time. Recent works on action recognition are based on CNNs (Han et al., 2017)(Lan et al., 2017) and learn the the features atomically, which leads to state-of-the-art results on available datasets.

Despite the significant progress made, full-body pose estimation from natural images remains a difficult and a largely unsolved problem due to numerous difficulties in real-life applications: the many degrees of freedom of the human body model, the variance in appearance, the changes in viewpoints, and lastly, an absence of data about an objects' shape. 3D data give a new important information which allows for improving posture recognition results. Depth-based pose estimation can be categorized into two classes.

Generative approaches (Ye and Yang, 2014)(Ganapathi et al., 2012) use a geometric or probabilistic human body model and estimate a pose by minimizing the distance between the human model and the input depth map. Human pose estimation is performed by optimizing the objective function for geometric model fitting by the means of variants of iterative closest point (Ganapathi et al., 2012) and graphical models (Li et al., 2014) or pictorial structures (Charles and Everingham, 2011). A recent method by Wang et al. (Wang et al., 2016) uses several handcrafted descriptors to recognize 5 distinct postures from the data obtained by a Kinect camera. Their algorithm is based on a simple 3D-2D projection method and the star skeleton technique. The final posture descriptor is composed of skeleton feature points together with a center of gravity. A pre-trained Learned Vector Quantization (LVQ) neural network is used for classification.

Discriminative approaches (Shotton et al., 2013)(Yub Jung et al., 2015) perform classification on a pixel level and attempt to detect instances of body parts. Shotton et al. (Shotton et al., 2013) trained a random forest classifier for body part segmentation from a single depth image and used Mean Shift (Comaniciu and Meer, 2002) to estimate joint locations. Chang et al. (Chang and Nam, 2013) propose a fast random-forest-based human pose estimation method, where classifier is applied directly to pixels of the segmented human depth image. Jung et al. (Yub Jung et al., 2015) used randomized regression trees and made their algorithm even faster by estimating the relative direction to each joint to avoid computationally demanding aggregating pixel-wise tree evaluations. The obtained skeleton data can later be used as the base for action recognition in videos as in recent Log-COV-Net method (Cavazza et al., 2017).

Most of the work on 3D pose estimation uses a single depth camera. The most successful examples of single view pose estimation are (Shotton et al., 2013)(Ye and Yang, 2014)(Yub Jung et al., 2015)(Chang and Nam, 2013) and most of them use randomized trees and shape context features for pixel-wise classification which leads to real-time solutions.

Lately, multi-view depth image based posture recognition approaches acquired the attention of researchers (Shafaei and Little, 2016)(Peng and Luo,

2016). The recent framework proposed by (Shafaei and Little, 2016) uses several Kinect sensors and a deep CNN architecture. Multi-view scenarios allow to reconstruct 3D point clouds in the reference space. The authors use curriculum learning (Bengio et al., 2009) to train the system on purely synthetic data. Curriculum learning modifies the order of the training procedure, gradually increasing the complexity of the instances, which hypothetically improves the convergence speed and the quality of the final local minima.

It is clear that the currently prevailing strategy is to use Machine Learning methods, specifically randomized trees (Shotton et al., 2013)(Yub Jung et al., 2015)(Tang et al., 2014), and a huge amount of training data. Modern posture recognition methods (Shotton et al., 2013)(Shafaei and Little, 2016) have shown to be both effective and efficient in real-time posture estimation. Similar, for the following action recognition from videos, hand-crafted methods were overshadowed by deep learning based methods (Lan et al., 2017).

This work introduces a new descriptor that estimates 3D human pose from a single point cloud. We are not attempting to out perform machine-learning based algorithms (Shotton et al., 2013)(Yub Jung et al., 2015), but mostly propose a simple alternative, which does not require a priori human body model. In contrast to (Wang et al., 2016), we do not use a descriptor for a given posture but aim to use a general 3D point cloud structure. Unlike other popular descriptors (Shotton et al., 2013)(Yub Jung et al., 2015) which use depth image features, our descriptor is based on a 3D structure and therefore can be used in a multi-camera scenario.

3 DESCRIPTOR

We propose a handcrafted compact and discriminative descriptor for a single point cloud. The most similar descriptor to ours is the Space-Time Occupancy Patterns method proposed by Vieira et al. (Vieira et al., 2012) for the task of action recognition. Similar to this work, we propose to divide the 3D space by a regular grid and base our descriptor on spatial occupancy information. However, in (Vieira et al., 2012) researchers compute the final descriptor vector by reassigning weights based on cells where motion occurred. We are concentrated on a description of each static frame in order to recognize the posture in it. Other differences include the method of 3D space partitioning and descriptor cell initialization. Our partitioning is inspired by the 3D partitioning for human recognition from 3D point clouds proposed in (Essmaeel et al., 2016). Vieira et al. specifically design their method for video sequences, taking the time dimension into account. We assume that every initial frame posture is more important and temporal information can be encoded later in the process depending on the specific application. For the gait analysis and action recognition, a Hidden Markov Model can be coupled with a descriptor to capture the temporal information.

To construct our descriptor for each depth map video frame, we perform the following steps. First, the 2D-3D transformation is done to obtain a point cloud in 3D space from a depth map. We use a standard equation for basic geometric transformations:

$$X = Z * \frac{(j - c_x)}{f_x}; \qquad Y = Z * \frac{(i - c_y)}{f_y}; \qquad Z = z \tag{1}$$

where X, Y, Z are the point coordinates in 3D, j and i are the pixel coordinates, and c_x, c_y, f_x and f_y are the intrinsic matrix parameters obtained by a calibration of the Kinect camera. Then the 3D spatial partitioning is performed. The center of gravity in 3D is calculated and projected to the ground plane:

$$C(X,Y,Z) = \frac{\sum_1^n (X,Y,Z)}{n} \tag{2}$$

where n is the total number of points in the point cloud. A 3D cylinder of varying dimensions with a base center in the computed centroid projection defines the space partitioning limits. The height and radius of the cylinder are varying to adjust for the height

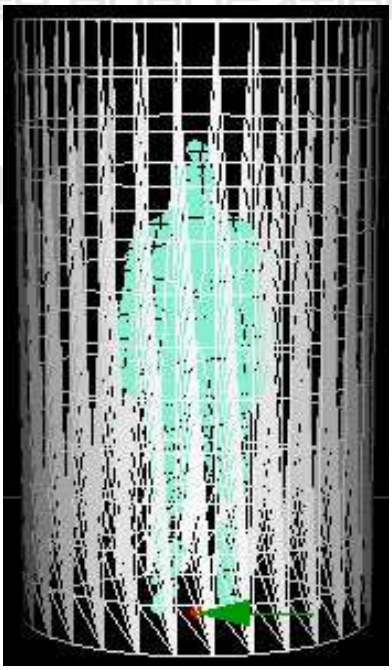

Figure 1: 3D spatial partitioning in 12 sections. Projected center of gravity is shown in red, fixed point view direction is shown by a green arrow.

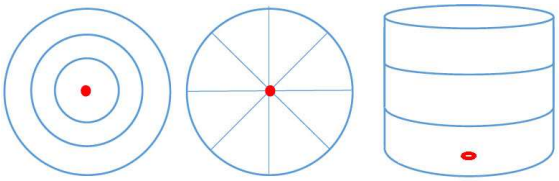

Figure 2: Descriptor spatial partitioning: 3 circles, 8 sectors, 3 sections. Projected center of gravity is shown in red.

of a person. The data about human body proportions ratio is used. A height of a person is estimated, simply via the minimum and maximum value calculated for the first static point cloud of a video sequence. To have an equal grid for all frames of a video sequence, the normal is fixed based on the viewing point. The partitioning in sectors starts from the same position for each video frame.

Figure 1 shows an example of 3D partitioning for one of the frames from MSR3D dataset. The only parameters of the descriptor are the number of sections, the number of sectors and the number of circles. An visualization of the descriptor parameters is shown in Figure 2. For this work, we use only a uniform space subdividing scheme and the cylinder volume partitioning is then performed as:

$$V = 2\pi RH \tag{3}$$

$$r_n = \frac{R}{n_c}; \qquad h_n = \frac{H}{n_h}; \qquad s_{angle} = \frac{360}{n_s} \tag{4}$$

where R and H are the fixed radius and height, r_n, h_n and s_{angle} are the circle and height intervals and the angle corresponding to each sector.

The final descriptor is obtained by calculating the number of points in each formed 3D cell i.e. the cell occupancy. The descriptor is normalized by the total number of points in the point cloud in order to compensate for possible noise or shape differences.

The OpenNI Framework (Consortium et al.,) is used by many 3D cameras and provides the user with automatic body recognition and skeleton joints extraction functionality. Therefore, we are not addressing the task of background subtraction in our work and assume that it is a prior step. For this paper, the data from an RGBD camera, where the human is located and the background is subtracted, were used to test the proposed descriptor.

Our descriptor design allows it to be used in a multiple camera views scenario to grant a more reliable and accurate pose description. For example, such partitioning was successfully employed earlier for human recognition from complete point clouds (Essmaeel et al., 2016) based on histograms of normal orientations.

4 TRAINING AND TESTING DATA

MSR Action3D Dataset (Li et al., 2010) was selected to perform the experiments and evaluate the proposed descriptor. This is one of the most used RGBD human action-detection and recognition datasets. It is also one of the first RGBD datasets capturing motions (dated 2010) and it contains a big amount of different actions performed by different persons. It consists of 20 action types performed by 10 subjects 2 or 3 times. The actions are: high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw an x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw. The resolution of the video is not very high, namely 320x240 and so is the frame rate, namely 15 fps. The data was recorded with a depth sensor similar to the Kinect device and contains color and depth video sequences. The sequences are presegmented for the background and foreground. An example of superimposed point clouds corresponding to 3 actions from MSR Action 3D dataset is shown in Figure 3. Skeleton joints data are also provided with a higher framerate than the depth maps. However, many joints are wrongly estimated, as can be seen in Figure 4. For our experiments, we had to further manually segment the dataset into key postures in 3D. There is no accurate database with full body human poses as depth maps publicly available, despite several works where the features which represent the posture are learned from real and synthetic examples (Shotton et al., 2013)(Ganapathi et al., 2010), neither the data nor the implementation of these methods are available. Recently a new multi-kinect posture dataset was published (Shafaei and Little, 2016), however, this one is huge and is not dedicated to the global pose estimation but body parts segmentation. Since we are not using any deep learning and proposing a handcrafted descriptor, we considered that a well-known and widely used MSR Action 3D will be sufficient to perform the test and training to show the capabilities

Figure 3: Three actions from MSR Action 3D dataset shown as point clouds: high arm wave, horizontal wave, golf swing.

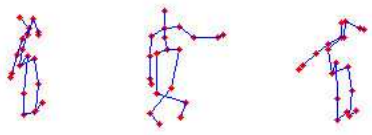

Figure 4: Examples of wrong skeleton estimation for MSR 3D dataset, actions 'High Arm Wave', 'Horizontal Arm Wave', 'Hammer'. A person is always facing the camera straight and his legs are not crossed.

Table 1: Postures selected from the MSR3D dataset.

	Posture	Training	Test
1	Staying relaxed	160	54
2	Forward Kick	102	50
3	Hand lifted 45°	29	13
4	Right hand up	137	64
5	Right hand to the left	80	71
6	Clap	59	25
7	Hands wide open	35	13
8	Pick from the ground	33	34
9	Half bend	80	53
10	Full bend	65	69
11	Right leg kick	60	40
12	Right leg kick on side	49	34
13	Throw from the back	134	78
14	Right hand up	42	28
15	Both hands left half bend	62	30
16	Both hands to the left half bend	62	30
17	Both hands to the right half bend	88	37
18	Throw from the front	54	33

and limitations of our method. In this work, we are aiming to perform a pose recognition without a skeleton aligning or human-body parts segmentation. The number of sequences for each action in MSR Action 3D dataset is between 27 and 30. We separated the data in a training and testing set, and selected between 3 to 7 key poses for each action. For the posture recognition test, 18 well distinguishable poses were selected. The resulting dataset structure is explained in Table 1. All the data acquired from person 3 were excluded from the dataset because half of the depth information was missing. Subjects 1-7 from the dataset are used for training and 8-10 for testing. The resulting dataset is not very big but corresponds to our goal to evaluate the descriptive capacities of the proposed solution.

5 EXPERIMENTS

We perform 3 series of experiments: unsupervised clustering of frames into k-postures in a video sequence, posture recognition in one action sequence and posture recognition for a set of postures. The average estimated time for the descriptor calculation (with the 2D-3D transformation performed beforehand) is 0.2 μs on a Intel C602 machine, which is compatible to the time of the extraction of feature vectors in (Wang et al., 2016).

5.1 Unsupervised K-means Clustering

A simplistic way to compare any two pose descriptors is to calculate a Euclidean distance between them. At first, we observed the dynamics of the distance changes on all the frames of a single video sequence from the dataset. Figure 5 visualizes the distance computations for the sequence 'Horizontal Wave' of MSR Action 3D dataset. There are 5 distinctive postures in this action. The result shows that there is a small distance between similar postures (i.e. consequent frames, frames in the beginning and the end of the sequence corresponding to the same 'neutral' posture). To exploit this trend further, a simple test with K-means is performed, which shows that the descriptor captures the posture difference well. Automatic key positions were obtained by performing the K-means clustering for 3 video sequences when one person is performing an action 3 times. The optimal number of basis K was estimated using the elbow method. Figure 6 shows the results of this experiment. Qualitative visual analysis shows that automatically detected poses correspond well with the 5 most different poses in the action 'Horizontal Wave' selected manually. These tests work well for each person performing a single action multiple times, but the test for the whole data gives worse results, probably due to the fact that the neutral posture is dominant in the dataset and people tend to perform similar actions differently. Hence, we obtain more intermediate clusters which do not correspond precisely to key-postures. Nevertheless, the obtained results are interesting enough to continue the tests and try to evaluate complete posture recognition based on the proposed descriptor.

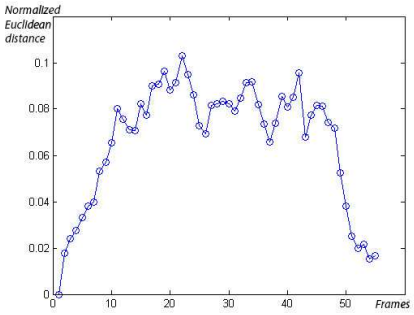

Figure 5: Pairwise descriptor distance. The video sequence starts and finishes by the same posture. The distance between consequent frames is smaller and distinct 'key' positions can be viewed as peaks of the graph.

Table 2: Classification results for 5 postures of the action 'Horizontal Wave' show good results in terms of precision.

Posture	initial	arm 45°	kick	arm front left	arm right
Precision	0.94	0.81	1	1	1
Recall	1	0.8	0.76	1	0.71
F-measure	0.97	0.82	0.86	1	0.83

5.2 Single Performance Action

A Support Vector Machine (SVM) classifier was trained, One vs All, in order to classify the postures, followed by 3-fold cross validation.

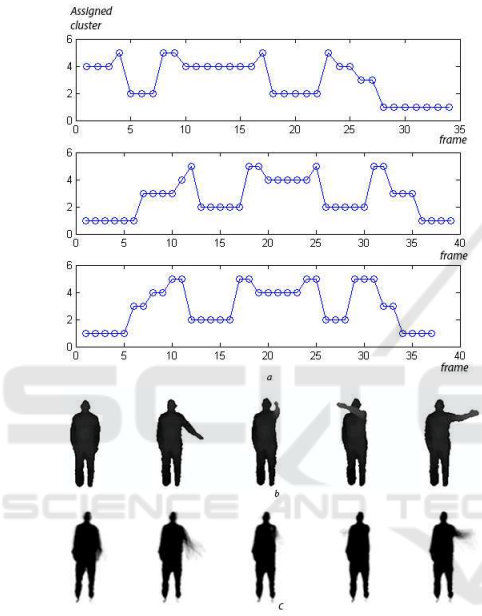

Figure 6: a) Three video sequences are shown as a succession of cluster centers. In first sequence person is starts to perform the action faster than in sequence 2 and 3; b) 5 key postures of the action 'Horizontal Wave' (selected manually); c) 5 clusters obtained automatically. Pixel values are averaged: the darker the color is, the more is the occurrence.

The F-measure, recall and precision were used to evaluate the performance of the classifier. The main criteria for our task is precision, but we included recall and F-measure parameters in order to evaluate a possibility to use the descriptor in a scenario where accurate retrieval of all postures is essential.

The results for each posture recognition for the action 'Horizontal Wave' are summarized in Table 2. Train and test data for this sequence were segmented manually according to the scheme introduced in the previous section. This simple test shows excellent results in terms of precision for all but one posture.

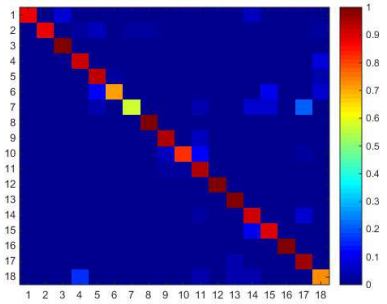

Figure 7: Confusion matrix for the SVM-based classification shows good results for all postures but one.

5.3 Set Retrieval Performance

The test for a single action posture estimation shows good results, hence we conducted an extended version of this test containing a bigger number of various postures. A full test for 18 postures was performed with an SVM. Feature vectors of the selected postures were used for training and testing. Figure 7 shows the confusion matrix for the classes obtained by the SVM. The descriptor parameters (number of sections, circles and sectors) were tuned for the best performance. We obtained the best results with 12 sections, 10 circles and 10 sectors, corresponding average precision is 0.94. The parameter tuning is straight-forward and shows that the different parameters combinations do not have much of an effect on performance. The main observation is that for postures selected the most important parameter is the number of sections which helps to separate the volume by vertical planes. Different combinations of parameters can give slightly better or worse results in terms of precision, recall and F-measure. Corresponding curves obtained for different parameters are shown in Figure 8.

The results show good performance in terms of precision which is excellent for simple postures. Our results are comparable with the results of (Wang et al., 2016) where authors are using only 5 distinct postures: standing, sitting, stooping, kneeling and lying. Of these, several postures are similar to ours, plus we are aiming at more complex and varied postures. The original dataset of (Wang et al., 2016) is not available, but we also performed a test with just 3 very different postures and a similar amount of training and testing data. As before, the training data and test data are formed from different subjects. Our postures are: staying, right-hand up, bending. The corresponding numbers of training and testing images are: 384/125, 246/125, and 98/103. With this small dataset we obtain excellent results in term of precision and recall, all the tests are assigned correctly. Our results and the results from (Wang et al., 2016) can not

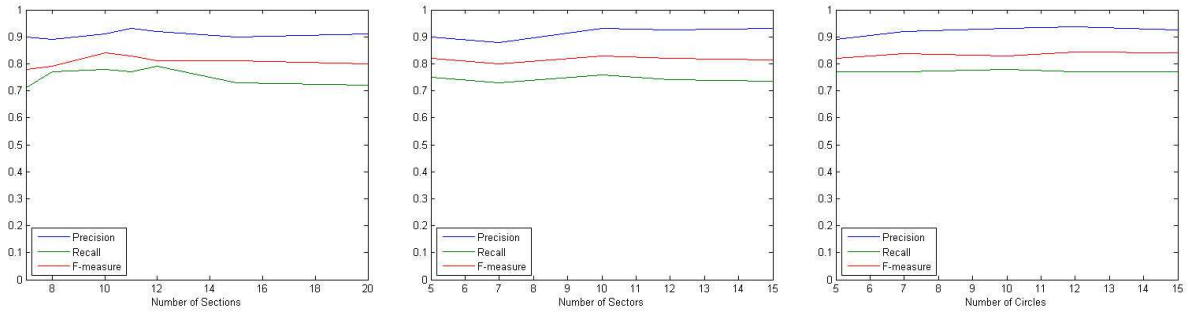

Figure 8: Tuning of the parameters. Precision, recall and F-measure curves for a) the number of section varies, sectors and circles fixed to 10; b) the number of sectors varies, sections and circles fixed to 10; the number of circles varies, sections and sectors are fixed to 10.

be directly compared, but this test gives an idea about the descriptor capabilities. Wang et al. test their posture recognition method on 80-100 depth images taken each for 8 persons. The recognition rate is also very high, with some minor errors (for example, for the first person the recognition rate is: 79/80, 99/100, 80/80, 80/80, 79/80). It should be mentioned, that (Wang et al., 2016) uses same subjects for testing and training, which is probably easier as we have shown in our tests from the previous section.

6 CONCLUSIONS

This paper shows that body pose may be adequately represented without joint estimation. The proposed descriptor can be used exclusively, or as an advantageous addition to tradition skeleton-joints estimation methods.

The introduced descriptor works well for capturing the 3D spatial arrangement of a point cloud structure. Experiments show that our method achieves competitive results compared to current hand crafted state of the art descriptors. Learned or trained descriptors may give superior performance but critically depend on the availability of large amounts of labeled data. Secondly, these architectures don't generalize outside their initial domain. Our algorithm is a simple and elegant solution, when joint information is not available or unreliable.

Example applications include action recognition and gait analysis. For the latter, the descriptor may be deployed for cycle event or symmetry detection and evaluation (Auvinet et al., 2015). Another possibility is to be able to divide a video along the time axis using posture information in the case of misalignment. Detected postures can be used to temporally align the data or as key-words describing the action.

There are number of open issues. The descriptor is noise sensitive, which becomes more apparent if part of the depth data is missing. Secondly, the Euclidean distance metric between two descriptor vectors currently excludes 3D spatial information. Semantically different postures can thus result in descriptor vectors that are near similar.

Future work will address these issues next to developing an application for real-time gait cycle event recognition.

REFERENCES

Agarwal, A. and Triggs, B. (2006). Recovering 3d human pose from monocular images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):44–58.

Auvinet, E., Multon, F., and Meunier, J. (2015). New lower-limb gait asymmetry indices based on a depth camera. *Sensors*, 15(3):4605–4623.

Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48. ACM.

Cavazza, J., Morerio, P., and Murino, V. (2017). When kernel methods meet feature learning: Log-covariance network for action recognition from skeletal data. *arXiv preprint arXiv:1708.01022*.

Chang, J. Y. and Nam, S. W. (2013). Fast random-forest-based human pose estimation using a multi-scale and cascade approach. *ETRI Journal*, 35(6):949–959.

Charles, J. and Everingham, M. (2011). Learning shape models for monocular human pose estimation from the microsoft xbox kinect. In *IEEE International Conference on Computer Vision Workshops (ICCV Workshops),*, pages 1202–1208.

Chen, C.-H. and Ramanan, D. (2017). 3d human pose estimation= 2d pose estimation+ matching. *Computer Vision and Pattern Recognition (CVPR)*.

Chen, X. and Yuille, A. L. (2014). Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, pages 1736–1744.

Chéron, G., Laptev, I., and Schmid, C. (2015). P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226.

Cippitelli, E., Gasparrini, S., Spinsante, S., and Gambi, E. (2015). Kinect as a tool for gait analysis: validation of a real-time joint extraction algorithm working in side view. *Sensors*, 15(1):1417–1434.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

Consortium, O. et al. Openni, the standard framework for 3d sensing. *URL as accessed on*. 2017-09-30.

Essmaeel, K., Migniot, C., and Dipanda, A. (2016). 3d descriptor for an oriented-human classification from complete point cloud. In *VISIGRAPP (4: VISAPP)*, pages 353–360.

Ganapathi, V., Plagemann, C., Koller, D., and Thrun, S. (2010). Real time motion capture using a single time-of-flight camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 755–762.

Ganapathi, V., Plagemann, C., Koller, D., and Thrun, S. (2012). Real-time human pose tracking from range data. In *European conference on computer vision*, pages 738–751. Springer.

Han, Y., Zhang, P., Zhuo, T., Huang, W., and Zhang, Y. (2017). Video action recognition based on deeper convolution networks with pair-wise frame motion concatenation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 8–17.

Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G. R., Konolige, K., and Navab, N. (2012). Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV (1)*, pages 548–562.

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., and Black, M. J. (2013). Towards understanding action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3192–3199.

Lan, Z., Zhu, Y., Hauptmann, A. G., and Newsam, S. (2017). Deep local video feature for action recognition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1219–1225.

Li, S., Liu, Z.-Q., and Chan, A. B. (2014). Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 482–489.

Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3d points. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 9–14.

Mentiplay, B. F., Perraton, L. G., Bower, K. J., Pua, Y.-H., McGaw, R., Heywood, S., and Clark, R. A. (2015). Gait assessment using the microsoft xbox one kinect: Concurrent validity and inter-day reliability of spatio-temporal and kinematic variables. *Journal of biomechanics*, 48(10):2166–2170.

Peng, B. and Luo, Z. (2016). Multi-view 3d pose estimation from single depth images. Technical report, Technical report, Stanford University, USA, Report, Course CS231n: Convolutional Neural Networks for Visual Recognition.

Pishchulin, L., Andriluka, M., Gehler, P., and Schiele, B. (2013). Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595.

Sarafianos, N., Boteanu, B., Ionescu, B., and Kakadiaris, I. A. (2016). 3d human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1–20.

Shafaei, A. and Little, J. J. (2016). Real-time human motion capture with multiple depth cameras. In *IEEE 13th Conference on Computer and Robot Vision (CRV)*, pages 24–31.

Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., and Moore, R. (2013). Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124.

Tang, D., Jin Chang, H., Tejani, A., and Kim, T.-K. (2014). Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3786–3793.

Tompson, J. J., Jain, A., LeCun, Y., and Bregler, C. (2014). Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807.

Vieira, A., Nascimento, E., Oliveira, G., Liu, Z., and Campos, M. (2012). Stop: Space-time occupancy patterns for 3d action recognition from depth map sequences. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 252–259.

Wang, W.-J., Chang, J.-W., Haung, S.-F., and Wang, R.-J. (2016). Human posture recognition based on images captured by the kinect sensor. *International Journal of Advanced Robotic Systems*, 13(2):54.

Wohlhart, P. and Lepetit, V. (2015). Learning descriptors for object recognition and 3d pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3109–3118.

Yang, Y. and Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1385–1392.

Ye, M. and Yang, R. (2014). Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2345–2352.

Yub Jung, H., Lee, S., Seok Heo, Y., and Dong Yun, I. (2015). Random tree walk toward instantaneous 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2467–2474.