Research on Detection and Modification System Based on Virtual Panorama

Jing Ruan, Yanlei Shang*
Beijing University of Posts and Telecommunications, Beijing, China

*Corresponding author e-mail: shangyl@bupt.edu.cn

Abstract. In recent years, Web VR has developed rapidly in the wave of VR with its easy accessibility and easy development, and the virtual panorama is particularly prominent. With its comprehensive and vivid display of things, it gradually replaces pictures as display media and information carriers. However, in today's highly competitive environment, it is necessary to improve the stereoscopic and 3D nature of virtual panorama rendering. This paper aims to combine the image detection with the virtual panorama to increase the stereoscopic and 3D nature of the virtual panorama.

1. Introduction
In recent years, Web VR technology has developed rapidly. With the popularity of HTML5, the performance of web pages has become more and more powerful. The emergence of WebGL defines a web-oriented graphics standard. Based on this, a series of APIs can use graphics hardware to perform graphics rendering. Since Web VR does not require expensive wearable devices, it can be easily accessed on the browser, which greatly promotes the spread of Web VR.

Among them, the virtual panorama technology is particularly prominent. The virtual panorama is also called the three-dimensional panoramic virtual reality. It is a real-life virtual reality technology based on panoramic images. The web-side can interactively view the restored real scenes, making the viewers immersive. The virtual panorama has low cost, and the work has a good navigation and interoperability. Its rich playback format is more suitable for network communication. It has been widely used in campus scenery, urban scenery, tourism landscape and digital products.

Although the virtual panorama has its obvious advantages, its single lack of vividness often requires developers to modify it. Therefore, this paper designs a modification system for virtual panoramas. It is modified by the object detection model to locate the modified area, which increases the vividness and richness of the virtual panoramic display and reduces the labor cost.

Figure 1. distortion in virtual panorama
At present, more applications of object detection technology in AR face detection and gesture recognition, which is modified by detecting the position of faces and facial features to achieve rich and interesting effects, but the application of technology in virtual panorama is very few. On the one hand, the object in the virtual panorama is not as large as the face in the camera. Secondly, the real scene reconstructed by the virtual panorama is a spherical area, and the original straight picture is in the virtual panorama. It is easy to see serious distortion in Figure 1.

In view of the above problems and the characteristics of the virtual panorama generated by the panorama and one-to-one correspondence, this paper designs the object detection system to obtain the detected virtual panoramic area by detecting the panoramic picture and the conversion formula of the panoramic view to the virtual panorama, avoiding distortion problems with virtual panoramic view images. Improve the recall rate and accuracy of test results by improving and designing object detection models and classification models. Adding modifiers through the converted 3D coordinates to complete the modification effect.

The structure of this thesis is as follows, the second description related work, the third chapter introduces the structure design and implementation method of the detection modification system, the fourth chapter introduces the experimental results and attention details of this paper, the fifth chapter will introduce the experimental summary and prospect.

2. Related word
At present, the mainstream virtual panoramic engine has pano2vr, krpano, three.js, etc. The virtual panoramic engine generates virtual panoramic roaming through the panoramic image stitched by 360-degree photos, so that it can be accessed through the browser on the mobile phone or the computer. Pano2vr was popular a few years ago, and it is powerful enough to generate simple roaming, but pano2vr is weak in the HTML5 era, lacking support for various new features, such as panoramic video, virtual reality and so on. Krpano has gradually emerged in recent years. Such as 720 cloud, the panoramic platform such as Tuyun is using krpano kernel. Krpano itself is a panoramic roaming production kernel plus a set of auxiliary tools, and krpano has its original xml programming language, which is convenient for users to modify the virtual panorama through programming. Three.js is a third-party written in JavaScript. Provides a lot of 3D display capabilities. It also provides the ability to make virtual panoramas, but the loading speed on the web side is relatively slow, so this experiment uses the popular krpano virtual panorama engine.

In the field of object detection for deep learning, a new and innovative structure is born every year. Detection refers to detecting the position and category of a specific object (such as a human face) in the image, from the early RCNN to FasterRcnn, RFCN. MaskRCNN, YOLO and later SSD, the SSD model is faster than the fastest YOLO model before, and the accuracy is comparable to FasterRCNN. The SSD model uses a convolution kernel to predict a series of default bounding boxes. The classification scores and offsets are predicted on the feature maps of different scales to improve the detection accuracy. In addition, end-to-end training is adopted, and the accuracy of the detection can be ensured even if the resolution of the image is relatively low.

This paper uses a SSD detection model with guaranteed rate and accuracy for object detection, and uses the krpano virtual panorama engine, which is easy to program and has a fast loading rate, to make a virtual panorama.

3. Method

3.1. Virtual panoramic detection modification system
In this paper, the virtual panoramic detection system is mainly divided into detection layer, filter layer and coordinate transformation modification layer. As shown in Figure 2. In the detection layer, we will construct and train the detection model for the characteristics of the panorama. The model structure, data preprocessing and migration learning method achieve high accuracy detection results in the case of small data volume. In the filter layer, we will design the classification model to filter out the detection error of the detection layer, due to the small amount of data. Under the model, the model inevitably appears to have a serious error detection condition, and the filter layer will ensure accurate detection.
detection of the error detection when the recall rate is high. In the virtual panorama decoration layer, we convert the previous detection coordinates into virtual panoramic 3D coordinates through the coordinate transformation model, and modify the coordinates through this coordinate. This method can avoid the problem that the image distortion in VR is difficult to detect and the detection result is difficult to accurately convert into the virtual panoramic three-dimensional coordinates.

3.2. Detection layer

The detection layer mainly comprises a slice module, an object detection module and a result integration module, where in the slice module cuts the image into a plurality of overlapping portions according to the characteristics of the panorama, so that the sliced image is more easily detected by the object position. The object detection module is constructed and trained to detect the model, and the integration module integrates all the detection results and removes the duplicate detection.

It is difficult to obtain a good effect by directly detecting the panorama. On the one hand, because the panorama is a fixed 1 to 2 size, it will be stretched to a size of 1 to 1 before the input detection model, which will cause a certain deformation of the object. On the other hand, the object of the panorama is too small in the panorama to be detected, so the slicing module divides the panorama into three overlapping images, as shown in Figure 3.

The sliced picture conforms to the input size of the model, and the detected object has a larger proportion in each picture, and the middle sliced picture avoids the case where the intermediate object is leak-detected due to the segmentation.

The object detection module uses a neural network model to perform object detection on the image. Here, the SSD detection model with faster rate and lower accuracy than other models is selected. Due to the lack of training data, this paper adopts the method of migration learning, and uses the pre-trained parameters of the coco dataset to initialize the model parameters. Since the underlying features of the pictures are similar to the combination of dotted lines, other data sets can be reused. We freeze the parameters of the bottom 3 blocks in the experiment and train the upper model parameters on our data set. Aiming at the problem that the model is easy to fit and the object distortion in the panorama is serious, the local response normalization layer proposed by Alexnet [4] is introduced in each block.
of the upper model, which can enhance the generalization ability of the model and enhance the distortion tolerance of model.

The integration module integrates the detection results of three slices. Since the three images have overlapping regions, the integration module inverts all the detection results back into the same coordinate system to calculate whether the detection results are seriously coincident. In the case of severe coincidence, only the result with the highest detection score is retained.

3.3. Filter layer
The filter layer is mainly composed of a classification model, which is different from the classifier in the detection model. The classification model of this layer will be more sensitive to the difference between the error detection and the object. The model depth and parameter quantity are more suitable for the data magnitude of the experiment, which can ensure the correctness. The classification test results, the classification model structure is shown in Figure 4.

![Figure 4. The architecture of classification model](image)

We can divide the model structure into an encoder, a classification part, and a loss function part. The encoder adopts a self-designed inception [5] structure to extract features through three channels of different depths of different receptive fields, which can take into account multiple types of features. The traditional structure of the convolutional layer and the pooled layer can further encode the feature to learn the high-dimensional features, and also enhance the distortion tolerance and translation invariance of the model. Finally, the fully connected layer is used as the classifier. It can be seen that the model contains a large number of BN layers, which can enhance the generalization ability of the model and prevent over-fitting, which is especially important for small data experiments. Because the positive and negative samples of this test are not balanced, the number of error detection picture training is less than that of correctly detecting pictures. Here, adopting Focal_loss [6] as the loss function can learn more misclassified pictures and increase the generalization ability of the model.

3.4. Virtual panorama decoration layer
The virtual panoramic decoration layer mainly includes a coordinate conversion module and a hotspot adding module, wherein the coordinate conversion module completes the conversion of the detected coordinates to the virtual panoramic coordinates, and the hotspot adding module generates the modified virtual panoramic view according to the virtual panoramic coordinates and the panoramic image.

The presentation of the virtual panorama is a spherical space with the camera angle as the center of the ball, so a spherical coordinate system can be established for the virtual panorama, for any point in the space, $R$ can be the distance from the center of the ball, $V_{th}$ is the left and right rotation angle, and $V_{ah}$ is the vertical rotation angle. For the object detection result, it is a rectangular area of a flat picture, which can be represented by the upper left corner $(X_{left}, Y_{left})$ and the lower right corner $(X_{right}, Y_{right})$. Since any point in the panorama has a hidden relationship with the virtual panorama, It can be understood that the virtual panorama is formed by connecting the panorama to the left and right and compressing the top and bottom, so it can be converted by the following formula to obtain coordinates $(R, V_{th}, V_{ah})$ and dimensions $(\text{Width}_{obj}, \text{Height}_{obj})$: 
\[ x_{\text{mid}} = x_{\text{left}} + \frac{x_{\text{right}} - x_{\text{left}}}{2} \]
\[ y_{\text{mid}} = y_{\text{top}} + \frac{y_{\text{bottom}} - y_{\text{top}}}{2} \]
\[ v_{\text{left}} = (x_{\text{left}} - \text{Width}_{\text{pic}} - 0.5) \times 360 \]
\[ v_{\text{top}} = (y_{\text{top}} - \text{Height}_{\text{pic}} - 0.5) \times 180 \]
\[ \text{Width}_{\text{obj}} = \frac{x_{\text{right}} - x_{\text{left}}}{\text{Width}_{\text{pic}}} \times R \]
\[ \text{Height}_{\text{obj}} = \frac{y_{\text{right}} - y_{\text{top}}}{\text{Height}_{\text{pic}}} \times R \]

Where \( \text{Width}_{\text{pic}} \) and \( \text{Height}_{\text{pic}} \) represent the size of the panorama, and \( R \) is a fixed value.

The hotspot modification module is responsible for generating the virtual panorama and adding hotspots according to the virtual panoramic coordinates. In this paper, the krpano virtual panorama engine is used to generate the virtual panorama, and the xml language created makes it easier to modify it by programming, and it has a fast loading load. Low memory features. The modification part is modified by the way that the dynamic hotspot covers the static object, for example, detecting the position of the screen in the virtual panorama, and using the video hotspot to fit the position of the video in the virtual panorama, so that the screen in the virtual panorama plays the video, which is more authentic and vivid.

4. Experiment and evaluation

4.1. Object detection module

The detection object of the object detection module is a screen, and the data set adopts more than 500 pictures which are manually labelled. The experiment is shown in Table 1.

| Experiment | Val. loss | Recall | Accuracy |
|------------|-----------|--------|----------|
| F-1        | 1.06      | 0.733  | 0.548    |
| F-2        | 1.06      | 0.8    | 0.595    |
| F-3        | 1.02      | 0.822  | 0.606    |
| F-4        | 1.03      | 0.822  | 0.63     |
| F-5        | 0.93      | 0.844  | 0.681    |

F-2 has more slicing operations than F-1. It can be seen that there is a higher recall rate after the slice, and F-3 adjusts the learning rate compared with F-2. Adam optimizer with a learning rate of 3e-4 if more effective. it is good. F- 5 added LRN layer to F-3. It can be seen that LRN layer improve the accuracy and recall rate which enhance the generalization ability of the model. Finally, we It can achieve a 90% detection recall rate and a 68% accuracy rate, where additional error detection is still unavoidable and is done by the filtering model.

4.2. Classification model

The classification model is designed to improve the accuracy of the detect results without reducing the recall rate. The training set has a total of 400 positive samples and 300 negative. Shown in Table 2.

| Experiment | Train accuracy | Val. accuracy |
|------------|---------------|--------------|
| C-1(conv2) | 0.949         | 0.652        |
| C-2(conv_4) | 0.978        | 0.962        |
| C-3(conv4_res) | 0.954      | 0.948        |
| C-4(conv4_incep) | 0.989   | 0.975        |
| C-5(conv4_incep_res) | 0.976   | 0.958        |
| C-6(conv4_incep_bni) | 0.984   | 0.979        |
It can be seen from the results that the 4-layer convolution structure and the Inception structure is more suitable for our experiments, and the batch normalization layer can increase the model generalization. After filtering the model, and the accuracy rate is improved from 68% to 98%.

4.3. Virtual panorama decoration model

The virtual panorama module mainly completes the generation of panoramas and the addition of modifications. The experimental results are shown in Figure 5.

![Figure 5. modified virtual panorama and pre-modification](image)

As can be seen from the comparison effect of the above figure, the hot video perfectly fits the screen in the picture, making the static TV become the dynamic video, increasing its vividness.

5. Conclusion

The virtual panorama technology has a more vivid and comprehensive display effect than the picture, and will be more widely used in the future. This paper takes the detection of the panorama to find the virtual panorama. In the modified area, more research may be conducted in the future.

The idea of using migration learning in this experiment has achieved a good recall rate in small data sets, but the accuracy still needs to be Ascension, the second-level classification model was designed to improve the accuracy. With the momentum of today's deep learning development, migration learning will be more perfect, easy to fit, and poor generalization problems will be improved.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants No. 61671081 and No. 61720106007, the Beijing Natural Science Foundation under Grant No. 4172042, and the 111 project (B18008), the Fundamental Research Funds for the Central Universities under Grant No. 2018XKJC01.

References

[1] Zhu C, Huang K, Li G. Automatic Salient Object Detection for Panoramic Images Using Region Growing and Fixation Prediction Model[J]. 2017.

[2] Liu L, Ouyang W, Wang X, et al. Deep Learning for Generic Object Detection: A Survey[J]. 2018.

[3] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[C]// European Conference on Computer Vision. Springer International Publishing, 2016:21-37.

[4] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[5] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[C]// Computer Vision and Pattern Recognition. IEEE, 2016:2818-2826.

[6] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):2999-3007.