Article

Pose Recognition of 3D Human Shapes via Multi-View CNN with Ordered View Feature Fusion

Hui Wang^{1,2}, Peng He¹ , Nannan Li^{3,*} and Junjie Cao^{4,5}

¹ School of Information Science and Technology, Shijiazhuang Tiedao University, Shijiazhuang 050043, China; wangh@stdu.edu.cn (H.W.); hpcalifornia@163.com (P.H.)

² Hebei Key Laboratory of Electromagnetic Environmental Effects and Information Processing, Shijiazhuang Tiedao University, Shijiazhuang 050043, China

³ Department of Information Science and Technology, Dalian Maritime University, Dalian 116086, China

⁴ School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China; jjcao@dlut.edu.cn

⁵ Key Laboratory for Computational Mathematics and Data Intelligence of Liaoning Province, Dalian University of Technology, Dalian 116024, China

* Correspondence: nannanli@dmlu.edu.cn

Received: 16 July 2020; Accepted: 14 August 2020; Published: 23 August 2020

Abstract: Rapid pose classification and pose retrieval in 3D human datasets are important problems in shape analysis. In this paper, we extend the Multi-View Convolutional Neural Network (MVCNN) with ordered view feature fusion for *orientation-aware* 3D human pose classification and retrieval. Firstly, we combine each learned view feature in an orderly manner to form a compact representation for orientation-aware pose classification. Secondly, for pose retrieval, the Siamese network is adopted to learn descriptor vectors, where their L_2 distances are close for pairs of shapes with the same poses and are far away for pairs of shapes with different poses. Furthermore, we also construct a larger 3D Human Pose Recognition Dataset (HPRD) consisting of 100,000 shapes for the evaluation of pose classification and retrieval. Experiments and comparisons demonstrate that our method obtains better results than previous works of pose classification and retrieval on the 3D human datasets, such as SHREC'14, FAUST, and HPRD.

Keywords: pose recognition; 3D human shape; multi-view CNN

1. Introduction

With the rapid development of 3D scanning and capturing technologies, an increasing amount of 3D human shape datasets have emerged, such as SHREC'14 [1], the statistical model shape [2], CAESAR [3], FAUST [4], and SURREAL [5], which usually consist of subjects in different poses. Meanwhile, some classical parametric modeling methods have been proposed for generating human subjects and poses, such as SCAPE [6] and SMPL [7]. The analysis and comparisons of classical parametric modeling methods of 3D human have been reported in surveys [8,9]. Pose is an atomic unit of gesture and action, which plays an important role in 3D human shape analysis. Thus, 3D human *pose classification and retrieval* have wide applications in computer vision and graphics [10], such as basketball or baseball shot recognition from a sports dataset and a dance pose recognition from a collection of ballet data.

However, most existing recognition approaches of 3D human datasets are focusing on shape retrieval whereby a deformable subject's shape is recognized regardless of its pose [11]. There are fewer works on 3D human pose classification and retrieval, i.e., recognizing a given pose regardless of the subject of the pose taken, such as Slama et al. [12] and Bonis et al. [13]. Furthermore, the previous shape classification and retrieval methods, such as Su et al. [14] and Pickup et al. [11],

are aimed at being orientation-invariant and cannot operate for pose recognition. Because pose is *translation-invariant* but *orientation-aware*, i.e., translating a human in a pose gets the same pose, but rotating the human would obtain a different pose demonstrated in Figure 1.

Figure 1. Rotating the left pose obtains a different one shown on the right.

In view of the above limitations of previous works, we improve the Multi-View Convolution Neural Networks (MVCNN) proposed by Su et al. [14] with a novel ordered view feature fusion to the orientation-aware 3D human pose classification and retrieval demonstrated in Figures 2 and 3, respectively. Our main technical contributions can be summarized as follows:

- Due to the max view-pooling in the original MVCNN being orientation-invariant, we combine the learned feature in each view in an orderly manner after the last convolutional layer and before the fully connected layers in the well known AlexNet [15] for pose recognition.
- For pose retrieval, the Siamese network [16–19] with two identical branches is used to learn descriptor vectors of shapes, where their L_2 distances in Euclidean space are close between pairs of shapes with the same pose and far away for pairs of shapes with different poses.
- Due to the existing 3D human datasets, such as SHREC'14 [1], the statistical model shape [2], CAESAR [3], and FAUST [4], only including hundreds of shapes, we construct a larger 3D Human Pose Recognition Dataset (HPRD), which includes 1000 subjects, with each subject in 100 poses, for the evaluation of pose classification and retrieval.

Our proposed method obtains ideal or almost ideal classification and retrieval results on the above datasets. Compared with the previous methods of pose classification [13] and retrieval [12], the proposed method also achieves better results. We will release the codes, pre-trained models, and the 3D human dataset of HPRD for future research.

Figure 2. The process of our MVCNN for orientation-aware pose classification, where the key is the ordered view feature fusion layer.

Figure 3. Pose retrieval: For the query 3D human shape (left one), our method retrieves the first six best shapes who have similar poses to that in the SHREC-RE dataset [1].

The rest of this paper is organized as follows. Related works are summarized in Section 2. In Section 3, our 3D human pose classification and retrieval based on the ordered view feature fusion multi-view CNN are introduced. Experimental results and performance evaluation of our approach are described in Section 4. Finally, we provide a conclusion and point out possible future works.

2. Related Works

Human pose recognition. As an atomic unit of action, pose is an important aspect of human communication. Accordingly, human pose recognition, i.e., classification and retrieval, has been the focus of many works in the recent past. However, most of the works regard the human pose of 2D images [20–22], videos [23], and multi-view videos [24]. There are fewer works for the pose of 3D shapes [12,13], where one [12] is only for pose classification and another [13] is only for pose retrieval. In this paper, we propose a unified end-to-end MVCNN framework for both 3D human pose classification and retrieval. It is noted that our work is different from the Human Pose Estimation (HPE), such as Ionescu et al. [25], Sarafianos et al. [26], and Luvizon et al. [27], which estimates the 2D or 3D pose of the human body in a given image or a sequence of images, while our work is to recognize shapes that have similar poses with a given one from the 3D human datasets.

Geometric deep learning. In the past several years, deep learning, in particular the Convolutional Neural Network (CNN), has shown an ability to learn powerful image features from large collections of examples. Different from images with a grid-based representation, 3D geometric shapes usually have a variety of complex representations. Thus, there is no universal framework for 3D geometric deep learning, such as multi-view CNN [14], voxel-based 3D CNN [28], and point-based CNN [29–31], graph or manifold based CNN [32]. Recent developments in 3D geometric deep learning from a representation perspective are summarized in [33]. Due to the 3D human pose that can be distinguished from certain key views, in this paper, we use the MVCNN framework [14], which is amenable to emerging CNN architectures and their derivatives, for pose classification and retrieval.

Multi-view CNN. The MVCNN framework uses a CNN to learn the feature of each projected 2D view image of the 3D model and fuses the view features using the max view-pooling for shape recognition [14]. The dominant sets are used for view clustering and pooling to improve the accuracy of 3D object recognition [34]. Multiple rendered views of shape regions in different scales are taken and processed through MVCNN for learning local shape descriptors [35]. The view images in the MVCNN are projected from the canonical form of the 3D shape for retrieval [36]. Unlike the above previous works mainly focusing on view generation and selection, we change the max view-pooling to a novel ordered view feature fusion for orientation-aware pose recognition.

Siamese network. The Siamese network proposed by Bromley et al. [16] is a pair of neural networks with the same architectures and weights for learning descriptors of images, where the distances of the descriptors are smaller for images within the same classes and larger for images in different classes. The Siamese network has wide applications in signature verification [16], face verification [17], learning patch descriptors [18,37], video tracking [19], sketch retrieval [38], and so on.

In this paper, we use the Siamese network to learn descriptors for 3D human pose retrieval, where the L_2 distances of the descriptors are close between pairs of shapes with the same poses and far away for pairs of shapes in different poses.

3. Methods

In this section, the MVCNN with ordered view feature fusion is proposed for orientation-aware 3D human pose classification in Section 3.1 and retrieval in Section 3.2.

3.1. Pose Classification

Multi-views of 3D human shape. The 3D human shapes are represented as triangular meshes, which are collections of points connected with edges forming triangles. The inputs of the MVCNN are multiple projected views of the 3D human shape. We use the classical Phong reflection model to generate the projected multi-views of triangular meshes, which are uniformly scaled to fit the viewing volume. To have a more three-dimensional message of a shape, we apply the second camera setup in the MVCNN [14], i.e., first constructing a regular dodecahedron with a total of 20 vertices and placing a virtual camera at the position of each vertex. Finally, the virtual camera takes each picture of the human body. It is worth noting that, different from the original MVCNN [14], the cameras do not rotate, so each shape renders 20 views.

Ordered view feature fusion MVCNN. In the original MVCNN for shape classification [14], the AlexNet [15] is used for processing the views, which consists of five convolutional layers $\text{Conv}_{1,\dots,5}$, some of which are followed by max-pooling layers, and three fully connected layers $\text{Fc}_{6,\dots,8}$ with a final softmax classification layer. Some dropout layers are added after the fully connected layers. In this network, the max view-pooling layer, which aggregates the information of multiple views into a single one, is placed after the last convolutional layer (Conv_5) and before the fully connected layers.

The above max view-pooling is orientation-invariant, which cannot handle the orientation-aware pose classification. We change the max view-pooling layer to the Ordered View Feature Fusion (OVFF); i.e., each view feature in the Conv_5 layer is treated as an ordered channel in the fused view feature shown in Figure 2. The input and output sizes of each layer are listed in Table 1.

Table 1. The input and output sizes of each layer of our MVCNN for pose classification, where the number of views is 20 and k is the number of pose classes.

Layer	Convolutional Kernel	Stride	Output Size
Input	-	-	$227 \times 227 \times 3$
Conv_1	$96 \times 11 \times 11$	4	$55 \times 55 \times 96$
Pool_1	3×3	2	$27 \times 27 \times 96$
Conv_2	$256 \times 5 \times 5$	1	$27 \times 27 \times 256$
Pool_2	3×3	2	$13 \times 13 \times 256$
Conv_3	$384 \times 3 \times 3$	1	$13 \times 13 \times 384$
Conv_4	$384 \times 3 \times 3$	1	$13 \times 13 \times 384$
Conv_5	$256 \times 3 \times 3$	1	$13 \times 13 \times 256$
Pool_5	3×3	2	$6 \times 6 \times 256$
OVFF	-	-	$20 \times 6 \times 6 \times 256$
Fc_6	-	-	4096
Fc_7	-	-	4096
Fc_8	-	-	k

3.2. Pose Retrieval

Network architecture. The Siamese architecture, shown in Figure 4, is composed of two branch networks sharing the same parameters. The network f of each individual branch is identical to the Multi-View CNN in Section 3.1, except that we replaced the final softmax classification layer with a d dimensional pose descriptor vector $f(s)$ for the shape s . The dropout layer of each branch has parameters that are randomly set to zero, which leads to the output of the two branches are not the same when two same 3D human shapes are passed to the Siamese network. Thus, we removed all of the dropout layers in the MVCNN after the fully connected layers in the AlexNet for each view.

Figure 4. Architecture of the Siamese network, where pairs of 3D human shapes are processed by two copies of the same MVCNN sharing the learned parameters W .

Loss. In the end, the two branches in the Siamese network are connected with a single loss layer. For the 3D human pose retrieval, we want the network to generate discriminative descriptor vectors that are close enough to each other for positive pairs (two shapes with the same pose) and far away at least by a minimum distance for negative pairs (two shapes in different poses). Similarly as [17,19], we employ the following margin contrastive loss:

$$l(s_i, s_j, y_{ij}) = y_{ij} D_{ij}^2 + (1 - y_{ij}) \max(0, \epsilon - D_{ij}^2), \quad (1)$$

where $y_{ij} \in \{0, 1\}$ indicates whether shapes s_i and s_j are within the same pose or not, ϵ is the minimum distance margin that pairs with different pairs should satisfy, and the L_2 distance D_{ij} between the descriptor vectors $f(s_i)$ and $f(s_j)$ is defined as follows:

$$D_{ij} = \|f(s_i) - f(s_j)\|_2. \quad (2)$$

The margin distance ϵ is set to be 10 for all the experiments in this paper. We find that replacing the other values of ϵ also obtains desirable results.

To express the above Siamese network more intuitively, we pick two models randomly from each pose in the SHREC-RE dataset [1] and show their 2D descriptors embedding in Figure 5, which demonstrates that shapes with the same or a similar pose are closer.

Figure 5. 2D descriptors embedding learned by the Siamese network of poses in the SHREC-RE datasets, where the same colors indicate the same poses.

4. Experiments

We evaluate the classification and retrieval methods on the following 3D human datasets: SHREC'14 [1] included the real and synthetic ones, FAUST [4], and our newly constructed datasets. The MVCNN for classification and retrieval are both pre-trained on ImageNet images from 1K categories, and then fine-tuned on all 2D views of the 3D human shapes in the training set. All of the training experiments are conducted on a desktop with a 16 GB NVIDIA Quadro GPU and 64 GB RAM.

4.1. 3D Human Datasets

SH-RE. The SHREC REal (SH-RE) dataset [1] is built from point-clouds contained within the Civilian American and European Surface Anthropometry Resource (CAESAR) [3] using the SCAPE (Shape Completion and Animation of PEople) model [6]. This dataset contains four hundred 3D human shapes represented as triangular meshes, consisting of 40 human subjects (20 males and 20 females), each in 10 poses. The ten poses and some representative subjects are demonstrated in Figures 6 and 7, respectively.

SH-SY. The DAZ Studio 3D modeling and animation software is used to create the SHREC SYnthetic (SH-SY) human dataset [1], which includes 15 different human subjects (5 males, 5 females, and 5 children), each in 20 different poses resulting in 300 human shapes in total. Some representative poses with different subjects are shown in Figure 8.

Figure 6. The ten human poses of a subject in the SHREC-RE dataset [1].

Figure 7. Ten representative human subjects of the same pose in the SHREC-RE dataset [1].

Figure 8. Fifteen representative human poses with different subjects in the SHREC-SY dataset [1].

FAUST. The FAUST [4] dataset is created by scanning human subjects with a sophisticated 3D stereo capture system. Ten subjects and 30 poses are included in this dataset, resulting in 300 human shapes. It should be noted that models have missing parts caused by occlusion and topological noise where touching body parts are fused together, shown in Figure 9.

HPRD. Since the above datasets only include hundreds of shapes, we also constructed a larger 3D Human Pose Recognition Dataset (HPRD) for better evaluation of our MVCNN based on SURREAL [5], which is a large-scale 3D human dataset with synthetically generated but realistic images of people. The synthetic 3D human shapes are created using the SMPL body model [7], whose parameters are fit by the MoSh [39] method given raw 3D MoCap marker data.

SURREAL contains more than 6 million frames of 3D human motions together with a ground truth pose, depth maps, and segmentation masks. We manually chose 100 motion frames as our human poses from SURREAL to have significant visual differences, where each pose is within 1000 subjects (500 males and 500 females), resulting in our HPRD with 100,000 shapes. Some representative poses with different subjects are demonstrated in Figure 10.

Figure 9. Fifteen representative poses in the FAUST dataset [4], where the blue rectangle regions contain large missing parts.

Figure 10. Some representative human poses with different subjects in our HPRD dataset.

4.2. Performance of Pose Classification

For each pose class in the SH-RE, SH-SY, and FAUST, 50% of shapes are randomly used for training, while the others are in the testing sets. In the HPRD, only 20% shapes in each pose class are randomly selected for training, while the other 80% shapes are used for testing.

The results of 3D human pose classification of the above datasets are recorded in Table 2, which demonstrates that our classification method obtains ideal or almost ideal results, i.e., almost 100% accuracy, on the SH-RE, SH-SY, and HPRD datasets. For the FAUST dataset, we obtained 88.33% classification accuracy. There are mainly two reasons for the lower accuracy. First, the training set is a bit smaller, with only five training shapes for each pose class. Second, the shapes in FAUST are real scans using a 3D stereo capture system, which has missing parts, topological noise, and non-manifold components, shown in Figure 9.

Table 2. Pose classification results of different datasets.

Dataset	Pose	Subject	Epoch	Accuracy (%)
SH-RE	10	40	300	100
SH-SY	20	15	300	99.38
FAUST	30	10	300	88.33
HPRD	100	1000	100	97

For the computational steps, there are enough training shapes in the HPRD dataset, so 100 epochs are enough to obtain desirable results, while 300 epochs are needed for the datasets of SH-RE, SH-SY, and FAUST. The classification losses with epochs on training and testing datasets of the SH-RE, SH-SY, and FAUST are shown in Figure 11.

Figure 11. Training and testing losses of pose classification on the SH-RE, SH-SY, and FAUST datasets.

Comparison with previous work. Our 3D human pose classification is compared with the work of Bonis et al. [13], where the bag-of-words pipeline is used based on topological persistence-based max pooling to summarize extracted features, and the final classification method is the well-known Support Vector Machine (SVM). They randomly used three shapes per pose class for the training set, two for the validation set, and five for the testing set. The mean accuracy of a hundred times experiments of curvature feature is 90%. For our approach, 12 shapes, i.e., 30% of total models, are randomly used for training in each pose class, and the other 28 shapes are used for testing. The classification accuracy is 100%, which is better than the SVM on hand-crafted features used by Bonis et al. [13].

Effectiveness of the ordered view feature fusion. To prove the effectiveness of the ordered view feature fusion layer, we constructed two 3D human datasets by rotating given shapes. The first dataset called RAP (rotating of a given pose) contains 1000 human subjects in a given pose and rotating 0, 90,

180, and 270 degrees along the z-axis of this pose, respectively. Therefore, the RAP dataset has four classes with a total of 4000 3D human models, shown in Figure 12. The second SH-RE-RO dataset is generated by 10 selected poses from the SH-RE dataset. Each selected pose class in the SH-RE generates 34 ($3 \times 11 + 1$) different orientation classes, using 30, 60, 90, ..., 330 degrees rotation along the x , y , and z axis, yielding a total of 340 (34×10) pose classes in the SH-RE-RO datasets.

Comparisons of our OVFF layer with traditional max and mean view-pooling on the above two datasets are recorded in Table 3, which demonstrates that the max and mean view-pooling fail in orientation-aware pose classification. This is because they ignore orientations of the views, while the proposed OVFF reflects the orientations in an orderly manner and obtains desirable results. Furthermore, we also try with different view permutation order in the OVFF, which gets the same results.

Table 3. Comparisons of our OVFF with traditional pooling layer for pose classification on the RAP and SH-RE-RO datasets.

Dataset	Pose	Subject	Accuracy(%)		
			Max View-Pooling	Mean View-Pooling	OVFF (Our)
RAP	4	800	25	25	100
SH-RE-RO	340	40	29.4	30.2	99.3

Figure 12. The given pose and its three rotated ones in the dataset RAP.

Certainty of the classification result. The information entropy $H(s) = -\sum_{i=1}^k p_i(s) \log_2(p_i(s))$ is used to measure the certainty of the classification result of each 3D human shape s , where k is the number of pose classes, and $p_i(s)$ is the probability that s belongs to the i -th pose class. If the value of $H(s)$ is higher, the classification of s is more uncertain. In the testing dataset of FAUST, there are 12 shapes with values of information entropy larger than 0.9, where 8 shapes of the above 12 shapes are classified incorrectly shown in Figure 13. The 12 shapes with the smallest values of information entropy are all classified correctly.

Figure 13. The 12 shapes with largest values of information entropy on the testing dataset of the FAUST, where orange colors indicate incorrect classification.

4.3. Performance of Pose Retrieval

For the pose retrieval experiments, we use the same training and testing splits on the above datasets as the classification in Section 4.2. In the training part, at each step, m_p positive pairs and m_n

negative pairs are used to minimize the loss in Equation (1). Given a testing shape s , the output of the MVCNN $f(s)$ is the final d dimensional pose descriptor vector, where the pose dissimilar measure of the shapes is the L_2 distances between their descriptor vectors defined in Equation (2). The default value of descriptor dimension is $d = 32$. In each step, the numbers of positive pairs and negative pairs are listed in Table 4.

Table 4. Pose retrieval results of different datasets.

Dataset	m_p	m_n	Step	NN(%)	FT (%)	ST (%)
SH-RE	10	10	500	100	100	100
SH-SY	20	40	500	100	97	100
FAUST	30	30	500	84	79	93
HPRD	50	50	3000	100	99.8	100

Evaluation metrics and results. We evaluate the retrieval results using the following statistical measures, which indicate the percentage of the top K matches that belongs to the same pose class as the query shape:

- Nearest Neighbour (NN): the percentage of the closest matches that belongs to the same class as the query (here, $K = 1$).
- First Tier (FT): the recall for the smallest K that could possibly include 100% of the models in the query class.
- Second Tier (ST): the same type of the result as FT, but a little less stringent (i.e., K is twice as big).

For the above measures, an ideal retrieval result gives a score of 100%, and higher values indicate better results. Results of the measures and training steps of the proposed pose retrieval method on the above datasets are recorded in Table 4. We plot the retrieval losses of the first 100 steps on training and testing datasets of the SH-RE, SH-SY, and FAUST in Figure 14, which shows that the training process converges quickly.

Figure 14. Training and testing losses of pose retrieval on the SH-RE, SH-SY, and FAUST datasets.

Some retrieval results of the SH-RE datasets are shown in Figure 3. Figure 15 demonstrates more retrieval results, where the query poses are not included in the HPRD dataset, and the proposed method can still effectively find the most similar poses in the HPRD dataset. The proposed retrieval method is robust to noises demonstrated in Figure 16, where the query shape is corrupted by Gaussian noises with the variance in 50% of the mean edge length of the shape.

Figure 15. Pose retrieval: For the query 3D human shapes (left one), our method retrieves the first six best shapes who have similar poses to the query one in the HPRD dataset.

Choices of dimension of the descriptor vector. We also try the other choices of dimension of the descriptor vector $d = 2, 16, 32, 64$ on the FAUST dataset in Figure 17. It can be seen that the $d = 32$ obtains the best results. Therefore, the default value of the descriptor dimension is $d = 32$ in this paper.

Comparison with previous work. We compare our deep learning retrieval method with the approach using the unsupervised Extremal Human Curve (EHC) descriptor [12] on the statistical shape database [2], which is challenging for human pose retrieval because it is a realistic shape database captured with a 3D laser scanner. There are 550 shapes in this dataset with 114 subjects in 35 poses, where the subject number of each pose class ranges from 4 to 111.

Figure 16. Given a noisy query shape, the proposed pose retrieval method is still successful.

Figure 17. Retrieval performance with different dimensions of descriptors on the FAUST dataset.

The comparison is performed on a subset of the above dataset consisting of 18 poses (p0, p1, p2, p3, p4, p5, p6, p7, p8, p9, p10, p11, p12, p13, p16, p28, p29, and p32), reported in Table 5. Several of the used representative poses are demonstrated in Figure 18. This demonstrates that our method obtains better pose retrieval results. However, it should be noted that the EHC-based method is unsupervised, while 50% of shapes are used for training, and the others are for testing in our retrieval approach.

Table 5. Pose retrieval comparisons of our method with the one in [12].

Approach	NN (%)	FT (%)	ST (%)
EHC 10 curves [12]	80.3	75.5	85.2
EHC 5 curves [12]	84.8	77.2	89.1
Our method	89.6	88.3	96.3

Figure 18. Ten used representative human poses with different subjects in the statistical shape database [2], where some artifacts are contained.

5. Conclusions and Future Works

The pose recognition of large 3D human datasets has wide practical applications, for example, in sport and dance. Inspired by the powerful success of deep learning and its generalization to 3D geometrical shapes in computer vision and graphics, the improved MVCNN is adopted for orientation-aware 3D human pose classification and retrieval in a united framework in this paper. Firstly, the AlexNet operating on multi-views with an orderly view feature fusion layer is used for 3D human pose classification. Secondly, the shape descriptor vectors are learned by the Siamese network for pose retrieval. Furthermore, we also construct a larger 3D human dataset called HPRD consisting of 100,000 shapes, which will be publicly released for future research. Experiments and comparisons demonstrate that the extended MVCNN framework obtains desirable results.

The proposed method assumes that the 3D human shapes are represented as triangle meshes. Thus, in the future, we plan to extend this framework to other shape representations, such as point

clouds or triangle soups. The accuracy of the proposed method has potential room for improvement on the FAUST dataset, which includes shapes with artifacts or large missing parts. We will investigate methods to overcome this limitation. Furthermore, our work regards the pose recognition of *static* 3D human shapes. We will extend the proposed method to the 3D human action recognition of *dynamic* shapes sequences using a recurrent neural network.

Author Contributions: Conceptualization, N.L.; formal analysis, N.L. and J.C.; funding acquisition, N.L.; investigation, H.W., P.H. and N.L.; methodology, H.W., P.H. and J.C.; project administration, H.W.; resources, H.W.; software, J.C.; validation, P.H.; visualization, P.H.; writing—original draft, H.W.; writing—review and editing, H.W., P.H., N.L. and J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This paper is supported by the National Natural Science Foundation of China (No. 61972267, 61802045, 61976041), the Natural Science Foundation of Hebei Province (No. F2018210100), the Youth Talent Support Program of Universities of Hebei Province (No. BJ2018003), the National Science and Technology Major Project (No. 2018ZX04041001-007, 2018ZX04016001-011), and the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems of Beihang University (No. VRLAB2020A04).

Acknowledgments: The authors sincerely thank the anonymous reviewers for their valuable suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Pickup, D.; Sun, X.; Rosin, P.L.; Martin, R.R.; Cheng, Z.; Lian, Z.; Bu, S. SHREC'14 Track: Shape retrieval of non-rigid 3D human Models. In Proceedings of the Eurographics Workshop on 3D Object Retrieval, Strasbourg, France, 6 April 2014; pp. 1–10.
- Hasler, N.; Stoll, C.; Sunkel, M.; Rosenhahn, B.; Seidel, H.P. A statistical model of human pose and body shape. In Proceedings of the Computer Graphics Forum, Budmerice, Slovakia, 23–25 April 2009; Volume 28, pp. 337–346.
- CAESAR. Available online: <http://store.sae.org/caesar> (accessed on 20 April 2020).
- Bogo, F.; Romero, J.; Loper, M.; Black, M.J. FAUST: Dataset and evaluation for 3D mesh registration. In Proceedings of the Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3794–3801.
- Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M.J.; Laptev, I.; Schmid, C. Learning from synthetic humans. In Proceedings of the Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 109–117.
- Anguelov, D.; Srinivasan, P.; Koller, D.; Thrun, S.; Rodgers, J.; Davis, J. SCAPE: Shape completion and animation of people. In Proceedings of the ACM SIGGRAPH, Anaheim, CA, USA, 24–28 July 2016.
- Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* **2015**, *34*, 1–16. [[CrossRef](#)]
- Michael, B.; Javier, R.; Gerard, P.M.; Federica, B.; Naureen, M. Learning human body shapes in motion. In Proceedings of the ACM SIGGRAPH 2016 Courses, New York, NY, USA, 24–28 July 2016; pp. 1–411.
- Cheng, Z.Q.; Chen, Y.; Martin, R.R.; Wu, T.; Song, Z. Parametric modeling of 3D human body shape—a survey. *Comput. Graph.* **2018**, *71*, 88–100. [[CrossRef](#)]
- Berretti, S.; Daoudi, M.; Turaga, P.; Basu, A. Representation, analysis, and recognition of 3D humans: A survey. *ACM Trans. Multimed. Comput. Commun. Appl.* **2018**, *14*, 1–36. [[CrossRef](#)]
- Pickup, D.; Sun, X.; Rosin, P.L.; Martin, R.R.; Cheng, Z.; Lian, Z.; Bu, S. Shape retrieval of non-rigid 3D human models. *Int. J. Comput. Vis.* **2016**, *120*, 169–193. [[CrossRef](#)]
- Slama, R.; Wannous, H.; Daoudi, M. 3D human motion analysis framework for shape similarity and retrieval. *Image Vis. Comput.* **2014**, *32*, 131–154. [[CrossRef](#)]
- Bonis, T.; Ovsjanikov, M.; Oudot, S.; Chazal, F. Persistence-based pooling for shape pose recognition. In Proceedings of the International Workshop on Computational Topology in Image Context, Marseille, France, 15–17 June 2016; pp. 19–29.
- Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3D shape recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 945–953.

15. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
16. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a “siamese” time delay neural network. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 29 November–2 December 1993; pp. 737–744.
17. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively with application to face verification. In Proceedings of the Computer Vision and Pattern Recognition, Hofgeismar, Germany, 7–9 April 2005; pp. 539–546.
18. Simo-Serra, E.; Trulls, E.; Ferraz, L.; Kokkinos, I.; Fua, P.; Moreno-Noguer, F. Discriminative learning of deep convolutional feature point descriptors. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 118–126.
19. Tao, R.; Gavves, E.; Smeulders, A.W. Siamese instance search for tracking. In Proceedings of the Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1420–1429.
20. Ferrari, V.; Marin-Jimenez, M.; Zisserman, A. Pose search: Retrieving people using their pose. In Proceedings of the Computer Vision and Pattern Recognition, Voss, Norway, 1–5 June 2009; pp. 1–8.
21. Eichner, M.; Marin-Jimenez, M.; Zisserman, A.; Ferrari, V. 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. *Int. J. Comput. Vis.* **2012**, *99*, 190–214. [[CrossRef](#)]
22. Jammalamadaka, N.; Zisserman, A.; Jawahar, C.V. Human pose search using deep networks. *Image Vis. Comput.* **2017**, *59*, 31–43. [[CrossRef](#)]
23. Jammalamadaka, N.; Zisserman, A.; Eichner, M. Video retrieval by mimicking poses. In Proceedings of the ACM International Conference on Multimedia Retrieval, Nara, Japan, 27–31 October 2012; pp. 1–8.
24. Pehlivan, S.; Duygulu, P. 3D human pose search using oriented cylinders. In Proceedings of the International Conference on Computer Vision Workshops, Kyoto, Japan, 27 September–4 October 2009; pp. 16–22.
25. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [[CrossRef](#)] [[PubMed](#)]
26. Sarafianos, N.; Boteanu, B.; Ionescu, B.; Kakadiaris, I.A. 3D human pose estimation: A review of the literature and analysis of covariates. *Comput. Vis. Image Underst.* **2016**, *152*, 1–20. [[CrossRef](#)]
27. Luvizon, D.C.; Picard, D.; Tabia, H. 2D/3D pose estimation and action recognition using multitask deep learning. In Proceedings of the Computer Vision and Pattern Recognition, Guangzhou, China, 23–26 November 2018; pp. 5137–5146.
28. Wang, P.S.; Sun, C.Y.; Liu, Y.; Tong, X. Adaptive O-CNN: A patch-based deep representation of 3D shapes. *ACM Trans. Graph.* **2018**, *37*, 1–11. [[CrossRef](#)]
29. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5099–5108.
30. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution on x-transformed points. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 820–830.
31. Hoang, L.; Lee, S.H.; Kwon, K.R. A 3D shape recognition method using hybrid deep learning network CNN-SVM. *Electronics* **2020**, *9*, 649. [[CrossRef](#)]
32. Bronstein, M.M.; Bruna, J.; LeCun, Y.; Szlam, A.; Vandergheynst, P. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Process. Mag.* **2017**, *34*, 18–42. [[CrossRef](#)]
33. Xiao, Y.P.; Lai, Y.K.; Zhang, F.L.; Li, C.; Gao, L. A survey on deep geometry learning: From a representation perspective. *Comput. Vis. Media* **2020**, *6*, 113–133. [[CrossRef](#)]
34. Wang, C.; Pelillo, M.; Siddiqi, K. Dominant set clustering and pooling for multi-view 3D object recognition. In Proceedings of the British Machine Vision Conference, London, UK, 4–7 September 2017.
35. Huang, H.; Kalogerakis, E.; Chaudhuri, S.; Ceylan, D.; Kim, V.G.; Yumer, E. Learning local shape descriptors from part correspondences with multiview convolutional networks. *ACM Trans. Graph.* **2017**, *37*, 1–14. [[CrossRef](#)]
36. Zeng, H.; Wang, Q.; Liu, J. Multi-feature fusion based on multi-view feature and 3D shape feature for non-rigid 3D model retrieval. *IEEE Access* **2019**, *7*, 41584–41595. [[CrossRef](#)]

37. Fried, O.; Avidan, S.; Cohen-Or, D. Patch2Vec: Globally consistent image patch representation. In Proceedings of the Computer Graphics Forum, Barcelona, Spain, 12–16 June 2017; Volume 36, pp. 183–194.
38. Mao, D.; Hao, Z. A novel sketch-based three-dimensional shape retrieval method using multi-view convolutional neural network. *Symmetry* **2019**, *11*, 703. [[CrossRef](#)]
39. Loper, M.; Mahmood, N.; Black, M.J. MoSh: Motion and shape capture from sparse markers. *ACM Trans. Graph.* **2014**, *33*, 220–233. [[CrossRef](#)]

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).