Hindawi
Complexity
Volume 2023, Article ID 9763495, 1 page
https://doi.org/10.1155/2023/9763495

Retraction

Retracted: Human Posture Recognition and Estimation Method Based on 3D Multiview Basketball Sports Dataset

Complexity

Received 19 December 2023; Accepted 19 December 2023; Published 20 December 2023

Copyright © 2023 Complexity. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Manipulated or compromised peer review

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] X. Song and L. Fan, "Human Posture Recognition and Estimation Method Based on 3D Multiview Basketball Sports Dataset," *Complexity*, vol. 2021, Article ID 6697697, 10 pages, 2021.

Hindawi
Complexity
Volume 2021, Article ID 6697697, 10 pages
https://doi.org/10.1155/2021/6697697

RETRACTED

Research Article

Human Posture Recognition and Estimation Method Based on 3D Multiview Basketball Sports Dataset

Xuhui Song[1] **and Linyuan Fan**[2]

[1]Department of Sports, Capital University of Economics and Business, Beijing 100070, China
[2]School of Statistics, Capital University of Economics and Business, Beijing 100070, China

Correspondence should be addressed to Linyuan Fan; fanlinyuan@cueb.edu.cn

Received 29 December 2020; Revised 7 February 2021; Accepted 22 February 2021; Published 3 March 2021

Academic Editor: Zhihan Lv

Copyright © 2021 Xuhui Song and Linyuan Fan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In traditional 3D reconstruction methods, using a single view to predict the 3D structure of an object is a very difficult task. This research mainly discusses human pose recognition and estimation based on 3D multiview basketball sports dataset. The convolutional neural network framework used in this research is VGG11, and the basketball dataset Image Net is used for pretraining. This research uses some modules of the VGG11 network. For different feature fusion methods, different modules of the VGG11 network are used as the feature extraction network. In order to be efficient in computing and processing, the multilayer perceptron in the network model is implemented by a one-dimensional convolutional network. The input is a randomly sampled point set, and after a layer of perceptron, it outputs a feature set of $n \times 16$. Then, the feature set is sent to two network branches, one is to continue to use the perceptron method to generate the feature set of $n \times 1024$, and the other network is used to extract the local features of points. After the RGB basketball sports picture passes through the semantic segmentation network, a picture containing the target object is obtained, and the picture is input to the constructed feature fusion network model. After feature extraction is performed on the RGB image and the depth image, respectively, the RGB feature, the local feature of the point cloud, and the global feature are spliced and fused to form a feature vector of $N \times 1152$. There are three branches for this vector network, which, respectively, predict the object position, rotation, and confidence. Among them, the feature dimensionality reduction is realized by one-dimensional convolution, and the activation function is the ReLU function. After removing the feature mapping module, the accuracy of VC-CNN_v1 dropped by 0.33% and the accuracy of VC-CNN_v2 dropped by 0.55%. It can be seen from the research results that the addition of the feature mapping module improves the recognition effect of the network to a certain extent

1. Introduction

The 3D reconstruction based on the depth map requires the input of RGB image and the corresponding depth image. Since the depth image contains the vertical distance of each spatial coordinate from the depth camera position, the difficulty of 3D reconstruction is greatly reduced. In addition, because a single depth image lacks the spatial information of the occluded part of the object, it is usually necessary to combine multiple depth images to model the three-dimensional structure. The single-view image of the object is affected by the observation angle and often loses a large amount of spatial information. The internal and external parameters of the camera and the data conversion in different spatial coordinate systems are required as a priori knowledge when reconstructing, which greatly reduces the actual algorithm application performance.

First, based on the superpixel segmentation, the single-view depth map obtained by the traditional multiview stereo is optimized and completed to improve the reconstruction effect of the multiview stereo in the untextured area, then the surface of the object is used on the basis of the initial 3D mesh obtained from the multiview stereo, and the lighting information is refined to improve the accuracy and completeness of surface reconstruction. This research proposes a superpixel-based plane fitting and space propagation

RETRACTED

algorithm that can improve the depth and density of a single view and proposes an iterative subdivision-based spatial variation albedo expression model to improve the surface reconstruction details of objects in actual scenes that are not global and constant. A new geometric constraint and shadow constraint joint optimization method improves the reconstruction performance of the algorithm. This method has better results than other existing multiview algorithms.

The theory and technology of human-machine coordination and natural interaction have broad application prospects in future smart factories. Wu introduced in detail the use of gesture recognition based on Kinect to enhance interactive performance. He proposed an improved optical flow method to obtain the direction and speed of the target movement. The smoothing parameters in traditional optical flow are replaced by variables. His research method is not novel enough and lacks practicality [1]. Nicolaou focuses on face and body image analysis, which is one of the most studied objects. He catalyzes the development of automated analysis of human behavior under uncontrolled "wild" conditions. We summarize the development of algorithms and systems for research methods, database collections and benchmarks, and machine analysis of human behavior, focusing on facial expressions, body gestures, speech, and various other sensors. Although his research proposed a gesture analysis algorithm, the research process lacked data [2]. Syed introduced how physical therapy rehabilitation can improve the functional ability of the disabled through range of motion exercise to improve the quality of life. The framework he proposed is "AR-NUI-REHAB-MDSS," which uses natural user interface- (NUI-) based physical therapy rehabilitation, a personalized exercise presentation and monitoring system for patients, and a mobile decision support system (MDSS) for therapists. Although he proposed AR, a less entertaining adjuvant therapy environment, the research lacks comparative data [3]. The Brito mission augmented reality (AR) platform is used in various applications. His experimental design compared two different optical tracking systems based on ARa-marked AR (MB) and unmarked AR (ML) for two types of interfaces: a tangible interface based on gesture recognition and a multimodal interface. Both AR technologies allow consumers to visually observe the function of sports shoes. Although he compared AR (MB) and AR (ML), the research method is too complicated [4].

This research uses some modules of the VGG11 network. For different feature fusion methods, different modules of the VGG11 network are used as the feature extraction network. In order to be efficient in computing and processing, the multilayer perceptron in the network model is implemented by a one-dimensional convolutional network. The input is a randomly sampled point set, and after a layer of perceptron, it outputs a feature set of $n \times 16$. Then, the feature set is sent to two network branches, one is to continue to use the perceptron method to generate the feature set of $n \times 1024$ and the other network is used to extract the local features of points. After the RGB basketball sports picture passes through the semantic segmentation network, a picture containing the target object is obtained, and the picture is input to the constructed feature fusion network model.

2. Human Body Gesture Recognition

2.1. Evaluation of the Reconstructed 3D Mesh Model. For 3D reconstruction tasks, given the reconstructed point cloud and the real point cloud of the scene, the quality of the reconstructed model can generally be evaluated by accuracy, completeness, and F score [5, 6]. For each point in the reconstructed model, find the point with the smallest Euclidean distance from the point in the real 3D point cloud model and calculate the number of points whose distance is less than the given threshold range. The accuracy is defined as the amount of points satisfying the conditions in the reconstructed model. The ratio of accuracy is the ratio of the total number of points in the model [7]. For each point in the real 3D point cloud model, calculate the point with the smallest Euclidean distance from the point in the reconstructed model. The completeness is defined as the number of points in the real 3D model that is less than the given distance threshold, accounting for the total number of points in the real model [8, 9]. Since the real results obtained by laser scanning are not necessarily complete, care should be taken to prevent the deviation of the results caused by the actual scene points that may be lost. This can be reflected in the qualitative analysis [10]. Let G be the point set of the real 3D model and R be the test point set of the model obtained by the reconstruction algorithm. For any point, R has

$$e_{r\longrightarrow G} = \min_{g\in G} |r - g|. \tag{1}$$

Given a distance threshold d, the accuracy of the reconstructed point set R is defined as the cumulative result of all points [11, 12]:

$$P(d) = \frac{100}{|R|} \sum_{r\in R} \left[e_{r\longrightarrow G} < d\right]. \tag{2}$$

Similarly, for any real 3D point $g \in G$, the distance between it and the point in the reconstructed model is defined as

$$e_{r\longrightarrow G} = \min_{r\in R} |g - r|. \tag{3}$$

The recall rate of the reconstructed model R for a given distance threshold d is defined as

$$R(d) = \frac{100}{|G|} \sum_{g\in G} \left[e_{g\longrightarrow R} < d\right]. \tag{4}$$

The definition of F score can be obtained [13] as follows:

$$F(d) = \frac{2P(d)R(d)}{P(d) + R(d)}. \tag{5}$$

2.2. Human Body Gesture Recognition. With the development of wireless technology and sensing methods, many studies have shown the successful use of wireless signals

RETRACTED

(for example, WiFi) to perceive human activities and thus realize a series of emerging applications, including intrusion detection, daily activity recognition, and gestures to vital signs monitoring, and the recognition of user recognition even includes more fine-grained motion sensing [14, 15]. It can be said that these applications can support various areas of smart home and office environments, including security protection, health monitoring/management, smart healthcare, and smart device interaction [16]. The movement of the human body will affect the propagation of wireless signals (such as reflection, diffraction, and scattering), which provides a great opportunity to capture the movement of the human body by analyzing the received wireless signals [17, 18]. Researchers take advantage of existing wireless links between mobile/smart devices (e.g., laptops, smartphones, smart thermostats, smart refrigerators, and virtual assistance systems) by extracting readymade signs [19]. Suppose there is a set of data to fit an estimation function formula:

$$h(x) = h\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 +, \ldots, \theta_n x_n. \tag{6}$$

This function can be reduced to the form of vector product [11, 12]:

$$h(x) = \sum_{i=0}^{n} \theta_i X_i = \theta^T X. \tag{7}$$

In order to evaluate whether the fitting effect of the parameter θ is optimal, a loss function is needed to express it, which is generally used in the model [20]:

$$J = \frac{1}{2}\sum_{i=1}^{m} \left(h_\theta(x_i) - y_i\right)^2. \tag{8}$$

Due to a series of problems such as the chaotic background environment, object occlusion, and perspective change, behavior recognition is still a challenging task [21, 22]. With the emergence of a large number of behavior recognition video datasets, the research on behavior recognition has encountered a serious computational burden [23]. How to extract effective features from the video is the key to solving the abovementioned difficult problems and designing a more effective behavior recognition framework.

2.3. 3D Convolutional Network. In order to increase the number of feature maps, two different convolution sets are applied to each position, and two sets containing 23 feature maps are generated in the C2 layer. S5 is obtained by using 3 × 3 downsampling on each feature map of the C4 layer. The C6 layer contains 128 feature maps with a size of 1 × 1, and each map is connected to all 78 feature maps in the S5 layer. All the trainable parameters in the model are initialized randomly and trained through the error back propagation algorithm. In the process of network training, high-level motion features containing long-term motion information are extracted from a large number of continuous video sequences and connected to the last hidden layer of the network as an auxiliary output unit, so that the network can learn a feature vector that is similar to the high-level feature, thereby achieving regularization of 3DCNN models [24].

$$I(\mu, v, 2k-1) = d^{\lambda}_{\tau+k-1}(\mu, v), \tag{9}$$

$$I(\mu, v, 2k) = d^{y}_{\tau+k-1}(\mu, v), \mu = [1; w], \tag{10}$$

where μ and h, respectively, represent the width and height of the video frame image. In practice, the displacement vector field component can take a positive value or a negative value, and it contains a relatively large range of motion. A motion in one direction may be a motion in the opposite direction. The position of the mth feature cube in the kth layer in the network structure is calculated by 3D convolution to obtain the neuron output:

$$v^{xyz}_{km} = f\left(b_{km} + \sum_{p=0}^{P_k-1}\sum_{q=0}^{Q_k-1}\sum_{r=0}^{R_k-1} w^{pqr}_{kmn}\mu^{(x+p)(y+q)(z+r)}_{(k-1)n}\right). \tag{11}$$

In the formula, v is the output at (x, y, z) of the kth layer. Similar to the 2D pooling layer, the commonly used 3D pooling layer sampling methods include maximum pooling, average pooling, and random pooling. The three-dimensional maximum pooling formula is

$$v_{x,y,z} = \max_{0\le i\le S_1, 0\le j\le S_2, 0\le k\le S_3}\left(\mu_{x\times s+i, y\times t+j, z\times r+k}\right). \tag{12}$$

In the formula, the input vector of the 3D pooling layer is μ, and the output after pooling is v. The N-dimensional feature vector is obtained from input through get_flow.

$$f_{[m\times k]} = \text{get\_flow}(\text{video}_n). \tag{13}$$

Take 10 consecutive optical flow images as an input unit to extract a feature. To ensure that it corresponds to the C3D feature image, the step size of the optical flow image input is 16. If the total number of video frames is Num, the number of features extracted for each video is

$$m_n = \frac{(\text{Num} - 16)}{16 + 1}. \tag{14}$$

Therefore, the total extracted temporal convolutional network feature size for video samples is

$$D = \sum_{n=1}^{13320} m_n \times 4096. \tag{15}$$

The constructed feature fusion network is shown in Figure 1.

3. Human Body Gesture Recognition Experiment

3.1. Point Cloud Feature Extraction. In order to be efficient in computing and processing, the multilayer perceptron in the network model is implemented by a one-dimensional convolutional network. The input is a randomly sampled point set, and after a layer of perceptron, it outputs a feature set of $n \times 16$. Then, the feature set is sent to two network branches, one is to continue to use the perceptron method to

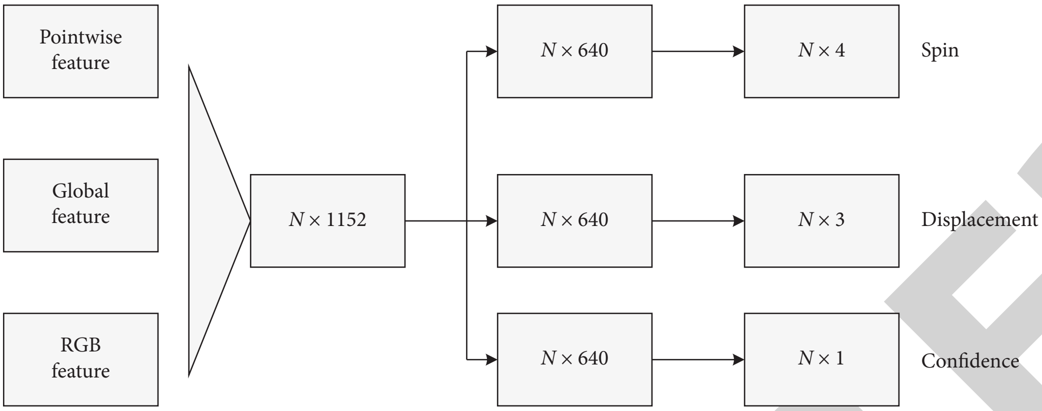

FIGURE 1: Feature fusion network constructed.

generate the feature set of $n \times 1024$, and after the maximum pooling, the global feature of 1×1024 is obtained. Another network is used to extract local features of points. In this network, in order to achieve the two-dimensional convolution used in similar processing pictures, it is first necessary to sample the high-dimensional space point set. Here, for each high-dimensional space point p in the point set, the K nearest neighbor algorithm is used to determine the surrounding area of the point. Transform these points into the local coordinate system with point p as the origin and then use the convolution operation.

3.2. RGB Feature Extraction. After the RGB basketball sports picture passes through the semantic segmentation network, a picture containing the target object is obtained, and the picture is input to the constructed feature fusion network model. The network is downsampling through a series of convolutional pooling operations. The downsampling mainly occurs in the pooling operation, and then the output feature map requires a series of upsampling operations. Upsampling includes double-sampling low-resolution feature maps. Linear interpolation allows it to be enlarged to a specified resolution and also includes convolution, which enables the network to learn the interpolated features, and the final output result is a 32-channel feature map of the original image resolution.

After feature extraction is performed on the RGB image and the depth image, respectively, the RGB feature, the local feature of the point cloud, and the global feature are spliced and fused to form a feature vector of $N \times 1152$. There are three branches for this vector network, which, respectively, predict the object position, rotation, and confidence. Among them, the feature dimensionality reduction is realized by one-dimensional convolution, and the activation function is the ReLU function.

3.3. Iterative Closest Point Method ICP. The posture estimation obtained above is the initial posture. In order to obtain a more accurate posture, this research uses the iterative closest point method to optimize the posture results obtained by the neural network model. The specific method is to transform the complete point cloud of the target object from the object-based coordinate system to the camera-based coordinate system according to the posture predicted by the network. At this time, the point cloud is very similar to the object point cloud restored by the semantic segmentation network segmentation. The iterative nearest point method is used to obtain the coordinate conversion relationship between the two, and finally, the 6D posture of the object is obtained by combining the above two steps.

The convolutional neural network framework used in this research is VGG11, and the basketball dataset ImageNet is used for pretraining. This research uses some modules of the VGG11 network. For different feature fusion methods, different modules of the VGG11 network are used as the feature extraction network. In this study, the VGG networks in the two cases are represented by different symbols. The structure of the feature extraction network is shown in Table 1.

3.4. Network Training Environment. Since the network in this experiment is trained under GPU, the system used in this experiment environment is Linux system, and the development environment used is Ubuntu16.03, CUDA 8.0, cuDNN v5.0, OpenCV 2.4.11, and CAFFE. And, the PC used in the experiment has higher configuration requirements.

The PC hardware configuration used in this experiment includes: Intel Core Quad-core i7-6700*k*, Cooler Master Core processor, ASUS GTX1080-8G graphics card, Gigabyte z170 motherboard, WD1T + Samsung 250G hard drive, and Kingston 32 GB memory. Dataset: the basketball dataset is divided into training dataset and test dataset. The training dataset contains 9537 basketball video clips, and the test dataset contains 3783 basketball video clips. In each long basketball video segment, 16 consecutive frames of basketball video segment are extracted without overlapping, and the size of the basketball video segment is normalized to 128×171. During training, the input data is randomly cropped to $16 \times 112 \times 112$.

Training process: when the data is being trained on the network, the size of the data will change accordingly according to the characteristics of different network layers. Edge processing is performed in the convolutional layer, and

RETRACTED

Table 1: Structure of feature extraction network.

Module	Structure	Size	Number of output channels	Proportion to original image
Convolutional layer 1	Conv2D	3 × 3	64	1
Convolutional layer 2	Maxpooling	2 × 2	64	1/2
Convolutional layer 3	Comv2D	3 × 3	128	1/2
Convolutional layer 4	Maxpooling	2 × 2	128	1/4
Convolutional layer 5	Conv2D	3 × 3	256	1/4
Convolutional layer 6	Conv2D	3 × 3	256	1/4
Fully connected layer 1	Dense	—	256	—
Fully connected layer 2	Dense	—	512	—

the size of the image will not change, but the number of filters in each convolutional layer changes, so the number of image channels will vary with the number of filters. In the pooling layer, the purpose is to downsample the image features, and the image feature size will be halved.

C3D network feature extraction: after network training, the C3D convolutional neural network model can be used as a feature extractor to perform feature extraction and analysis of other basketball video clips.

4. Human Body Posture Recognition Analysis

4.1. Comparison with Cutting-Edge Algorithms. Three-dimensional object recognition at home and abroad is mainly divided into three main directions: point cloud-based methods, voxel-based methods, and multiview-based methods. In this section, we will compare these methods and analyze the results in detail. A total of 16 comparison methods were selected in the experiment. Among them, there are 5 methods based on point clouds, namely, SO-Net, PointNet, PointNet++, KCNet, and ECC; there are 4 methods based on voxels, namely, VRN, 3D-GAN, VoxNet, and 3D ShapeNets; methods based on multiple views: there are 7 types, namely, RotationNet, MHBN, DominantSet, SeqViews2SeqLables.GVCNN, MVCNN-MultiRes, and MVCNN. Among them, the method with better effect than VC-CNN is marked with underscore, and the best effect under each method type is marked with bold numbers, and the data that cannot be obtained is marked with "–." It can be seen from Table 2 that although ModelNet10 is a subset of ModelNet40, different frameworks perform slightly differently on these two datasets. For example, RotationNet trained with 12 views performs better on ModelNet40, but it is not as good as many other methods on ModelNet10. The accuracy of MHBN on the two datasets is not much different. In general, the ModelNet40 dataset is more challenging than ModelNet10. The comparison result with the leading edge algorithm is shown in Table 2.

Table 2: Comparison results with cutting-edge algorithms.

Method type	Method name	ModelNet40 (%)	ModelNetl0 (%)
Voxel	3D-GAN	91.30	93.60
	VRN	83.30	91.00
	3D shape nets	83.00	92.00
Point cloud	SO-Net	91.90	95.70
	PointNet	91.00	94.40
	KCNet	83.20	90.00
Multi-view	MHBN(RGB+depth)J	94.70	95.00
	Ours_v1(VC–CNN _v1)	93.07	94.60
	Ours_v2(VC–CNN_v2)	93.47	95.04

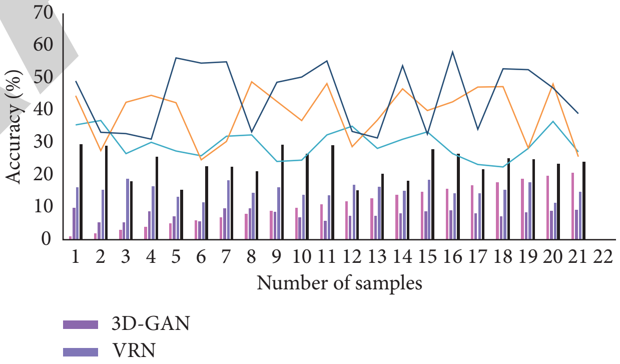

Figure 2: Comparison results with voxel-based methods.

4.2. Comparison with Voxel-Based Methods. The comparison result with the voxel-based method is shown in Figure 2. In the two neural network frameworks proposed in this study, the recognition effect of convolution fusion on feature response graph is better than that of convolution fusion on high-dimensional features. On both datasets, the accuracy of VC-CNN_v2 is 0.4% higher than VC-CNN_v2. In terms of network structure, VC-CNN_v2 uses the maximum pooling layer in the range of 7 × 7, and the number of hidden units in the fully connected layer is less, and the network complexity is lower than VC-CNN_v1. At the same time, the view loses spatial information after extracting the fully connected layer of the network through the features of VC-CNN_v1, which makes it difficult for the network to learn the correlation between adjacent views during the training process. The best performance in Figure 2 is the VRN network structure. VRN provides experimental results in two cases, one is VRN-Ensemble that uses integrated strategy, and the other is VRN that does not use integrated strategy. The integration strategy has greatly improved the recognition effect of VRN. On ModelNet40, the accuracy rate has increased by 4.2%. The accuracy of 3D-GAN, VoxNet and 3D ShapeNets on ModelNet40 is not more than 90%. In the ModelNet10 dataset, although most of the voxel-based methods as well as the multiview method do not perform on the dataset, the

RETRACTED

RETRACTED

VRN network with integrated strategy still surpasses some multiview method networks. However, after removing the integration strategy, the accuracy of the VRN network is only 93.6%, which is 1.44% lower than the method of this study. In the ModelNet40 dataset, the VRN using the integrated strategy is 2.03% more accurate than VC-CNN_v2. However, when not in use, VRN is 2.17% lower than VC-CNN_v2.

The comparison result with the point cloud-based method is shown in Figure 3. The point cloud-based method has better overall performance than the voxel-based method. Although ECC and PointNet do not exceed 90% on ModelNet40, these two methods are, respectively, proposed in point cloud recognition segmentation and point cloud convolution design. PointNet++ encodes the local features on the basis of PointNet, which improves the accuracy by 1.3%. The best performer is SO-Net, with an accuracy rate of 93.4%. On the one hand, this method uses self-organizing mapping (SOM) instead of kd number to point cloud coding. On the other hand, SO-Net uses 5000 points to describe the data in the experiment and increases the amount of training data. On ModelNet10, SO-Net is 0.66% higher than VC-CNN_v2, but on ModelNet40, the accuracy rate is about the same as VC-CNN_v2. In general, considering the number of training samples and the complexity of the network framework, the accuracy of the convolution-based feature fusion framework on ModelNet10 in this study is higher than that of most point cloud-based methods in the table.

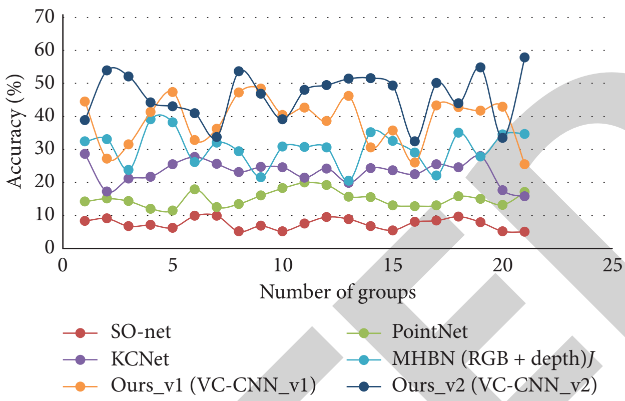

FIGURE 3: Comparison with point cloud-based methods.

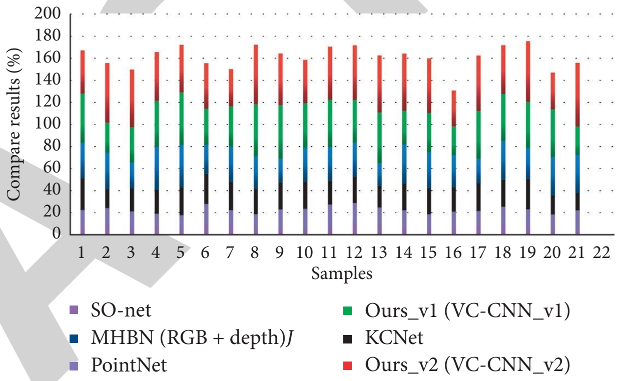

FIGURE 4: Comparison results with multiview-based methods.

The comparison result with the multiview-based method is shown in Figure 4. In the overall recognition rate, the method based on multiview has certain advantages compared with the above two methods. Among them, MVCNN is the baseline of this research method, and the accuracy rate on ModelNet40 is 90.1%. MVCNN-MultiRes uses multiresolution views for data enhancement, which increases the accuracy of MVCNN by 1.3%. RotationNet, MHBN, and Dominant have all carried out data enhancement to a certain extent. For MHBN and Dominant, adding the depth information of the three-dimensional object projection improves the recognition accuracy by about 1%. For RotationNet, 80 views are used as training input for a three-dimensional object, which improves the recognition accuracy to 97.37%. However, in the study, the RotationNet with 12 views has a recognition accuracy of 93.84%. On ModelNet10, RotationNet is 4.42% higher than VC-CNN_v2 in accuracy, but the number of views is 5 times the number of training views in this study. The expansion of the training set reduces the overfitting phenomenon of RotationNet. When only 12 views are used for training, the accuracy of RotationNet is only 90.65%, which is 4.4% lower than VC-CNN_v2 in this study.

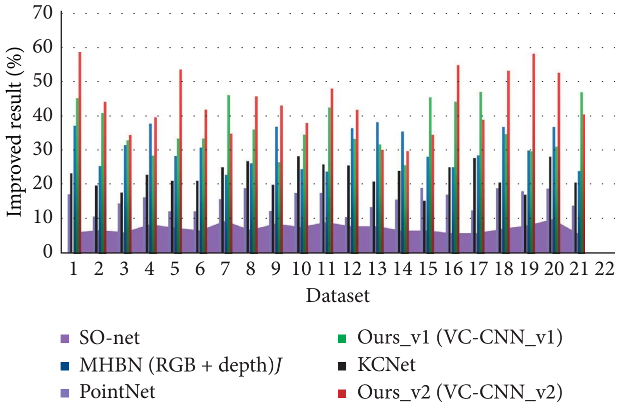

FIGURE 5: Accuracy results after the change.

For different input numbers, corresponding changes are made to the convolution fusion operation in the network. The accuracy result after the change is shown in Figure 5. In order to ensure that the basic viewing angles corresponding to the group-level features are the same, for the input of 8 basketball sports views, the two adjacent views are divided into 4 groups. In the first feature fusion, the size of the convolution kernel is set to 2 × 1. For the input of 12 views, the adjacent 3 views are used as a group, and the size of the first feature fusion convolution kernel is 3 × 1. The settings of the two views are the same in the second feature fusion, VC-CNN_v1 convolution kernel. The size of is set to 4 × 1, and the size of the VC-CNN_v2 convolution kernel is set to 2 × 2. The selected dataset is ModelNet40, and the results are shown in Table 3. It can be seen from Table 3 that the results of 8 views are not much different from those of 12 views, and

Table 3: The result of selected dataset ModelNet40.

Perspective mode	Number of views	VC-CNN_vl (%)	VC-CNN_v2 (%)
Ring	8	92.18	92.26
	12	92.42	92.73

the accuracy difference is kept within 0.5%. This shows that a small increase in the number of views will not have much impact on the recognition effect. At the same time, the contents of 8 views and 12 views overlap, and the images are relatively similar. Therefore, the prediction ability of the network trained with 12 views is not much different from that of 8 views. However, compared with the perspective model proposed in this study, the recognition accuracy of the two networks has decreased.

This study has done a comparative study on ModelNet40 with different height views. A circular viewing angle setting is adopted, and the number of viewing angles is 8. The experimental results are shown in Table 4. It can be seen from Table 4 that when the viewing angle points are distributed on the coordinate plane, the recognition of objects by the network is greatly affected. The accuracy of VC-CNN_v1 dropped by 1.44%, and the accuracy of VC-CNN_v2 dropped by 2.07%. The reason why VC-CNN__v1 is less affected is that VC-CNN_v1 fusion features are high-dimensional information, and the weights can be learned from the training set, while VC-CNN_v2 is more affected by the spatial structure on the view, and the larger changes between adjacent views make it difficult for the network to capture its relevance. The perspective selection strategy proposed in this study avoids the above two problems. Three heights are covered in the 16 viewing angle settings, which makes the views have a certain degree of distinction and avoids the problem of self-occlusion caused by the complex structure of the object. At the same time, the refined perspective points after multiple iterations of the algorithm will not be distributed on the coordinate plane, but will only be infinitely close to the coordinate plane, avoiding the problem that the rectangular object is completely blocked by a certain side. The comparative study results are shown in Figure 6.

Table 4: Experimental results.

Viewing angle parameter θ	VC-CNN_v1	VCC-NN_v2
60°	92.18	92.26
90°	90.74	90.19

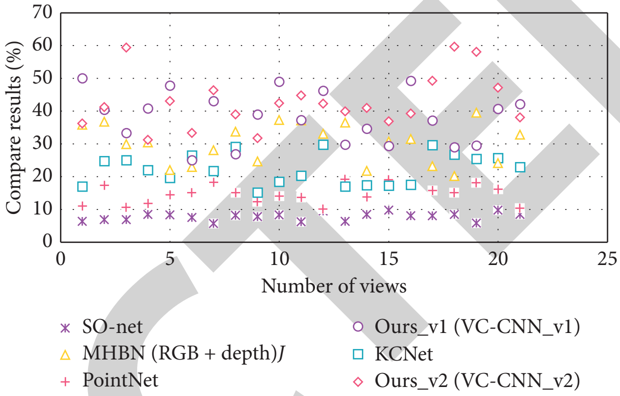

Figure 6: Comparative study results.

Table 5: Recognition accuracy rate of the feature mapping module on ModelNet10.

Feature mapping module	VC-CNN_v1 (%)	VCC-NN_v2 (%)
Yes	94.60	95.04
No	94.27	94.49

4.3. Feature Mapping Module. Table 5 shows the recognition accuracy rate of the feature mapping module on ModelNet10. This section mainly analyzes the proposed view feature mapping module. Since the clustering of views depends on the spatial distribution of perspectives, the content of adjacent views contains descriptions of different perspectives of the partial structure of three-dimensional objects. Based on the coincidence of the three-dimensional structure, we hope that the feature mapping module can learn the transformation between features from a fixed perspective conversion and can map to a local description feature. In order to verify the improvement brought by the feature mapping module to the recognition accuracy of the network, the experiment tested the performance of the network on ModelNet10 after the feature mapping module was removed. It can be seen from the table that, after removing the feature mapping module, the recognition accuracy of the two networks has decreased. The accuracy of VC-CNN_v1 dropped by 0.33%, and the accuracy of VC-CNN_v2 dropped by 0.55%. It can be seen from the experimental results that the addition of the feature mapping module has a certain improvement on the recognition effect of the network.

At the same time, in order to verify the dependence of the network on similar views, in the training phase of the experiment, the order of the input views is shuffled so that there is no similar relationship between the views, and the recognition effect of the two networks at this time is judged. The recognition effects of the two networks are shown in Table 6. The database selected for the experiment is ModelNet10. It can be seen from Table 6 that when the order of the test views is disturbed, the recognition effect of the two networks has decreased. Among them, the accuracy rate of VC-CNN_v1 dropped by 1.1%, and VC-CNN_v2 was affected relatively more, dropping by 1.98%. Unlike MVCNN, the network structure does not have the order invariance of views. After the views are randomly scrambled, the views within the groups are not similar, and the corresponding perspective changes between the groups are different. In the case of constraints on the mapped features, the mapping module is difficult to generalize the distribution of view features in the feature space, and it is difficult to mine the local structure information of three-dimensional objects under the perspective conversion. The second reason why

RETRACTED

Table 6: The recognition effect of the two networks.

View order	VC-CNN_v1 (%)	VCC-NN_v2 (%)
Choose strategy by perspective	94.60	95.04
Random scramble	93.50	93.06

the network is more sensitive is that the feature fusion module is more advanced in the network structure and is more sensitive to the spatial structure of the image. And, VC-CNN_v1 is to learn the correlation between high-dimensional features. After the basketball sports view passes through the first 10 layers of VGG11, the features tend to be abstract descriptions of objects, so the impact is less.

4.4. Feature Fusion Module. The classification effect of VC-CNN_v1 on ModelNet10 with different convolution channels is shown in Figure 7. In the feature fusion module of VC-CNN_v2, the convolution process is the maximum response value on the feature response graph, which is the same feature dimension as the feature extraction network. So, we follow the setting of the last convolution layer of VGG11 and set the number of convolution channels for feature fusion to 512. However, in the VC-CNN_v1 network, the feature fusion operation combines the one-dimensional features of the fully connected layer into a matrix, which changes the dimension of the features, so the number of convolution channels does not have a suitable reference value. We did a preliminary search on the number of convolutional channels on ModelNet10. As can be seen from Figure 7, the accuracy is the lowest when the number of channels is 1. Because the feature fusion in the group uses convolution with the same weight, it is difficult to learn the relationship between the perspectives in different directions with a single convolution kernel. With the increase of the number of convolution kernels, the accuracy rate also gradually rises and tends to be stable. Considering the complexity of the model and the consumption of training time, this study sets the number of feature fusion convolution channels of VC-CNN_v1 to 32. It can be seen from the experimental results that the accuracy of the feature fusion has been greatly improved compared with the 3D convolutional network. The second weight setting strategy has the highest accuracy of weighted feature fusion. Among them, the accuracy of serial feature fusion is slightly higher than that of the first weight setting strategy, but the feature dimension is doubled after serialization, which increases the feature size, so it takes more time to run.

The characteristic response after the convolution layer is shown in Figure 8. This research also analyzes the feature response map output by the VC-CNN_v2 convolutional layer. In the experiment, in order to eliminate the influence of the feature mapping module on the feature map, the feature extraction module of VC-CNN_v2 did not add the

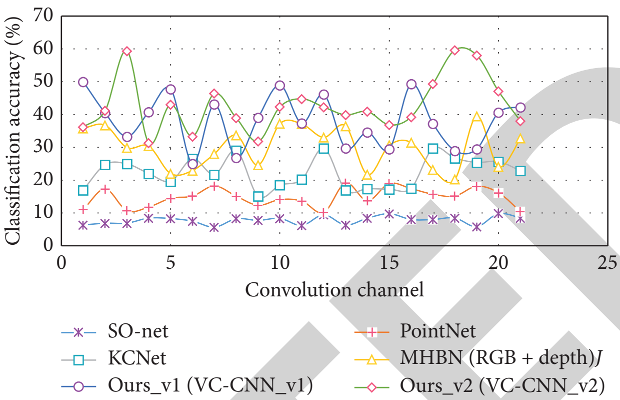

Figure 7: The classification effect of VC-CNN_v1 on ModelNet10 with different number of convolution channels.

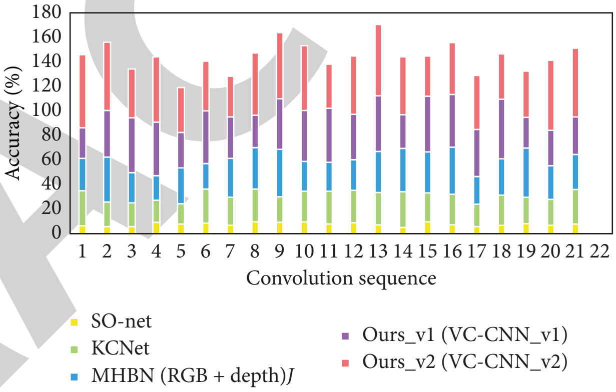

Figure 8: Feature response map after convolutional layer.

feature mapping operation. Figure 8 shows the output feature map of the 8th convolutional layer of VGG11. VGG11 is pretrained on ImageNet and fine-tuned with the training set of ModelNet40. The figure shows the characteristic response of adjacent views on two channels. Since the rendered view has no background, after the RELU activation function of the multilayer convolutional layer, the surrounding response of the feature map is 0, and many parts of the response are suppressed. Because the view background is clean, the number of pixels in the feature map is small, which avoids the interference of the maximum pooling by the background information. It can be seen from the response distribution of the feature maps that there is a certain similarity between the feature maps, and the position of the response within the set is not much different. This shows that the same area between adjacent views is simultaneously focused by the network.

This research evaluates the proposed network on a three-dimensional object retrieval task. The results of the study are shown in Table 7. The dataset used in the research is ModelNet40. The network is trained on ModelNet40

RETRACTED

RETRACTED

Table 7: Research results.

Method name	ModelNet40 (%)	ModelNet10 (%)
Deep Pano	76.81	84.18
3D ShapeNets	49.20	68.3
Ours_v1(VC–CNN_v1)	77.02	84.62
Ours_v2(VC–CNN_v2)	85.53	87.26

through classification, and then the trained network extracts features from the objects on the test set.

5. Conclusion

After the RGB basketball sports picture passes through the semantic segmentation network, a picture containing the target object is obtained, and the picture is input to the constructed feature fusion network model. The network is downsampling through a series of convolutional pooling operations. The downsampling mainly occurs in the pooling operation, and then the output feature map requires a series of upsampling operations. Upsampling includes double-sampling low-resolution feature maps. After feature extraction is performed on the RGB image and the depth image, the RGB feature, the local feature of the point cloud, and the global feature are spliced and merged to form a feature vector. There are three branches for this vector network, which, respectively, predict the object position, rotation, and confidence. Among them, the feature dimensionality reduction is realized by one-dimensional convolution, and the activation function is the ReLU function.

The posture estimation obtained above is the initial posture. In order to obtain a more accurate posture, this research uses the iterative closest point method to optimize the posture results obtained by the neural network model. The specific method is to transform the complete point cloud of the target object from the object-based coordinate system to the camera-based coordinate system according to the posture predicted by the network. At this time, the point cloud is very similar to the object point cloud restored by the semantic segmentation network segmentation. The iterative nearest point method is used to obtain the coordinate conversion relationship between the two.

The convolutional neural network framework used in this research is VGG11, and the basketball dataset ImageNet is used for pretraining. This research uses some modules of the VGG11 network. For different feature fusion methods, different modules of the VGG11 network are used as the feature extraction network. In this study, the VGG networks in the two cases are represented by different symbols.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare no conflicts of interest.

Authors' Contributions

All authors have read the manuscript and approved the final version of the manuscript.

Acknowledgments

This work was supported by Fund for Reserve Academic Leader 2020-2022 granted by Capital University of Economics and Business and Special Fund for Fundamental Scientific Research of the Beijing Colleges in CUEB granted by Capital University of Economics and Business.

References

[1] Y. Wu, Q. Yang, and X. Zhou, "An improved method of optical flow using human body-following wheeled robot," *International Journal of Modeling, Simulation, and Scientific Computing*, vol. 10, no. 2, pp. 1950003.1–1950003.12, 2019.

[2] M. A. Nicolaou, S. Zafeiriou, I. Kotsia, G. Zhao, and J. Cohn, "Editorial of special issue on human behaviour analysis "in-the-Wild"," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 4–6, 2019.

[3] A. F. S. J. Syed and S. Shankar, "AR using NUI based physical therapy rehabilitation framework with mobile decision support system," *Journal of Global Information Management*, vol. 26, no. 4, pp. 36–51, 2018.

[4] P. Q. Brito and J. Stoyanova, "Marker versus markerless augmented reality. Which has more impact on users?" *International Journal of Human–Computer Interaction*, vol. 34, no. 7–9, pp. 819–833, 2018.

[5] J. Liu, H. Liu, Y. Chen, Y. Wang, and C. Wang, "Wireless sensing for human activity: a survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1629–1645, 2020.

[6] I.-J. Ding and Z.-G. Wu, "Two user adaptation-derived features for biometrical classifications of user identity in 3D-sensor-based body gesture recognition applications," *IEEE Sensors Journal*, vol. 19, no. 19, pp. 8432–8440, 2019.

[7] X. Wang, Y. Guo, Y. Guo, J. Li, and H. Liu, "Two gesture-computing approaches by using electromagnetic waves," *Inverse Problems & Imaging*, vol. 13, no. 4, pp. 879–901, 2019.

[8] A. Khelalef, F. Ababsa, and N. Benoudjit, "An efficient human activity recognition technique based on deep learning," *Pattern Recognition and Image Analysis*, vol. 29, no. 4, pp. 702–715, 2019.

[9] C. Selvarathi, "Human computer interaction using hand gesture recognition," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 2, pp. 1674–1677, 2020.

[10] A. A. Liu, Y. Shi, W. Z. Nie et al., "View-based 3D model retrieval via supervised multi-view feature learning," *Multimedia Tools and Applications*, vol. 77, no. 1, pp. 1–15, 2018.

[11] K. Liu and G. Kang, "Multiview convolutional neural networks for lung nodule classification," *International Journal of Imaging Systems and Technology*, vol. 27, no. 1, pp. 12–22, 2017.

[12] C. Ozcinar, E. Ekmekcioglu, J. Ćalić, and A. Kondoz, "Adaptive delivery of immersive 3D multi-view video over the Internet," *Multimedia Tools and Applications*, vol. 75, no. 20, pp. 12431–12461, 2016.

[13] L. U. Yonghua, "Multi-view based neural network for semantic segmentation on 3D scenes," *Ence China Information Ences*, vol. 62, no. 12, pp. 1–3, 2019.

[14] C. Rubino, M. Crocco, and A. Del Bue, "3D object localisation from multi-view image detections," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1281–1294, 2018.

[15] X. Cheng, N. Ikoma, M. Honda, and T. Ikenaga, "Multi-view 3D ball tracking with abrupt motion adaptive system model, anti-occlusion observation and spatial density based recovery in sports analysis," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 100, no. 5, pp. 1215–1225, 2017.

[16] N. I. Ratyal, I. Taj, U. I. Bajwa et al., "Pose and expression invariant alignment based multi-view 3D face recognition," *KSII Transactions on Internet and Information Systems*, vol. 12, no. 10, pp. 4903–4929, 2018.

[17] P. Ma, J. Ma, X. Wang, L. Yang, and N. Wang, "Deformable convolutional networks for multi-view 3D shape classification," *Electronics Letters*, vol. 54, no. 24, pp. 1373–1375, 2018.

[18] C. Xiao, R. Qin, X. Xie et al., "Individual tree detection and crown delineation with 3D information from multi-view satellite images," *Photogrammetric Engineering and Remote Sensing*, vol. 85, no. 1, pp. 55–63, 2018.

[19] K. Zempo, T. Kurahashi, K. Mizutani et al., "Analysis of retinal images for retinal projection type super multi-view 3D head-mounted display," *Electronic Imaging*, vol. 2017, no. 5, pp. 188–193, 2017.

[20] R. Böheim, C. Freudenthaler, and M. Lackner, "Gender differences in risk-taking: evidence from professional basketball," *IZA Discussion Papers*, vol. 7, no. 3, pp. 281–306, 2016.

[21] D. C. Bishop and C. Wright, "A time-motion analysis of professional basketball to determine the relationship between three activity profiles: high, medium and low intensity and the length of the time spent on court," *International Journal of Performance Analysis in Sport*, vol. 6, no. 1, pp. 130–139, 2017.

[22] H. Manner, "Modeling and forecasting the outcomes of NBA basketball games," *Journal of Quantitative Analysis in Sports*, vol. 12, no. 1, pp. 31–41, 2016.

[23] K. Tsunoda, H. Mutsuzaki, K. Hotta, Y. Shimizu, N. Kitano, and Y. Wadano, "Correlation between sleep and psychological mood states in female wheelchair basketball players on a Japanese national team," *Journal of Physical Therapy Science*, vol. 29, no. 9, pp. 1497–1501, 2017.

[24] L. Santos, J. Fernández-Río, B. Fernández-García, M. D. Jakobsen, L. González-Gómez, and O. E. Suman, "Effects of slackline training on postural control, jump performance, and myoelectrical activity in female basketball players," *Journal of Strength and Conditioning Research*, vol. 30, no. 3, pp. 653–664, 2016.

RETRACTED