Multimedia Systems (2017) 23:325–332
DOI 10.1007/s00530-015-0485-2

SPECIAL ISSUE PAPER

Convolutional deep learning for 3D object retrieval

Weizhi Nie[1] · Qun Cao[1] · Anan Liu[1] · Yuting Su[1]

Published online: 28 October 2015

Abstract In recent years, with the development of 3D technologies, 3D model retrieval has become a hot topic. The key point of 3D model retrieval is to extract robust feature for 3D model representation. In order to improve the effectiveness of method on 3D model retrieval, this paper proposes a feature extraction model based on convolutional neural networks (CNN). First, we extract a set of 2D images from 3D model to represent each 3D object. SIFT detector is utilized to detect interesting points from each 2D image and extract interesting patches to represent local information of each 3D model. X-means is leveraged to generate the CNN filters. Second, a single CNN layer learns low-level features which are then given as inputs to multiple recursive neural networks (RNN) in order to compose higher order features. RNNs can generate the final feature for 2D image representation. Finally, nearest neighbor is used to compute the similarity between different 3D models in order to handle the retrieval problem. Extensive comparison experiments were on the popular ETH and MV-RED 3D model datasets. The results demonstrate the superiority of the proposed method.

Keywords 3D model retrieval · Deep learning · CNN · RNN · X-means

✉ Anan Liu
anan0422@gmail.com

[1] School of Electronics Information Engineering, Tianjin University, Tianjin, China

1 Introduction

With the development of computer vision and computer equipment, a large scale of 3D models have been generated for industrial design, virtual game, multimedia living and human interaction, which has created an urgent demand for effective 3D model retrieval. The effective 3D model retrieval method can improve the utilization of virtual model and save the human cost. Thus, content-based 3D model retrieval methods have become a hot topic in recent years.

Many methods have been proposed to handle 3D model retrieval problem [10, 14, 21, 22, 35]. Ankerst et al. [1] proposed an optimal selection of 2D views from a 3D model, which focuses on numerical characteristics obtained from the 3D model representative features. Shih et al. [28] proposed the Elevation Descriptor (ED) feature, which is invariant to translation and scaling of 3D models. However, it is not suitable for the 3D model that consists of a set of 2D images. Ansary et al. [2] proposed a Bayesian model-based method for 3D object search, which utilizes X-means [3] to select characteristic views and applies the Bayesian model to compute the similarity between different models. These methods focus on the similar measurement between different models while ignore the low-level feature representation of 3D model.

Obviously, a robust low-level feature can effectively improve the performance of methods in many researching fields. For example, Liu et al. [18] proposed an interesting method for human action recognition based on HCRF [27]. In this paper, the author discussed the performance of different features based on the same classification. Obviously, more robust feature can bring better recognition results. In recent years, deep learning method is a hot topic in

feature learning, which can learn a robust feature for different objects by iterations. This conclusion is also proved by many prior works [10, 33]. Richard et al. [29] proposed a model based on convolutional neural networks (CNN) to learn feature for 3D object classification. The classic SVM is utilized to handle classification problem. The performance of feature learned by CNN is obviously better than other traditional methods. Liu et al. [19] proposed accurate content-based PET images retrieval that utilizes high-level ROI features with deep learning architecture, which also proves the effectiveness of deep learning in feature learning.

Based on these conclusions, we proposed an effective 3D model retrieval method based on CNN architecture. First, SIFT detector is utilized to detect the interesting points in order to extract interesting patches. X-means is leveraged to train the CNN filters. Second, a single CNN layer learns low-level features which are then given as inputs to multiple, fixed-tree RNNs in order to compose higher order features. RNNs can generate the final feature for 2D image representation. Finally, Nearest Neighbor is used to compute the similarity between different 3D models in order to handle the retrieval problem.

The main contributions of this paper are summarized as follows:

- We leveraged the SIFT detector to extract effective patches for feature learning. This design can reduce the redundancy of patches and guarantee the usefulness of each patch.
- We utilized the X-means to train the CNN filters. It is hard to make sure the size of CNN filter in deep learning process. X-means can get the best size of classification according to different training samples, which can effectively improve the performance of CNN model.
- This paper is organized as follows. The framework of this work will be illustrated in Sect. 2. we will detail the processes of feature learning based on CNN in Sect. 2. Experimental results and discussion are provided in Sect. 3. Finally, the conclusion is stated in Sect. 4.

2 Related work

In this section, we will review the representative work in the field of 3D model retrieval [4, 8, 13, 17, 26, 32] and deep learning. 3D model retrieval methods are mainly classified into two categories [6]: (1) geometry-based techniques, (2) view-based techniques [3, 23, 34, 37].

1. Geometry-based techniques. This approach applied shape distribution, shape histogram, and other 3D spatial information to represent a 3D model. For example, Ankerst et al. [1] introduced 3D shape histograms as intuitive feature vectors. The histograms were based on the partitioning of the space. Vandeborre et al. [31] proposed the use of completed 3D information. The 3D objects were represented as mesh surfaces and 3D shape descriptors were used. Zhao et al. [36] proposed a novel depth image-based method to handle action retrieval problem. Hilaga et al. [15] developed a matching technique to compute the similarity between two models based on the shape matching of their multi-resolutional Reeb Graphs. Murugappan et al. [24] compared 3D objects by applying graph-matching techniques to match their skeletons.
2. View-based techniques have attracted more attention in recent years because these techniques can utilize many proven techniques of digital images process. Wang et al. [34] applied group sparse coding to represent the image set of each object, and utilized the reconstruction error to compute the similarity between the query model and the candidate model. The authors [7] proposed a novel 3D model descriptor, spatial structure circular descriptor (SSCD), that can convert the structure information of a 3D model into 2D images and applied traditional image processing methods for 3D model retrieval. A learning method different from the existing pair-wise object distance measure is introduced in [11]. In this method, the hypergraph structure was employed to learn the object relevance. Daras et al. [5] proposed a novel compact multi-view descriptors (CMVDs) for 3D model representation. Camera arrays were set at the 18 vertices of a 32-hedron to capture the CMVD, where the multiple views were uniformly distributed.

All of these methods focus on the view generation methods and similarity measurement while ignore the low-level feature of 2D images. In recent years, deep learning provides an effective tool for feature learning. Sun et al. [30] proposed a novel deep identification-verification features based on deep convolutional networks, which can effectively solve the face identification and recognition. The experimental result also proved the performance of deep learning. Gao et al. [12] proposed an automatic feature learning grade nuclear cataracts based on deep learning. Neverova et al. [25] presented a method for gesture detection and localization based on multi-scale and multimodal deep learning. Many methods have proved the effectiveness of feature learning based on deep learning, which was the motivation of this paper.

3 The framework

The framework of this work includes three steps. (1) Filter training, in this step, we applied SIFT detector to detect

interesting region and extracted patches from each 2D image. Then, the unsupervised learning method X-means is used to train a set of CNN filters, which is the biggest different between our approach and traditional R-CNN methods. (2) Based on the above trained filters, we leveraged R-CNN model to learn feature for each 2D image from 3D objects. (3) Nearest Neighbor is used to compute the similarity between the query model and candidate model, which is utilized to generate the final retrieval result. We will detail these steps in the next subsections.

3.1 Filter training

In this step, we applied X-means to learn filters which will be used in the convolution. First, SIFT detector is utilized to detect interesting points. We extract patches according to these interesting points as center points. Then, these patches are normalized and whitened. X-means is used to cluster these patches. Finally, the centers of these clustering can be seen as the filters for feature learning.

3.2 Feature learning based on convolutional neural networks

Based on the filters learned by the above step, we utilized R-CNN method to extract the final feature for each image. Here, the CNN architecture is chosen for its translational invariance properties. The first step is to convolve all of filters over the input image in order to extract the final features. We convolve each image of size d_i with K square filters of size d_p, resulting in K filter responses. The final dimensionality will be $d_i - d_P + 1$. Then, we average pool them with square regions of size d_l and a stride size of s in order to obtain a pooled response equal to $r = (d_i - d_l)/s + 1$. So the output X of the first layer is a $K \times r \times r$ dimensional 3D matrix.

Then, we define a block to be a list of adjacent column vectors which are merged into a parent vector which are merged into a parent vector $p \in R^K$. Then, the square blocks are leveraged for convenience. The size of these blocks are $K \times b \times b$. In each block, we have b^2 vectors. The computing function is Eq. 1.

$$p = f\left(W\left[x_1, \ldots, x_{b^2}\right]'\right), \tag{1}$$

where the parameter matrix $W \in R^{K \times b^2 K}$, f is a nonlinearity such as tan$h$. Equation 1 is utilized in all blocks of vectors in X with the weight W. There will be $(r/b)^2$ parent vectors p forming a the matrix P_1. The vectors of P_1 will be merged in blocks again resulting in matrix P_2. This procedure continues until only one parent vector remains. The 3D matrix X is used as input to a number of RNNs. The output will be a K-dimensional vector. We concatenate all of outputs to generate a NK-dimensional vector as the feature of image.

3.3 Similarity measurement

Based on the high-dimension feature learned by our approach, Euclidean distance is utilized to compute the similarity between different 2D images. The reason of this design is that the goal of this paper is to demonstrate the performance of high-dimension feature. Thus, we applied the simple similarity measure method in order to highlight the role of feature. The similarity can be computed by Eq. 2

$$S_1\left(v_i, v_j\right) = 1 \Big/ \sqrt{\left(\varphi(v_i) - \varphi\left(v_j\right)\right)^2}, \tag{2}$$

where v_i and v_j represent two different views from 3D model, $\varphi(\cdot)$ represents the feature mapping function. Then, we used the Eq. 3 to calculate the similarity score between different 3D models.

$$S_2(Q, M) = \min_{i,j}^{n,m} \left\{S_1\left(v_i, v_j\right)\right\}, \tag{3}$$

where n represents the number of 2D images in model Q, m represents the number of 2D images in model M. The retrieved model with the highest similarity score can be achieved by $M^* = arg \max_{M_i \in \widetilde{M}} S_2(M_i, Q)$. $\widetilde{M}$ represents the candidate model set, M_i is the candidate model, Q is the query model.

4 Experiments

4.1 Dataset

The proposed method was evaluated on following datasets:

- ETH dataset [16]: The ETH dataset is a real-world 3D object multi-view database. It contains 80 objects categorized into 8 classes, and each class has 10 objects. There are 41 multiple views for each object spaced evenly over the upper viewing hemisphere, and all positions for cameras are determined by subdividing the faces of an octahedron to the third recursion level.
- The National Taiwan University 3D Model Database(NTU) [3]: The NTU dataset contains 500 objects in total. Virtual cameras are employed to capture initial views for 3D objects. The camera array contains 60 cameras, which are set on the vertices of a polyhedron with the same structure with Buckminsterfullerene (C60). Therefore, there are 60 views for each object.

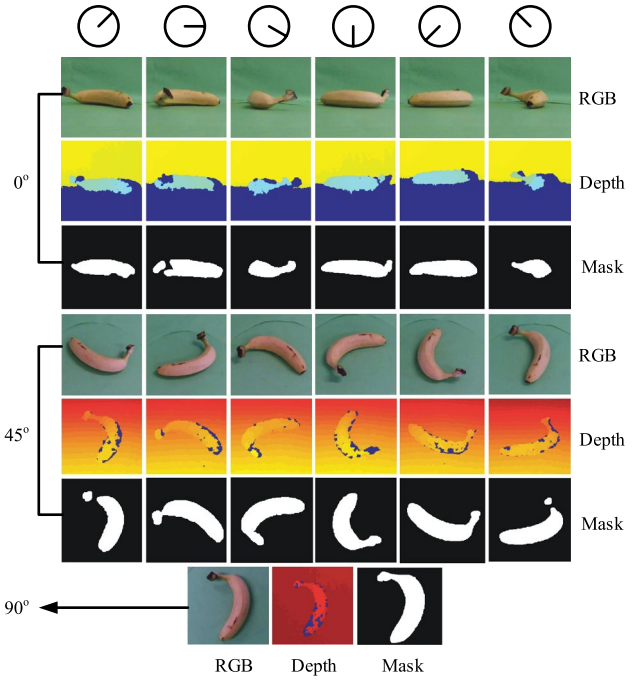

Fig. 1 RGB image and depth image of banana object

- The Multi-view RGB-D Object Dataset (MV-RED)[1]: MV-RED dataset is a real-world object dataset with multi-view and multimodal information. The MV-RED dataset consists of 505 objects belonging to 61 categories, which are selected apple, cap, scarf, cup, mushroom, toy and so on. For each object, both RGB and depth information were recorded. Some example views are provided in Fig. 1. The dataset is made publicly available so as to enable rapid progress based on this promising technology.

4.2 Evaluation

In order to evaluate the proposed method, the following criteria are employed as the evaluation measures of the retrieval performance.

- Nearest neighbor (NN). It is the percentage of the closest match models belonging to the querys category.
- First tier (FT). It is the recall for the first K relevant match samples, where K is the cardinality of the querys category.
- Second tier (ST). It is the recall for the first 2K relevant match samples, where K is the cardinality of the querys category.
- F-measure (F). It is a synthetical measurement of precision and recall for a fixed number of retrieved results.
- Discounted cumulative gain (DCG). It is a statistical measure that assigns higher weights to relevant results occupying the top-ranking positions.
- Average Normalized Modified Retrieval Rank (ANMRR). It measures the rank performance given a ranking list, which considers the ranking information of relevant objects among the top retrieved objects.
- Precision–recall curve (PR). It is a crucial indicator that shows the relationship between the precision and the recall.

4.3 Comparison to CNN and CNN-RNN

In this study, we utilized SIFT detector to detect interesting points and extract patches and leveraged X-means to train CNN filters for feature learning. In order to demonstrate the superiority of this proposed method. The CNN and CNN-RNN [20, 29] methods are selected for comparison. MV-RED dataset is selected as the evaluation dataset. The experimental results are shown in Fig. 2.

Figure 2a shows the precision–recall curves on MV-RED dataset by different methods. Figure 2b shows the performances by different methods on MV-RED dataset. From these experimental results, the proposed method outperforms CNN and CNN-RNN, which means that our design of filter training can extract more robust feature and improve the performance of the final feature. In this process, the dimension of feature is 4096, 8192, 6400 of CNN, CNN-RNN, and our approach, respectively. From this parameter, we can find that higher dimension feature cannot bring better retrieval results. One appropriate dimension is more suitable for object representation.

4.4 Comparison to traditional features

In order to demonstrate the effectiveness of high-dimension feature, we compare the proposed method with some traditional low-level features such as Zernike moment, HSV and HoG. MV-RED dataset is selected as the evaluation dataset. The experimental results are shown in Fig. 3.

Figure 3a shows the precision–recall curves on MV-RED by different methods. Figure 3b shows the performances by different methods on MV-RED dataset. From these experimental results, our approach obviously outperforms other retrieval results based on other low-level features, which fully demonstrate the effectiveness of deep learning in feature learning. Here, HSV is not utilized in NTU dataset. The reason is that NUT dataset is a virtual 3D model dataset. All of 3D models do not include color information, which leads that it is hard to extract effective feature to represent each 3D model. Thus, HSV is not used in NTU dataset to test the performance of the proposed method.

[1] http://media.tju.edu.cn/mvred/dataset1.html.

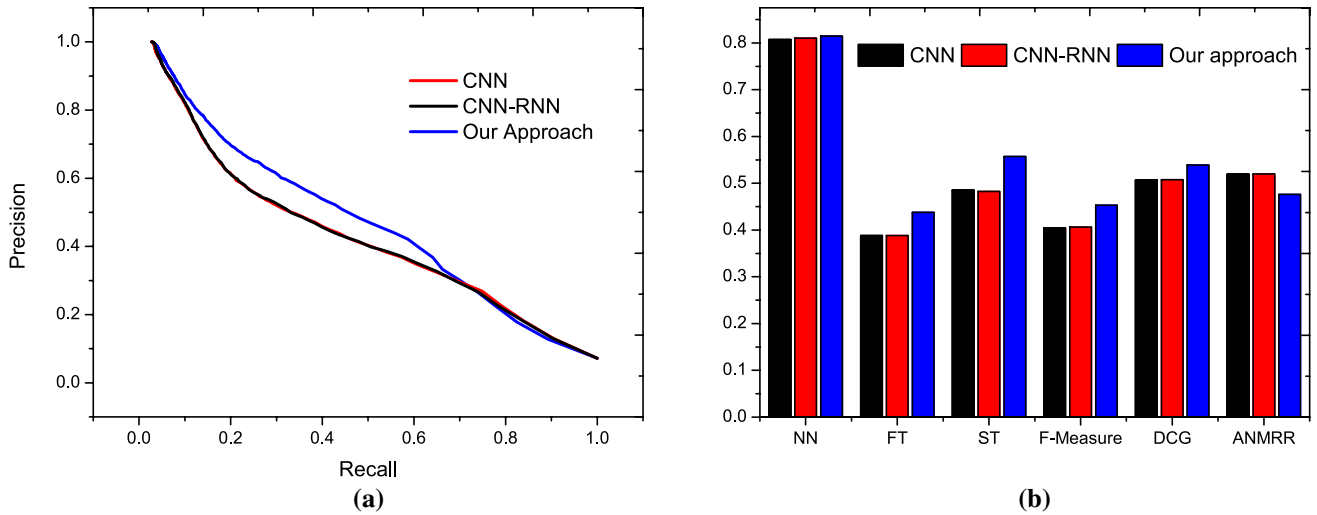

Fig. 2 The experimental results on MV-RED

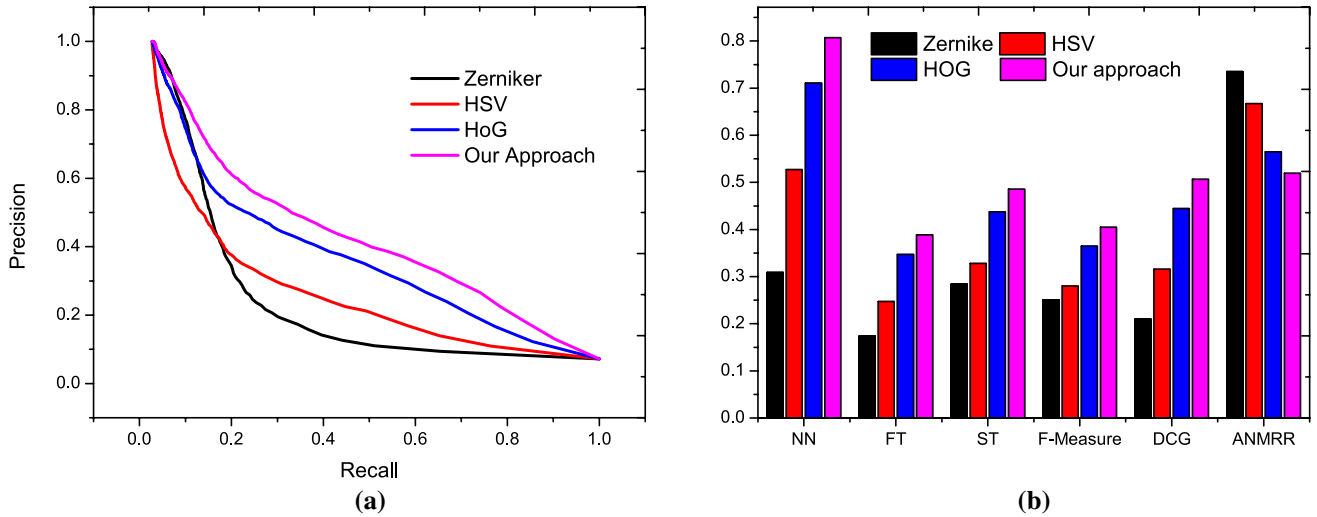

Fig. 3 The experimental results on MV-RED

4.5 Comparison to other methods

In order to evaluate the performance of the proposed method, some state-of-the-art methods are selected for comparison:

- Adaptive views clustering (AVC) [2]: The adaptive views clustering method provides an optimal selection of 2D views from a 3D model and a probabilistic Bayesian method for 3D model retrieval from these views. The characteristic views selection algorithm is based on an adaptive clustering algorithm and uses statistical model distribution scores to select the optimal number of views.
- Camera constraint-free view-based (CCFV) method [10]: In the camera constraint-free view method, each object is represented by a free set of views, which implies that these views can be captured from any direction without camera constraints. For each query object, all query views are clustered to generate the view cluster, which is then used to build the query models. For a more accurate 3-D object comparison, a positive matching model and a negative matching mode are individually trained using positive and negative matched samples, respectively. The CCFV model is generated on the basis of the query Gaussian models by combining the positive matching model and the

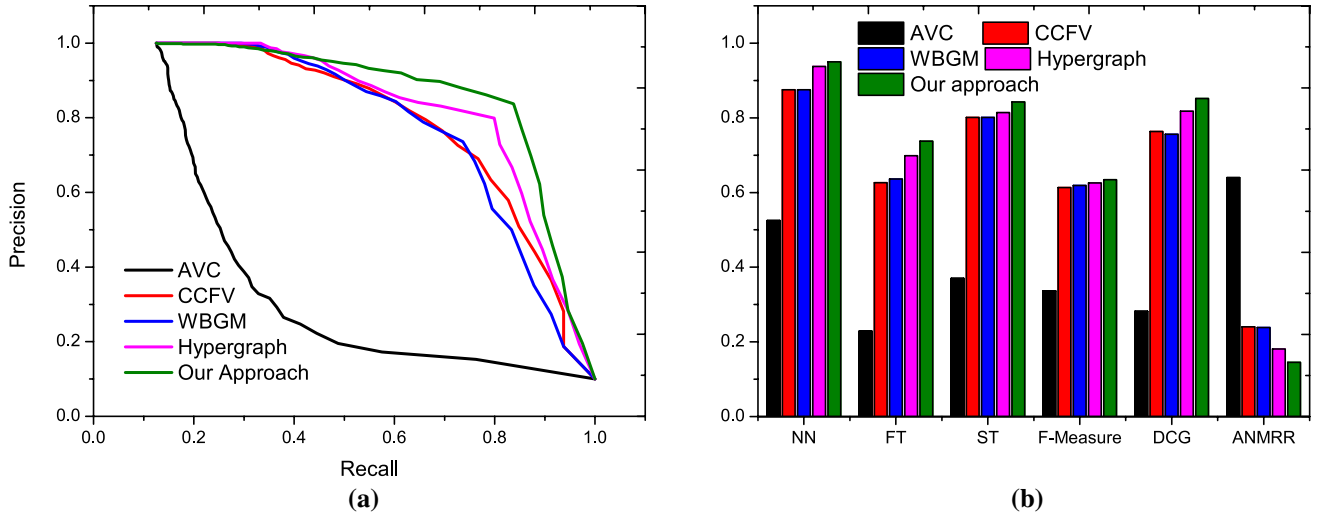

Fig. 4 The experimental results on ETH

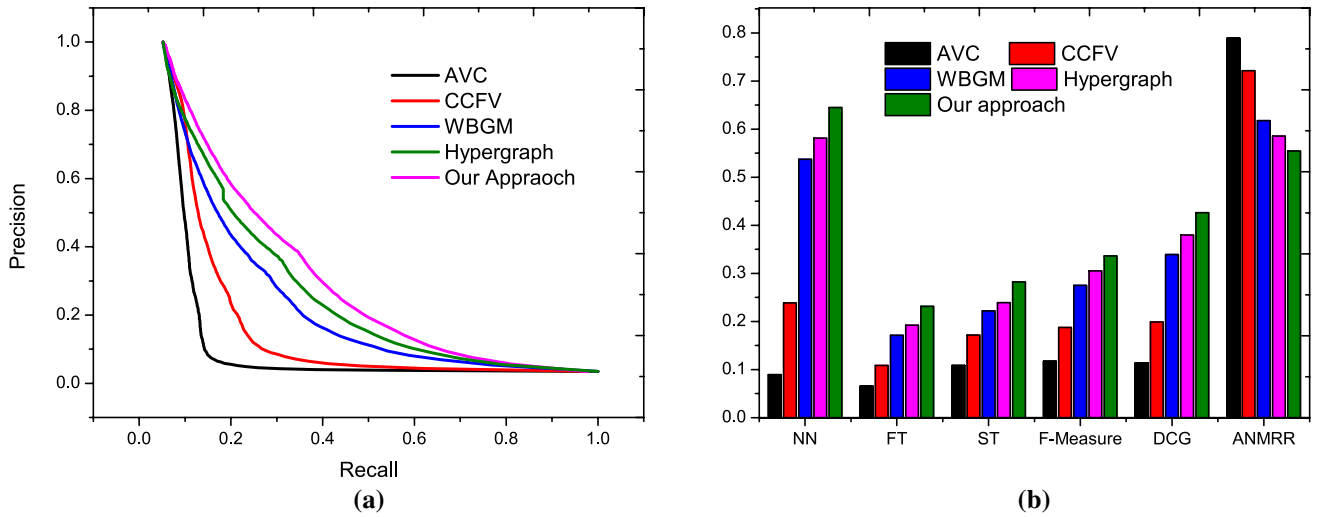

Fig. 5 The experimental results on ETH

negative matching model. The CCFV method removes the constraint of static camera array setting for view capturing and can be applied to any view-based 3D object database.

- Bipartite graph matching (WBGM) [9]: In this method, weighted bipartite graph matching is employed for comparison between two 3D models. Each 3D model is represented by a set of 2D views. Representative views are selected from the query model and the corresponding initial weights are provided. These initial weights are further updated based on the relationship among these representative views. The weighted bipartite graph is built with these selected 2D views and the matching result is used to measure the similarity between two 3D models.
- Hypergraph [11]: This method proposed a hypergraph analysis approach to address retrieval problem by avoiding the estimation of the distance between objects. The author constructed multiple hypergraphs for a set of 3D objects based on their 2D images. In this hypergraph, each vertex is an object, and each edge is a cluster of views. Therefore, an edge connects multiple vertices.
- The experimental results are shown in Figs. 4, 5 and 6. Figures 4a, 5a and 6a, respectively, show the precision–recall curves on ETH, NTU and MV-RED by different methods. Figures 4b, 5b and 6b show the performances by different methods on these three datasets, respectively. From these experimental results, the proposed method outperforms the other comparative methods, which also demonstrate the superiority of our approach.

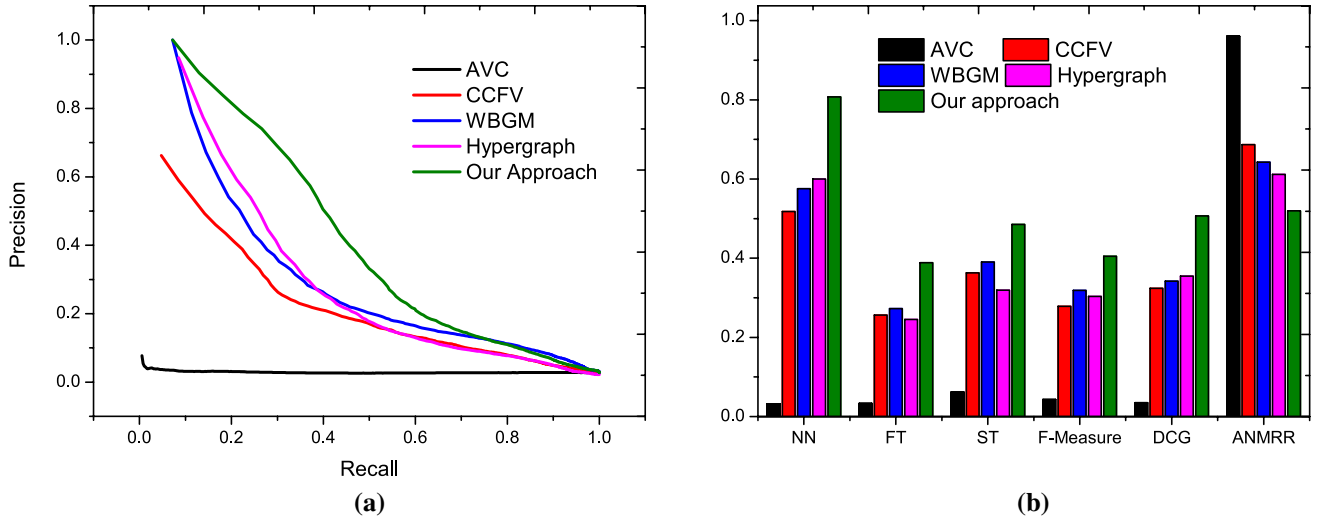

Fig. 6 The experimental results on MV-RED

5 Conclusion

In this paper, we proposed a feature extraction model based on convolutional neural networks. In this process, we leveraged the SIFT detector to extract effectiveness patches which can reduce the redundancy of patches. Then, X-means is used to train CNN filters, which significantly improved the robustness of feature. In this paper, we first utilized deep learning for 3D model retrieval. The extensive experiments on ETH and MV-RED demonstrate the superiority of this method.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China (61472275, 61170239, 61303208), the Tianjin Research Program of Application Foundation and Advanced Technology (15JCYBJC16200), and the Grant of Elite Scholar Program of Tianjin University (2014XRG-0046).

References

1. Ankerst, M., Kastenmüller, G., Kriegel, H.-P., Seidl, T.: 3D shape histograms for similarity search and classification in spatial databases. In: Advances in spatial databases, pp. 207–226. Springer, Berlin (1999)
2. Ansary, T.F., Daoudi, M., Vandeborre, J.P.: A bayesian 3-d search engine using adaptive views clustering. IEEE Trans. Multimed. **9**(1), 78–88 (2007)
3. Chen, D.-Y., Tian, X.-P., Shen, Y.-T., Ouhyoung, M.: On visual similarity based 3D model retrieval. Comput. Graph. Forum **22**(3), 223–232 (2003)
4. Chen, J.-Y., Lin, C.-H., Hsu, P.-C., Chen, C.-H.: Point cloud encoding for 3D building model retrieval. IEEE Trans. Multimed. **16**(2), 337–345 (2014)
5. Daras, P., Axenopoulos, A.: A 3D shape retrieval framework supporting multimodal queries. Int. J. Comput. Vis. **89**(2–3), 229–247 (2010)
6. Gao, Y., Dai, Q.: View-based 3D object retrieval: challenges and approaches. IEEE MultiMed. **21**(3), 52–57 (2014)
7. Gao, Y., Dai, Q., Zhang, N.Y.: 3D model comparison using spatial structure circular descriptor. Pattern Recognit. **43**(3), 1142–1151 (2010)
8. Gao, Y., Wang, M., Zha, Z.J., Tian, Q., Dai, Q., Zhang, N.: Less is more: efficient 3-D object retrieval with query view selection. IEEE Trans. Multimed. **13**(5), 1007–1018 (2011)
9. Gao, Y., Dai, Q., Wang, M., Zhang, N.: 3D model retrieval using weighted bipartite graph matching. Image Commun. **26**(1), 39–47 (2011)
10. Gao, Y., Tang, J., Hong, R., Yan, S., Dai, Q., Zhang, N.Y., Chua, T.S.: Camera constraint-free view-based 3-D object retrieval. IEEE Trans. Image Process. **21**(4), 2269–2281 (2012)
11. Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q.: 3-D object retrieval and recognition with hypergraph analysis. IEEE Trans. Image Process. **21**(9), 4290–4303 (2012)
12. Gao, X., Lin, S., Wong, T.Y.: Automatic feature learning to grade nuclear cataracts based on deep learning. In: Computer Vision—ACCV 2014, pp. 632–642. Springer, Switzerland (2015)
13. Gao, Z., Zhang, H., Liu, A.A., Xu, G., Xue, Y.: Human action recognition on depth dataset. Neural Comput. Appl. (2015). doi:10.1007/s00521-015-2002-0
14. Guo, Y., Sohel, F., Bennamoun, M., Wan, J., Lu, M.: A novel local surface feature for 3D object recognition under clutter and occlusion. Inf. Sci. **293**, 196–213 (2015)
15. Hilaga, M., Shinagawa, Y., Komura, T., Kunii, T.L.: Topology matching for fully automatic similarity estimation of 3d shapes. In *SIGGRAPH*, pp. 203–212 (2001)
16. Leibe, B., Schiele, B.: Analyzing appearance and contour based methods for object categorization. In: Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on, vol. 2, pp. II–409. IEEE (2003)
17. Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., Burtscher, M., Fu, H., Furuya, T., Johan, H., et al.: Extended large scale sketch-based 3D shape retrieval. Eurograph. Assoc. **73**(4), 128–139 (2014)
18. Liu, A., Han, D.: Spatiotemporal sparsity induced similarity measure for human action recognition. JDCTA **4**(8), 143–149 (2010)
19. Liu, S., Liu, S., Cai, W., Che, H., Pujol, S., Kikinis, R., Fulham, M., Feng, D.: High-level feature based pet image retrieval with deep learning architecture. J. Nucl. Med. **55**(supplement 1), 2028–2028 (2014)

20. Liu, A., Su, Y., Nie, W., Yang, Z.: Jointly learning multiple sequential dynamics for human action recognition. PLoS ONE **10**(7), 1–21 (2014). doi:10.1371/journal.pone.013088
21. Liu, A., Su, Y., Jia, P., Gao, Z., Hao, T., Yang, Z.: Multiple/single-view human action recognition via part-induced multi-task structural learning. IEEE Trans. Cybern. **45**(6), 1194–1208 (2015)
22. Liu, A., Wang, Z., Nie, W., Su, Y.: Graph-based characteristic view set extraction and matching for 3D model retrieval. Inf. Sci. **320**, 429–442 (2015)
23. Liu, A., Nie, W., Su, Y., Ma, L., Hao, T., Yang, Z.: Coupled hidden conditional random fields for RGB-D human action recognition. Signal Process. **112**, 74–82 (2015)
24. Murugappan, S., Liu, H., Ramani, K.: Shape-it-up: hand gesture based creative expression of 3D shapes using intelligent generalized cylinders. Comput. Aided Des. **45**(2), 277–287 (2013)
25. Neverova, N., Wolf, C., Taylor, G.W., Nebout, F.: Multi-scale deep learning for gesture detection and localization. In: Computer Vision-ECCV 2014 Workshops, pp. 474–490. Springer (2014)
26. Paquet, E., Rioux, M., Murching, A.M., Naveen, T., Tabatabai, A.J.: Description of shape information for 2-D and 3-D objects. Signal Process. Image Commun. **16**(1–2), 103–122 (2000)
27. Richter, R.M., Mulvany, M.J.: Comparison of hCRF and oCRF effects on cardiovascular responses after central, peripheral, and in vitro application. Peptides **16**(5), 843–849 (1995)
28. Shih, J.L., Lee, C.H., Wang, J.T.: A new 3D model retrieval approach based on the elevation descriptor. Pattern Recognit. **40**(1), 283–295 (2007)
29. Socher, R., Huval, B., Bath, B., Manning, C.D., Ng, A.Y.: Convolutional-recursive deep learning for 3D object classification. In: Advances in Neural Information Processing Systems. In: NIPS, pp. 665–673 (2012)
30. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: IEEE conference on computer vision and pattern recognition (CVPR), 2014 IEEE pp. 1891–1898 (2014)
31. Vandeborre, J.P., Couillet, V., Daoudi, M.: A practical approach for 3D model indexing by combining local and global invariants. In: 3DPVT, pp. 644–647 (2002)
32. Vranic, D.V.: An improvement of rotation invariant 3D-shape based on functions on concentric spheres. ICIP **3**, 757–760 (2003)
33. Wang, F., Lin, L., Tang, M.: A new sketch-based 3D model retrieval approach by using global and local features. Graph. Models **76**(3), 128–139 (2014)
34. Wang, X., Nie, W.: 3D model retrieval with weighted locality-constrained group sparse coding. Neurocomputing **151**, 620–625 (2015)
35. Xu, Q., Liu, Y., Li, X., Yang, Z., Wang, J., Sbert, M., Scopigno, R.: Browsing and exploration of video sequences: a new scheme for key frame extraction and 3D visualization using entropy based Jensen divergence. Inf. Sci. **278**, 736–756 (2014)
36. Zhao, S., Yao, H., Yang, Y., Zhang, Y.: Affective image retrieval via multi-graph learning. In: Proceedings of the ACM international conference on multimedia, MM '14, Orlando, FL, USA, November 03–07, 2014, pp. 1025–1028 (2014)
37. Zhou, J.L., Zhou, M.Q., Geng, G.H.: 3D model retrieval based on distance classification histogram. Appl. Mech. Mater. **733**, 931–934 (2015)