Sensor Review

A novel remote sensing image retrieval method based on visual salient point features

Xing Wang, Zhenfeng Shao, Xiran Zhou, Jun Liu,

A novel remote sensing image retrieval method based on visual salient point features

Xing Wang
School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China
Zhenfeng Shao and Xiran Zhou
State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, and
Jun Liu
Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing, China

Abstract
Purpose – This paper aims to present a novel feature design that is able to precisely describe salient objects in images. With the development of space survey, sensor and information acquisition technologies, more complex objects appear in high-resolution remote sensing images. Traditional visual features are no longer precise enough to describe the images.
Design/methodology/approach – A novel remote sensing image retrieval method based on VSP (visual salient point) features is proposed in this paper. A key point detector and descriptor are used to extract the critical features and their descriptors in remote sensing images. A visual attention model is adopted to calculate the saliency map of the images, separating the salient regions from the background in the images. The key points in the salient regions are then extracted and defined as VSPs. The VSP features can then be constructed. The similarity between images is measured using the VSP features.
Findings – According to the experiment results, compared with traditional visual features, VSP features are more precise and stable in representing diverse remote sensing images. The proposed method performs better than the traditional methods in image retrieval precision.
Originality/value – This paper presents a novel remote sensing image retrieval method based on VSP features.

Keywords Image retrieval, Image key points, Remote sensing images, Visual attention models

Paper type Research paper

1. Introduction

The development of space survey, sensor and information acquisition technologies has led to higher spatial resolution of remote sensing images and an exponential growth in remote sensing data volume (Li *et al.*, 2012). However, as a result of the limited data processing and analysis capacity, mass remote sensing image data organization and management lag far behind the rising speed of remote sensing image data. Therefore, high-efficiency remote sensing image retrieval is viewed as one major critical solution to the large-scale applications of remote sensing images.

Remote sensing image retrieval is the extension and utilization of content-based image retrieval (CBIR) techniques in the fields of remote sensing applications (Wang and Song, 2013). In the CBIR field, most studies focused on building visual features, which can be included into spectral (colour) features, texture features, shape features and salient point features (Datta *et al.*, 2008). Owing to global image features, traditional colour and texture features are not fine enough to describe complex objects in remote sensing images, especially when comparing with salient point features (Chen *et al.*, 2011; Xu *et al.*, 2010). In addition, as a result of the disadvantages in spatial complexity, the traditional shape features can only be applied for simple and easy-recognition natural image retrieval (Faloutsos *et al.*, 1994). Therefore, we focus on salient point features in this paper.

As a typical salient point feature, as being invariant to image rotation, scale range and illumination variance, the scale-invariant feature transform (SIFT) feature is used largely in CBIR. Smith and Harvey (2011) proposed an approach to set noisy document retrieval through SIFT features. Their works proved that the SIFT-based approach outperformed traditional text-based approaches for noisy text. Lee *et al.* (2012) developed Pill-ID, an automatic retrieval system for drug pill images based on imprint, colour and shape information, encoded as a SIFT and multi-scale local binary pattern feature vector, a three-dimensional histogram and

The current issue and full text archive of this journal is available at **www.emeraldinsight.com/0260-2288.htm**

Sensor Review
34/4 (2014) 349–359

[DOI 10.1108/SR-03-2013-640]

The authors thank the anonymous reviewers for their comments and suggestions. This research is supported in part by the National Basic Research Program of China (No. 2010CB731800); National Science and Technology Specific Projects (No. 2012YQ16018505 and No. 2013BAH42F03); National Natural Science Foundation of China (No. 61172174); Program for New Century Excellent Talents in University (No. NCET-12-0426); the Fundamental Research Fund for the Central Universities (No. 201121302020008) and Program for Luojia young scholars of Wuhan University.

invariant moments, respectively. Wang and Hong (2012) indicated a two-step progressive trademark retrieval method through global and local features' descriptions based on Zernike moments and SIFT features. Newsam and Yang (2007) built a comparative experiment using SIFT feature full descriptor and quantized descriptor, applied in IKONOS remote sensing image retrieval. From the above research works, it can be concluded that the SIFT feature can provide a good performance in CBIR. However, the land cover classes of remote sensing images are usually diverse and complex. So when we use the SIFT feature in remote sensing image retrieval, it should be noted that only the points located in salient regions or regions of interest (ROIs) are valid to represent the objects of interest. Hence, salient region extraction should be an important preprocessing stage in remote sensing image retrieval.

In another research field, visual attention modelling has been a hot research issue in image processing and analysis, as it can provide effective predictions on human visual focus and visual salient regions' extraction (Borji and Itti, 2013). Bao *et al.* (2011) used a visual attention model to improve image retrieval performance through computing regional saliency and analysing the relation between low-level visual features and regional significance. Liang *et al.* (2010) built their image retrieval method based on salient-SIFT features, which are located in the human ROI and extracted using an Itti visual attention model (Itti *et al.*, 1998a, 1998b) and Canny edge algorithm. Li *et al.* (2011) were inspired by Itti's model and proposed a novel perspective to retrieval partial-duplicate images with contented-based saliency region and interest points. Huang *et al.* (2011) proposed a selective attention-driven model for general object recognition and image understanding. Acharya and Vimala Devi (2011) extracted the image ROI using the Itti and Stentiford models, and built image retrieval based on a colour feature vector and several ROI feature parameters. Wang *et al.* (2011) created airport detection in remote sensing images using a visual attention mechanism. All of the above research works focused mostly on saliency computation or ROI extraction based on the classical Itti model. However, the Itti model cannot provide adequate precision in salient region extraction when the objects take a large portion of images (Marques *et al.*, 2007), which creates an application restriction for this model in remote sensing image analysis. In view of this, we use an alternative model, graph-based visual saliency (GBVS, Harel *et al.*, 2006), which is powerful in predicting human eye fixations on multi-resolution images, to extract salient regions in this paper.

From all above, considering the advantages of SIFT feature and GBVS model in high-resolution remote sensing image analysis, a visual salient point (VSP) feature is proposed for remote sensing image retrieval in this paper. The VSP feature is built based on image key point detection and a visual attention model to represent local salient features in remote sensing images. As a result of filtering out the point features located in background regions, the VSP feature can describe the objects in image more precisely and efficiently. Therefore, the image retrieval results in accord with human visual perception can be achieved through similarity measurements among these features.

2. Image key point detection and description

Image key point in remote sensing image refers to a significant local feature in image representation and analysis. The descriptor extracted from a key point neighbourhood can effectively represent local information. Popular image key point detectors include the Moravec detector, Harris detector, Harris-Laplace detector, affine-invariant Harris-Affine detector, Smallest Univalue Segment Assimilating Nucleus (SUSAN) detector and SIFT algorithm proposed by Lowe in 2004 (Lowe, 2004). With holding invariance in image rotation, scale range and illumination variance, SIFT has found increasingly wider utilization in image analysis. The SIFT algorithm is exploited in this work for image key point detection and description.

There are four major steps in SIFT feature extraction (Lowe, 2004): first, scale-space extreme detection is designed to identify potential interest points that are invariant to scale and orientation; second, accurate key point localization is set to determine location and scale of stable key points, and unstable key points are rejected; third, orientation assignment is carried out to each key point location based on local image gradient directions, for providing invariance to image rotation; and fourth, the key point descriptor is extracted based on local image gradients, which are measured at the selected scale in the region around each key point.

To generate the scale space of an image, image sampling should be exploited during scale-space extreme detection. The Gaussian function has proven to be the only possible scale-space kernel under a variety of reasonable assumptions (Koenderink, 1984; Lindeberg, 1994), and it can be formulated as:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma} e^{-(x^2+y^2)/2\sigma^2} \tag{1}$$

Where σ stands for the Gaussian distribution standard deviation. The scale space of an image, $L(x, y, \sigma)$, can be achieved using the convolution of a variable-scale Gaussian with an input image $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2}$$

The difference-of-Gaussian (DOG) scale space is then exploited to detect stable key point locations in the scale space. The DOG function can be expressed as:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{3}$$

Where k is a constant factor for the difference between two nearby scales. In the DOG scale space, one pixel point will be selected as a local extreme only if it is the local maxima or minima compared with 26 pixels around (eight points in the same scale and nine points in two adjacent scales separately). All of these local extreme points are assembled as a set of candidate key points. The final key points in the image can be achieved by filtering unstable candidate key points. In addition, through orientation assignment for every key point according to neighbourhood gradient directions, SIFT features can be rotation-invariant. Finally, every single

detected key point holds three parameters involving the location, scale and orientation. A corresponding descriptor can be extracted to represent the local feature.

3. Visual attention model

Recently, visual attention/visual salient feature-based image analysis has been a popular research issue (Borji and Itti, 2013). The Itti model (Itti *et al.*, 1998a, 1998b), which was created based on feature integration theory (Treisman and Gelade, 1980), has been one of the most representative models in this field. This model used three feature channels: colour, intensity and orientation. In this model, an input image was subsampled into a Gaussian pyramid first, and each pyramid level was decomposed into channels for red, green, blue, yellow, intensity and local orientations. From these channels, centre-surround "feature maps" for different features were constructed and normalized. Then these feature maps were linearly summed and normalized once more to yield the "conspicuity maps". Finally, conspicuity maps were linearly combined once more to generate the saliency map, which represented the saliency value of each pixel in the image. However, this model works well only when the object appears small, with the relative object size to a maximum of 5 per cent (Marques *et al.*, 2007). In urban high-resolution remote sensing images, object contour appears clearly and often covers a significant proportion of the image area, making object detection imprecise when using the Itti model.

The leading models of visual saliency may be organized into three stages: feature extraction, activation and normalization/combination (Harel *et al.*, 2006). In feature extraction, most classic methods used biologically inspired filters to extract feature maps. In the second stage, activation maps are obtained by subtracting feature maps at different scales. In the last stage, normalization/combination is accomplished in one of three ways:

1. a normalization scheme based on local maxima (Itti *et al.*, 1998a, 1998b);
2. an iterative scheme based on convolution with a DOG filter; and
3. a non-linear interactions approach that divides local feature values using weighted averages of the surrounding values in a way that is modelled to fit psychophysics data (Itti *et al.*, 1998a, 1998b).

Based on the classic visual attention models, Harel *et al.* (2006) proposed a new bottom-up visual saliency model called GBVS. In this model, a scale-space pyramid was first derived from image features: colour, intensity and orientation (similar to the Itti model). Then, a fully connected graph over all grid locations of each feature map was built. Weights between two nodes were assigned proportional to the similarity of feature values and their spatial distance. Assuming M was a feature map of an input image I, the dissimilarity between two positions (p_1, q_1) and (p_2, q_2) in the feature map, with respective feature values $M(p_1, q_1)$ and $M(p_2, q_2)$, was defined as:

$$d((p_1, q_1)\,||\,(p_2, q_2)) = \left|\log \frac{M(p_1, q_1)}{M(p_2, q_2)}\right| \tag{4}$$

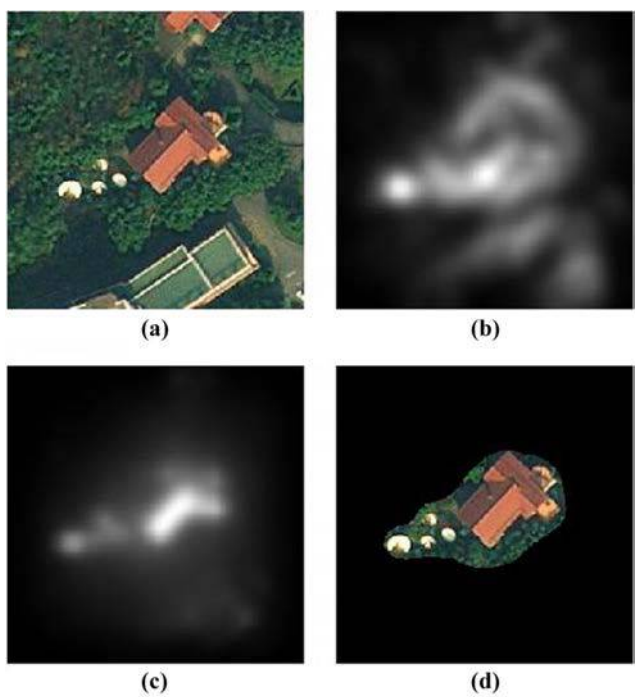
The directed edge from node (p_1, q_1) to node (p_2, q_2) was then assigned a weight proportional to their dissimilarity and their distance on lattice M:

$$\left.\begin{aligned} w((p_1, q_1), (p_2, q_2)) &= d((p_1, q_1)\,||\,(p_2, q_2)) \bullet f(p_1 - p_2, q_1 - q_2) \\ f(a, b) &= \exp\left(-\frac{a^2 + b^2}{2\delta^2}\right) \end{aligned}\right\} \tag{5}$$

Where δ was a free parameter set by experience. The resulting graphs were treated as Markov chains by normalizing the weights of the outbound edges of each node to 1 and by defining an equivalence relation between nodes and states, as well as between edge weights and transition probabilities. Their equilibrium distribution was adopted as the activation and saliency maps. In the equilibrium distribution, nodes that were highly dissimilar to surrounding nodes would be assigned large values. The activation maps were finally normalized to emphasize conspicuous detail, and then combined into a single overall saliency map.

To clarify the efficiency of the two visual attention models, saliency maps from Itti and GBVS are compared in Figure 1. In comparison between Figure 1(b) and 1(c), it is obvious that GBVS holds a significant advantage in structure information extraction of salient objects. As shown in Figure 1(d), the salient region of the image can be achieved from GBVS with a saliency threshold of 85 per cent. Therefore, we adopt GBVS for computing saliency map in this paper.

Figure 1 Comparison of saliency maps from different visual attention models



Notes: (a) Original image of Worldview-2; (b) saliency map from Itti; (c) saliency map from GBVS; (d) salient region of the image (with a threshold of 85 per cent)

4. VSP feature-based remote sensing image retrieval

4.1 VSP feature extraction

There are three major issues in visual salient feature-based image retrieval: salient region detection, feature descriptor determination and similarity measurement of features. The SIFT algorithm is used first to extract the image key points and their descriptors. Generally, about 2000 stable key points can be extracted from a typical image with size of 500 × 500 (Lowe, 2004). However, not all of these key points are necessary to represent object features in image retrieval. In fact, the key points located in background regions are helpless to represent the objects precisely, but also increase the computation workload in image retrieval. Only the key points located in salient regions or ROIs are valid to represent the objects of interest. Therefore, key points extracted by SIFT should be filtered through image salient regions. GBVS is exploited to generate the saliency map and the saliency threshold is set by experience to detect salient regions in images. These key points in image salient regions are then extracted and defined as VSPs. The SIFT descriptors of these VSPs are normalized and combined into a VSP feature matrix of the image. Given a remote sensing image I, all SIFT feature vectors are normalized to unit length and collected into a set $\{SP_l | l = 1, 2, ..., n\}$, where $SP_l = (H_1, H_2, ..., H_{128})^T$ is a 128-dimensional column vector, n is the quantity of VSPs, then VSP feature matrix of the image can be shown as:

$$F = [SP_1, SP_2, ..., SP_n] \tag{6}$$

VSPs of a Worldview-2 remote sensing image are shown in Figure 2. Comparing Figure 2(b) and 2(d), it is obvious that all VSPs are located around the salient object after filtering key points using the image salient region. Hence, it can be concluded that the VSP feature proposed in this paper holds better salient object representation in remote sensing images.

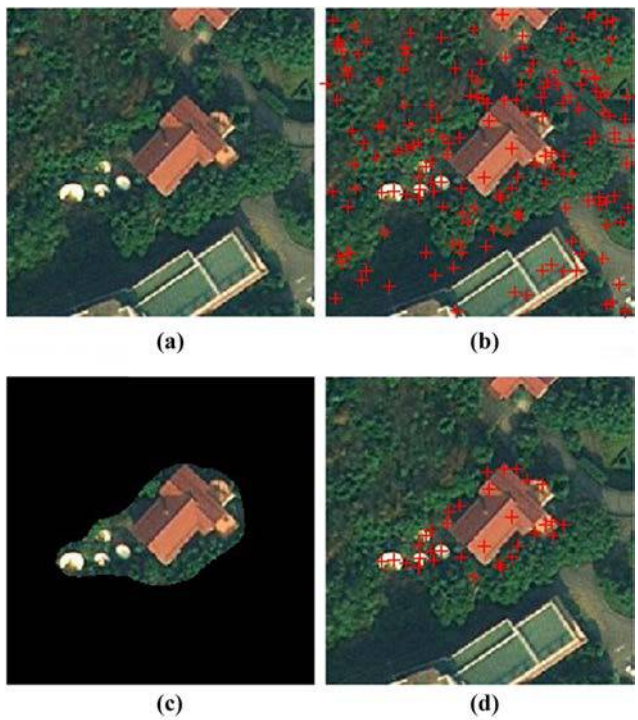
4.2 Similarity measurements of VSP features

To achieve an efficient similarity measurement of VSP features, three measurements are compared in this paper. Assuming the query image is I_q, the VSP feature matrix of I_q is F_q and the quantity of VSPs in I_q is n_q; candidate image is I_c, the VSP feature matrix of I_c is F_c and the quantity of VSPs in I_c is n_c. According to research operated by Lowe (2004), VSP feature matching can be accomplished using the nearest neighbour distance ratio. In general, the more matching point pairs two images share, the more similar they are. Therefore, similarity between I_q and I_c can be formulated as:

$$S_1 = n'/n_q \tag{7}$$

Where n' is the quantity of matching VSP pairs between I_q and I_c. In addition, similarity between two VSPs can also indicate the similarity of local regions in images. A smaller difference between two VSPs corresponds to a higher similarity of the local regions in two images. Assuming $ed_{i,j}$ is the Euclidean distance between the i-th column vector SP_i in F_q and the j-th column vector SP_j in F_c, min $(ed_{i,j}, j = 1, 2, ..., n_c)$ can be seen as the distance from SP_i to the VSP feature matrix F_c, which takes values between 0 and $\sqrt{2}$ in theory. Therefore, similarity between I_q (F_q) and I_c (F_c) can also be expressed as:

Figure 2 Distribution of SIFT key points and visual salient points in image



Notes: (a) Original image of Worldview-2; (b) distribution of SIFT key points; (c) salient region of the image; (d) distribution of VSPs

$$S_2 = 1 - \frac{\sqrt{2}}{2n_q}\sum_{i=1}^{n_q} \min\,(ed_{i,j}, j = 1, 2, ..., n_c) \tag{8}$$

Besides that, dissimilarity between VSP feature vectors can also be measured using the discrete Kullback–Leibler distance. The expression is shown as follows:

$$kl(SP_i, SP_j) = \sum_{r=1}^{128} (H_i^{(r)} - H_j^{(r)}) \log \frac{H_i^{(r)}}{H_j^{(r)}} \tag{9}$$

Where $H_i^{(r)}$ and $H_j^{(r)}$ stand for the r-th value in feature vector SP_i and SP_j separately. As equation (9) contains a logarithmic operation, it is a time-consuming process. The first-order approximation distance χ^2 can be exploited to improve the computation efficiency. Its formula is:

$$\chi^2(SP_i, SP_j) = \sum_{r=1}^{128} \frac{(H_i^{(r)} - H_j^{(r)})^2}{H_i^{(r)} + H_j^{(r)}} \tag{10}$$

As a result, the similarity between I_q and I_c can be computed as follows:

$$S_3 = 1 - \frac{1}{2n_q}\sum_{i=1}^{n_q} \min\,(\chi^2(SP_i, SP_j), j = 1, 2, ..., n_c) \tag{11}$$

4.3 Remote sensing image retrieval algorithm based on VSP features

After discussing the solutions for salient region detection, feature descriptor determination and similarity measurement, we use the SIFT algorithm to extract image key points and their descriptors. Meanwhile, the saliency map is accomplished using the GBVS model to separate the salient image regions from the background. Then, the key points located in the salient regions are extracted as VSPs to build VSP feature matrices. Finally, similar images can be sorted using the similarities, which are determined by these VSP feature matrices. The flow chart of the image retrieval algorithm is illustrated in Figure 3. Operations enclosed by the solid and dashed lines compose the feature extraction stages for the query image and remote sensing image database, respectively. The feature extraction for remote sensing image database can be completed offline.

5. Experiment and analysis

Our experimental databases were composed of two Worldview-2 image databases and an aerial image database. The two Worldview-2 image databases were derived from two Worldview-2 remote sensing images, whose multispectral images and panchromatic images had been fused using the ERDAS IMAGINE software previously. So the two Worldview-2 images both had a spatial resolution of 0.5 m. The first image covers a Hangzhou scene with a size of 7,632 × 9,808. The second image covers a Zhengzhou scene with a size of 8,740 × 11,644. The two source images were cut into non-overlapped sub-images with a size of 256 × 256 to build the two image databases, that included 1,170 and 1,610 sub-images, respectively. Thirty typical building and vegetation images each were selected from the Hangzhou image database. Thirty typical road images were selected from the Zhengzhou image database. These 90 images composed the query image set. In the experiments, building and vegetation image retrievals were carried out using the Hangzhou image database, and road image retrievals were carried out using Zhengzhou image database. The retrieval performance was evaluated using the average precision of each query image category and the total average precision.

The aerial image database consisted of images of 21 land-use classes selected from aerial orthoimagery with a pixel resolution of 30 cm (Yang and Newsam, 2010). Large images were downloaded from the USA Geological Survey (USGS) National Map of 20 selected US regions. Hundred images measuring 256 × 256 pixels were manually selected for each of the follow 21 classes: agricultural, airplane, baseball diamond, beach, buildings, chaparral, dense residential, forest, freeway, golf course, harbour, intersection, medium density residential, mobile home park, overpass, parking lot, river, runway, sparse residential, storage tanks and tennis courts. In the experiments, each image of the 21 classes got a retrieval result from the entire aerial image database, then

Figure 3 Flow chart of image retrieval

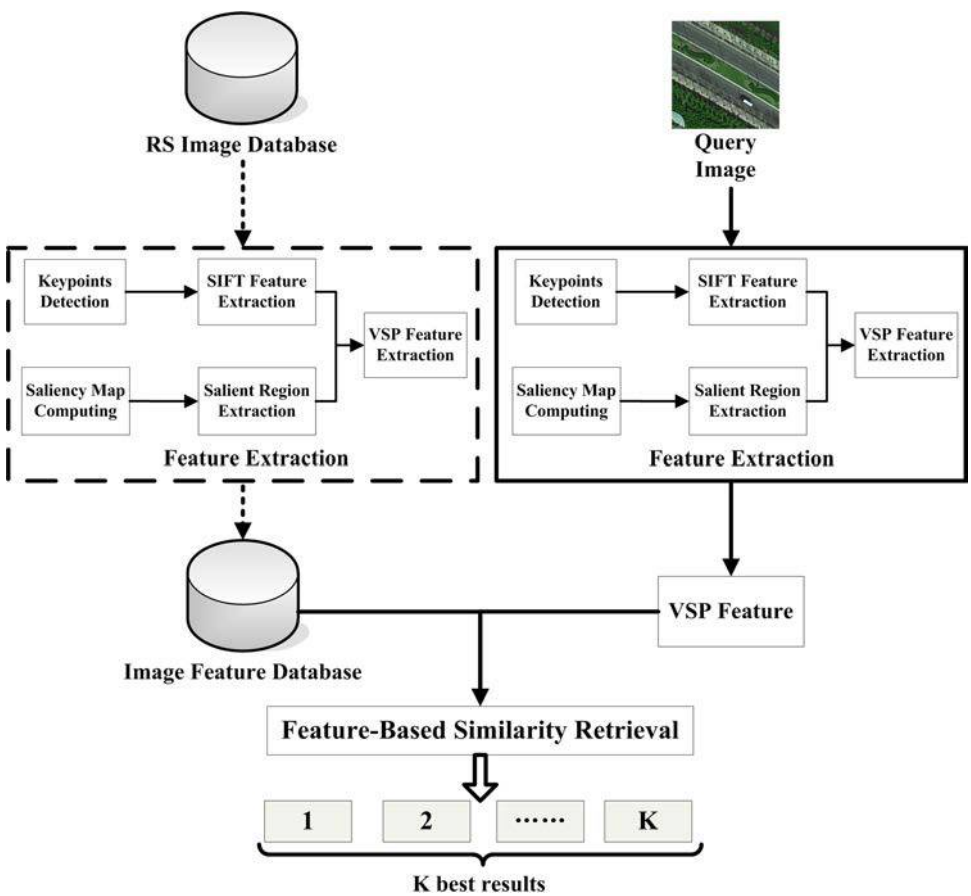


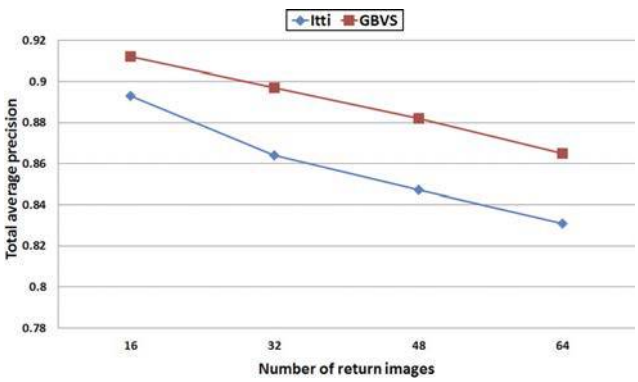
Table I Retrieval performance comparison between Itti and GBVS using Worldview-2 image databases

Query image category	Building				Vegetation				Road			
Number of return images	16	32	48	64	16	32	48	64	16	32	48	64
Average precision (%)												
Itti	**88.5**	85.9	84.4	83.2	98.5	98.2	97.2	95.8	80.8	75.8	72.4	70.4
GBVS	88.3	**86.3**	**85.5**	**83.5**	**99.4**	**99.2**	**98.8**	**98.3**	**85.8**	**83.5**	**80.2**	**77.6**

precision in top 100-ranked returns and recall in top 200-ranked returns of each retrieval result were calculated. The retrieval performance was evaluated using the average precision and recall of all the 21 classes.

To solve the salient region detection, feature descriptor determination and similarity measurement of features, our experiments were divided into three groups. Group 1 extracted the image salient regions using two visual attention models, Itti and GBVS, to compare the retrieval performance under different visual attention models. Group 2 described the images using three kinds of features: VSP features, colour moments and spectral histograms in dual-tree complex wavelet transform (DTCWT) domain, to compare the retrieval performance between the VSP features and traditional features. Group 3 compared the retrieval performance under the three similarity measurements mentioned in Section 4.2.

Figure 4 Total average precision comparisons between Itti and GBVS using Worldview-2 image databases



5.1 Comparison of different visual attention models

In this experiment, the Itti and GBVS models were exploited to extract image salient regions. The saliency threshold was set to 80 per cent, namely, 20 per cent of the image area was viewed as the salient image region.

The retrieval performances of the two visual attention models using the Worldview-2 image databases are illustrated in Table I. Average precisions in top 16-ranked returns, top 32-ranked returns, top 48-ranked returns and top 64-ranked returns for the three query image categories are recorded. The comparisons on total average precision between Itti and GBVS using the Worldview-2 image databases are shown in Figure 4. The retrieval performances of the two visual attention models using the aerial image database are illustrated in Table II.

Table I shows a minor disparity occurs in retrieval performances between the two visual attention models. In general, the GBVS model works a little better than the Itti model. However, when comparing the differences in average precision in the three categories, the GBVS model works significantly better than the Itti model in road image retrieval. This happens for the following reasons:

- Itti model works well in small salient object extraction, but exhibits inadequate precision when facing large image salient regions.
- The object distributions in building images and vegetation images are relatively homogeneous; causing the salient region detection affects less in the image feature extraction.
- In road images, objects cover a large proportion and are irregularly distributed. In addition, the contrast between objects and the background (green field, lake, i.e.) is remarkable. Therefore, the salient region detection has a great influence on the image feature extraction.

Table II Retrieval performance comparison between Itti and GBVS using aerial image database

	Precision (%)		Recall (%)			Precision (%)		Recall (%)	
Class	**Itti**	**GBVS**	**Itti**	**GBVS**	**Class**	**Itti**	**GBVS**	**Itti**	**GBVS**
Agricultural	**24.5**	20.3	**35.9**	30.2	Intersection	6.9	**34.5**	11.4	**47.5**
Airplane	3.9	**5.6**	5.4	**8.1**	Medium residential	3.5	**5.1**	8.6	**14.6**
Baseball diamond	**19.7**	19.7	28.0	**29.7**	Mobile home park	15.1	**21.9**	28.0	**36.5**
Beach	26.6	**30.9**	39.4	**47.8**	Overpass	7.6	**20.4**	13.8	**34.9**
Buildings	6.8	**8.7**	11.7	**15.0**	Parking lot	31.7	**41.1**	64.1	**67.5**
Chaparral	88.2	**88.8**	96.3	**96.5**	River	12.8	**18.9**	21.0	**31.2**
Dense residential	5.6	**10.4**	10.9	**17.4**	Runway	25.8	**31.9**	37.4	**46.8**
Forest	20.1	**23.3**	77.2	**80.2**	Sparse residential	2.0	**2.5**	4.2	**5.5**
Freeway	37.7	**51.8**	53.2	**71.1**	Storage tanks	2.8	**3.5**	4.1	**5.5**
Golf course	3.9	**8.5**	6.6	**12.8**	Tennis court	6.4	**11.6**	11.4	**19.8**
Harbour	48.4	**64.9**	65.7	**78.9**	**Overall**	19.03	**24.96**	30.20	**37.98**

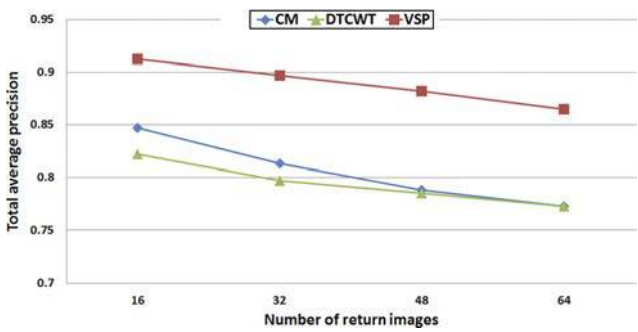
Table III Retrieval performance comparison of different image features using Worldview-2 image databases

Query image category	Building				Vegetation				Road			
Number of return images	16	32	48	64	16	32	48	64	16	32	48	64
Average precision (%)												
CM	**92.7**	90.0	89.2	88.5	82.3	77.5	74.0	72.7	79.2	76.5	73.1	70.5
DTCWT	92.5	**90.5**	**90.1**	**89.0**	91.7	90.5	89.7	88.7	62.3	57.9	55.9	54.3
VSP	88.3	86.3	85.5	83.5	**99.4**	**99.2**	**98.8**	**98.3**	**85.8**	**83.5**	**80.2**	**77.6**

As shown in Figure 4, the two models both appear to decline in total average precision when return images increase, but the GBVS model generally performs better than the Itti model.

Table II also shows an overwhelming superiority for the GBVS model in contrast with the Itti model in retrieval performance. When comparing the retrieval performance in precision, the GBVS model performs better on almost all the classes, except agricultural and baseball diamond. In a more detail comparison, it can be found that the GBVS model performs significantly better than the Itti model on the freeway, harbour, intersection, overpass and parking lot classes. It is consistent with the performance comparison in road image retrieval using the Worldview-2 image databases in Table I, and also for the same reasons. When comparing the retrieval performances in recall, the superiority of the Itti model is only on the agricultural classes.

Figure 5 Total average precision comparisons of different image features using Worldview-2 image databases



For all above performance comparisons in this experiment, it can be concluded that the GBVS model is more appropriate for salient region extraction in remote sensing image retrieval.

5.2 Comparison of different image features

In this experiment, colour moments (CM), spectral histograms of DTCWT (DTCWT) and VSP features are exploited to represent image features. The retrieval performances of the three kinds of features using the Worldview-2 image databases are illustrated in Table III. The average precisions in top 16-ranked returns, top 32-ranked returns, top 48-ranked returns and top 64-ranked returns for the three query image categories are recorded. The comparisons on total average precision of the three kinds of features using the Worldview-2 image databases are shown in Figure 5. The retrieval performances of the three kinds of features using the aerial image database are illustrated in Table IV.

According to Table III, CM and DTCWT perform better than VSP features in building image retrieval. As some VSPs are located in building shadow regions, whose features are similar to those ones located in the shadow regions of vegetation, some vegetation images unexpectedly appear in the retrieval results, causing the average precision to decrease. However, in vegetation image retrieval, the VSP features work perfectly with the average precision staying above 98 per cent, which is significantly better than that of other features.

As shown in Figure 5, the three kinds of features all appear to decline in total average precision when return images increase, but the VSP feature always performs the best of all.

According to Table IV, the VSP feature performs better than the other features on the baseball diamond, beach,

Table IV Retrieval performance comparison of different image features using aerial image database

	Precision (%)			Recall (%)				Precision (%)			Recall (%)		
Class	**CM**	**DTCWT**	**VSP**	**CM**	**DTCWT**	**VSP**	**Class**	**CM**	**DTCWT**	**VSP**	**CM**	**DTCWT**	**VSP**
Agricultural	**24.6**	21.1	20.6	35.9	**38.9**	30.2	Intersection	10.1	15.7	**34.5**	17.1	24.3	**47.5**
Airplane	13.2	**25.3**	5.6	22.2	**33.7**	8.1	Medium residential	14.6	**24.5**	5.1	24.8	**38.9**	14.6
Baseball diamond	11.4	13.3	**19.7**	18.5	20.9	**29.7**	Mobile home park	18.2	**32.1**	21.9	29.0	**52.8**	36.5
Beach	20.2	13.6	**31.0**	31.2	20.6	**47.8**	Overpass	11.3	13.2	**20.4**	17.5	23.0	**34.9**
Buildings	**18.8**	14.4	8.7	**29.0**	22.7	15.0	Parking lot	12.9	25.7	**41.1**	20.6	35.8	**67.5**
Chaparral	61.7	58.8	**88.8**	82.2	90.1	**96.5**	River	14.1	18.4	**18.9**	22.7	27.9	**31.2**
Dense residential	11.1	**19.1**	10.4	18.7	**31.1**	17.4	Runway	10.6	13.9	**31.9**	16.9	17.8	**46.8**
Forest	41.5	**58.7**	23.3	56.7	79.9	**80.2**	Sparse residential	9.2	**17.5**	2.5	14.9	**28.3**	5.5
Freeway	10.6	10.3	**51.8**	17.0	13.3	**71.0**	Storage tanks	9.2	**18.6**	3.5	14.3	**30.3**	5.5
Golf course	13.9	**17.7**	8.5	21.7	**25.9**	12.8	Tennis court	9.3	**18.0**	11.6	16.0	**29.2**	19.8
Harbour	43.3	32.9	**64.9**	55.4	36.2	**78.9**	**Overall**	18.56	22.99	**24.96**	27.71	34.37	**37.98**

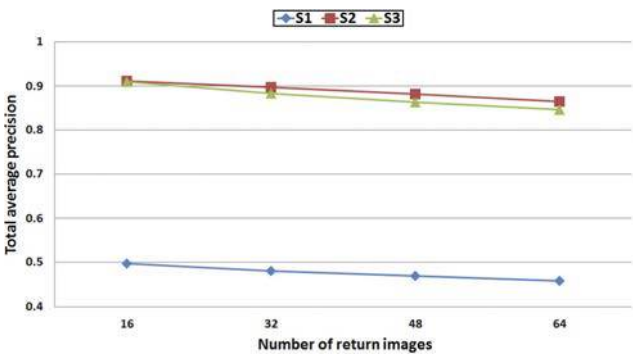
Table V Retrieval performance comparison of different similarity measurements using Worldview-2 image databases

Query image category	Building				Vegetation				Road			
Number of return images	16	32	48	64	16	32	48	64	16	32	48	64
Average precision(%)												
S_1	81.5	80.9	78.2	77.6	14.0	10.8	9.4	9.3	53.8	52.2	51.7	50.1
S_2	88.3	86.3	85.5	83.5	**99.4**	**99.2**	**98.8**	**98.3**	**85.8**	**83.5**	**80.2**	**77.6**
S_3	**90.6**	**87.6**	**86.0**	**84.4**	98.5	98.0	97.4	96.8	84.0	79.3	75.4	72.5

chaparral, freeway, harbour, intersection, overpass, parking lot, river and runway classes. In a more detail comparison, it can be found that the VSP feature holds a remarkable superiority on the chaparral, freeway, harbour, parking lot and runway classes. However, the VSP feature performs poorly on the airplane, building, dense residential, golf course, medium density residential, sparse residential and storage tanks classes. In addition, it is interesting that the VSP feature holds a precision of 23.3 per cent and a recall of 80.2 per cent on the forest class, which means a lot of correct returns are ranked between 100 and 200. Finally, when comparing the retrieval performances over all the 21 classes, the VSP feature is still the best of the three kinds of features.

For all above performance comparisons in this experiment, it can be summarized that the VSP feature is the best of the three kinds of features for remote sensing image retrieval.

Figure 6 Total average precision comparisons of different similarity measurements using Worldview-2 image databases



5.3 Comparison of different similarity measurements

To achieve an efficient similarity measurement of VSP features, three similarity measurements mentioned in Section 4.2 are exploited in this experiment. The retrieval performances of the three similarity measurements using the Worldview-2 image databases are illustrated in Table V. The average precisions in top 16-ranked returns, top 32-ranked returns, top 48-ranked returns and top 64-ranked returns for the three query image categories are recorded. The comparisons on total average precision of the three similarity measurements using the Worldview-2 image databases are shown in Figure 6. The retrieval performances of the three similarity measurements using the aerial image databases are illustrated in Table VI.

According to Table V, S_1 performs worst in this experiment. As the amount of VSPs in a single image is usually less than 100, the number of images in the two remote sensing image databases is more than 1,000; it is obvious that the discrimination of S_1 is not fine enough to distinguish similar images from so many dissimilar images. In addition, S_3 works a little better than S_2 in building image retrieval, and S_2 works a little better than S_3 in vegetation and road image retrieval.

As shown in Figure 6, the three similarity measurements all appear to decline in total average precision when the number of return images increases; however, S_2 and S_3 always outperform S_1 , and S_2 always performs the best of all.

According to Table VI, S_1 performs best on the airplane, baseball diamond, building, golf course, sparse residential and storage tanks classes; S_2 performs best on the agricultural and chaparral classes; and S_3 performs best on all the rest 13 classes. When comparing the retrieval performance over all the

Table VI Retrieval performance comparison of different similarity measurements using aerial image database

Class	Precision (%)			Recall (%)			Class	Precision (%)			Recall (%)		
	S_1	S_2	S_3	S_1	S_2	S_3		S_1	S_2	S_3	S_1	S_2	S_3
Agricultural	11.0	**20.6**	20.3	26.2	**31.0**	30.2	Intersection	6.4	28.0	**34.5**	12.9	40.4	**47.5**
Airplane	**31.0**	4.4	5.6	**43.9**	6.2	8.1	Medium residential	4.4	2.7	**5.1**	7.7	7.3	**14.6**
Baseball diamond	**25.2**	19.9	19.7	**40.4**	28.3	29.7	Mobile home park	6.5	13.3	**21.9**	13.1	26.0	**36.5**
Beach	10.9	25.4	**30.9**	22.3	39.7	**47.8**	Overpass	15.5	17.3	**20.4**	26.6	29.6	**34.9**
Buildings	**16.5**	7.8	8.7	**28.5**	13.3	15.0	Parking lot	7.4	31.8	**41.1**	12.5	62.9	**67.5**
Chaparral	1.2	**88.9**	88.8	2.6	**96.9**	96.5	River	6.2	14.5	**18.9**	10.4	24.8	**31.2**
Dense residential	6.4	6.7	**10.4**	13.0	12.5	**17.4**	Runway	15.2	30.1	**31.9**	27.1	43.8	**46.8**
Forest	2.5	18.6	**23.3**	5.3	76.5	**80.2**	Sparse residential	**4.1**	1.6	2.5	**9.5**	3.6	5.5
Freeway	12.0	47.9	**51.8**	22.6	66.6	**71.1**	Storage tanks	**10.5**	2.9	3.5	**17.0**	4.4	5.5
Golf course	**14.1**	6.5	8.5	**23.3**	10.1	12.8	Tennis court	8.4	9.5	**11.6**	14.0	15.8	**19.8**
Harbour	28.2	46.2	**64.9**	42.6	64.0	**78.9**	**Overall**	11.60	21.16	**24.96**	20.06	33.50	**37.98**

Figure 7 Retrieval results by VSP features in the Worldview-2 image databases

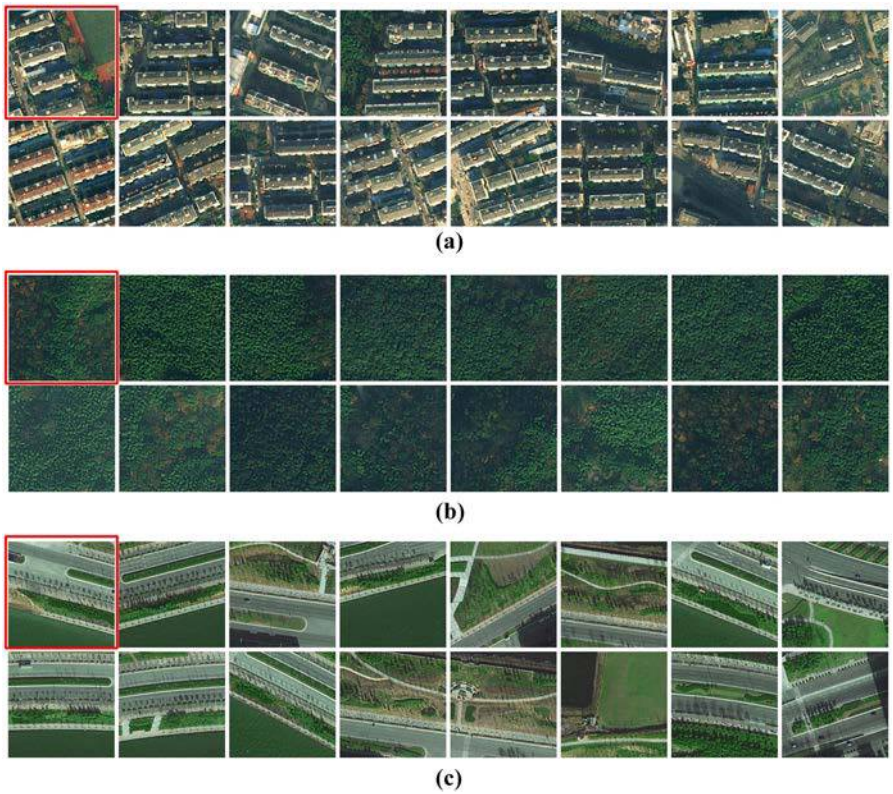
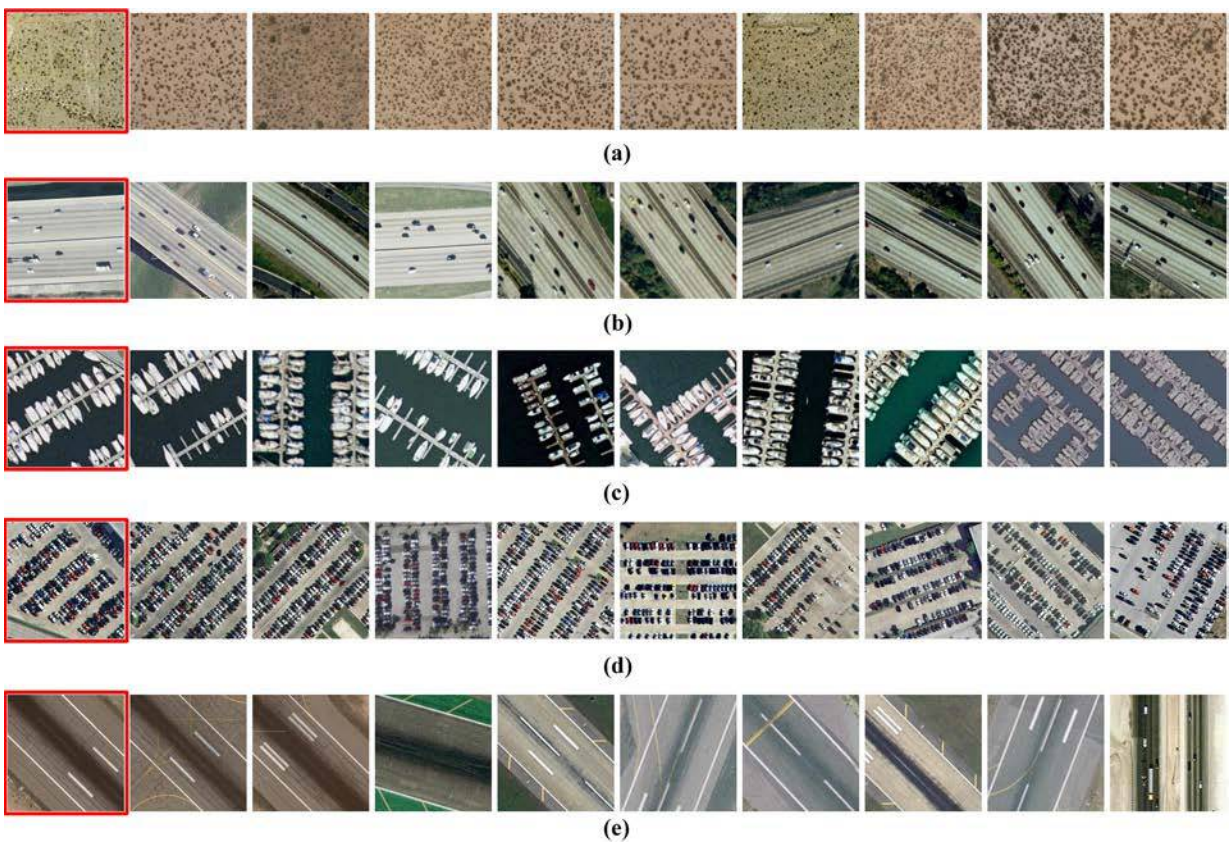


Figure 8 Retrieval results by VSP features in the aerial image database



21 classes, it can be found that S_1 performs worst in general, and S_3 performs a little better than S_2 on most classes.

For all above performance comparisons in this experiment, it can be concluded that S_2 and S_3 can both perform well on similarity measurement of the VSP features in remote sensing image retrieval, and which one of them performs the best depends on the image database.

VSP feature-based retrieval results for three typical images in the two Worldview-2 image databases are shown in Figure 7. As the images enclosed by red boxes show, the three query images are a building image, a vegetation image and a road image separately. The top 16-ranked returns for the three queries were recorded and the precisions are 100 per cent, 100 per cent and 93.75 per cent, respectively. VSP feature-based retrieval results for five typical images in the aerial image databases are shown in Figure 8. As the images enclosed by red boxes show, the classes of the five query images are chaparral, freeway, harbour, parking lot and runway, respectively. The top 10-ranked returns for the five queries were recorded. All the returns of the five queries are correct, except the tenth one (a freeway image) in the runway image retrieval.

6. Conclusion

This paper proposed a novel remote sensing image retrieval method based on VSP features. The three major issues in remote sensing image retrieval using visual salient features were proposed and discussed in depth. A VSP feature was defined based on image key point detection and a visual attention model to represent salient objects in remote sensing images. Because of filtering out many point features located in image background regions, the VSP feature could describe the salient objects more precisely and efficiently. A remote sensing image retrieval algorithm based on VSP features was designed. Two Worldview-2 image databases and an aerial image database were used to evaluate the retrieval performance of the algorithm. In the experiments, different factors in retrieval performance were analysed and a comparison of retrieval performances between the VSP features and traditional features was built. The experimental results showed that the proposed algorithm performs well in salient region detection, feature descriptor determination and similarity measurement in remote sensing image retrieval. The VSP feature-based algorithm outperformed the traditional methods in retrieval precision. Our further research focuses on dimension reduction for the VSP feature and efficiency optimization of the retrieval algorithm while maintaining good retrieval performance.

References

Acharya, S. and Vimala Devi, M.R. (2011), "Image retrieval based on visual attention model", *Proceedings of International Conference on Communication Technology and System Design*, *Coimbatore*, Elsevier, Oxford, pp. 542-545.

Bao, H., Feng, S.H., Xu, D. and Liu, S.Y. (2011), "A novel saliency-based graph learning framework with application to CBIR", IEICE Transactions on Information and Systems, Vol. E94-D No. 6, pp. 1353-1356.

Borji, A. and Itti, L. (2013), "State-of-the-art in visual attention modeling", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35 No. 1, pp. 185-207.

Chen, L.J., Yang, W., Xu, K. and Xu, T. (2011), "Evaluation of local features for scene classification using VHR satellite images", *Proceedings of Joint Urban Remote Sensing Event*, *Munich*, IEEE Computer Society, Piscataway, NJ, pp. 385-388.

Datta, R., Joshi, D., Li, J. and Wang, J.Z. (2008), "Image retrieval: ideas, influences, and trends of the new age", *ACM Computing Surveys*, Vol. 40 No. 2, pp. 51-60.

Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D. and Equitz, W. (1994), "Efficient and effective querying by image content", *Journal of Intelligent Information Systems*, Vol. 3 Nos 3/4, pp. 231-262.

Harel, J., Koch, C. and Perona, P. (2006), "Graph-based visual saliency", paper presented at the *Advances in Neural Information Processing Systems (NIPS)*, *Vancouver*, *4-7 December*, available at: http://books.nips.cc/papers/files/nips19/NIPS2006_0897.pdf (accessed 16 October 2013).

Huang, T.J., Tian, Y.H., Li, J. and Yu, H.N. (2011), "Salient region detection and segmentation for general object recognition and image understanding", *Science China Information Sciences*, Vol. 54 No. 12, pp. 2461-2470.

Itti, L., Braun, J., Lee, D.K. and Koch, C. (1998a), "Attentional modulation of human pattern discrimination psychophysics reproduced by a quantitative model", paper presented at *The Advances in Neural Information Processing Systems (NIPS)*, *Denver*, *30 November-5 December*, available at: http://books.nips.cc/papers/files/nips11/0789.pdf (accessed 16 October 2013).

Itti, L., Koch, C. and Niebur, E. (1998b), "A model of saliency-based visual attention for rapid scene analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20 No. 11, pp. 1254-1259.

Koenderink, J.J. (1984), "The structure of images", *Biological Cybernetics*, Vol. 50 No. 5, pp. 363-370.

Lee, Y.B., Park, U., Jain, A.K. and Lee, S.W. (2012), "Pill-ID: matching and retrieval of drug pill images", *Pattern Recognition Letters*, Vol. 33 No. 7, pp. 904-910.

Li, D.R., Tong, Q.X., Li, R.X., Gong, J.Y. and Zhang, L.P. (2012), "Current issues in high-resolution earth observation technology", *Science China-Earth Sciences*, Vol. 55 No. 7, pp. 1043-1051.

Li, L., Wu, Z.P., Zha, Z.J., Jiang, S.Q. and Huang, Q.M. (2011), "Matching content-based saliency regions for partial-duplicate image retrieval", *Proceedings of IEEE International Conference on Multimedia and Expo*, *Barcelona*, IEEE, New York, NY.

Liang, Z., Fu, H., Chi, Z. and Feng, D. (2010), "Salient-SIFT for Image Retrieval", *Proceedings of 12th International Conference on Advanced Concepts for Intelligent Vision Systems in Sydney*, *Springer Verlag*, Heidelberg, pp. 62-71.

Lindeberg, T. (1994), "Scale-space theory: a basic tool for analysing structures at different scales", *Journal of Applied Statistics*, Vol. 21 No. 2, pp. 225-270.

Lowe, D.G. (2004), "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, Vol. 60 No. 2, pp. 91-110.

Marques, O., Mayron, L.M., Borba, G.B. and Gamba, H.R. (2007), "An attention-driven model for grouping similar images with image retrieval applications", *EURASIP Journal on Advances in Signal Processing*, Vol. 2007.

Newsam, S. and Yang, Y. (2007), "Geographic image retrieval using interest point descriptors", *Proceedings of 3rd International Symposium on Visual Computing*, *Lake Tahoe*, Springer Verlag, Heidelberg, pp. 275-286.

Smith, D. and Harvey, R. (2011), "Document retrieval using SIFT image features", *Journal of Universal Computer Science*, Vol. 17 No. 1, pp. 3-15.

Treisman, A.M. and Gelade, G. (1980), "A feature-integration theory of attention", *Cognitive Psychology*, Vol. 12 No. 1, pp. 97-136.

Wang, M. and Song, T.Y. (2013), "Remote sensing image retrieval by scene semantic matching", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 51, No. 5, pp. 2874-2886.

Wang, X., Wang, B. and Zhang, L.M. (2011), "Airport detection in remote sensing images based on visual attention", *Proceedings of 2011 International Conference on Neural Information Processing in Shanghai*, Springer Verlag, Heidelberg, pp. 475-484.

Wang, Z.H. and Hong, K. (2012), "A novel approach for trademark image retrieval by combining global features and local features", *Journal of Computational Information Systems*, Vol. 8 No. 4, pp. 1633-1640.

Xu, S., Fang, T., Li, D.R. and Wang, S.W. (2010), "Object classification of aerial images with bag-of-visual words", *IEEE Geoscience and Remote Sensing Letters*, Vol. 7 No. 2, pp. 366-370.

Yang, Y. and Newsam, S. (2010), "Bag-of-visual-words and spatial extensions for land-use classification", *Proceedings of ACM International Conference on Advances in Geographic Information Systems*, *San Jose*, Association for Computing Machinery, New York, NY, pp. 270-279.

Corresponding author

Zhenfeng Shao is the corresponding author and can be contacted at: shaozhenfeng@whu.edu.cn