Hindawi
Complexity
Volume 2021, Article ID 5556899, 10 pages
https://doi.org/10.1155/2021/5556899

Research Article

Commodity Image Classification Based on Improved Bag-of-Visual-Words Model

Huadong Sun,[1,2,3] **Xu Zhang**,[1,2] **Xiaowei Han**,[1,2] **Xuesong Jin**,[1,2] **and Zhijie Zhao**[1,2,3]

[1]*School of Computer and Information Engineering, Harbin University of Commerce, Harbin 150028, China*
[2]*Heilongjiang Provincial Key Laboratory of Electronic Commerce and Information Processing, Harbin University of Commerce, Harbin 150028, China*
[3]*North-East Asia Service Outsourcing Research Centre, Harbin University of Commerce, Harbin 150028, China*

Correspondence should be addressed to Huadong Sun; kof97_sun@163.com

Received 4 February 2021; Revised 1 March 2021; Accepted 4 March 2021; Published 17 March 2021

Academic Editor: Abd E.I.-Baset Hassanien

With the increasing scale of e-commerce, the complexity of image content makes commodity image classification face great challenges. Image feature extraction often determines the quality of the final classification results. At present, the image feature extraction part mainly includes the underlying visual feature and the intermediate semantic feature. The intermediate semantics of the image acts as a bridge between the underlying features and the advanced semantics of the image, which can make up for the semantic gap to a certain extent and has strong robustness. As a typical intermediate semantic representation method, the bag-of-visual-words (BoVW) model has received extensive attention in image classification. However, the traditional BoVW model loses the location information of local features, and its local feature descriptors mainly focus on the texture shape information of local regions but lack the expression of color information. Therefore, in this paper, the improved bag-of-visual-words model is presented, which contains three aspects of improvement: (1) multiscale local region extraction; (2) local feature description by speeded up robust features (SURF) and color vector angle histogram (CVAH); and (3) diagonal concentric rectangular pattern. Experimental results show that the three aspects of improvement to the BoVW model are complementary, while compared with the traditional BoVW and the BoVW adopting SURF + SPM, the classification accuracy of the improved BoVW is increased by 3.60% and 2.33%, respectively.

1. Introduction

With the increasing of e-commerce, online shopping has become the main way for the public to buy goods. In order to provide a better shopping experience for users to quickly browse and search for goods, a good commodity image classification system also highlights its importance. In the past, the image classification method based on manual tagging has not met the actual needs [1]. How to use image processing, computer vision, pattern recognition, and machine learning to realize the classification of commodity images has great research and commercial value.

The process of image classification includes feature extraction, classifier training, and classification. In the mature development of classifier, image feature extraction often determines the final effect. Whether in the field of image recognition, image retrieval, image classification, or image ranking, feature extraction is very important [2, 3]. At present, the image feature extraction mainly includes the underlying visual feature and the intermediate semantic feature [4, 5]. The underlying visual features mainly refer to the color, texture, and shape of the image. By extracting the underlying features, the high semantics of the image can be inferred and the image classification can be realized. The commonly used underlying visual features are as follows.

Color feature is one of the most commonly used underlying features, which has the advantages of simple extraction, rotation, translation and scale invariance, and clear physical meaning. Color is also an important basis for commodity image classification and retrieval. Different types

of commodity images show great differences in color proportional distribution and color spatial distribution. The color histogram algorithm is proposed for the first time by Swain and Ballard [6], which is an intuitive expression of image content. By counting the frequency of different color pixels in the image, the composition of color is reflected. Strieker and Orengo et al. [7] proposed the color moment, which does not need vector quantization and directly accumulates statistics on each channel, and then only nine values are needed to describe the feature information. However, the recognition accuracy of this method is low, so it is necessary to combine other methods to characterize the image. Naushad et al. [8] constructed three probability histograms for each color component, which were then divided into numbers of several valid intervals and calculated statistics such as standard deviation, skewness, and kurtosis from each interval which were used as image color features.

Texture feature is based on the gray level statistics, describing the smoothness, roughness, and appearance law, reflecting the structure information of the image and the spatial distribution of the gray level. Texture features, as inherent properties of object surfaces, are of great significance in the classification of commodity images. Ojala et al. proposed LBP operator [9, 10] for texture classification. This method has small computational complexity and multiscale and rotation invariant properties and is widely used in texture retrieval. With the depth study of LBP rotation invariance, other forms, such as LBP variance and global matching [11], complete model [12] of LBP, joint distribution [13] of simulated Gaussian mixed local patterns, and modified LBP [14], have been effectively applied in texture classification.

Shape can distinguish the different objects in the image more intuitively; shape features are usually related to the target which the user is interested. Shanmugavadivu et al. [15] used fuzzy-object-shape to capture the shape of the object, improving the accuracy of boundary information, and provided an approximation measure of the object by a conventional shape. Wu et al. [16] proposed a new algorithm to calculate the rotation invariant of Tchebichef moment, and translation and scale invariance of Tchebichef moments are achieved by prealigning the image into a standard image. The proposed descriptor is compared with radial Tchebichef moment, and two kinds of Zernike moment experimental results show that the proposed shape features are robust to deformations generated by image shape rotation and scaling. Sokic et al. [17] proposed an improved Fourier descriptor method, which is capable of extracting Fourier descriptors in condition of translation, scaling, rotation, and starting point changes.

Commodity images are rich in color, shape, and texture information. Such underlying features provide simple representation of images based on physical level. Once there is a large change in the class or there is a significant background interference, the classification accuracy will be reduced. The intermediate semantics of the image acts as a bridge between the underlying features and the high semantics of the image, which can make up for the semantic gap to a certain extent robustness. As a typical intermediate semantic representation method, the bag-of-visual-words (BoVW) model [18] has received extensive attention in image classification. However, the traditional BoVW loses the location information of local features, and its local feature descriptors always lack the expression of color information.

In recent years, deep learning has become a research hotspot in the field of machine learning and artificial intelligence. As a feature learning method, deep learning has achieved good results in image classification. The AlexNet [19] model is the beginning of CNN widespread concern, followed by many innovative classification structure models, such as GoogLeNet [20], Inception v3 [21], ResNet [22], and so on. However, the structure of the CNN is complex, and it takes a lot of time in the calculation process, and the network lacks the necessary interpretability.

In this paper, the improved BoVW for the commodity image classification according to the shortcomings of traditional BoVW and the characteristics of commodity images is studied, which explore a more reasonable local feature description and add a description of location information to the model.

2. Principle of Bag-of-Visual-Words Model

The bag-of-visual-words (BoVW) model is a natural extension of the bag-of-words (BoW) model from natural language processing field to image processing field [18]. The principle of BoVW is described as follows: dividing the image into small pieces and then clustering similar chunks into visual words, BoVW counts the frequency of these visual words in the image, represented in the form of a histogram. Generally, image local features are used to compare words in the BoW model, such as SIFT [23] and SURF [24]. Since the importance of each visual word to different categories of images is different; hence, image classification can be performed by BoVW combined classifier (such as SVM). The schematic diagram of BoVW is illustrated in Figure 1.

By using the BoVW model to represent the image and obtain the global histogram representation of the image, there are five steps:

Step 1. Automatically detect the key points of the image and search for local regions.

Step 2. Feature extraction of local regions: According to the specific application considerations, the uniqueness of the features, the complexity of the extraction algorithm, and the effect of the selection features, the local feature extraction algorithm is used to extract local features from images.

Step 3. Visual dictionary construction: Generally speaking, a part of the image from different categories is selected from the image library to form the training image set, and its local features are extracted. Then, all the local feature vectors of the training image are defined as visual words by proper redundancy processing. The usual processing method is to cluster all the local feature vectors of the training image and define the cluster

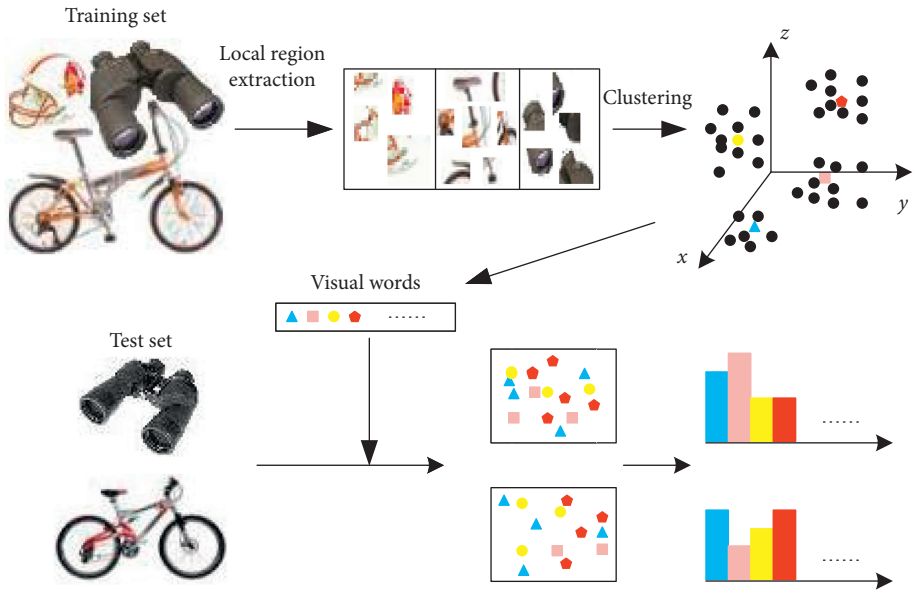

Figure 1: Schematic diagram of BoVW.

centre as a visual word. All visual words form a visual dictionary for histogram representation of images.

Step 4. Feature vector encoding: The BoVW model uses vector quantization technology to encode feature extraction of local regions. The result of vector quantization is to quantify the local feature vector of the image into the visual word which is the most similar. The vector quantization process is actually a search process. Usually, the nearest neighbor search algorithm is adopted to search for the most matching visual words with the local feature vectors of local regions.

Step 5. Use a visual word histogram to represent the image: After all the local feature vectors of an image are quantized, the frequency of each visual word in the visual dictionary can be counted, and a histogram about the visual word can be obtained. Its essence is the global statistical result of quantization coding obtained in the previous step. It is a numerical vector composed of visual word index order. This vector is the final representation of the image.

3. Improved Bag-of-Visual-Words Model

The traditional BoVW loses the location information of local features, and its local feature descriptors mainly focus on the texture or shape information of local regions but lack the expression of color information. In this paper, the BoVW model will be improved. Technical details include (1) multiscale local region extraction; (2) local feature description by SURF + CVAH; (3) dictionary generation and feature coding; and (4) diagonal concentric rectangular pattern.

3.1. Multiscale Local Region Extraction (MLRE). Multiscale local region extraction is the first step in BoVW, including multiscale key point extraction, location mapping, and region division. The methods to extract key points are Harris operator, Fast operator, and SUSAN operator. These algorithms are very strict in the selection of key points, so they are very suitable for image matching. However, when applied to BoVW, too strict position selection will lead that some effective regions cannot be extracted sufficiently, resulting in insufficient information extraction of local regions. Hence, the multiscale feature of wavelet transform is used to solve this problem, and the local region extraction algorithm suitable for the BoVW model is explored.

3.1.1. Selection of Multiscale Key Points. As we know, four subimages at the current layer can be obtained after wavelet decomposition of two-dimensional images including a rough sub-band (low frequency component); and there are detailed sub-bands (high frequency component) in three directions: horizontal, vertical, and diagonal. Each rough sub-band can continue to do the next level of decomposition.

Here, wavelet decomposition is used to carry on the multiscale analysis to the image, which realizes the multiscale key point extraction, as is shown in Figure 2. The process is as follows:

(1) In order to obtain more key points, the commercial image grayscale value and double up-sampling processing are adopted.

(2) Multilayer wavelet decomposition to up-sampled image is carried on, and the number of decomposition layers is 3.

(3) High frequency sub-bands' coefficients are normalized, and candidates are selected according to coefficients in condition that the coefficients of three high frequency channels are greater than 0.1 in the same coordinates.

(4) Nonmaximum suppression is carried on to all candidates, and then the key points in the

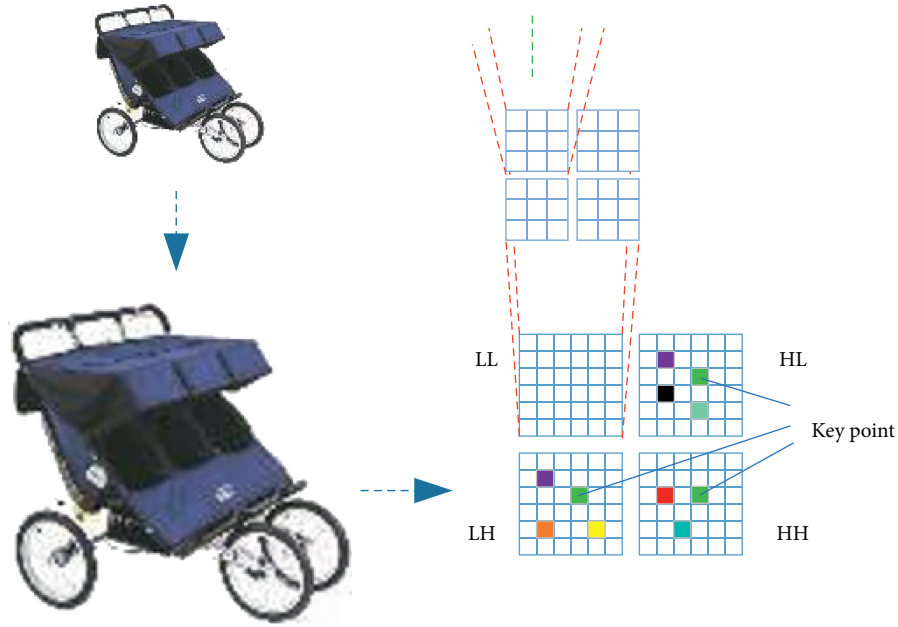

Figure 2: Schematic diagram of multiscale key point extraction.

corresponding scale are obtained. The detailed process is as follows. The matrix block of size 5×5 is delimited with the centre of candidate point. The value of each position of matrix block is the sum of the response values of three wavelet high frequency sub-bands' coefficients in the same coordinates. If the value of the candidate point is the maximum value of the region, it is retained as the key point.

3.1.2. Location Mapping and Local Region Extraction. Because the feature points are detected at different wavelet layers and each layer is carried on, the image is reduced by a quarter. In order to describe the local region features with more abundant information, it is necessary to map the coordinates to the original map. Location mapping relationship is described as

$$\begin{cases} X = x^L \cdot \left(2^{(L-1)} - k \cdot 2^{(L-2)}\right), \\ Y = y^L \cdot \left(2^{(L-1)} - k \cdot 2^{(L-2)}\right), \end{cases} \quad k = \begin{cases} 0, & L = 1, \\ 1, & L > 1, \end{cases} \tag{1}$$

where L is the number of wavelet decomposition layers, x and y are the position coordinates of key point detected on the scale, and X and Y are the coordinate positions corresponding to the original image.

It is worth noting that in image matching, the region size of key points needs to be calculated accurately, but in the BoVW model, the scale parameter is only a measure of the local region size. Therefore, it is not necessary to calculate the accurate scale by interpolation. The scale parameter is determined according to the following formula:

$$\sigma = 1.6 * 2^{L-1}, \tag{2}$$

where L is the number of wavelet decomposition layers, and 1.6 is the initial scale. Then, we can define local region as the square with length 20σ, whose centre is key point. However, another need to be determined is the main direction of local region, which will be introduced in detail in Section 3.2.1.

3.2. Local Feature Description by SURF + CVAH. The local feature description requires the balanced uniqueness of the balance. Uniqueness is so strong to distinguish between the two features which originally belong to the same visual word, while uniqueness is so weak to cause the visually distinct regions to be regarded as the same in the visual dictionary, which is obviously inappropriate. In addition to the uniqueness or differentiability, the local feature description generally requires the following properties: rotation invariance, illumination invariance, and noise resistance.

Color, texture, and shape, as three underlying features, play different roles in describing images. In the commodity image, the color of commodity is changeable, and even the same product has many colors, which leads to the fact that some color descriptors, such as color histogram and color moment, are not suitable. On the other hand, the texture of commodity image is relatively single, and there are many similar texture features between different commodities. Therefore, it is not enough to describe the commodity image by texture to satisfy the balance of descriptor's uniqueness. Therefore, this paper uses the series of texture and color features to solve this problem.

3.2.1. Speeded Up Robust Features (SURF). SURF can achieve a good local region texture description, which can be divided into two processes. (1) Main direction confirmation: in order to achieve the rotation invariance of features, we need to confirm the main direction of feature points. The direction is chosen around the feature point in a small area of scale. (2) Region feature generation: the feature region is

rotated to the main direction, and the texture information of local region is extracted.

In the SURF algorithm, the main direction is determined by the response value of Haar wavelet which is applied to multiple subdomains in the circular scale region. Specifically, the circular scale region with a radius of 6σ is taken as the local region of key point, and the sum of the horizontal and vertical Haar wavelet response values of all pixel points in the 60° sector region is counted. Then, the sector rotates at an interval of 11.46° and counts the sum of wavelet response values in the region after rotation. The wavelet response values of all sampling directions are obtained by pushing it until the sector region is rotated whole circle. The direction of the sector with the largest response value is selected as the main direction of key point. The schematic diagram is shown in Figure 3.

After the main direction selected, the coordinate system of image is rotated in the main direction so that the coordinate axis is consistent with the main direction. Then, $20\sigma \times 20\sigma$ rectangular areas are taken around the key point and this region is divided into 16 rectangular subregion blocks whose size is $5\sigma \times 5\sigma$. Next, calculating the Haar wavelet response value of subregion blocks, we can obtain the sum of response values in horizontal direction $\sum dx$, the sum of response values in vertical direction $\sum dy$, the sum of the absolute values in horizontal direction $\sum |dx|$, and the sum of the absolute values in vertical direction $\sum |dy|$. So, we can get the feature vector $(\sum dx, \sum dy, \sum |dx|, \sum |dx|)^T$ from one subregion block. Finally, after all feature vectors of 16 subregion blocks are concatenated in the same order, the SURF vector whose size is 64×1 can be obtained. The schematic diagram is illustrated in Figure 4.

3.2.2. Color Vector Angle Histogram (CVAH). A variety of methods are used to measure the difference between two kinds of colors in RGB color space, and the most commonly used distance measurement method is Euclidean distance. The Euclidean distance calculation method is simple and easy, whose characteristic is rotation invariance. However, the RGB color model is not uniform space, and its visual differences can be hardly are reflected by Euclidean distance, which exposes the shortcoming of Euclidean distance. Therefore, using angles to measure color differences is a good choice. In the RGB space, CVA represents the angle between the RGB color vector of two adjacent pixels, as shown in Figure 5.

The formula for calculating the color vector angle is as follows:

$$\theta = \arccos\left(\frac{r_1 r_2 + g_1 g_2 + b_1 b_2}{\sqrt{r_1^2 + g_1^2 + b_1^2}\sqrt{r_2^2 + g_2^2 + b_2^2}}\right), \quad (3)$$

where (r_1, g_1, b_1) is the color vector of a pixel in RGB space, (r_2, g_2, b_2) is that of adjacent pixels, and θ is the color vector angle between the two pixels.

The implementation process is as follows. In the neighborhood centered on the key points, according to formula (3), calculate the color vector angle between each pixel point and the key point in the local region. Then, quantify the color vector angle uniformly, count the number of pixels in each interval segment, and obtain the color vector angle histogram (CVAH). The color vector angle reflects the color difference of each pixel point in the local region relative to the central key point. Generally speaking, the color vector angle is between 0° and 90°. Here, the quantization step is chosen as 0.5°, so that the dimension of color vector angle histogram is 180×1.

The final local region feature description vector can be obtained by splicing the 64-dimensional SURF vector with the 180-dimensional CVAH, which is 244-dimensional and can effectively describe the color, shape, and texture information of the local region.

3.3. Dictionary Generation and Feature Coding. After local features extraction, each feature represents a local region, and visual words can be obtained by the clustering algorithm. This is a typical dictionary generation and coding problem. The detail is as follows.

T_i is the local feature vector, whose size is $D \times 1$; that is, $T_i \in R^{D\times 1}$. Obviously, if the feature descriptor introduced in Section 3.2 is used, T_i is made of SURF and CVAH splicing, that is; $D = 244$. Suppose there are N local feature vectors corresponding to all local regions from all training images, the set composed by all local feature vectors can be written as

$$T = [T_1, T_2, \ldots, T_N], \quad T \in R^{D\times N}. \quad (4)$$

Then, K-means quantization can be described by the following formula:

$$\begin{aligned} &\min_{U,V} \quad \sum_{i=1}^{N} \left\| T_i - V \cdot u_i \right\|^2, \\ &\text{s.t.} \quad \text{Card}(u_i) = 1, \quad \left\| u_i \right\|_1 = 1, \; u_i \geq 0. \end{aligned} \quad (5)$$

In formula (5), V is codebook (dictionary), which contains K visual words (cluster centers), that is,

$$V = [V_1, V_2, \ldots, V_K], \quad V \in R^{D\times K}, \quad (6)$$

where column vector $V_i (i = 1, 2, \ldots, K)$ is visual word and $V_i \in R^{D\times 1}$.

In formula (5), u_i is the code which is generated by using codebook V to encode local feature vector T_i and u_i is a column vector of K dimension, $u_i \in R^{K\times 1}$, while $\text{Card}(u_i) = 1$ and $\|u_i\|_1 = 1$ can ensure that one element of u_i can be taken as 1 and the other elements of u_i are all 0. The encoding matrix of all local feature vectors in the training set can be described as

$$U = [u_1, u_2, \ldots, u_N], \quad U \in R^{K\times N}. \quad (7)$$

In the quantization process, the BoVW model assigns a local feature vector to a unique visual word closest to it. The index of the only nonzero element in the code u_i indicates the cluster centre to which the local feature vector T_i belongs. Such an optimization problem can be transformed

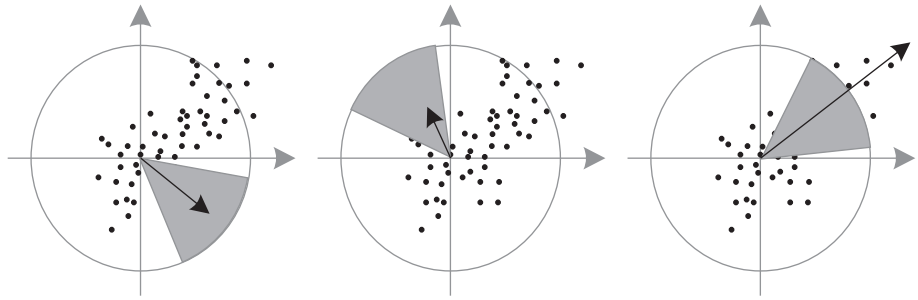

Figure 3: Main direction selection.

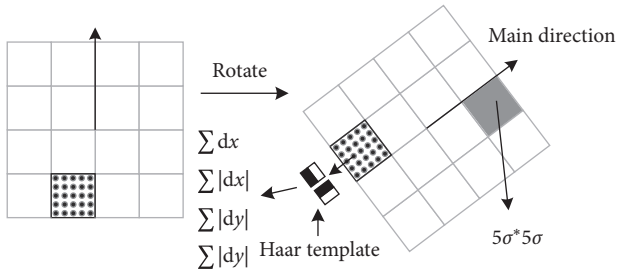

Figure 4: Region feature generation.

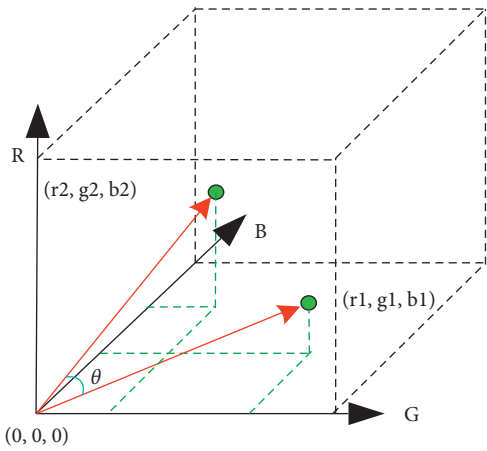

Figure 5: Color vector angle.

into a matrix decomposition problem of the encoding matrix U.

To an image, the final feature is the number of each visual words appearing in the whole dictionary. After normalization, the frequency of each visual word can be obtained. Suppose the local feature vector of an image corresponds to the first M column vectors of the matrix T, then the visual word histogram of this image is counted as

$$H = \frac{1}{M}\sum_{i=1}^{M} u_i. \tag{8}$$

The visual word histogram is a K-dimensional column vector, which can be used as the middle semantic expression of the image. The visual word histograms of all pictures in the training set can be obtained, and the classifier can be trained to guide the classification of the test images.

3.4. Diagonal Concentric Rectangular Pattern (DCRP). The traditional BoVW model discards the sequential relationship of feature descriptors, and the global histogram is completely missing the information of local features, which limits its representation ability. Spatial pyramid matching (SPM) was proposed to make up for its deficiency. In SPM, considering the spatial information, the image is divided into several blocks (sub-blocks), and the features of each sub-block are counted separately. Finally, the features of all blocks are spliced together to form the complete features, which are the distribution of image feature at different resolutions, and can obtain the local information of the image. During the partitioning process, SPM adopts a multiscale method to make its structure present a hierarchical pyramid shape. The SPM model is suitable for most scenes, but it also lacks particularity, which is not the best representation in the case of spatial distribution with characteristic law. Therefore, we propose diagonal concentric rectangular pattern (DCRP) which is suitable for describing the spatial distribution characteristics of commodity images.

It is necessary to analyze the characteristics of commodity images in e-commerce platforms. Firstly, commodity images of the same e-commerce platform always have a uniform resolution; that is, the row and column size is equal or close to equal. Secondly, the background of commodity images is always monotonous, which can be clearly divided into foreground region with a large amount of commodity information and monochromatic background region with noise points. Thirdly, the commodity target position in commodity images is generally in the middle.

Taking into account the spatial distribution characteristics of commodity images and drawing lessons from SPM ideas, we propose diagonal concentric rectangular pattern (DCRP), as shown in Figure 6. The scale definition of DCRP is slightly different from that of SPM. Scale 0 refers to the whole commodity image. Scale 1 refers to the central square region composed of four small squares, and the surrounding region composed of four trapezoids. Scale 2 refers to the 4 small squares and 4 trapezoids. Then, there are total $1+2+8=11$ blocks in all scales of DCRP. The final representation vector of the image is 11K dimension, which is less than 21K dimensions of SPM, which greatly reduces the computation. DCRP effectively introduces the location information of local features in commodity images, which improves the representation ability of BoVW.

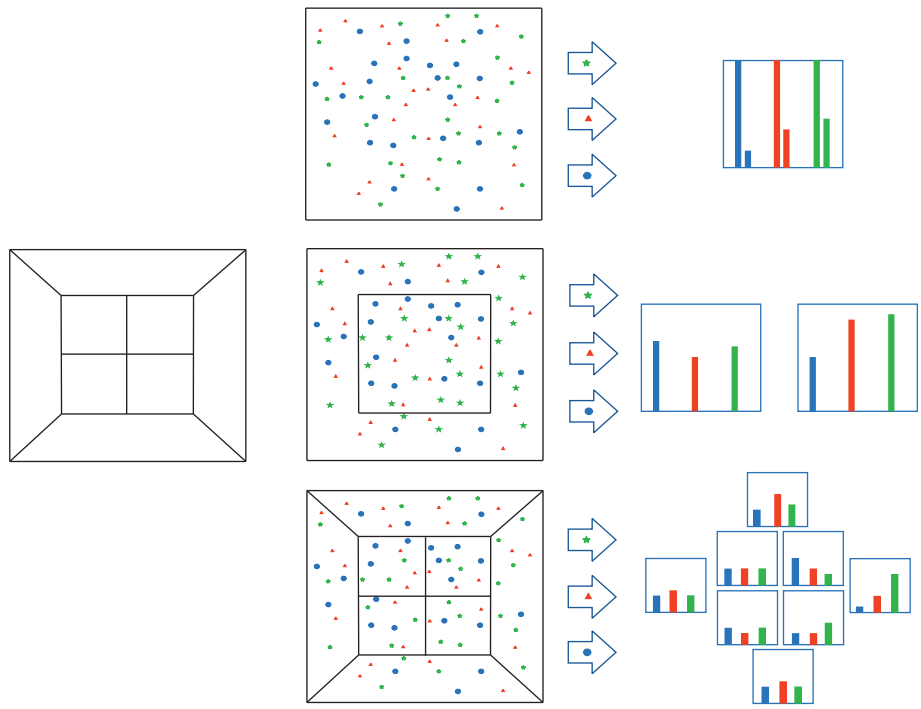

Figure 6: Schematic diagram of diagonal concentric rectangular pattern.

4. Experimental Results and Analysis

Experimental dataset is the Microsoft PI100, which has 100 kinds of commodity images, and each class has 100 images, a total of 10000 images. In the experiment, we select 10 kinds of commodity images. For each kind of images, 30 pictures are selected as the training set and 70 as the test set. SVM (support vector machine) is adopted to classify. Because the clustering centers obtained by the K-means algorithm are not completely consistent for each run, the classification accuracy is calculated by averaging the results of 10 runs. Experimental simulation environment is as follows: CPU: Intel Core i5-5200U; memory: 4G; simulation platform: Matlab R2018.

4.1. Effects of Multiscale Local Region Extraction on Classification. In Figure 7, the red curve is the result of BoVW with SIFT, the green is that of SURF, and the black one is that of SURF with the proposed multiscale local region extraction (for short, MLRE) algorithm. It can be seen clearly, with the increase in the number of visual words in BoVW, the classification effect is incremental. When the number of visual words tends to 1000, classification accuracy tends to be close to the limit. When the number of visual words is 1000, the accuracy of SURF (MLRE) is 87.12%, 1%, and 15.1% higher than that of SURF and SIFT, respectively. It explains that MLRE proposed here is effective to commodity image classification.

4.2. Comparison of Different Descriptors of Local Region. Figure 8 shows the contrast of different local area descriptors, where the classification effect is incremental with

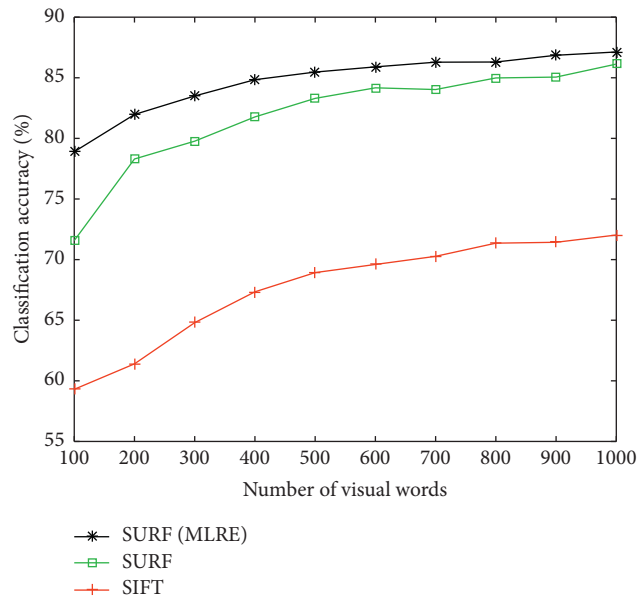

Figure 7: Effects of multiscale local region extraction on classification.

the increase in the number of visual words in the BoVW model. When the number of word packets is 900, without MLRE, the accuracy of BoVW using SURF as local feature description is 85.06%, while that of the BoVW model adopting SURF + CVAH is 86.80%, and there is 1.74% increase. On the other hand, when MLRE is used, the accuracy of BoVW using SURF is 86.86, while that of the BoVW model adopting SURF + CVAH is 88.69%, increased by 1.83%. As can be seen from the above results, adding CVAH

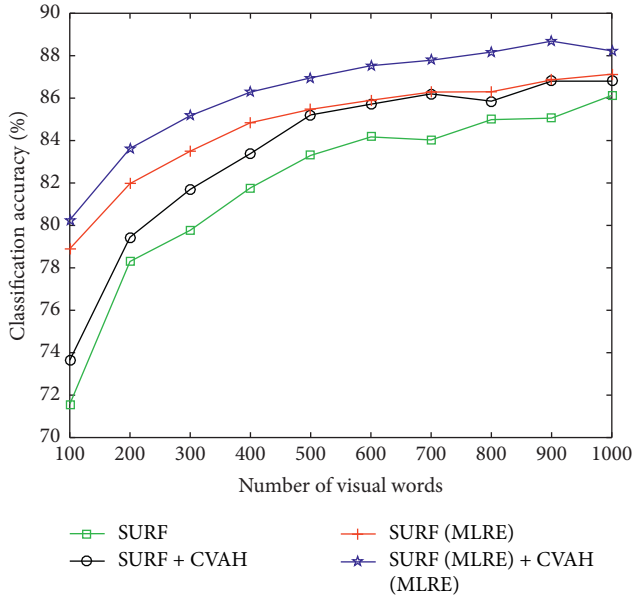

Figure 8: Comparison of different descriptors of local region.

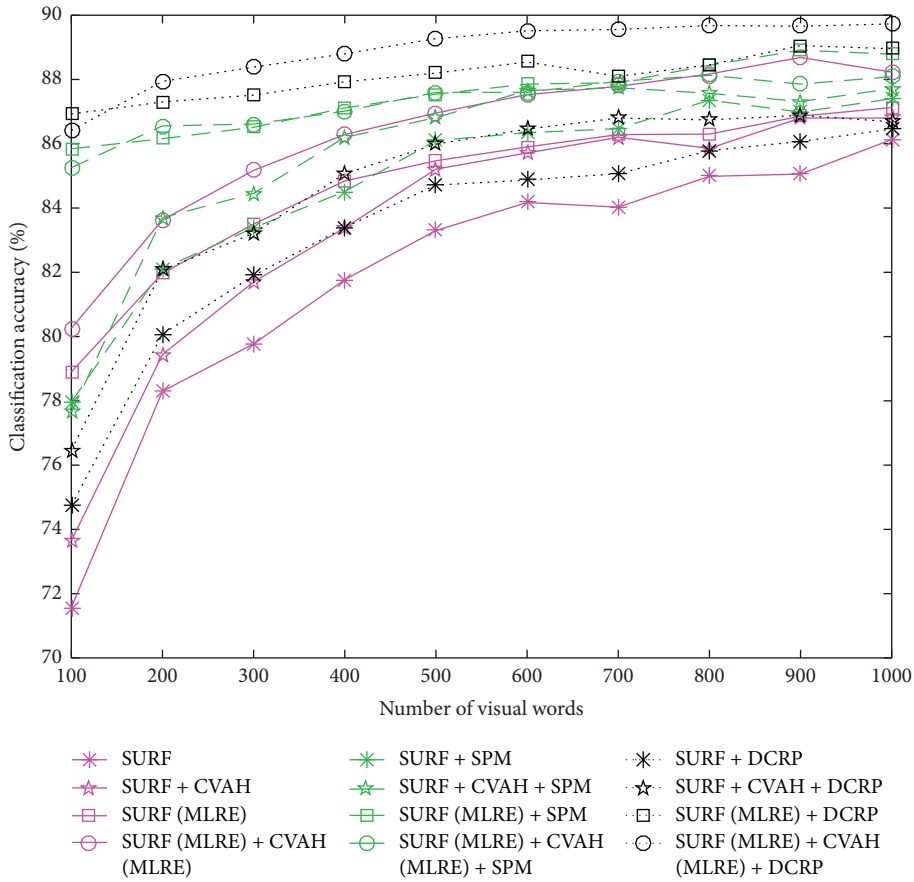

Figure 9: Effects of different spatial patterns on classification.

to local feature description is beneficial to the improvement of classification accuracy. And it can also be seen that the classification accuracy of MLRE and CVAH is 3.63% higher than that of SURF alone.

4.3. Effects of Different Spatial Patterns on Classification. Figure 9 shows the effects of different patterns on classification, where the four pink lines represent the classification accuracy of tradition BoVW without any spatial pattern, the

Table 1: Comparison with other image classification methods on Microsoft PI100 dataset.

Method	Classification accuracy (%)
LBP + SVM	81.29
HOG + SVM	84.71
BoVW (SURFC + CLBP) + SVM [25]	85.66
BoVW (SIFT) + SPM + SVM [26]	87.32
Proposed algorithm	**89.73**

four green lines represent those of BoVW using SPM patterns, and the four black lines are the those of BoVW using the proposed DCRP patterns. Overall, the BoVW using SPM pattern is better than tradition BoVW without any space pattern, and the BoVW using DCRP pattern is better than that of SPM pattern. It shows that for commodity images, the introduction of DCRP spatial pattern is beneficial to the improvement of classification accuracy. Besides, as seen from the diagram, the improved BoVW, which adopts MLRE, local feature description of SURF + CVAH, and DCRP spatial patterns, has the highest classification accuracy 89.73%. It shows that the three aspects of improvement to the BoVW model are complementary. Compared with the traditional BoVW and the BoVW adopting SURF + SPM, the classification accuracy of the improved BoVW is increased by 3.60% and 2.33%, respectively.

4.4. Comparison with Other Image Classification Methods. In order to further show the effectiveness of the proposed algorithm (improved BoVW), the classification effect of other algorithms is compared here, as shown in Table 1. LBP and HOG (histograms of oriented gradients) are the method based on the global underlying visual features, and the other three approaches all employ BoVW to describe the intermediate semantic feature. As can be seen from Table 1, the classification effects of the methods using BoVW are better than those based on the global underlying visual features. In [25], the local feature is extracted by SURFC + CLBP, but no spatial model is adopted, which results to loss of spatial information. In [26], the local feature is extracted by SIFT, and the SPM model is also employed. In the proposed method, the local feature is described by SURF (MLRE) + CVAH (MLRE), and the DCRP model is explored to compensate spatial information. Because of good local feature description and spatial pattern, compared with other four methods, the classification accuracy of the proposed method is increased by 8.44%, 5.02%, 4.07%, and 2.41%, respectively.

5. Conclusions

As a subtopic of image processing in the field of e-commerce, commodity image classification should not only solve the common problems in general image classification but also make targeted improvement according to the characteristics of commodity image. To the disadvantage that the traditional BoVW model cannot effectively describe the characteristics of commodity images, this paper proposes an improved BoVW model, which realizes multiscale local region extraction, adopts SURF + CVAH local feature description to add color information, and explores diagonal concentric rectangular pattern to supply spatial information. Experimental results show that the improved BoVW is very suitable for commodity image classification. As providing discriminative features, the improved BoVW proposed in this paper can be also used in image retrieval systems [27]. The subsequent work is to further develop reasonable local feature descriptors and further simplify the model to reduce the computational cost.

Data Availability

The Microsoft PI100 used to support the findings of this study is open dataset, which can be downloaded from: https://pan.baidu.com/s/15nVkXkw06GoFxqs1fVtrVQ using password t29E.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This paper was supported by the Heilongjiang Provincial Natural Science Foundation of China (LH2020F008) and Young Innovative Talents Support Project of Harbin University of Commerce (2020CX08).

References

[1] S. Chen and Z. An, "Image retrieval based on image entropy and regional expansion," *International Journal of Control and Automation*, vol. 9, no. 6, pp. 403–410, 2016.

[2] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, p. 1, 2019.

[3] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, vol. 23, no. 5, pp. 2019–2032, 2014.

[4] J. Yu, D. Tao, M. Wang, and Y. Rui, "Learning to rank using user clicks and visual features for image retrieval," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. 767–779, 2015.

[5] M. Hidajat, "Annotation based image retrieval using GMM and spatial related object approaches," *International Journal of Control and Automation*, vol. 8, no. 8, pp. 399–408, 2015.

[6] M. J. Swain and D. H. Ballard, "Color indexing," *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[7] M. A. Stricker and M. Orengo, "Similarity of color images," *Proceedings of SPIE*, vol. 2420, pp. 381–392, 1995.

[8] N. Varish and A. K. Pal, "Content based image retrieval using statistical features of color histogram," in *Proceedings of the 2015 3rd International Conference on Signal Processing, Communications and Networking*, pp. 1–6, Chennai, India, March 2015.

[9] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[10] M. Pietikäinen, T. Ojala, and Z. Xu, "Rotation-invariant texture classification using feature distributions," *Pattern Recognition*, vol. 33, no. 1, pp. 43–52, 2000.

[11] Z. Guo, L. Zhang, and D. Zhang, "Rotation invariant texture classification using LBP variance (LBPV) with global matching," *Pattern Recognition*, vol. 43, no. 3, pp. 706–719, 2010.

[12] Z. Guo, L. Zhang, and D. Zhang, "A completed modeling of local binary pattern operator for texture classification," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1657–1663, 2010.

[13] H. Lategahn, S. Gross, T. Stehle, and T. Aach, "Texture classification by modeling joint distributions of local patterns with Gaussian mixtures," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1548–1557, 2010.

[14] S. Fekri-Ershad, "Developing a gender classification approach in human face images using modified local binary patterns and tani-moto based nearest neighbor algorithm," *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 12, no. 4, pp. 1–12, 2019.

[15] P. Shanmugavadivu, P. Sumathy, and A. Vadivel, "FOSIR: fuzzy-object-shape for image retrieval applications," *Neurocomputing*, vol. 171, pp. 719–735, 2016.

[16] H. Wu and S. Yan, "Computing invariants of Tchebichef moments for shape based image retrieval," *Neurocomputing*, vol. 215, pp. 110–117, 2016.

[17] E. Sokic and S. Konjicija, "Phase preserving fourier descriptor for shape-based image retrieval," *Signal Processing: Image Communication*, vol. 40, pp. 82–96, 2016.

[18] Wikipedia. Bag-of-words model in computer vision [EB/OL]. http://en.wikipedia.org/wiki/Bagof-words_model_in_coniputer_vision.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Convolutional neural networks," *Advances in Neural Imagenet Classification with Deep Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[20] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9, Boston, MA, USA, June 2015.

[21] C. Szegedy, V. Vanhoucke, S. Ioffe, and J. Shlens, "Rethinking the inception architecture for computer vision," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, Las Vegas, NV, USA, June 2016.

[22] K. He, X. Zhang, S. Ren et al., "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[23] Q. Meng and Z. Lv, "An improved SIFT algorithm for image registration based realization of the vision figure," *International Journal of Control and Automation*, vol. 9, no. 6, pp. 51–58, 2016.

[24] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[25] Z. Zhang and Z. Ju, "Multi-feature fusion fruit and vegetable image classification based on bag of feature model," *Electronic Science and Technology*, vol. 31, no. 1, pp. 1–6, 2019.

[26] H. Zhang, S. Liu, B. Zhang, and J. Wang, "Natural scene recognition based on spatial pyramid integrated bag-of-visual-words model," *Journal of Shanghai Jiaotong University*, vol. 50, no. 6, pp. 902–909, 2016.

[27] N. T. Bani and S. Fekri-Ershad, "Content-based image retrieval based on combination of texture and colour information extracted in spatial and frequency domains," *The Electronic Library*, vol. 37, no. 4, pp. 650–666, 2019.