TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 09/14 pp93–102
DOI: 10.26599/TST.2018.9010113
Volume 25, Number 1, February 2020

Web3D Learning Framework for 3D Shape Retrieval Based on Hybrid Convolutional Neural Networks

Wen Zhou*, Jinyuan Jia, Chengxi Huang, and Yongqing Cheng

Abstract: With the rapid development of Web3D technologies, sketch-based model retrieval has become an increasingly important challenge, while the application of Virtual Reality and 3D technologies has made shape retrieval of furniture over a web browser feasible. In this paper, we propose a learning framework for shape retrieval based on two Siamese VGG-16 Convolutional Neural Networks (CNNs), and a CNN-based hybrid learning algorithm to select the best view for a shape. In this algorithm, the AlexNet and VGG-16 CNN architectures are used to perform classification tasks and to extract features, respectively. In addition, a feature fusion method is used to measure the similarity relation of the output features from the two Siamese networks. The proposed framework can provide new alternatives for furniture retrieval in the Web3D environment. The primary innovation is in the employment of deep learning methods to solve the challenge of obtaining the best view of 3D furniture, and to address cross-domain feature learning problems. We conduct an experiment to verify the feasibility of the framework and the results show our approach to be superior in comparison to many mainstream state-of-the-art approaches.

Key words: Web3D; sketch-based model retrieval; Convolutional Neural Networks (CNNs); best view; cross-domain

1 Introduction

With rapid development both in 3D technologies and Virtual Reality (VR), the potential for virtual furniture design in the web browser has attracted attention. This new concept of 3D plus VR furniture design depends to some extent on technologies of 3D image retrieval to achieve a better user experience and permit new forms of human-computer interaction. The traditional method of text-based retrieval, requiring manual image annotation, is a tedious and difficult task, especially with the explosive growth of storage capacities. Besides its tedium, a defect of text based image retrieval is becoming increasingly apparent, which is its subjective and biased nature. Therefore, Sketch Based 3D Model Retrieval (SBMR) has attracted significant interest in the field of image retrieval. However, SBMR still presents a major problem of how to efficiently acquire and accurately represent the descriptor of a hand-drawn sketch. There also remains a difficulty in the asymmetry in dimensions between a sketch and a 3D model; that is to say, how to obtain the best view for a represented 3D model.

There have been many notable previous contributions to this research field. Kato et al.[1] proposed a Query by Visual Example (QVE) method, while Niblack et al.[2] developed the Query by Image and Video Content

• Wen Zhou is with School of Computer and Information, Anhui Normal University, Wuhu 241002, China. E-mail: w.zhou@ahnu.edu.cn.
• Jinyuan Jia is with School of Software Engineering, Tongji University, Shanghai 201804, China. E-mail: jyjia@tongji.edu.cn.
• Chengxi Huang is with College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China. E-mail: 1710051@tongji.edu.cn.
• Yongqing Cheng is with School of Engineering and Computer Science, University of Hull, Hull, HU6 7RX, UK. E-mail: Y.cheng@hull.ac.uk.
∗ To whom correspondence should be addressed.

Manuscript revised: 2018-07-02; accepted: 2018-07-20

system (QIVC). Under their methods, the user submits an "example image" similar to what they were seeking, and the system retrieves and displays a list of images in the database that are visually similar to the user's query. However, it was sometimes difficult to find appropriate example images to submit as a query. Therefore, an alternative approach for querying an image retrieval system was explored, which was for the user to simply draw what they had in mind, which further represented an intuitive means of communication between a user and the system. This led to a new kind of query method known as sketch-based retrieval, which gave rise in turn to several new problems. According to Hu and Collomosse[3], a key challenge for sketch-based retrieval is how to overcome the ambiguity inherent in a sketch, which contains less information than an example image—perhaps only the contour and some pixels or strokes. Compounding this difficulty are the very large differences caused by the varying drawing ability level of different users. In addition, one salient characteristic of a sketch is the stroke orientation. Orientation is a characteristic that has been exploited widely and is achieving good results in tasks like object recognition and object categorization. Furthermore, due to the lack of features in a sketch, it demands the use of many more robust descriptors to exploit the relationship between a sketch and images or view images of a shape.

The contributions of this paper can be summarized as follows.

(1) We propose an innovative method to obtain the best view for a shape, providing a feasible alternative for solving the asymmetry problem between sketch and shape.

(2) We present a hybrid networks schema to finish the task of extracting features and classification. In practice, a lack of information is a basic characteristic of a sketch, and it is often difficult to obtain better features via deep neural networks. In addition, the learning method for obtaining the best view for a shape greatly depends on good training samples; however, these samples need to be selected based on features of the sketch and view image.

(3) We present two Siamese networks for shape retrieval with similarity metric method to measure the relationship between their output features. Furthermore, this proposed method can be extended into new methods of human-computer interaction for VR applications.

The outline of this paper is as follows: in Section 2, we present related work on image retrieval methods; in Section 3, we put forward our innovative framework; in Section 4, we explain our framework, including the training and testing pipelines, and describe our hybrid Convolutional Neural Network (CNN) based learning method for selecting the best view; the experimental results, evaluation, and comparison with other approaches are presented in Section 5; and Section 6 concludes the work and looks forward to future research.

2 Related Work

In recent years, 3D model retrieval has been a popular research topic in the fields of computer graphics, information retrieval, and pattern recognition. However, with rapid growth in the scale and diversity of 3D model data, the identification, retrieval, reuse, and re-modeling of 3D data have become common issues of concern to designers, engineers, and researchers.

Relatively complete sketch-based 3D model retrieval systems are currently found in the systems outlined in Refs. [4–7]. There are two key points in SBMR system: the 2D transformation and the extraction of the sketch features to a 3D model. The quality of these two steps directly determines the accuracy of the search results. Eitz et al.[6] realized a sketch-based image retrieval algorithm using bag-of-words and Histogram of Orient Gradients (HOG) methods. However, these methods did not conduct any pre-processing operations before retrieval, which meant that the result could be affected by ambiguous strokes in the sketch or by poor drawing abilities leaving the sketch not accurately expressing the user's purpose. Funkhouser et al.[8] proposed a 3D model retrieval engine that supports the switch between 3D and 2D, with a method based on spherical harmonics. Therefore, Li et al.[9] proposed a pre-processing operation before retrieval; this would check the user's hand-drawn sketch and display a possible sketch matching the user's demand. To date, eight benchmarks have been formed for sketch retrieval: a standard set of line drawings proposed by Snograss and Vanderwart[10] in 1980, the line drawing benchmark of Cole et al.[11] in 2008, the sketch dataset of Saavedra and Bustos[12] in 2010, the sketch-based 3D model retrieval benchmark of Yoon et al.[13] in 2010, the sketch-based shape retrieval and sketch recognition benchmarks of Eitz et al.[6,7] in 2011 and 2012, respectively, and the large-scale SHREC'2013 track benchmark presented by Li et al.[9] in 2013. These benchmarks all played important roles in the research and application of sketch-based retrieval.

Saavedra and Bustos[12] introduced an improved descriptor, Histograms of Edge Local Orientations (HELO). HELO takes a cell-wise strategy; therefore, it seemed highly appropriate for representing sketch-like images. Saavedra[14] also proposed a Soft computation variant of HELO (S-HELO); it computes cell orientations in a soft manner using bilinear and tri-linear interpolation, and takes spatial information into account. Finally, it computes an orientation histogram using weighted votes from the estimated cell orientations.

Dalal and Triggs[15] presented the HoG descriptor, which can capture edges of gradient structures. Translations and rotations make very little difference with a small local spatial or orientation bin size. However, due to the fact that the HOG descriptor follows a pixel-wise strategy, the representation of a sketch image always produces many zeroes in the final histogram, since a sketch tends to be sparse by nature. Fu et al.[16] also introduced an improvement on the HOG descriptor, namely, the Binary HOG descriptor (BHOG). This can compute the feature vectors more quickly than the HOG descriptor while also taking up less memory.

On the other hand, a cross-domain CNN approach has recently been successfully used in sketch-based 3D retrieval. For example, training two Siamese cross-domain CNNs[17] can obtain excellent accuracy. However, these methods do not focus on how to obtain the best view image, only imposing minimal assumptions when choosing views for the whole dataset (e.g., 3D models in the dataset are upright). In addition, Pyramid Cross-Domain Neural Networks (PCDNNs)[18] were proposed to conduct the learning process between sketch and shape. PCDNN relies on a multi-layer pyramid; however, the number of pyramid layers required to correctly represent a shape or sketch is hard to determine.

In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton won the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) with an error rate of 16%, a huge improvement on the best models from the first two years of the challenge, which achieved error rates of around 28% and 26%. The solution contained several aspects of deep learning that are now standard in any deep learning implementation. It was the first time that a CNN architecture, the famous AlexNet[19], had beaten other methods by a large margin. The VGG group were runners up in ILSVRC 2014 with a 16-layer architecture named VGG-16[20], and a 152-layer-deep convolutional neural network from Microsoft named ResNet[21] won ILSVRC 2015 with an error rate of only 3.6%, which is better than the perceived human error rate of 5%–10%. However, these networks were designed for images, not sketches. In this paper, VGG-16 and AlexNet are used to conduct the task of classification and extracting features. We did not select a deeper network, such as ResNet, because a sketch is in fact a lack of features, merely consisting of several abstract strokes. Because of this characteristic of sketches, it is difficult to produce better results from very deep networks to justify the requirement for more training and/or predication time.

3 Proposed Framework

The framework is mainly comprised of two parts: one is a hybrid CNN-based learning pipeline for obtaining the best view, and the other is a CNN-based learning pipeline for shape retrieval. An overview of the proposed framework can be seen in Fig. 1.

In the pipeline of obtaining the best view for a shape, three tasks must be performed, i.e., collecting training samples, training network for fit-related parameters, and testing or predicting the label of view images. We can only obtain good view images, not the best view image; in order to further obtain the best view for a shape, a ranking operation is performed based on these good view images.

In addition, in the pipeline of shape retrieval, training the network is a key step to obtaining the related network parameters. Siamese networks are built to conduct cross-domain learning tasks. A similarity metric is utilized to measure the relation between the output features of the two Siamese networks. In this pipeline, the training samples are based on pairing a sketch and its best view image.

The following section describes the proposed method in more detail.

4 Framework Description

In this section, a CNN-based learning method is proposed to perform the two key tasks of our framework: shape retrieval and determining the best view for a shape.

4.1 Hybrid CNN-based best view for shape

Eitz et al.[7] adopted a Support Vector Machine (SVM) classifier to categorize multi-view images, projected

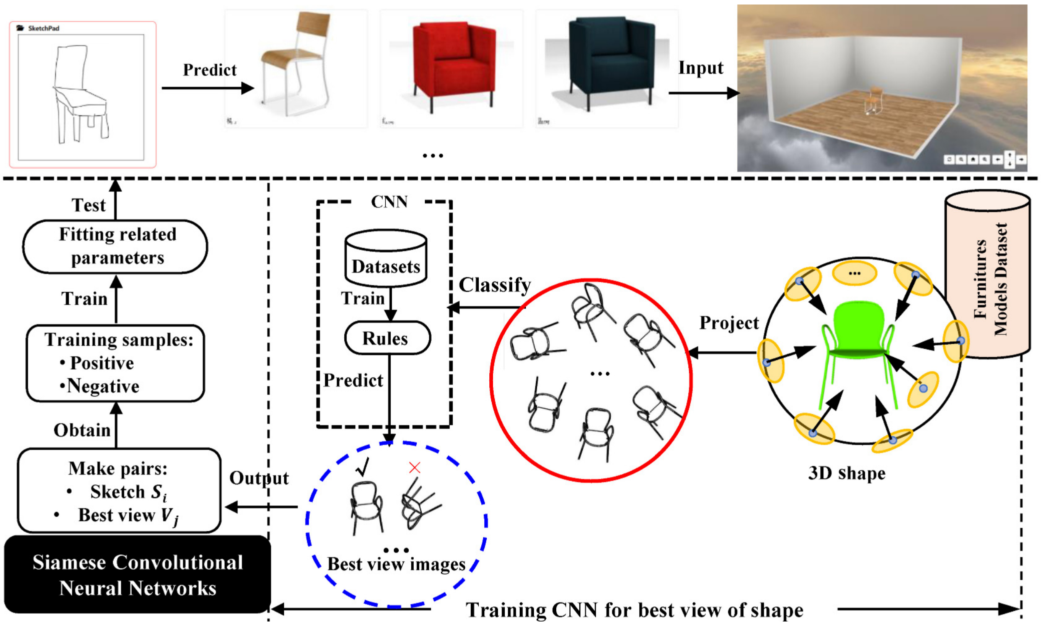


Fig. 1 Overview of proposed framework.

from the model from many different viewpoints. Under this method, three hundred cameras were uniformly placed on the bounding sphere of a model, so that it could be projected into multiple view images. However, a large number of view images are bad, so we need to train an intelligent classifier to categorize the view images and eliminate the negative interference that bad viewpoint images have on our retrieval results. This method is adopted by Zhao et al.[22] to acquire the best-view images for a shape. In this paper, a hybrid CNN-based supervisor learning method is used to obtain the best view of a shape. Moreover, the AlexNet convolutional neural network and VGG-16 CNN architecture are used. AlexNet has obtained good results in image classification, and the addition of VGG-16 CNNs allows for the extraction of features with structures more complex than AlexNet. The detail of our adopted CNNs is presented in Section 4.3. Our proposed method for obtaining the best view can be seen in Fig. 2.

Figure 2 shows the process of the learning method for obtaining the best view of a shape. The steps taken to obtain the best view of a shape are as follows.

Step 1. Obtain pairs set We obtain a pair set of sketches S and furniture models H from a dataset, such as SHREC'2013. Then, we project the i-th shape H_i into N different view images, represented as the term $V_i = \{t \in [0, T-1] | v_i^t\}$. In this way, a pair set can be generated, represented as $P_i = \{0 \leqslant t \leqslant N-1 | (v_i^t, s_j^k)\}$, where the term s_j^k represents the k-th sketch in the j-th category of sketch dataset S. In the Eitz sketch dataset that proposed by

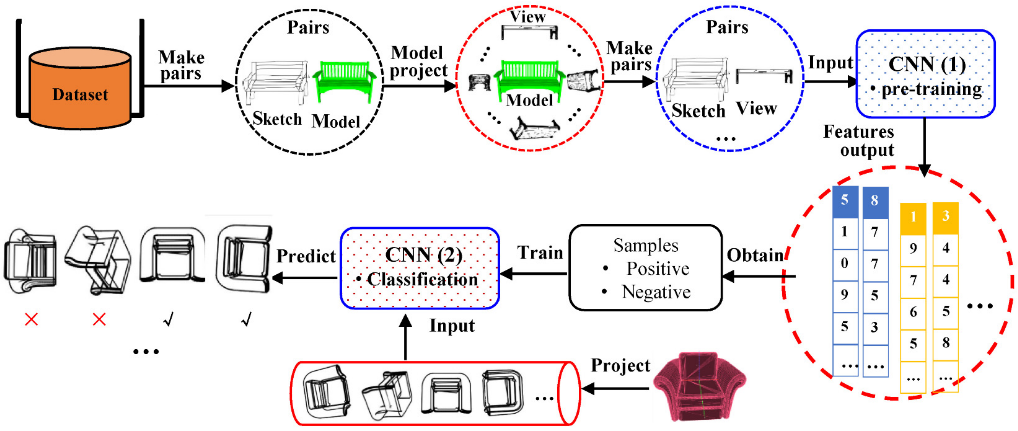


Fig. 2 Overview of best view for shape based on two hybrid CNNs.

Eitz et al.[23], there are 20 000 sketches in 250 different categories, i.e., there are 80 sketches in each category.

Step 2. Collecting samples We must obtain many different positive and negative samples in order to train the network. In general, the bad view images are often believed to be the negative samples, whereas the good view images are seen as positive samples. However, we cannot know what is a good or bad view image for a given shape; at least, no related views dataset exists. Therefore, we assume that the hand-drawn sketch is an example of a good view. Based on this assumption, a similarity metric between the sketch and view image is proposed to obtain the good views and bad views for a shape.

A VGG-16 CNN is used to extract the features of every pair. Our aim is to obtain positive and negative samples; there are always positive and negative pairs between a sketch and the projected view images.

Step 3. Similarity metric In order to measure the similarity relationship of each pair, we define a function (Eq. (1)) to represent their relation.

$$S_{cnn}(x, y) = \sum_{i} \exp\left(-\frac{d_{euc}^2(f_i(x), g_i(y))}{2\sigma^2}\right) \tag{1}$$

where the terms x and y represent sketch and view image, respectively, and the terms f and g denote the extracted features of the sketch and view image, respectively, by the network. The term σ is a constant value; in this paper, we set it to 0.2. The function d_{euc} is the Euclidean distance equation.

A pre-trained VGG-16 CNN is used to extract images features, to obtain more accurate results. VGG-16 CNNs with pre-training are known to obtain good results in image feature extraction tasks.

Step 4. Determine positive and negative samples We define a decision function to determine which samples are positive and which are negative. In addition to the above similarity metric method, a probability function is needed to classify samples into positive and negative categories. For all view images $V_i = \{0 \leqslant t < T | v_i^t\}$ for shape H_i, and M sketches $S_j = \{0 \leqslant k < M | s_j^k\}$ from the j-th category of the sketch dataset, the similarity probability between any view image $v_i^t \in V_j$ and any sketch $s_j^k \in S_j$ is calculated as in Eq. (2). Based on this, the decision function is then defined as Eq. (3).

$$P_{cnn}(s_j^k, v_i^t) = \frac{S_{cnn}(s_j^k, v_i^t) - \min\limits_{0 \leqslant k < M}(s_j^k, v_i^t)}{\max\limits_{0 \leqslant k < M}(s_j^k, v_i^t)} \tag{2}$$

$$D(v_i^t) = \begin{cases} 1, & \text{if } \exists s_j^k \in S_j, P_{cnn}(s_j^k, v_i^t) \geqslant 0.9; \\ 0, & \text{if } \forall s_j^k \in S_j, P_{cnn}(s_j^k, v_i^t) \leqslant 0.1; \\ -1, & \text{otherwise} \end{cases} \tag{3}$$

where the sample v_i^t should be removed when $D(v_i^t) = -1$. Additionally, it is a positive simple if $D(v_i^t) = 1$ and a negative one if $D(v_i^t) = 0$.

Step 5. Train a CNN classifier According to the above decision function, we can obtain the view image v_n^i as a positive sample or negative sample.

These samples can be used to train our AlexNet. In particular, AlexNet needs to set its parameters as they are very important to predicting the label of input data hereafter.

Step 6. Ranking the views In order to add to the diversity of best views, we have to remove good views projected from nearby positions. In fact, our method can obtain many good views, but they are mostly similar. Therefore, we adopt the Intersect of Unions (IoUs) method to remove these similar view images. In addition, a repressive function is defined to decrease the scores of these good views; the repressive function follows as Eq. (4).

$$\Delta(x) = \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{4}$$

The ranking method is based on Eq. (5).

$$\Theta_t = \Psi_t + \Delta\Big(\sum_{\{v_i^q | v_i^q \in V_i\}} IoUs(v_i^t, v_i^q)\Big) \tag{5}$$

where the term $\Psi_t \in [0, 1]$ represents the prediction value of the view image v_i^t. Based on the SoftMax function, we can obtain the value Ψ_t. Besides, the term V_i is the set of view images projected by the shape H_i. Finally, using the Mean-Shift algorithm to rank the view images, the best view images can be obtained.

4.2 Architecture of hybrid CNNs

In this paper, hybrid CNN networks are used to perform the corresponding tasks, such as classification and extracting features. The AlexNet and VGG-16 networks have been successfully used to conduct related jobs, obtaining good results in the fields of image classification and recognition. A hybrid of different CNNs are used to improve the efficiency of training samples; the task of VGG-16 in this paper is to extract features, while AlexNet has achieved good results in classification tasks and is used to classify features. The structure of VGG-16 can be seen in Fig. 3, while the simpler architecture of AlexNet can be seen in Fig. 4.

The Sigmoid function is used as the activation function in the CNNs. Due to the fact that sketches and

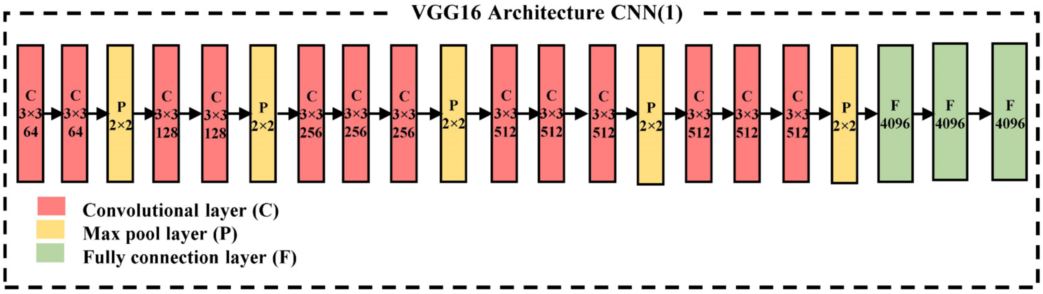


Fig. 3 Architecture of VGG16.

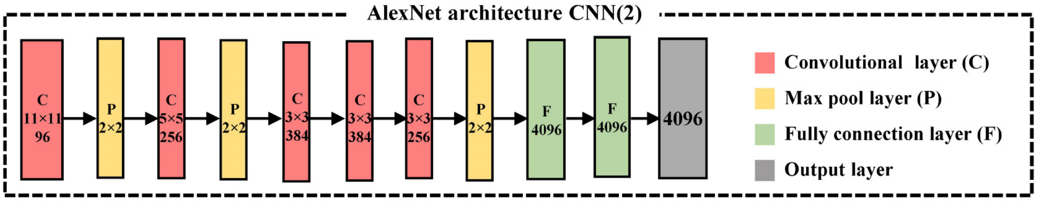


Fig. 4 Architecture of Alex Net CNN.

view images feature a lack of salient rich information, the simpler structure of AlexNet can obtain good results as a classifier. In this paper, the cost function of networks is based on the cross-entropy function and the Adam optimizer is used to minimize the cost function. Therefore, in the training stage, the backpropagation method is used to set the related parameters of networks. In the testing stage, the labeling of input data is based on the feed-forward method.

4.3 Training Siamese CNNs for shape retrieval

The best view for a shape can be obtained using the methods outlined above in Section 4.1. In this section, we perform shape retrieval between sketches and view images. In particular, we build pair relationships between sketches and view images. The framework also presents a kind of Zero-Shot learning schema. In addition, Wang et al.[17] was successful with shape retrieval based on two Siamese cross-domains CNNs; a similarity metric method is used to measure the similarity relationship between sketches and views, as was proposed by Chopra et al.[24] for performing face verification; therefore, we also adopt this model to measure the sketch and view relationship. Two VGG-16 CNNs with pre-trained Siamese networks are adopted; one, the sketch CNN, is to acquire the sketch features, while the other, the view CNN, is to acquire the best view images for a shape.

Based on the backpropagation algorithm, with the optimizer minimizing the cost function of the network, the parameters of VGG-16 can be adjusted to obtain better results.

Meanwhile, the pre-training of the CNNs aims to accelerate the training procedure. This is important for very large samples. The training process of the proposed learning framework can be seen in Fig. 5. The Adam

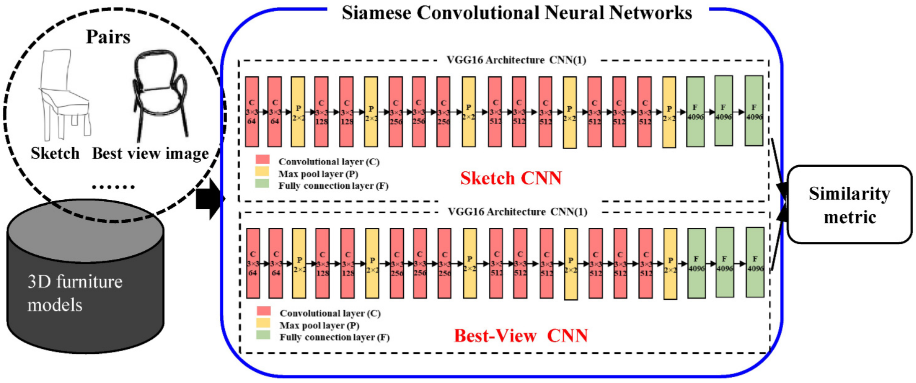


Fig. 5 Training process of Siamese CNN-based learning method for Shape retrieval. Similarity metric (see Eq. (7)) is used to measure the relation between the output features of two Siamese CNNs.

optimizer and cross entropy cost function are also used in training the network. The similarity metric method is shown in Eq. (6), which can be used to measure the relation between the output features of the sketch CNN and view CNN,

$$\Psi(s_i, v_j; b) = (1-b)\alpha D_{Man}^2 + b\beta \mathrm{e}^{\gamma D_{Man}} \qquad (6)$$

where the term b is the binary similarity label, that is, $b = 0$ or $b = 1$; the term D_{Man} is the Manhattan distance between the sample s_i and v_j. Moreover, the terms α, β, and γ are experimental values, we set them to 5, 0.1, and -0.277, respectively. In order to enhance the retrieval performance, we fusion different pairs of features, including sketch and sketch, and sketch and view. Therefore, the fusion model can be represented as in Eq. (7).

$$\Psi(s_i, v_j; b) = \Psi(s_i, s_j; b) + \Psi(s_i, v_k; b) \qquad (7)$$

A similarity matrix can be obtained from Eq. (7); we can then directly use this matrix to finish the retrieval task.

4.4 Testing CNNs for shape retrieval

This section presents details of how to test Siamese CNNs for Shape retrieval. In the previous section, the parameters of the networks were obtained; the predication of networks can now be rapidly conducted. In fact, we conduct once retrieval, the network needs to perform N (the term N represents the number of best view images for a shape) predications, i.e., the input sketch s_i needs to form a pair with every shape in the dataset while, at the same time, for every shape, there are several best view images.

Figure 6 presents an overview of testing for shape retrieval; the process consists of three steps as follows.

Step 1. Network predication For any model H_i, its best view images can be represented in mathematical form as follows: $V_i = \{0 \leqslant k < n | v_i^k\}$; in this paper, the term n is set to 3. From the Siamese networks conducting the related predication, there exists a view image v_i^k in the term V_i with its label predication $y(v_i^k, s_0) = 1$, then the shape H_i belongs to the retrieval result. Besides, the predication value z of shape H_i can be represented as $Z = \max(y_{pred}^{(0)}, y_{pred}^{(1)})$, where the term $y_{pred}^{(i)}$ represents the predication value of the i-th unit in the output layer of whole networks. We can be sure that $y = \mathrm{argmax}(y_{pred}^{(0)}, y_{pred}^{(1)})$.

Step 2. Output retrieval result For a model H_j, if there exists a view image v_j^k, the predication label $y(v_j^k, s_0) = 1$, then we can output this model as a retrieval result.

Step 3. Ranking retrieval result Based on the predictation value z, the ranking operation is performed to obtain a better retrieval result. For a model H_j, the term $z(H_j) = \max\limits_{0 \leqslant k < n} z(s_0, v_j^k)$. Finally, we obtain the retrieval result.

5 Experiments

In this section, we experimentally validate and test our proposed framework. Our experiments operate on: (1) the SHREC'2013 dataset[9], which is one of the most well-known 3D model datasets and includes all TU-Berlin sketch datasets; (2) the dataset proposed by Taiwan University composed of 10119 3D models; and (3) the dataset of sketches from Eitz et al.[23], containing

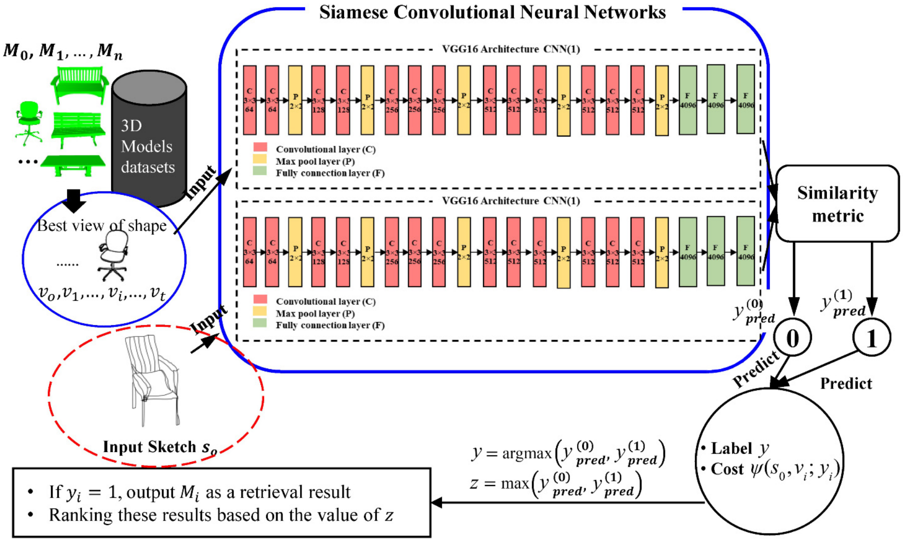

Siamese Convolutional Neural Networks
$M_0, M_1, \ldots, M_n$
3D Models datasets
Best view of shape
......
$v_o, v_1, \ldots, v_i, \ldots, v_t$
Input
Input
Input Sketch s_o
VGG16 Architecture CNN(1)
VGG16 Architecture CNN(1)
C 3×3 64
P 2×2
F 4096
Convolutional layer (C)
Max pool layer (P)
Fully connection layer (F)
Similarity metric
$y_{pred}^{(0)}$
$y_{pred}^{(1)}$
0
1
Predict
Predict
• Label y
• Cost $\psi(s_0, v_i; y_i)$
$y = \mathrm{argmax}(y_{pred}^{(0)}, y_{pred}^{(1)})$
$z = \max(y_{pred}^{(0)}, y_{pred}^{(1)})$
• If $y_i = 1$, output M_i as a retrieval result
• Ranking these results based on the value of z

Fig. 6 Test process for shape retrieval.

20 000 query sketches.

The method presented in this paper has been implemented using C++ program and Python 3.5 languages and is executed on PC under Windows 10 OS, Intel core I7-7700HQ processor, 8 GB memory size. Besides, Google tensorflow open source framework is used in this paper.

5.1 Experiment for best view of shape

For best view selection, we demonstrate that our approach achieves results that are competitive with other state-of-the-art methods, specifically perceptually-based best view classifer[7], SVM-based learning[22], and web image driven methods[25]. In particular, the Area Under the Curve (AUC) computed from the precision-recall curve of a retrieval result is often applied to evaluate retrieval performance. Noting that bad viewpoints would badly hamper the retrieval performance, we compared our method with the others; the result is shown in Fig. 7.

As Fig. 7 shows clearly, our method has an advantage in terms of the AUC indicator. This advantage arises from the capacity of our method to minimize the number of best views, which is no doubt useful for improving both the performance of retrieval, and also the response time. In our experiment, when the indicator of AUC arrives at 0.25, the number of best views obtained by our method is 6, which is less than that obtained by the other methods. The results of the comparison are shown in Table 1.

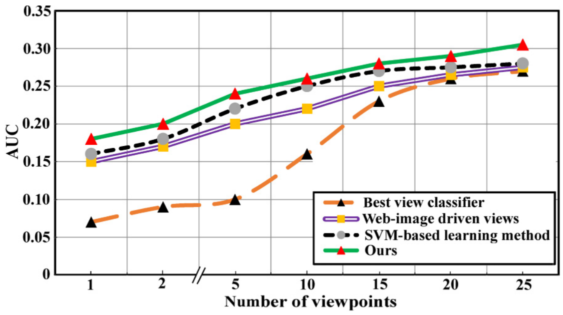


Fig. 7 Compared figure over AUC criterion between ours and others.

Table 1 Number of views when AUC is larger than overpass 0.2 in sketch-based shape retrieval.

Method	AUC	Number of views
Perceptually best view classifier	0.22	15
Web-image driven views	0.23	13
SVM-based learning method	0.24	7
Ours	0.25	6

5.2 Experiment for shape retrieval

For training and testing of the dataset, we evaluate our method in comparison with PHOG[7], SHELO[14], HOG[15], BHOG[16], ORB, and SIFT. The experiment result based on the precision-recall curve can be seen in Fig. 8. Our method is shown to be feasible and robust, with performance overtaking state-of-the-art alternatives. The experimental results indicate that our method achieves its design aims. In order to better demonstrate the performance of our proposed method, we compared it with other state-of-the-art CNN methods, namely, Siamese Cross-Domain CNN (CDCNN)[17], and Multi-View CNN (MVCNN)[26]. The result can be seen in Fig. 9.

Our method outperforms the others on the precision-recall curve criterion, mainly because the best view is acquired. Our network is also more complex than Siamese CNNs[17], because the number of view images is fixed in CDCNN methods. In fact, the number of best views varies with different shapes, and having fixed numbers of views for shapes consumes computing time and gives the opportunity for collecting bad views. By obtaining the best view for a shape, our approach decreases computing time and improves performance.

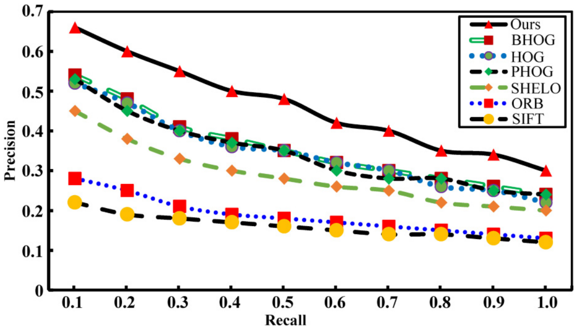


Fig. 8 PR curve criterion on NTU dataset.

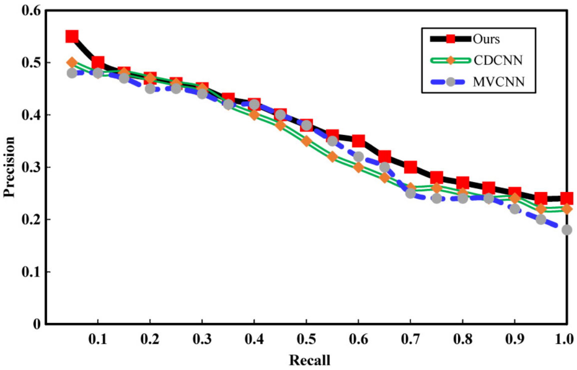


Fig. 9 PR curve criterion on SHREC'2013 dataset.

5.3 Discussion

Based on the framework proposed above we execute sketch-based shape retrieval, and present three examples in Fig. 10.

Our framework is able to retrieve models accurately. However, there are some mistakes in our retrieval result, which occur where there are small inter-class differences separating a correct model from an incorrect one. To some extent, a better similarity metric function can decrease this kind of mistake. Furthermore, sketches are by nature abstract and ambiguous, which gives rise in itself to incorrect retrieval results. For sketch-based retrieval, therefore, a small number of wrong retrieval results are completely acceptable, and our proposed framework is highly feasible.

6 Conclusion

In this paper, we proposed a hybrid CNN-based learning framework for sketch-based retrieval, including the use of a hybrid of CNNs for best view selection and Siamese CNNs for shape retrieval. The hybrid learning algorithm used to acquire the best view image of a shape makes use of the VGG-16 and AlexNet CNNs. The learning framework has been adopted to obtain the relation between a sketch and a view. A fusion model can then be formed to complete the retrieval task. Finally, the comparison result shows that our proposed framework is feasible and achieves results that are superior to a range of previous methods.

Acknowledgment

The authors appreciate the comments and suggestions of all anonymous reviewers, whose comments helped significantly improve this paper. This work was supported in part by the Fundamental Research Funds for the Central Universities in China (No. 2100219066) and the Key Fundamental Research Funds for the Central Universities in China (No. 0200219153).

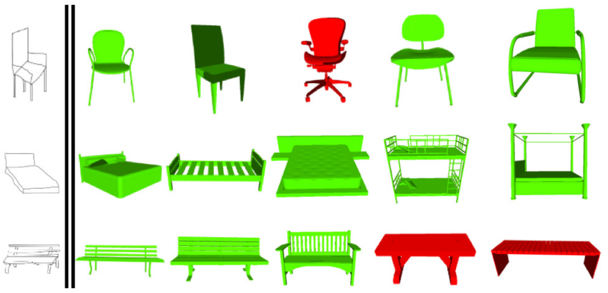


Fig. 10 Three examples of sketch-based 3D furniture retrieval on SHREC'2013 dataset (the red is wrong result).

References

[1] T. Kato, T. Kurita, N. Otsu, and K. Hirata, A sketch retrieval method for full color image database-query by visual example, in *Proc. 11th IAPR Int. Conf. Pattern Recognition*, Hague, the Netherlands, 1992, pp. 530–533.

[2] C. W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin, QBIC project: Querying images by content, using color, texture, and shape, in *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, San Jose, CA, USA, 1993, pp. 173–187.

[3] R. Hu and J. Collomosse, A performance evaluation of gradient field HOG descriptor for sketch based image retrieval, *Comput. Vis. Image Underst.*, vol. 117, no. 7, pp. 790–806, 2013.

[4] Y. J. Liu, X. Luo, A. Joneja, C. X. Ma, X. L. Fu, and D. W. Song, User-adaptive sketch-based 3-D CAD model retrieval, *IEEE Trans. Autom. Sci. Eng.*, vol. 10, no. 3, pp. 783–795, 2013.

[5] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, An evaluation of descriptors for large-scale image retrieval from sketched feature lines, *Comput. Graph.*, vol. 34, no. 5, pp. 482–498, 2010.

[6] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, Sketch-based image retrieval: Benchmark and bag-of-features descriptors, *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 11, pp. 1624–1636, 2011.

[7] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa, Sketch-based shape retrieval, *ACM Trans. Graph.*, vol. 31, no. 4, p. 31, 2012.

[8] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs, A search engine for 3D models, *ACM Trans. Graph.*, vol. 22, no. 1, pp. 83–105, 2003.

[9] B. Li, Y. Lu, A. Godil, T. Schreck, M. Aono, H. Johan, J. M. Saavedra, and S. Tashiro, SHREC'13 Track: Large scale sketch-based 3D shape retrieval, in *Proc. 6th Eurographics Workshop on 3D Object Retrieval*, Girona, Spain, 2013, pp. 89–96.

[10] J. G. Snodgrass and M. Vanderwart, A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity, *J. Exp. Psychol.: Hum. Learn. Memory*, vol. 6, no. 2, pp. 174–215, 1980.

[11] F. Cole, A. Golovinskiy, A. Limpaecher, H. S. Barros, A. Finkelstein, T. Funkhouser, and S. Rusinkiewicz, Where do people draw lines? *ACM Trans. Graph.*, vol. 27, no. 3, p. 88, 2008.

[12] J. M. Saavedra and B. Bustos, An improved histogram of edge local orientations for sketch-based image retrieval, in *Proc. 32nd DAGM Symp. Pattern Recognition Symp.*, Darmstadt, Germany, 2010, pp. 432–441.

[13] S. M. Yoon, M. Scherer, T. Schreck, and A. Kuijper, Sketch-based 3D model retrieval using diffusion tensor fields of suggestive contours, in *Proc. 18th Int. Conf. Multimedia*, Firenze, Italy, 2010, pp. 193–200.

[14] J. Saavedra, Sketch based image retrieval using a soft

computation of the histogram of edge local orientations (S-HELO), in *Proc. 2014 IEEE Int. Conf. Image Processing*, Paris, France, 2014, pp. 2998–3002.

[15] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 886–893.

[16] H. Y. Fu, H. G. Zhao, X. W. Kong, and X. B. Zhang, BHoG: Binary descriptor for sketch-based image retrieval, *Multimed. Syst.*, vol. 22, no. 1, pp. 127–136, 2016.

[17] F. Wang, L. Kang, and Y. Li, Sketch-based 3D shape retrieval using convolutional neural networks, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 1875–1883.

[18] F. Zhu, J. Xie, and Y. Fang, Learning cross-domain neural networks for sketch-based 3D shape retrieval, in *Proc. 30th AAAI Conf. Artificial Intelligence*, Phoenix, AZ, USA, 2016, pp. 931–941.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet classification with deep convolutional neural networks, in *Proc. 25th Int. Conf. Neural Information Processing Systems*, Lake Tahoe, NV, USA, 2012, pp. 1097–1105.

[20] K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv: 1409.1556, 2014.

[21] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, Deep residual learning for image recognition, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016, pp. 770–778.

[22] L. Zhao, S. Liang, J. Y. Jia, and Y. C. Wei, Learning best views of 3D shapes from sketch contour, *Vis. Comput.*, vol. 31, nos. 6–8, pp. 765–774, 2015.

[23] M. Eitz, J. Hays, and M. Alexa, How do humans sketch objects? *ACM Trans. Graph.*, vol. 31, no. 4, p. 44, 2012.

[24] S. Chopra, R. Hadsell, and Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Diego, CA, USA, 2005, pp. 539–546.

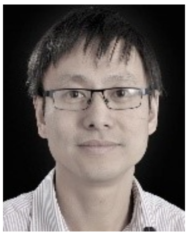
[25] H. Liu, L. Zhang, and H. Huang, Web-image driven best views of 3D shapes, *Vis. Comput.*, vol. 28, no. 3, pp. 279–287, 2012.

[26] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, Multi-view convolutional neural networks for 3D shape recognition, in *Proc. 2015 IEEE Int. Conf. Computer Vision*, Santiago, Chile, 2016, pp. 945–953.

Wen Zhou received the PhD degree from Tongji University in 2018. Since 2018, he has been in the School of Computer and Information, Anhui Normal University, Wuhu, China, where he is currently a lecturer. He is a member of IEEE and Chinese Computer Federation (CCF). His research interests include sketch-based retrieval, WebVR visualization, and machine learning.

Jinyuan Jia received the PhD degree from the Hong Kong University of Science and Technology in 2004. Since 2007, he has been with School of Software Engineering, Tongji University, Shanghai, China, where he is currently a professor. He is an ACM member, senior member of Chinese Computer Federation, and senior member of Chinese Steering Committee of Virtual Reality. His research interests include computer graphics, CAD, geometric modeling, Web3D, Mobile VR, game engine, digital entertainment, computer simulation, and peer-to-peer distributed virtual environment.

Chengxi Huang received the BS degree from Tongji University in 2015. He is pursuing the PhD degree in the Computer Science Department, Tongji University, Shanghai, China. His research interests include image processing, image reconstruction, data fusion and three-dimensional visualization, and machine learning.



Yongqiang Cheng received the BEng and MSc degrees from Tongji University, China and the PhD degree from University of Bradford, UK in 2001, 2004, and 2010, respectively. He joined the University of Hull in 2014 as a lecturer and is currently a senior lecturer with the Department of Computer Science and Technology at the University of Hull, UK. Before this, he worked as postdoctoral research fellow in Future Ubiquitous Networking lab in School of Engineering and Informatics, University of Bradford, from 2010 to 2014. He has very significant experience of working on large-scale projects (TSB, EU FP7 funded) and has led industrial collaborations as a PI on digital health technologies for over four years. He has published over 50 papers in high-impact scientific journals. He is also a member of the Chinese Automation and Computing Society in UK. His research interests include digital healthcare technologies, artificial intelligence, control theory and applications, embedded system, secure communication, and data mining.