Simple Salience-Driven Bag of Visual Words Models for Remote Sensing Scene Classification

Lipeng Ji[1, a], Xiaohui Hu[2, b], Beijia Hu[3, c] and Mingye Wang[1, d]

[1]School of automation science & electrical engineering, Beihang University, Beijing 100191, China;

[2]Institute of Software Chinese Academy of Science, Beijing 100190, China;

[3]School of mathematics & systems science, Beihang University, Beijing 100191, China.

[a]jlp_1987@163.com, [b]hxh@iscas.ac.cn, [c]hubeijiabuaa@163.com, [d]marcel0829@126.com

Keywords: remote sensing scene classification, Itti model, GBVS, salience-driven BoVW model.

Abstract. In order to improve the accuracy of remote sensing scene classification, this paper proposes to integrate visual saliency into bag of words model. In this paper, the color histogram (CH), Scale Invariant Feature Transform (SIFT) and local binary pattern (LBP) methods are used to extract the features of the original remote sensing image firstly. Then, Itti model and Graph-based Visual Saliency (GBVS) algorithm are used to analyze the salient region separately. Finally, salience-driven Bag of Visual Words (BoVW) models are established by features which are both from the original images and filtered by salient regions. Experiments show that the salience-driven BoVW models can improve the accuracy of remote sensing scene classification obviously than without salience-driven..

1. Introduction

The remote sensing images contain very rich semantic information, which include global spatial structure information, local target with relative position information and high-level semantic information such as scene semantics and behavior semantics. The scene classification has always been an important issue in the field of remote sensing. With the increasing resolution of the remote sensing images, the details of the object are more obvious, and at the same time, the sense classification becomes more difficult. It is a new challenge for scene classification in remote sensing field.

Psychophysical and physiological evidence indicates that the visual system of primates and humans has evolved a specialized processing focus, which is directed to particular locations in the visual field [1]. And this is called “attentive” mode. Based on the “attentive” mode, researchers have proposed a lot of visual saliency models. Itti[2]model and Craph-based Visual Saliency (GBVS)[3]model are two typical model. In [4], a series of quantitative indicators are adopted to evaluate five significant visual saliency models, and the conclusion is that Itti model and GBVS model can highlight the salient region better than the other three models.

Bag of Visual Words (BoVW) [5] model is widely used in classification since it is proposed. In this paper, Itti and GBVS models are integrated into BoVW model for remote sensing scene classification. 10 groups of different scene semantic remote sensing images are used for experiment. And the experiment results show salience-driven BoVW models can improve classification accuracy.

2. Related Work

2.1 Bag of Visual Words.

Bag of Visual Words model is a number of occurrences of a vocabulary of image features, and an image can be conducted as a document. Furthermore, the definition of "words" is obligatory, which in images is it also needs to be defined. To apply the BoVW model, the following steps are taken into consideration: feature extraction and description, construct visual dictionary and training classifier.

The implement of image classification based on BoVW model can be described as follow:

Step 1: Feature extraction and description. Given an image, the main task is to extract global and local features to describe the image. The color histogram (CH), local binary pattern (LBP) and Scale Invariant Feature Transform (SIFT) descriptors are applied to achieve this process in this paper. CH feature is characterized by 120-dimesion vector. LBP feature is characterized by 59-dimesion vector, while SIFT feature is characterized by 128-dimesion vector.

Step 2: Construct visual dictionary. The k-means clustering algorithm [6] is applied to cluster a large number of feature points obtained in Step 1. The cluster center is defined as visual word, all the words are combined as visual dictionary and the size of dictionary is the number of words. In traditional methods, images are commonly represented as the histograms of visual words.

Step 3: Training classifier. Support Vector Machine (SVM)[7] is one of the most popular classifier and its implementation is simple in various classifiers. Besides, the main goal of SVM is to find a decision plane for separating between a set of objects which have different class memberships. SVM is initially applied to two-class classification problems and now it has been gradually used to solve multi-class high-dimension classification problems. It can be described as the following optimization problem:

$$\min_{w,\xi}\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i\right\} \quad (1)$$

Subject to the constraints: $y_i(w \cdot x_i - b) \geq 1 - \xi_i,\ \xi_i \geq 0, i = 1,2,\dots,n$. where w is the vector of coefficients, C is the capacity constant, b is bias and denotes a constant, ξ_i is the vector which represents parameters for handling nonseparable data. The index i labels N training cases and $y \in \pm 1$ represents the class labels.

2.2 Itti Model.

Attention model Itti was proposed to simulate the human visual system under natural environment. In Itti model three feature channels were extracted: the intensity channel, the color channel, and the orientations channel. Each feature was obtained by a set of linear center-surround operations akin to visual receptive fields. The processing of Itti model mainly contains two steps: extraction of visual features and establishment of saliency map.

Step 1: Extraction of visual features. Many visual features that have been found to influence visual salience in the primate brain, but in Itti model, these visual features are intensity I, red color R, green color G, blue color B, yellow color Y and orientations. The intensity and color features are obtained by following formulas:

$$\mathrm{I} = (\mathrm{r} + \mathrm{g} + \mathrm{b})/3 \quad (2)$$

$$\mathrm{R} = \mathrm{r} - (\mathrm{g} + \mathrm{b})/2 \quad (3)$$

$$\mathrm{G} = \mathrm{g} - (\mathrm{r} + \mathrm{b})/2 \quad (4)$$

$$\mathrm{B} = \mathrm{b} - (\mathrm{r} + \mathrm{g})/2 \quad (5)$$

$$\mathrm{Y} = \mathrm{r} + \mathrm{g} - 2(|\mathrm{r} - \mathrm{g}| + \mathrm{b}) \quad (6)$$

where r、g、b are red, green, blue color channels of the original image. Four preferred orientations $\theta = \{0°, 45°, 90°, 135°\}$ are chosen and obtained by oriented Gabor filters.

For a given image what attracts human's attention is not visual feature itself but "contrast", and center-surrounding structure is akin to visual receptions fields, so center-surround (c-s) difference is used to calculate each feature. The center-surround difference between two maps, denoted"Θ"below, is obtained by interpolation of the coarser scale to the finer scale, followed by point-by-point subtraction. The final intensity features are defined as $I(c,s) = |I(c)\Theta I(s)|$, final orientation features are defined as $(c,s,\theta) = |O(c,\theta)\Theta O(s,\theta)|$, while final color features are denfined as $RG(c,s) = |(R(c) - G(c))\Theta(G(s) - R(s))|$ and $BY(c,s) = |(B(c) - Y(c))\Theta(Y(s) - B(s))|$, where RG and BY are red/green and blue/yellow color opponencies.

Step 2: Establishment of saliency map. In Itti model, 42 feature maps (6 maps for intensity, 12 maps for color and 24 maps for orientation) are obtained throught across-scale. Itti proposed "Iterative localized interactions" to combine the above 42 feature maps into 3 conspicuity maps (intensity, color and orientation). Each of these three so-called "conspicuity maps" are needed to be normalized into a fixed range [0,1], in order to eliminate modality-dependent amplitude differences. Itti thought that similar features compete strongly for conspicuity map, while different conspicuity map contribute

independently and equally to the saliency map. So the final saliency map was created by summing the three conspicuity maps averagely.

2.3 GBVS Model.

GBVS model use Markovian approach to establish saliency map, which is different form Itti model. The steps of GBVS model can be described as follow:

Step 1: Extraction of visual features. This step is similar to Itti model step 1.

Step 2: Establishment of saliency map. Suppose that a feature map M: $[\mathrm{n}]^2 \to R$ is already obtained from the original image. And then define $d((i,j)||(p,q)) \triangleq log\,|\frac{M(i,j)}{M(p,q)}|$ as the dissimilarity of $M(i,j)$ and $M(p,q)$. This is a natural definition of dissimilarity: simply the distance between one and the ratio of two quantities, measured on a logarithmic scale. Consider now the fully-connected directed graph G_A, obtained by connecting every node of the lattice M, labelled with two indices $(\mathrm{i},\mathrm{j}) \in [\mathrm{n}]^2$, with all other $n-1$ nodes. The directed edge from node (i,j) to node (p,q) will be assigned a weight $w_1((i,j),(p,q)) = d((i,j)||(p,q)) \times F(i-p, j-q)$, where $F(a,b) = \exp(-\frac{a^2+b^2}{2\sigma^2})$ (σ is a free parameter). Thus, the weight of the edge from node (i,j) to node (p,q) is proportional to their dissimilarity and to their closeness in the domain of M. Define a Markov chain on G_A by normalizing the weights of the outbound edges of each node to 1, and drawing an equivalence between nodes & states, and edges weights & transition probabilities. The saliency map is the result of an activation measure which is derived from pairwise contrast.

3. Salience-Driven BoVW Models

In this paper, saliency visual methods are integrated into BoVW model for remote sensing scene classification. The modeling processing can be shown in Figure 1.

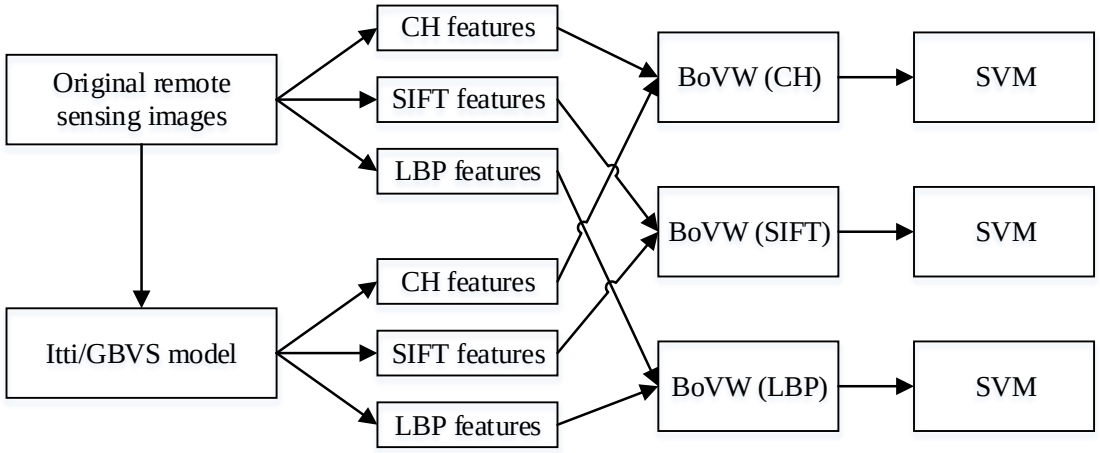


Fig. 1 Salience-driven BoVW model

The implement of remote sensing sense classification based on salience-driven BoVW models can be described as follow:

Step 1: Feature extraction and description. CH, SIFT and LBP descriptors are applied to achieve the features separately.

Step 2: Features extraction form saliency images. Using Itti model or GBVS model to get the saliency images of original remote sensing images separately. Then extraction CH, SIFT and LBP features from these saliency images separately. The final features are formed by adding the features extracted in Step1 together.

Step 3: Construct visual dictionary. Construct visual dictionary by features obtained in Step 2.

Step 4: Classification. SVM is used as the training classifier in this paper.

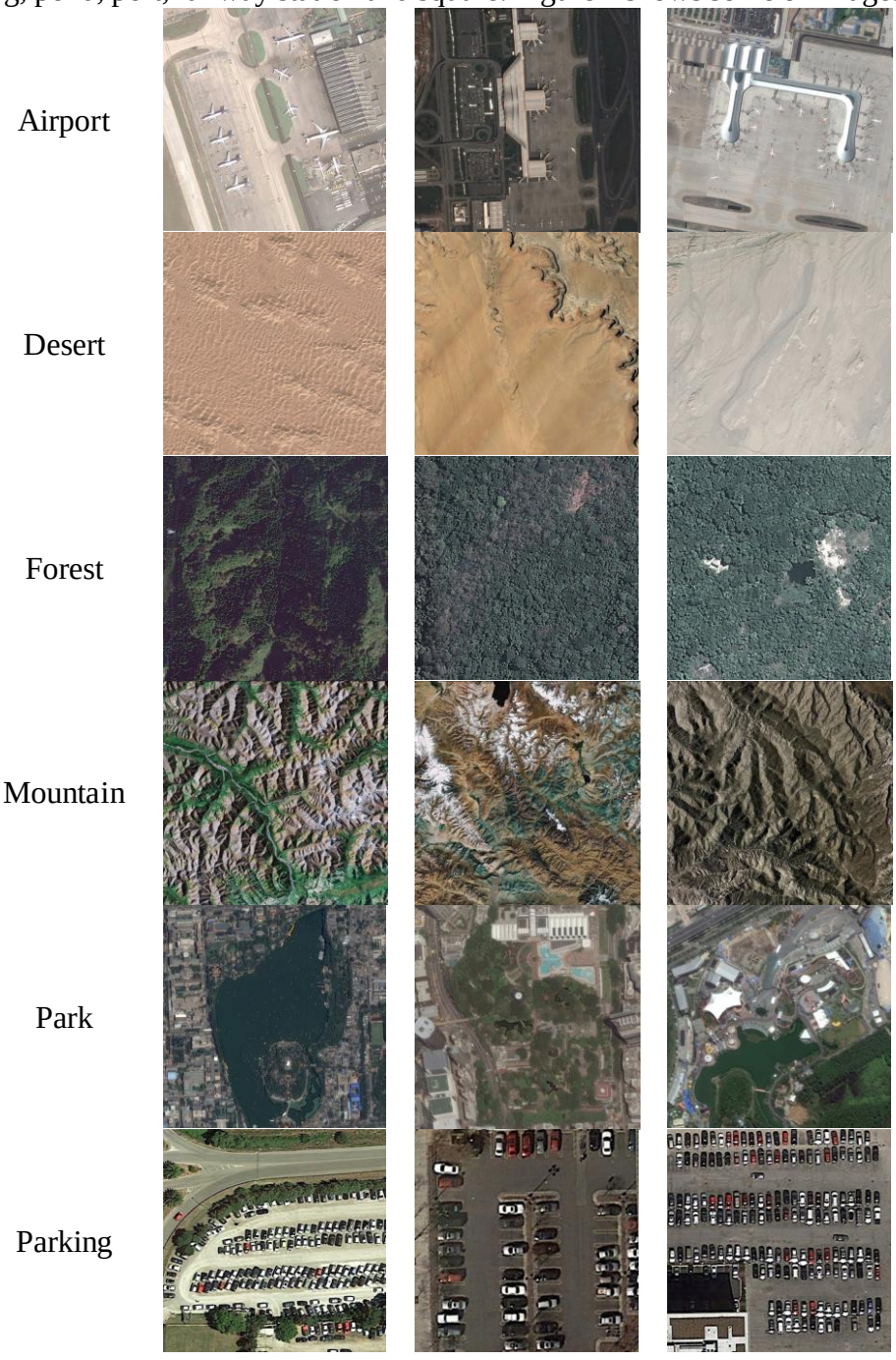
In this paper, the saliency images are seen as important as the original remote sensing images. However, it's not the state-of-art in saliency-driven models. There is a ratio between the original and saliency images showing in formula below:

$$I(i,j) = \alpha I(i,j) + \beta M(i,j) \quad (7)$$

where $I(i,j)$ and $M(i,j)$ are the original image and saliency image, while α and β show the ratio between the original and saliency images, which range in [0,1]. If you want to get a higher accuracy of remote sensing scene classification, you can choose a better ratio for your dataset. While in this paper, both α and β are equal to 1.

4. Experiments

The dataset used in this paper is collected form Google Earth, which contains 2000 remote sensing images labeled into 10 typical scene categories. There are 200 images in each scene type, and each image has a size of 600×600 pixels. The types of these images are airport, desert, forest, mountain, park, parking, pond, port, railway station and square. Figure 2 shows some of images from the dataset.



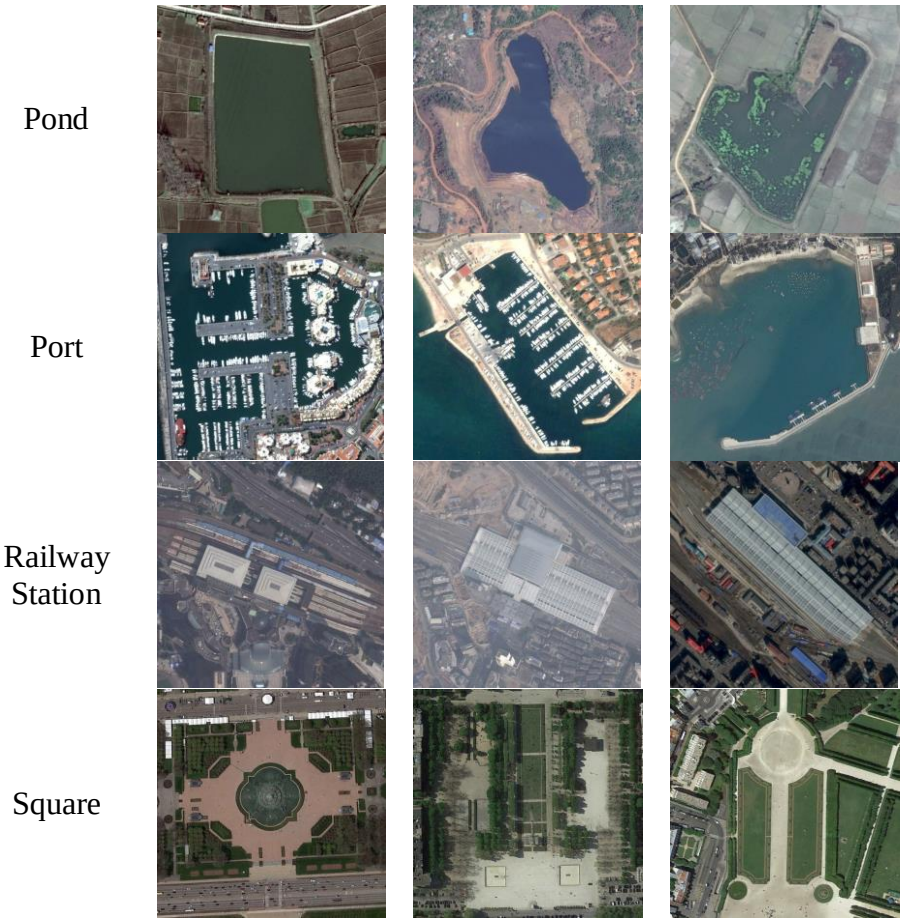


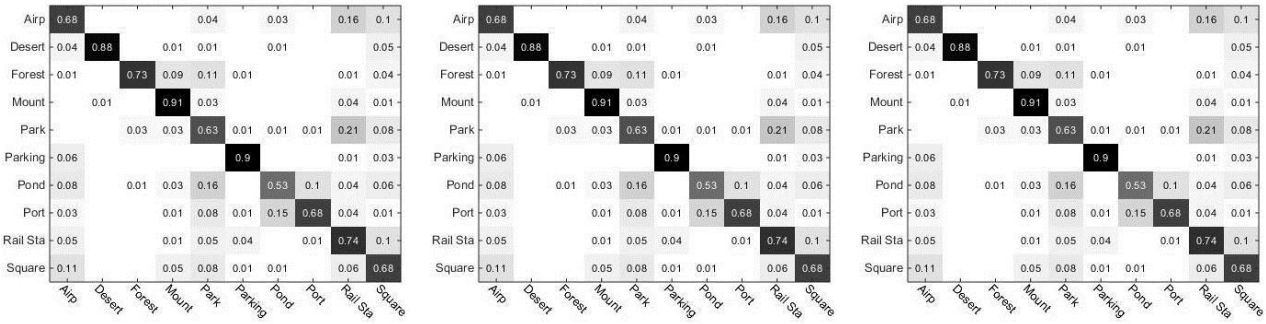
Fig. 2 Images randomly selected form the dataset

In this paper， one group experiments was conducted only by CH, LBP and SIFT features separately, the other group was conducted by adding Itti model to these three features separately, and the last group experiments was carried out by adding GBVS model to these features separately. The dataset was divided into training dataset and test dataset. The ratio of the training dataset in total dataset is 0.2, 0.4, 0.6 and 0.8 respectively. So, there are 36 results of the classification experiments. The accuracies of the experiments were shown in table 1.

Table 1. Accuracy results of the experiments (%)

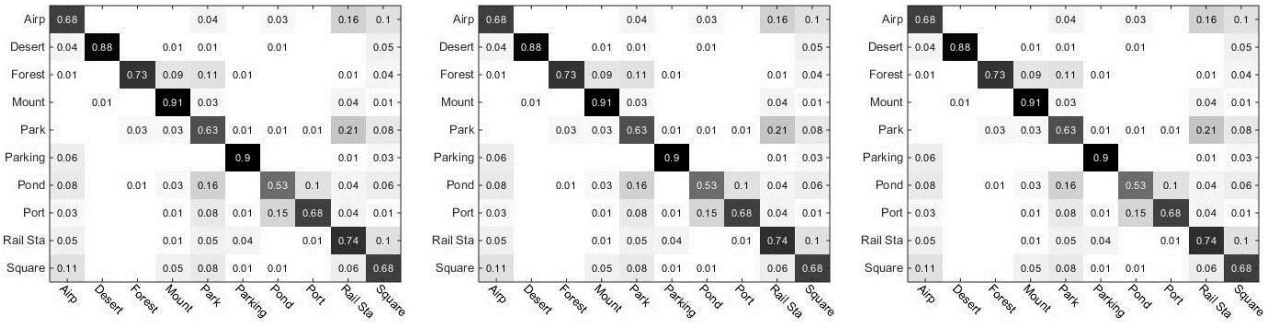
Model \ Training ratio		0.2	0.4	0.6	0.8
CH features	basic	69.125	72	73.25	75.5
	Itti	71.125	74.4167	75.375	76.75
	GBVS	71.125	74.3333	75.125	77
LBP features	basic	75.5	81.5833	82.125	83.25
	Itti	76.4375	81.8333	82.375	84
	GBVS	75.75	81.8333	82.5	85
SIFT features	basic	78.9375	85.5	87.375	87
	Itti	80.375	86.1667	87.5	87.5
	GBVS	80.75	86.8333	87.5	88.25

Considering the length of the paper, only a few confusion matrices are given out, when the condition is that the ratio of training dataset in total dataset is equal to 0.6.



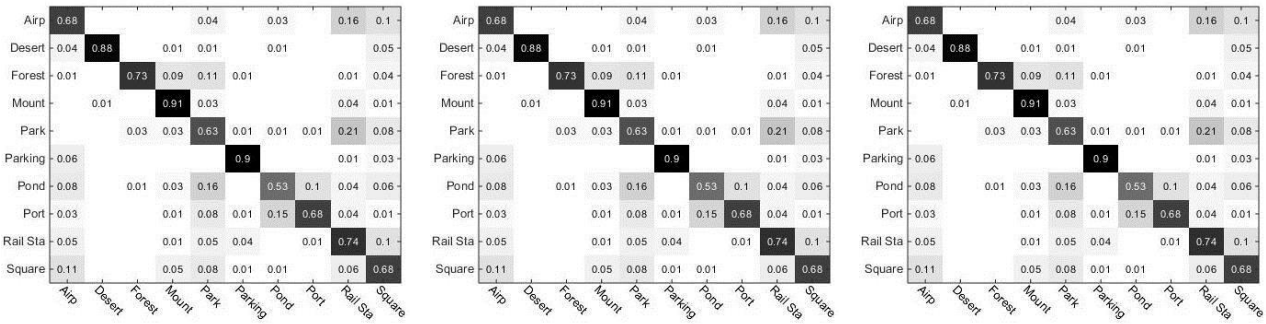
(a)Basic CH BoVW model (b)CH+Ittii BoVW model (c)CH+GBVS BoVW model

Fig. 3 Confusion matrix obtained by CH/ CH+Ittii/ CH+GBVS BoVW model



(a)Basic LBP BoVW model (b)LBP+Ittii BoVW model (c)LBP+GBVS BoVW model

Fig. 4 Confusion matrix obtained by LBP/ LBP+Ittii/ LBP+GBVS BoVW model



(a)Basic SIFT BoVW model (b)SIFT+Ittii BoVW model (c)SIFT+GBVS BoVW model

Fig. 5 Confusion matrix obtained by SIFT/ SIFT +Ittii/ SIFT +GBVS BoVW model

It is obvious that, the salience-driven BoVW models show better performances in remote sensing scene classification, in spite of the salience-driven method is by Itti model or GBVS model. While there is no significant difference between the two salience-driven BoVW models. Although the overall classification accuracy is improved by saience-driven BoVW model, the classification accuracy is likely to deteriorate for a particular category, and the details can be seen from the confusion matrices. For example, in Figure 3 (a) and (c), the self-recognition rate of railway station category is reduced from 0.74 to 0.71 when GBVS salience-driven is integrated. The result also shows that SIFT feature is the best choice in classification, while CH feature is the worst choice. And the reason why CH feature shows worst maybe causing by the reflectivity of the remote sensing images. Anyway, the overall classification accuracy is improved.

5. Conclusion

In order to improve the accuracy of remote sensing scene classification, this paper proposes salience-driven BoVW models. Experiments are conducted by 36 types, which consider the different extraction features, the training data ratio, and salience-driven method. The results show that the salience-driven BoVW models can improve the accuracy of remote sensing scene classification obviously than without.

References

[1]. Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology. Vol. 4 (1985) No. 4, p. 219-227.
[2]. Itti L, Koch C, Niebur E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis & Machine Intelligence. Vol. 20 (1998) No. 11, p. 1254-1259.
[3]. Harel J, Koch C, Perona P. Graph-Based Visual Saliency. Proceedings of the 19th International Conference on Neural Information Processing Systems. Cambridge, MA, USA, 2006, p. 545-552.
[4]. Ma R N, Tu X P, Ding J D, et al. To Evaluate Salience Map towards Popping out Visual Objects. Acta Automatica Sinica. Vol. 38 (2012) No. 5, p. 870-876.
[5]. Zhao R, Grosky W I. Narrowing the semantic gap - improved text-based web document retrieval using visual features. IEEE Transactions on Multimedia. Vol. 4 (2001) No. 2, p. 189-200.
[6]. Likas A, Vlassis N, Verbeek J J. The global k-means clustering algorithm. Pattern recognition. Vol. 36 (2003) No. 2, p. 451-461.
[7]. Cortes C, Vapnik V. Support-Vector Networks. Machine Learning. Vol. 20 (1995) No. 3, p. 273-297.