Scene classification based on visual feature channels and SVM

LI Ming-zhao [*1], WANG Zhi-yan [*2]

[*1] Jilin International Studies University, Jingyue Street 3658, Changchun City, Jilin Province
E-mail: limingzhao@jisu.edu.cn
[*2] Jilin International Studies University, Jingyue Street 3658, Changchun City, Jilin Province
E-mail: wangzhiyan@jisu.edu.cn

Abstract. Nowadays, scene classification is widely used as the pre-classification of target detection and behaviour detection, and it is an essential part of computer vision. In this paper, an improved scene classification method is proposed. This method firstly uses the naturalness in the spatial envelope model to classify scenes into base classes. It then classifies scenes more carefully through the visual feature channel model to obtain specific scene categories. Simulation experiments show that, under the exact condition of the SVM classifier, the improved scene classification method in this paper is better than the traditional scene classification method for scene classification. On this basis, the parameters of the SVM classifier are optimized for specific data sets.

Keywords: Scene classification, visual feature channels, support vector machines

1. INTRODUCTION

With the rapid development of computer technology and multimedia equipment, the number of electronic images and videos is growing at an unpredictable rate. The database and applications that depend on these images and videos are also increasing day by day. Therefore, relying on the traditional manual classification and labelling method becomes impractical because it needs a lot of human resources. Therefore, how to use computers to classify and label images and videos automatically is a primary research direction in computer vision.

Scene classification is an essential branch of computer vision image classification. The concept of scene description and understanding was further clarified at the 2006 MIT Workshop on Scene Understanding, which also pointed out that scene classification is a new and promising research direction [1]. Its main applications are in four main areas: image/video retrieval, computer vision tasks, mobile robotics, and image enhancement. (1) Image/video retrieval: In scene classification, image/video retrieval is the most direct application in recent years. With the rapid development of the Internet and electronic devices, image/video databases are also overgrowing. Therefore, an effective scene classification method can help to better store and browse the image/video database. Most image/video database search technologies appearing on the Internet are based on keywords rather than image content and video keyframe. Therefore, a reliable direct search method based on image content and video keyframe is one of the primary uses of scene classification at present [2].

(2) Computer vision tasks: another significant use of scene classification is the pre-detection of target and behavior detection. The background of this demand is also based on the growing image/video library, so in computer vision, recognizing a target has become very difficult. At present, the most common approach is to use scene classification to divide the scene that may contain a target and then carry out further target recognition. For example, if you want to detect a target containing "horse", you might want to search for a scene with a semantic label such as "grassland". This method can not only improve the overall workload of target recognition but also improve the robustness of target recognition.

(3) Mobile robots: the three most urgent problems to be solved for the next generation of mobile robots are positioning, matching and navigation. The study of these tasks hinge on a central question: Where is the "we"? For now, most robots rely on hardware, such as sonar, lasers and distance sensors to solve such problems. Usually, these solutions are very effective for indoor spaces with spatial regularity, but for outdoor environments with varying environments, the robustness and efficiency of these methods quickly diminishes. Therefore, with the help of computer vision, it is a significant difficulty in scene classification to overcome the changes of illumination, background and Angle.

(4) Image enhancement: In recent years, with the deepening of scene classification research, some scholars and experts also apply scene classification results to image enhancement. Many images that need to be enhanced will adopt different image enhancement algorithms according to different scenes to achieve the best image enhancement effect.

2. ALGORITHM MODEL

2.1. Degree of naturalness

To discuss the naturalness, we have to introduce the theory of envelopment model of airspace proposed by Oliva et al. The spatial envelopment model is a method that ignores the segmentation and processing of objects

and uses some specified spectral and rough local information to estimate the reliable global spatial scene properties of the scene. The representation of the scene can be characterized by the envelopment set of the airspace, which provides a reasonable description of the scene and its semantic classification.

The model contains five characteristics: naturalness, openness, roughness, extensibility and ruggedness. One of the most important and the main consideration of this paper is the nature. The structure of the scene can strictly distinguish between the artificial environment and the natural environment. Continuous horizontal and vertical lines predominate in artificial scenes, while most natural scenes have textured areas and wavy contours. Therefore, scenes with high naturalness can usually be found in the natural landscape, while scenes with more vertical and horizontal edges have low naturalness [3].

In the absence of prior region demarcation and independent object recognition, scene classification is mainly based on the structural differences among scene categories. The global and local energy spectra, which define the image-based feature space generated by each scene, are not very direct to human observation. This part aims to estimate the naturalness in the spatial envelope model from the two spectral representations and use this naturalness to divide the scene into natural or artificial scenes.

In the case of simple linear regression, attribute S can be estimated by the global spectral feature V of the image, as shown below:

$$s = v^T d = \sum^{N_G} v_i d_i = \iint A(f_x, f_y)^2 DST(f_x, f_y) df_x df_y$$

And DST can be expressed by the following formula:

$$DST(f_x, f_y) = \sum^{N_G} d_i \psi_i(f_x, f_y)$$

$\psi_i(f_x, f_y)$ is the K-L basis of the energy spectrum, which can be learned by using linear discriminant analysis, and the parameter d with the best classification rate can be obtained by this method.

As can be seen from the above equation, the naturalness can be obtained from the spectral dot product of the amplitude of the spectrum discriminant template and the image. In this paper, 400 images of natural and artificial scenes were randomly selected from an image library with 2688 images as training images, and DST(fx,fy) was calculated, as shown in Figure 1.

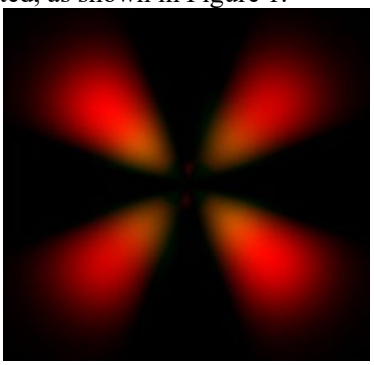

Fig.1 The DST of the Natural Degree

2.2. GIST model based on visual feature channels

At present, the research core of fast scene classification mainly focuses on extracting effective GIST from an image, using features of different fields to calculate its overall features, but at the same time, its rough spatial information should be considered. However, the GIST model starting from the mature ITTI significance model [4] is a relatively common scene GIST model at present. The GIST model uses the salience model's color, direction, and intensity features to calculate the overall features of the graph. At the same time, the rough spatial information is completed by using the block of the feature map. The specific process is shown in Figure 2.

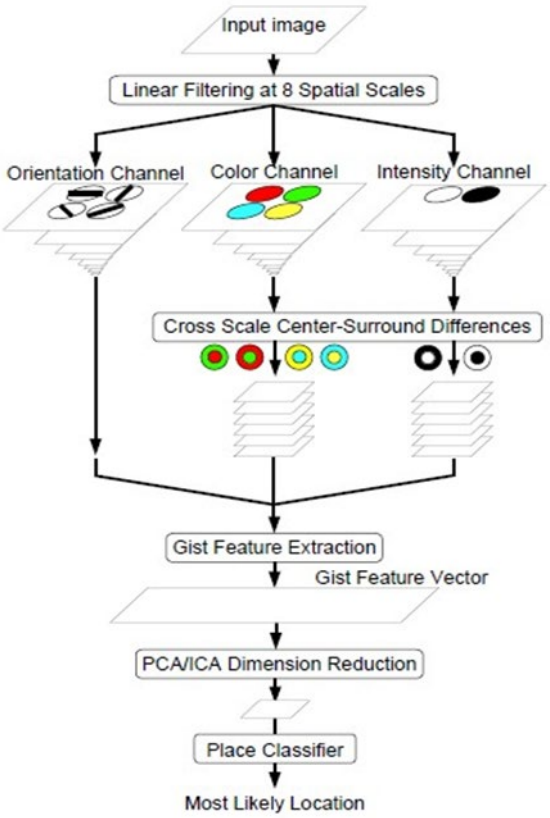

Fig.2 The Gist Model Based on Visual Feature Channel

In the GIST model based on visual feature channels [5], the first step is the same as visual cortex feature extraction in the ITTI saliency model. In other words, an input image on a complex spatial scale can be filtered into many low-level visual feature channels, such as color channels, density channels, direction channels, flicker channels, and motion channels, all of which are found in the visual cortex. Some channels have a lot of children, such as color channel including the color of many different types, the Angle of the direction of channels with different directions, by the same token, sports channel contains information such as the direction, each subchannel has a pyramid filter output, 9 layers in the proportion of the horizontal and vertical direction from 1:1 (0) to 25 June (tier 8) step by step a narrow, a 5*5 Gaussian smoothing is also applied between adjacent layers. In each subchannel I, the model performs a center-around mechanism in the filtered output graph Oi(s) (in biological vision, it is common to compare the

central position of the image with the neighborhood position), where s is the different scales in the pyramid. This results in a feature map Mi(c,s), given the "central" scale c and the "surrounding" scale s. The implementation here uses C =2,3,4, and S = C + D, where D =3,4. The scale difference (manipulation) between the two images is obtained by scaling up the major scale image to the corresponding pixel aberration around the scale image size.

For color and intensity channels:

$$M_i(c,s) = \left|O_i(c) \Theta O_i(s)\right| = \left|O_i(c) - Interp_{s-c}(O_i(s))\right|$$

Therefore, feature graphs of each feature type were calculated at six scales of 2-5,2-6,3-6,3-7,4-7 and 4-8 respectively, and these feature graphs were compared according to the center-surrounding mechanism.

For directional channel information, degree in four directions (Θ= 0,45, 90, 135 degree) and four scales (C = 0,1, 2,3) Apply the additive filter to the input grayscale images respectively to get 16 subchannels. The expression for the direction channel is as follows:

$$M_i(c) = Gabor(\theta_i, c)$$

Since the additive filter itself is already central-centered around contrast, no further processing is done here.

After the low-level center-around feature calculation, the GIST vector needs to be extracted from each corresponding subchannel. Then the whole picture is divided into 4*4 grid regions, and the eigenmatrices within the region are averaged.

For color and feature channels, the 16-dimensional vector of each feature image can be obtained by the following formula:

$$G_i^{k,l}(c,s) = \frac{1}{16WH} \sum_{u=\frac{kW}{4}}^{\frac{(k+1)W}{4}-1} \sum_{v=\frac{lH}{4}}^{\frac{(l+1)H}{4}-1} [M_i(c,s)](uv)$$

There, k and l are indexes in horizontal and vertical directions respectively, and W and H are the width and height of the feature map. The GIST vector is obtained by summing up the pixels of the feature map of the sub-region and dividing by the total number of pixels. Similarly, it can also be calculated from the direction characteristic diagram. Figure 3 illustrates this process.

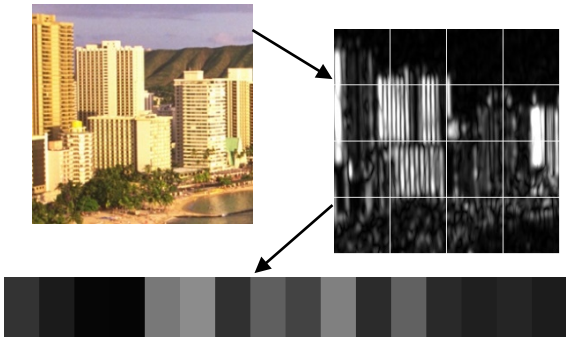

Fig.3 the Extraction Process of the Gist Vector

In this way, GIST vector of 544 dimensions can be obtained from 34 features, and then the dimensions of GIST vector can be reduced by using principal component analysis. Finally, SVM classifier [6] is used to obtain the final classification results.

2.3. Algorithm improvement

In view of the GIST model based on visual channels, the author found through a lot of experiments that beaches in natural scenes are easy to be confused with highways in artificial scenes, while there is also a lot of confusion between forest scenes and city center scenes, and the error rate of mountain scenes and street scenes is also very high. Finally, the suburbs of the outdoor scene and the skyscrapers of the indoor scene were also nearly half wrong with each other. That is to say, one category within the natural scene may easily diverge from another category within the artificial scene, but there are no false positives or omissions between the categories within the natural scene and the artificial scene.

It is suggested from the foregoing analysis that the GIST model based on feature channels can obtain better classification results in natural/artificial scenes. Then, in view of this problem, the classification process of GIST model based on feature channels is improved: The author airspace envelopment model combined with the Gist model based on visual characteristics of the channel, before the visual cortex characteristics of image are extracted, using the naturalness in the airspace envelopment model, the eight outdoor scene is divided into two categories: natural scene and artificial scene, in the good natural - classification under the condition of artificial scene, you can get a better scene semantic category classification, In this way, the negative impact of scene confusion between the two on the final classification can be reduced. The specific process is shown in Figure 4.

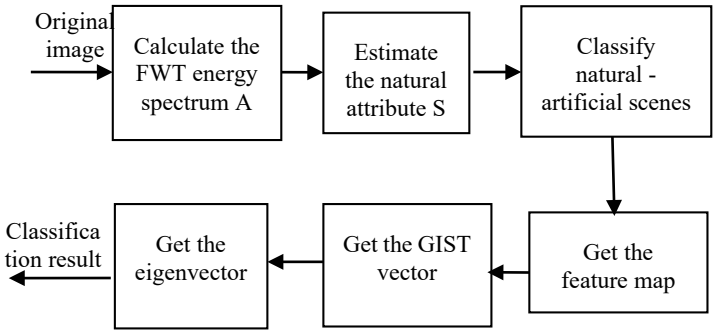

Fig.4 the Improved Classification Process of Gist Model

The process can be summarized as follows:

(1) Input the original image and Fourier transform the image to obtain the energy spectrum A (global energy spectrum/local energy spectrum) of the image;

(2) Integrate the energy spectrum A of the image with the naturalness discrimination template that has been trained to obtain the natural attribute S of the image;

(3) Divide the picture into natural scene/artificial scene by using natural attribute S;

(4) After processing the original image, 6 intensity feature images, 12 color feature images and 18 direction feature images were obtained;

(5) Extract the GIST feature vector from the feature graph obtained in step (4);
(6) The feature vector of the image is obtained by dimensionality reduction of GIST vector;
(7) Classify images in the category of natural scene/artificial scene by using feature vector to get the final semantic classification result.

3. EXPERIMENT

3.1. Scene image library

The scene image library selected for the simulation experiment in this paper is the image library proposed by Oliva and Torralba [7], which has a total of 2688 real scene pictures. Among them, there are 1472 pictures of natural scenes, which are: beach (360), forest (328), mountain (374), countryside (410); In addition, there are 1216 images of artificial scenes, namely: highways (260), city centers (308), streets (292), and skyscrapers (356) (the numbers in brackets are the number of images contained in the scene). It is worth noting that the number of images in the eight categories in the image library is not the same, and all images are 256*256 color images.

3.2. Process of experiment

According to the improved GIST model, the data structure of each image should include: global naturalness S1, local naturalness S2, feature vector G, natural-artificial scene label L1, natural-artificial scene prediction label LP1, semantic category label L2, semantic category prediction label LP2. The simulation experiment is mainly composed of training stage and testing stage. The overall process of the experiment is as follows:
(1) 400 natural pictures and 400 artificial pictures were randomly selected from the image library, and the remaining 1888 pictures were used as test pictures.
(2) According to the content mentioned in Section 2.1, the natural discriminant spectral template was obtained by using training pictures.
(3) The energy spectrum A of each image is obtained by Fourier transform processing of all images, and the naturalness S of each image is obtained by integrating it with the naturalness obtained by (2).
(4) Using the naturalness S of training pictures and the semantic labels of natural-artificial scenes, the training model M1 is obtained;
(5) Test the images according to the model trained in (4), and use the naturalness S of itself and the semantic label LP1 of natural-artificial scene.
(6) Then, according to LP1, feature extraction is carried out on the training images to obtain GIST vector and feature vector.
(7) The training model M2 was obtained by using the feature vector and semantic category label L2 of the training images.
(8) Finally, according to the training model M2 and their feature vectors and semantic category label L2, the predicted label LP2 is obtained for the test images, and the final classification results are obtained.

3.3. Natural - artificial scene classification

The GIST model proposed in this paper further classifies semantic categories based on the results of natural-artificial scene classification. Therefore, the natural-artificial scene classification with good recognition rate and stability is necessary for the GIST model proposed in this paper. For the reasons mentioned above, the author classifies natural-artificial scenes according to four different situations, which are mainly differentiated by energy spectrum segmentation and different ways of randomly selecting training images. In

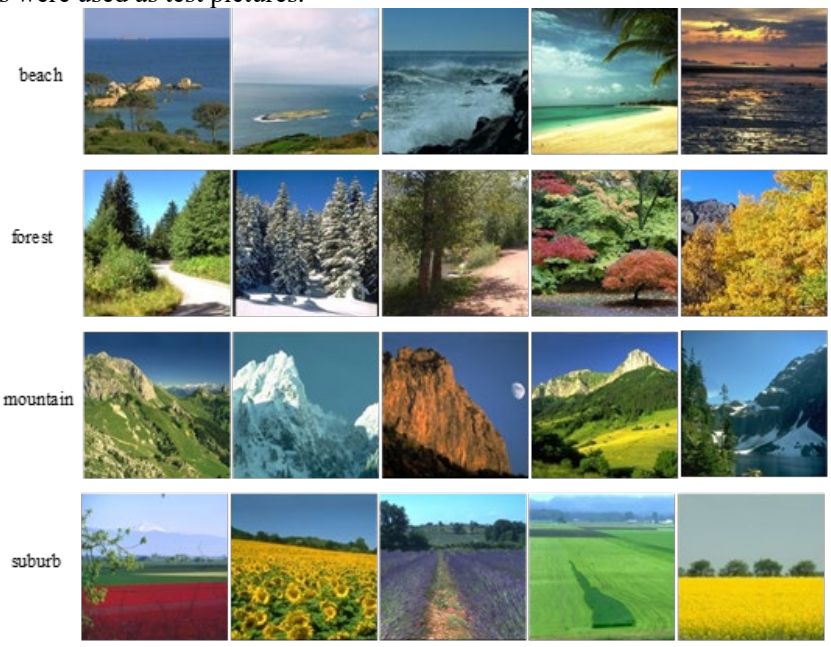

Fig.5 The Image Library of the Natural Scene

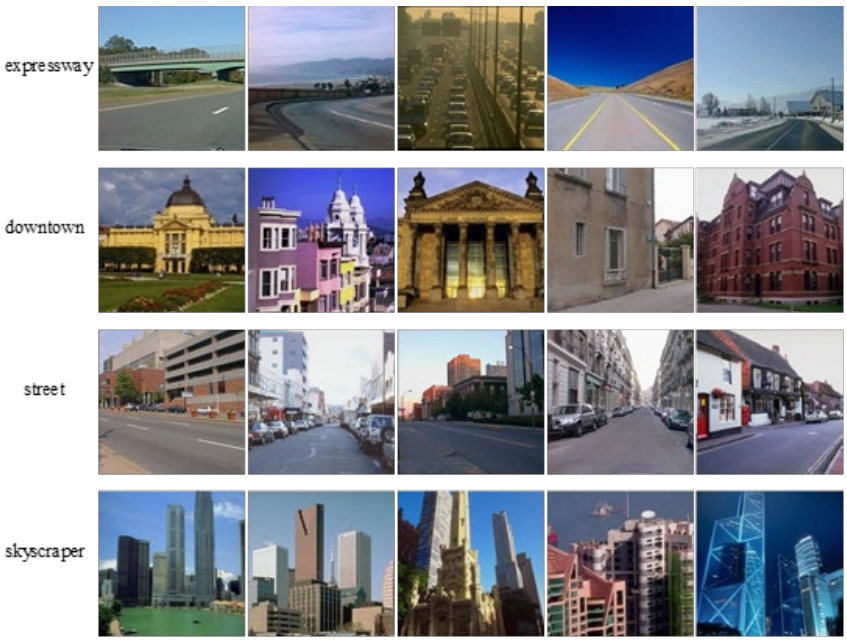

Fig.6 The Image Library of the Man-made Scene

these four situations, the author unified 800 training pictures (400 for natural scenes and 400 for artificial scenes) and 1888 test pictures (2688 for all pictures, and all pictures except training pictures were used for test pictures each time). The classifiers used are all SVM classifiers (where the parameters of SVM are the best parameters under the classification). In a large number of repeated experiments, the author took the results of 10 consecutive classifications under four conditions respectively.

The proposition P1 is: the estimated naturalness S in the global energy spectrogram. Proposition P2 is: the natural attribute s estimated in the local energy spectrum of 4*4. Q1 proposition is: 400 natural scene pictures and 400 artificial scene pictures are randomly selected from 2688 scene pictures. Q2 proposition is: 100 scene pictures are selected from 4 natural scenes and 4 artificial scenes respectively, as 400 natural scene pictures and 400 artificial scene pictures. From these four propositions, it can be seen that propositions P1 and P2 represent two comparative situations to obtain attribute S, while Q1 and Q2 represent two ways to select training pictures. In Figure7, the yellow curve represents P1&Q1, the blue curve represents P1&Q2, and the green and red curves represent P2&Q1 and P2&Q2, respectively. As can be seen from the figure, the natural attribute S estimated in the 4*4 local spectral energy figure can not only obtain a good classification result, but also maintain a very stable classification result. For the green and red curves under the same P2 proposition, it can be concluded that a more balanced selection of scene pictures can obtain better classification results, and the natural-artificial scene recognition rate in this case is stable at 100%. This is due to the selection of a semantic category from the eight images library scenario is different in the number of images, this can't guarantee when random nature - artificial scene image can have enough pictures for covering the small number of semantic category scene pictures, so good recognition results no equilibrium under the selection of recognition results.

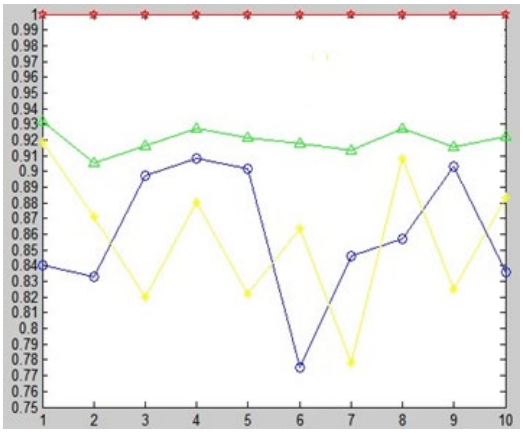

Fig.7 Comparison of recognition rate between natural and artificial scene classification

3.4. Experimental results and analysis

The image library proposed by Oliva and Torralba divides the scene into eight categories: beach, forest, mountain, suburb, expressway, downtown street, and skyscraper. In this part, the author can prove that the GIST model proposed in this paper can classify the image library according to its semantic category and obtain good classification results.

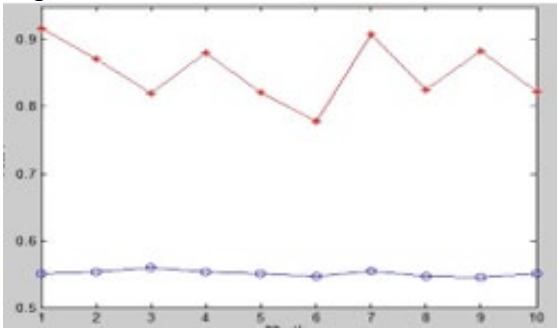

Fig.8 The Semantic Classification Comparison of the Scenes

Experiment as stated earlier, the process of global energy use first on the way to extract the naturalness of s overall division as the classification of the scene, and then through the visual features of three channels: the color, intensity and direction, to get in front of the "natural" and "artificial scene", the final classification results are obtained. In this paper, the recognition rate distribution of 10 continuous samples of this experiment is listed. At the same time, this paper will also give the confusion matrix of these eight categories. In order to verify the rationality of the data, the author did not list a single confusion matrix, but gave the corresponding confusion matrix of the previous 10 consecutive samples to sum and take the average matrix. In Figure8, the blue curve represents the recognition rate obtained by direct classification of 10 GIST vectors, while the red curve represents the sampling distribution map of the recognition rate obtained by using the GIST model proposed in this paper. As can be seen from the figure, the result of feature classification using GIST vector after natural-artificial scene classification is obviously better than that obtained by direct classification using GIST vector. However, the stability of the red curve is obviously worse than that of the blue curve, and the distribution space and fluctuation degree of the vertical axis of the red curve are similar to the classification recognition rate curve of the natural-artificial scene. Table 1 provides an explanation of the above results. The 4*4 square matrix on the upper left and the 4*4 square matrix on the lower right of the table are both diagonal matrices, which indicates that the four inner semantic categories can be completely separated through visual feature channels under the condition that the natural-artificial scene is classified correctly. The number of square matrixes 0 on the lower left is very small, while the number of square matrices 0 on the upper right is relatively large, which indicates that (False+ and False-) naturalness S mentioned above is easy to misreport artificial scenes as natural scenes, while natural scenes are difficult to be misclassified into artificial scenes. Theoretically, naturalness s mainly through judging direction advantage of natural or artificial scene, artificial there are a lot of images in the scene is dominated by the vertical - horizontal direction, such as unlimited extension of streets and roads, hovering up skyscrapers and so on, and the natural scene is difficult to have a level - the scene of the dominance of vertical direction. In the confusion matrix, it is worth noting that the row and column of the highway are easily confused with the beach scene in the natural scene. As can be seen from the scenes in Figure 5 and Figure 6, the beach scene is very similar to the highway scene in terms of color and orientation dominance, which is the main reason for the confusion between them.

Table.1 The confusion matrix based on the global natural degree and virtual feature channel model

category	beach	suburb	mountain	suburb	expressway	downtown	street	skyscraper
beach	258.6	0	0	0	2.8	0	0	0
forest	0	227.7	0	0	0.4	0	0	0
mountain	0	0	296.4	0	4.7	0	0.1	0
suburb	0	0	0	309.4	1.3	0	0	0
expressway	132.4	0	0.1	11.6	14.2	0	0	0
downtown	35	0.8	0.7	0	0	169.1	0	0
street	16.3	9.2	14.1	14.9	0	0	141.4	0
skyscraper	32.9	0	3.9	0	0	0	0	223.1

Table.2 The confusion matrix based on the localized natural degree and virtual feature channel model

category	beach	suburb	mountain	suburb	expressway	downtown	street	skyscraper
beach	262	0	0	0	0	0	0	0
forest	0	229	0	0	0	0	0	0
mountain	0	0	276	0	0	0	0	0
suburb	0	0	0	313	0	0	0	0
expressway	0	0	0	0	162	0	0	0
downtown	0	0	0	0	0	212	0	0
street	0	0	0	0	0	0	195	0
skyscraper	0	0	0	0	0	0	0	225

4. CONCLUSION

In this paper, a scene classification method based on natural and visual feature channels is proposed. This method ignores the steps of region segmentation and object recognition, and obtains the GIST vector of the scene image from the global features of the scene, thus simplifying the steps of scene classification. At the same time, the comparative experiment shows that this method is better than other scene classification methods for Oliva and Torralba image libraries, so it has a certain theoretical value and application significance.

Although the method in this paper has a high recognition rate, the time cost of the algorithm is worth further investigation for fast scene classification based on global features. The method proposed in this paper not only extracts the naturalness from the Fourier spectrum, but also extracts the GIST vector of the visual feature channel from the original image, which is much more complicated than the general method of feature extraction, so the time consumption is much higher than the cost of the ordinary fast scene classification method. In addition, the method in this paper has a big drawback that it is only applicable to outdoor scenes. This is because the naturalness S in the airspace envelopment model has a good effect on the classification of natural-artificial scenes, but the classification of indoor or outdoor scenes needs further research.