Cognitive Computing and Internet of Things, Vol. 43, 2022, 1–7
https://doi.org/10.54941/ahfe1001832

Adaptive Weighted 3D Object Image Inference Model Based on Image Complexity

Yueqi Liu[1], Pu Meng[1], Zhuoyue Diao[1], Xin Meng[1], Liqun Zhang[1], and Xiaodong Li[2]

[1]School of Design, Shanghai Jiao Tong University, 800 Dong Chuan Road, Shanghai 200240, China
[2]China National Gold Group Gold Jewelry Co., Ltd., 1 Liuyin Park South Street, Beijing 100011, China

ABSTRACT

The research on product style classification based on CNN (Convolutional Neural Networks) is very active, but the data used to train CNN are often single-view images of 3D objects, which will lead to the loss of unpredictable object feature information and does not match the real scene. It reduces the quality of the model training. This paper proposes an adaptive weighted CNN model based on image complexity. Using CNN extract features from images of different perspectives of 3D objects, and the final inference results are obtained by weighting according to the complexity of those images. The 3D objects inference model in this paper is more in line with the cognitive process of the audience, and can improve the quality of style inference.

Keywords: Image complexity, CNN, Cognitive inference, 3D Model recognition

INTRODUCTION

With the development of the economy and society, people's demand for products is no longer only at the functional level, people will pursue the inner emotional experience that products can bring (Maruca, Regina Fazio, 2006), and users' personalized and emotional needs are also more and more paid attention to. The appearance of the product is the key factor that makes consumers' cognition, which will make consumers form the initial impression and idea of the product, and then affect the subsequent purchase decision. Therefore, it is particularly important for designers to accurately understand the cognitive judgment that the output of the design will bring to consumers. Designers will incorporate their usual design language into the design, to convey certain emotions to users, but the user's cognition often has some deviation from the designer's language. Therefore, it is necessary to establish a system that can accurately predict the method of consumer cognition that helps the product to be successful in the subsequent commercial competition.

In the field of research on the relationship between user cognition and products, scholars have carried out a lot of exploration: Taking the bus handle as an example, Li Lanyou et al. (2019) abstracted the design genes

of representative samples through AHP and semantic difference method, and constructed the double-helix structure model of product genes and users' perceptions of products through Kansei Engineering; Fu Yetao et al. (2011) obtained the cognitive data of subjects on animation characters through questionnaire method, focus group method, and perceptual image evaluation experiment, and then established a user's cognitive mapping model through factor analysis to provide reference for designers; Su Jianning et al. (2004) established a method to describe the morphological characteristics of bicycles using morphological decomposition qualitative description and morphological delineation quantitative description, and then established a mapping model of perceptual image and bicycle shape.

However, in previous studies, most of the researchers performed manual feature extraction of products based on knowledge in their respective fields, which is often subjective to a certain extent, which will inevitably result in the loss of part of the original data, and users' perceptions of products are often subjective, so it is difficult to effectively establish a mapping model between user cognition and product (Zhu Bin, 2018). In recent years, deep learning technology has made great progress. Compared with traditional research methods that rely on knowledge in specific fields to extract artificial features, deep learning can automatically extract and model data sets. In response to such a problem, Zhu Bin (2018) proposed an intelligent product design method based on deep learning, that is, firstly, an image dataset of single-view product images is established through cognitive experiments, and a mapping relationship between products and intentions is established through convolutional neural networks.

However, the single-view image is an incomplete representation of the 3D product information. First, the single-view image is used to represent the entire 3D product, resulting in a large loss of 3D object feature information; The product image data set established by cognitive experiments is not reliable enough. In fact, when consumers get a three-dimensional product, they often need all-around and multi-angle observation to produce more accurate cognitive judgments. Therefore, it is necessary to find a better representation method for three-dimensional objects, and then to establish a more accurate consumer cognitive model.

At present, scholars in the computer field have done a lot of research on how to use deep learning models to deal with 3D models. Yang Jun, Wang Shun, Zhou Peng (2019) proposed a 3D model recognition and classification algorithm based on deep voxel convolutional neural networks, which uses voxelization technology to convert 3D polygonal mesh models into voxel matrices. Maturana et al. (2015) proposed a novel network, VoxNet, to represent 3D models in the form of grids. Based on the concept of "entropy", Wang Ya (2020) proposed a three-view representation method for three-dimensional objects and then constructed an adaptive weighted ensemble convolutional neural network model to weight the three views of the model based on the entropy value. B. Shi et al. (2015) proposed to convert each 3D object into a panoramic view, and then perform a cylindrical projection of the model around its main axis to learn classification

through CNN (Convolutional Neural Networks). Sinha et al. (2016) proposed to directly convert 3D models into planar "geometry", enabling CNNs to directly learn 3D structures. Zanuttigh et al. (2017) first input the six-view depth images of three-dimensional objects into multiple convolutional neural networks respectively, and finally, a linea into a compact image, which in turn provides better performance for subsequent CNN classification computations.

At present, in the field of 3D object representation, methods based on multi-view or point cloud and voxelization have made rapid progress, but each method has its own shortcomings. There are many methods to represent 3D objects in the form of point clouds, and it is difficult to distinguish them. For this study, the acquisition cost of point cloud data of 3D models is high and it is very difficult to obtain. The form of voxels also has the problem of feature loss, and both methods consume a lot of computing resources. Therefore, this study will avoid the above two methods of characterizing 3D objects. In the multi-view method, due to the inherent defects of such methods, a large amount of feature information will be lost. Therefore, it is necessary to adopt a characterization method that can cover as much 3D object information as possible. Taking a ring shot of a three-dimensional object can cover almost all the information, and then build a more accurate consumer cognitive inference model.

METHODS

First, establish a 3D product image dataset. Collect three-dimensional object samples, select the image evaluation vocabulary corresponding to the three-dimensional object samples, and mark the three-dimensional object samples through cognitive experiments to form a three-dimensional product image data set. Secondly, the 3D product image data set is dimensionally reduced by rendering software, and each 3D model is shot 360° around, and then the 2D representation of the 3D product image data set is obtained. Finally, the product image inference model is constructed, followed by data preprocessing, model training, and validation set testing.

ESTABLISHMENT OF PRODUCT IMAGE DATASET

Taking into account the richness of 3D objects, the difficulty of collection, and the familiarity, the sofa is selected as the 3D object sample in this article. The sofa model comes from a public dataset Modelnet10 of Princeton University. There are 810 sofa models in Modelnet10, and 600 of them are selected as training set and validation set in this study, and the ratio of the training set and validation set is 8:2.

Since the focus of this paper is to explore a product image inference model that is more in line with audience cognition, the establishment of the perceptual image space for product evaluation is not the focus of this paper, two pairs of perceptual words are used as two dimensions of the evaluation of 3D model samples, namely D1: Comfort (uncomfortable-comfortable), D2: appearance (Plain-Ornate).

IMAGE EVALUATION EXPERIMENT

In this experiment, 18 masters majoring in design were selected as subjects to participate in the image evaluation experiment of the sofa model, and the ratio of males to females was 1:1.

This paper randomly divides 600 model samples into 6 parts, each part contains 100 sofa models and assigns corresponding numbers. For the six-part sofa model, six questionnaires were made, and the six questionnaires all used the five-point semantic difference method to complete the image evaluation of each sofa sample. Each sample in the scale is evaluated by two pairs of intention words, taking the Plain-Ornate pair of intention words as an example, the 0 means neither plain nor ornate, −2 for plain, 2 for ornate. Each subject needs to watch the 3D shape of each sample on another device, and score the intention of each sample according to their subjective feelings in the questionnaire. Each questionnaire has 3 subjects participating in the scoring.

The subjects in this experiment basically completed the product image evaluation experiment within 30 minutes. The duration of the experiment was moderate, and the subjects' attention was relatively concentrated, so the experimental results had high accuracy. The average score calculated by each sample on the two dimensions of D1 and D2 determines the category to which the sample belongs in each dimension. For example, if the score of sample 1 on the dimension D1 is greater than 0, it is considered that sample 1 is comfortable, otherwise is uncomfortable. According to the calculation results, the experimental samples are divided into four categories, C1: (uncomfortable - plain), C2: (uncomfortable - ornate), C3: (comfortable - plain), C4: (comfortable - ornate). The numbers of the four categories are 218, 77, 151, 154 respectively.

CONSTRUCTION OF PRODUCT IMAGE INFERENCE MODEL

Due to the three-dimensional and spatial nature of the 3D model, a single-view image will lose a lot of information, and it is impossible to obtain the characteristics of the entire three-dimensional object. Therefore, consider using the rendering software to shoot the three-dimensional object. In order to obtain the multi-view image of each sample, first of all, Each model is imported into Keyshot, and through Keyshot's camera animation settings, the number of steps for each 3D object is 12°, that is, the rendering camera takes a 360° ring around the 3D object, and the image is rendered every 12°. Each model obtains 30 images from different perspectives so that the obtained representation of the 3D object can cover more information of the 3D object, and then more comprehensively characterize the 3D object to simulate the process of observing a three-dimensional product in reality. After all the 3D samples of the training set and the validation set are dimensionally reduced, the final product image data set used to train the model proposed in this paper is formed.

The data set is randomized into the training set and the validation set according to the ratio of 8:2, containing 480 and 120 samples respectively. The validation set is used to verify the accuracy of the model obtained after training. Then, after a series of resizing, cropping, and normalizing

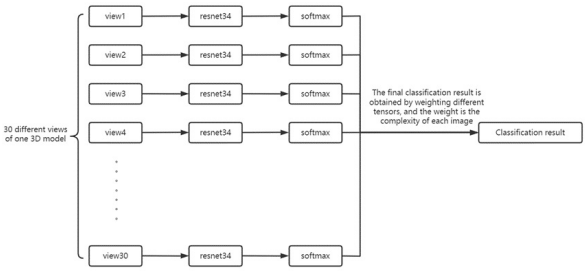

Figure 1: The structure of the 3D product image recognition model.

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
conv4_x	14×14	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Figure 2: Five Resnet residual network structures.

data preprocessing operations, it can enter the neural network model for calculation.

The structure of the 3D product image inference model based on image complexity weighting proposed in this paper is shown in Figure 1:

In this paper, Resnet is selected as the basic network part of the product intention recognition model to form the entire product image inference model. Compared with other network models, Resnet introduces a residual block structure to solve the problem of gradient disappearance and degradation as the network depth increases. The degradation problem is that when the network becomes deeper and deeper, the training accuracy rate changes will level off, but the training error will become larger. Therefore, this paper will choose the Resnet network as the basic network part of the product image inference model in this paper.

The five common Resnet residual network structures are shown in the figure, and the depths are 18, 34, 50, 101, and 152 respectively. The five network structures are firstly a 7x7 convolutional layer, followed by a 3x3 maximum pooling layer, followed by the superposition of residual blocks. The parameters and number of residual blocks of different networks are shown in Figure 2. The residual block is followed by a global average pooling layer, which can well prevent overfitting, is more robust, and strengthens the consistency of feature maps and categories.

In this paper, considering the computing performance and accuracy, Resnet34 is selected as the basic part of the model. After the data set is prepared, the model can be trained by the method of transfer learning. Since images from different perspectives have different characteristics, each Resnet34

needs to be trained separately. The output of Resnet34 corresponding to each perspective will be input into the Softmax classifier, and then a tensor will be output, that is, the image of this perspective belongs to the Probabilities of the four classes.

Since the amount of information contained in the images of each perspective is different, perspectives with greater information content tend to have a greater impact on the audience's cognitive judgment of the product. Therefore, it is necessary to perform weighted fusion of the tensors output by the oftmax layer according to the amount of information from different perspectives, so as to simulate the process of users observing and making cognitive judgments on a 3D product in reality, and obtain the final image inference of the 3D product. In this paper, the method of calculating the complexity of images from different perspectives is used to measure the amount of information contained in each perspective image. The selected method refers to the article(Donderi, 2006). This method has been proved that the calculation results are positively correlated with the audience's visual perception. After calculating the complexity of the images from different perspectives, the results obtained from the thirty perspectives output of the sample can be weighted and calculated, and then the classification results can be obtained.

RESULTS

By training the product image inference model proposed in this paper, the intention of three-dimensional objects can be classified. After training, the model can achieve an accuracy of 74.67% in the four classifications of the validation set, so it can be considered that the three-dimensional object proposed in this paper can be a product image recognition model that has a certain accuracy. The accuracy of the model proposed in this paper is 74.67% in the validation set, and there is still a lot of room for improvement.

CONTRIBUTION AND DISCUSSION

All in all, the inference model of product image obtained by previous research using single-view images of 3D products as training samples is not the same as the process of users observing 3D products in reality and making cognitive judgments, and will lose a lot of 3D products. For the features contained in the product, the multi-view adaptive weighted 3D product image inference model based on complexity proposed in this paper is more in line with the reality of user cognition, reduces the information loss of 3D products, and improves the quality of 3D product image inference.

The reasons why the accuracy rate needs to be improved may be: First, for the Plain-Ornate intention word, different subjects have certain differences in the cognition of the same product, so it has a certain impact on the recognition accuracy rate. Second, some samples are difficult to judge whether they are comfortable or uncomfortable, ornate or plain. In addition, using 30 images from different perspectives to represent a 3D object, although the information loss of the 3D object itself is reduced, it may cause a large amount of information redundancy in the data set, consume a lot of unnecessary computing resources, and greatly prolong the training of the

model. time. Since this paper is a preliminary exploration of a 3D product image inference method that is more in line with the perceived reality of the audience, it does not consider how to reduce information redundancy, which is also a point that needs to be considered in subsequent research.

REFERENCES

Donderi, D. C. (2006). An information theory analysis of visual complexity and dissimilarity. Perception, 35(6), pp. 823–835.

Fu, Y.T., Luo, S.J. and Zhou, Y.X. (2011). Jiyu ganxing yixiang de dongman juese xingxiang pingjia[Image Evaluation of Anime Characters Based on Perceptual Imagery]. Zhejiang daxue xuebao(gongye ban) (09),1544-1552+1570. doi:CNKI:SUN:ZDZC.0. 2011–09–007.

Li, L.Y., Lu,J.G. and Zhang, J.D. (2019). Jiyu chanpin jiyin de ganxing yixiang sheji[Perceptual image design based on product genes]. Nanjing gongye daxue xuebao (ziran kexue ban), 2019, v.41; No.192(01):pp. 71–78+88.

Maruca, R.F. (2000) 'Mapping the World of Customer Satisfaction', Harvard Business Review, 78(3), 30, available: https://link.gale.com/apps/doc/A63035676/AONE?u=anon~{}673cf3fe&sid=googleScholar&xid=df34c977 [accessed 13 Feb 2022].

Maturana, D., and Scherer, S. (2015). Voxnet: A 3d convolutional neural network for real-time object recognition. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 922–928). IEEE.

Shi, Baoguang and Bai, Song and Zhou, Zhichao and Bai, Xiang, (2015). DeepPano: Deep Panoramic Representation for 3-D Shape Recognition. IEEE Signal Processing Letters, p. 2339–2343.

Sinha, A., Bai, J., and Ramani, K. (2016). Deep learning 3D shape surfaces using geometry images. In European conference on computer vision (pp. 223–240). Springer, Cham.

Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. (2015). Multi-view convolutional neural networks for 3d shape recognition. In Proceedings of the IEEE international conference on computer vision (pp. 945–953).

Su,J.N.,Jiang,P.Y.,Zhu,B. and Li,H.Q. (2004). Ganxing gongxue jiqi zai chanpin sheji Zhong de yingyong yanjiu[Kansei engineering and its application in product design]. Xian jiaotong daxue xuebao(01), 60–63. doi:CNKI:SUN:XAJT.0. 2004–01–015.

Wang, Y. (2020).Jiyu 3Djuanji shenjing wangluo de sanwei moxing shibie ji jiansuo yanjiu(Shuoshi xuewei lunwen,Changchun gongye daxue). https://t.cnki.net/kcms/detail?v=so7-vcwoEmeeJ1OO8eSRU5CIsyhuS9T8pzzPUKxmJhNKQNAPwWV8gU08Pzw7Gmtp0CeQqozu432AnSmIvbvZYTGkqLiACgQA_i_KK3n22m3vt0wR9Nhtgw==&uniplatform=NZKPT

Yang, J., Wang, S. and Zhou, P. (2019). Jiyu shendu tisu juanji shenjing wangluo de sanwei moxing shibie fenlei[3D model recognition and classification based on deep voxel convolutional neural network]. Guangming xuebao(04),314–324. doi:.

Zanuttigh, P., and Minto, L. (2017). Deep learning for 3d shape classification from multiple depth maps. In 2017 IEEE International Conference on Image Processing (ICIP) (pp. 3615–3619). IEEE.

Zhu, B. (2018).Jiyu shendu xuexi de chanpin qinggan hua zhineng sheji [Product emotional intelligent design based on deep learning] (Shuoshi xuewei lunwen,Zhejiang daxue). https://t.cnki.net/kcms/detail?v=so7-vcwoEmeIVqIN_4s4s-9vmMLflUaoN949-wJol-VYjUK2Meo19fPvXgcciljbUjYX9jBI5oFv98aggJAo4NffiLc5uklHgsHZRYhL5PyfGgW_ceNDPQ==&uniplatform=NZKPT