Dense v.s. Sparse: A Comparative Study of Sampling Analysis in Scene Classification of High-Resolution Remote Sensing Imagery

Jingwen Hu\textsuperscript{1,2}, Gui-Song Xia\textsuperscript{1}, Fan Hu\textsuperscript{1,2}, Liangpei Zhang\textsuperscript{1}

\textsuperscript{1} LIESMARS, Wuhan University, Wuhan, China
\textsuperscript{2} SPL, EIS, Wuhan University, Wuhan, China
\{jingwenhu, quisong.xia, fanhu, zlp62\}@whu.edu.cn

Abstract

Scene classification is a key problem in the interpretation of high-resolution remote sensing imagery. Many state-of-the-art methods, e.g. bag-of-visual-words model and its variants, the topic models as well as deep learning-based approaches, share similar procedures: patch sampling, feature description/learning and classification. Patch sampling is the first and a key procedure which has a great influence on the results. In the literature, many different sampling strategies have been used, e.g. dense sampling, random sampling, keypoint-based sampling and saliency-based sampling, etc. However, it is still not clear which sampling strategy is suitable for the scene classification of high-resolution remote sensing images.

In this paper, we comparatively study the effects of different sampling strategies under the scenario of scene classification of high-resolution remote sensing images. We divide the existing sampling methods into two types: dense sampling and sparse sampling, the later of which includes random sampling, keypoint-based sampling and various saliency-based sampling proposed recently. In order to compare their performances, we rely on a standard bag-of-visual-words model to construct our testing scheme, owing to their simplicity, robustness and efficiency. The experimental results on two commonly used datasets show that dense sampling has the best performance among all the strategies but with high spatial and computational complexity, random sampling gives better or comparable results than other sparse sampling methods, like the sophisticated multi-scale key-point operators and the saliency-based methods which are intensively studied and commonly used recently.

1 Introduction

Scene classification of high-resolution remote sensing (HRRS) imagery, i.e. classifying HRRS images into categories with different semantic information according to their content, has recently drawn great attention in remote sensing image interpretation \cite{1–9}. It aims to cross the barrier between low-level visual features and high-level semantic information to interpret HRRS imagery, and can be widely used for many applications, such as urban planning, environmental monitoring, etc.

1.1 Problem and motivation

In recent years, tremendous investigations have been devoted to scene classification of HRRS imagery, see e.g. \cite{1–9}. Though, the methods of scene classification involve many different methodologies, such as the bag-of-visual-words models \cite{3, 7, 10–12}, the topic models \cite{4, 13, 14} as well as deep learning-based methods \cite{2, 9}, it is clear that one can summarize them into three main components: patch sampling, feature description/learning and classification, each of which has a great influence on the final performance. In the literature, most works focus on the later two different
components, e.g. designing/learning efficient scene features and classifiers [1–9] while the influence of sampling strategies is underestimated and thus few work has been devoted to the sampling step. However, one should notice following facts:

- **Sampling is an essential step for scene classification of HRRS imagery.** In image analysis, sampling is concerned with the selection of a subset of pixels from an image to estimate characteristics of the whole image. It is essential mainly due to the fact that the volume of image data is often too large to be processed and a small yet representative set of data is demanded. For instance, without sampling, it is intractable or even infeasible to train a bag-of-visual words model or deep learning-based classifier for scene classification [5,9] with a computer of advanced configurations, as the volume of HRRS imagery usually amount to serval gigabytes.

- **Sampling strategies are crucial to the performance of scene classification of HRRS imagery.** The main goal of sampling is to select a representative subset of the whole data. The subsequent procedures directly process the sampled data while leaving out the information of all the unselected ones, thus, how to sample the image data containing the most descriptive and discriminative information for classification has a great influence on the final results. Good samples can guarantee the classification performance and achieve huge improvement on the spatial and computational complexity at the same time, while non-representative samples will reduce the performance a lot. Thus in the scenarios of scene classification of HRRS imagery, it is crucial to pursue a sampling strategy which balances the classification accuracy and the spatial as well as computational complexity. However, it is not clear yet how to choose sampling strategies.

- **The lack of study on sampling strategies to scene classification of HRRS imagery.** It is worth noticing that different sampling strategies have been used by recent scene classification approaches, e.g. dense sampling by [5,12], random sampling by [9] and saliency-based one in [2]. To our knowledge, there is however few comparative study between these strategies.

- **Sampling strategies in natural scenes can not be copied to scene classification of HRRS imagery.** Note that sampling strategies have been better understood in natural images, e.g. some comparisons on sampling strategies have been made in natural image classification [15,16]. However, it can not be copied to the scenarios of HRRS imagery, as there are big differences between them, e.g. in the shooting angles and occlusions. For instance, natural images are mainly taken by cameras with manual focus even auto-focus capabilities and it thus makes the natural images tend to be center-biased, which is obviously not the case of remote sensing images that are often taken overhead.

Thus, it is of great importance and interest to study the influence of the different sampling strategies on the performance of scene classification of HRRS images, so as to give some instructions for later works when they need to choose a sampling strategy.

### 1.2 Objective and contributions

This paper aims to investigate and quantitatively compare different sampling strategies that can be used for scene classification of HRRS images. Especially, for the sake of simplicity, we fix the other three components, *i.e.* **feature learning, and classifier**, of scene classification and mainly test the influence of sampling strategies.

In order to compare different sampling strategies under the same condition, we choose the standard bag-of-visual-words model to construct a unified testing scheme embedded with various sampling strategies. Although the bag-of-visual-words model do not necessarily perform better than other state-of-the-art methods, it is very efficient, robust and easy to implement. Meanwhile, since the sampling step is an independent procedures, the comparison results achieved with bag-of-visual-words model under such a scheme can be extend to other methods, such as topic models and deep learning-based approaches. We perform our experiments on two commonly used datasets on the topic of scene classification in HRRS imagery.
Our work is distinguished by following contributions:

- We study and compare the effects of many different sampling strategies on scene classification performance of HRSS imagery, the results of which are instructive for following works on scene classification of HRSS imagery. In contrast with previous works which choose a certain sampling strategy but concentrate on developing discriminative feature descriptions or classifiers, our work is the first time to comparatively study different sampling strategies, and the result can be used to improve the performance of previous works.

- We have intensively studied the performances of various saliency-based sampling strategies, that are recently proposed and highlighted to be useful for classification, on different kinds of scenes. Our experiments shows that saliency-based sampling is mainly helpful for object-centered scenes but does not work for scenes with complex textures and geometries.

The rest of the paper is arranged as follows. In section 2, we briefly introduced the related works on different sampling strategies. In section 3, we describe our method for comparing different sampling strategies in detail. The experiment results are shown and analyzed in section 4 and the last is the conclusion of our work.

2 Related work

As mentioned before, sampling is a key step for building up an intelligent system of image classification or recognition. This section tries to provide a brief review of the tremendous investigations on this topic.

The volume of remote sensing images, especially HRSS images, is often too huge to be processed directly. A sampling step is usually required to select an informative and representative subset of the images for subsequent analysis. In the literature, various sampling strategies have been used in the processing of HRSS images [1–9, 12]. For instance, in order to build up a bag-of-words model to achieve better performance in land-use classification, Yang et al. [12] densely sampled the patches of images and adapted the spatial pyramid match kernel method [17] by incorporating the relative spatial information instead of the absolute one. Recently, deep learning-based methods show superiority on learning descriptive features for HRSS images [5] and many studies have been devoted to this direction, where sampling is a crucial step. For example, Hu et al. [9] proposed to use randomly sampled patches of images for unsupervisedly learning features for scene classification, while Zhang et al. [2] used a context-aware saliency detection model to sample the salient regions and then learned features based on them. Observe that different sampling strategies have been used, but few comparative study of these sampling strategies has been made and how to efficiently use the various sampling strategies in the processing of HRSS images is still not clear.

In contrast with the case of remote sensing images, the sampling strategies have been widely studied in the field of natural image processing. Dense sampling is the most widely used for its simplicity, see e.g. [11,17–19], where patches are uniformly sampled with a step (e.g. 4 pixels) in an image. However, for a common image size of 256 × 256 pixels, there are 4096 patches to be sampled for every 4 pixels, which is still very huge meanwhile high redundant for large scale datasets. Thus, sparse sampling methods have been used alternatively. For instance, with the development of various key-point detectors [20–22], which have shown great abilities in finding interest regions, some researchers have used them to sample the patches [10,23–26] belonging to the interest regions, which are informative and can help a lot in saving the memory space [10]. However, it has been proven that such key-point detectors cannot improve the classification performance [27,28], as they were not designed for classification problems but for image matching applications. Surprisingly, it has been found that random sampling, another sparse sampling method, is often more discriminant than key-point based ones [15,16]. While, as the name implies, random sampling seems to be not robust due to the randomness. Thus, some researchers have used saliency detection as a sparse sampling method recently [29–31], which aims to find the salient regions in accordance with human visual attention mechanism.
An early comparative study of the various sampling strategies in natural image classification has been made in [15], which found that dense sampling gave the best performance by using a bayesian hierarchical model. In addition, Nowak et al. [16] has compared the different sparse sampling methods and come to the conclusion that random sampling outperforms the sophisticated multi-scale interest detectors for large numbers of samples.

There is a lack of comparative study on the sampling strategies for HRRS images as the work in [15, 16] for natural images. Moreover, due to the big difference between HRRS images and natural images mentioned before, one cannot directly borrow the conclusions from natural images to use for HRRS images. In addition, saliency detection [32–35] has recently become a hot topic in image analysis and many works have used them as sparse sampling strategies, e.g. [2, 29–31]. However, how these sampling methods help the classification task and how to use them properly is still a problem. This paper thus attempts to provide a comparative study of the various sampling strategies for the classification of HRRS images.

3 Method for the comparative study

This section describes the method for making the comparative study of different sampling strategies.

3.1 The scheme of comparative study

In order to compare different sampling strategies in the case of scene classification for HRRS images, we use the bag-of-visual-words model as the scene classification approach, for its simplicity, robustness and efficiency. The flowchart of the comparative study scheme is illustrated in Fig. 1, which consists of three main steps:

- **Sampling** : selecting a compact but representative subset of images. Observe that this sampling step is in fact analogy to the sampling of signals, which is often guided by Nyquist-Shannon theorem [36], while in image classification it is more complicated. Given an image, a set of local image patches is selected via different sampling methods (e.g. dense, random, or saliency-based sampling strategies) to form a representative and informative subset of the image. Compared with the original image, the sampled subset of patches is more compact and with less spatial complexity. This step is the core part of our work, which the comparative studies are devoted to.

- **Feature learning** : learning feature descriptions of the images using sampled patches. Each patch in the sampled set is first characterized by certain feature descriptors, such as the vector of concatenated pixels, SIFT [21] or histogram of gradients (HoG) [37] or other texture [38, 39] descriptors. In our work, we use SIFT as the patch descriptor. Based on such patch description, one then can learn a feature representation of the image, e.g. by using bag-of-words model or using recently proposed deep learning method [2, 5, 9]. In our case, we use the bag-of-words model for feature learning. More precisely, we first apply a k-means clustering algorithm to the characterized sampled patches to form a dictionary of visual words, and then quantify the patches against the dictionary to obtain a histogram-like global feature. The resulting global feature vector is finally normalized as a characterization of an image.

- **Classification** : assigning a semantic class label to each image. This step is after finding a feature representation of images and usually relies on some trained classifier, e.g. support vector machine (SVM). As this part is out of our study in this paper, in our experiments, we use the simple linear SVM as our classifier.

As the main goal of this work is to investigate the effects of sampling strategies to the scene classification in HRRS images, we keep the feature learning and classification fixed and only vary the sampling approaches in the whole procedure. While, one should note that the studied sampling strategies and the whole procedure are of general setting and also applicable to other feature learning and classification approaches.
Figure 1: The flowchart of the comparative study scheme. It consists of three main steps. _Sampling_ is used to select a compact but representative subset of images. _Feature learning_ is to learn feature descriptions of the images using the sampled data. _Classification_ is for assigning a semantic class label to each image. Here the _feature learning_ step is based on the bag-of-visual-words model, but note that the studied sampling strategies and the whole procedure are of general setting and are also applicable to other feature learning and classification approaches.

| sampling strategy | method and description |
|-------------------|------------------------|
| dense             | sample local patches using a sliding window |
| random-           | randomly sample local patches [16] |
| keypoint-         | SIFT [21]: compute the DoG response map |
|                   | GBVS [41]: construct a graph from the feature map to compute saliency |
|                   | SR [42]: compute the spectral residual to produce saliency map |
|                   | Itti [40]: integrate color, intensity and orientation features across scale to generate saliency map |
|                   | SEG [33]: use a statistical framework and local feature contrast to compute the saliency measure |
|                   | RCS [34]: compute saliency over random rectangular regions of interest |
|                   | AIM [43]: quantify the Shannon’s self-information as the saliency measure |
| sparse            | FT [32]: eliminate fine texture details and noise in the frequency domain to produce saliency map |
|                   | SUN [44]: compute saliency based on the Bayesian framework using the statistics of natural images |
|                   | PISA [35]: compute saliency using color and structure contrast measures with spatial priors |

### 3.2 Involved sampling strategies

Given an image $f : \Omega \rightarrow \mathbb{R}^d$, with $\Omega = \{0, 1, \cdots, M - 1\} \times \{0, 1, \cdots, N - 1\}$ and $d$ as the number of channels, e.g. $d = 3$ for RGB images, the sampling is finding a subset $S$ of $\Omega$, such that

$$S := \{p \mid p \in \Omega, f(p) \text{ is informative}\}.$$

Observe that $S \subseteq \Omega$ and, especially, $S = \Omega$ corresponds to the so-called _dense sampling_. The content of $S$ depends on the definition of the term _informative_. For instance, if one takes _informative_ image content as that around key local feature points in an image, then it indicates _keypoint-based sampling_, while if one regards the _informative_ as image structures that are salient enough in an image, then it corresponds to the saliency-based sampling.

The sampling strategies adopted in this paper is illustrated in Table 1, and each of them will be reviewed in the following of this section.
3.2.1 Dense sampling

Dense sampling corresponding to $S = \Omega$, also called grid sampling [15, 16], is popular used for its simplicity. To be brief, dense sampling is to select all the possible local patches from the images by a fixed-size sliding window. Theoretically, the patches should be sampled pixel by pixel of the image, meaning that the sliding step size is set to be 1 pixel. However, this will result in heavy spatial complexity. Thus, in reality, the size of the sliding window is often empirically set to be a quarter or half of the side length of the sliding window in practical, which help a lot in reducing spatial complexity and keep the image information.

3.2.2 Sparse sampling

Sparse sampling is in contrast with the dense one, which can be roughly divided into three types: keypoint-based sampling, random sampling and saliency-based sampling.

Keypoint-based sampling  It defines the informative content of images as the part around local key points, based on the fact that keypoint detectors [20–22] is powerful to localize informative local points in images. This sampling method depends a lot on the keypoint detectors it used. In our comparison, we choose the most widely used detector, SIFT detectors [21], owing to its robustness and effectiveness. This algorithm actually computes the Difference-of-Gaussian (DoG) response in scale space and the informative points are those where the DoG responses take local maximum/minimum.

Random sampling  Random sampling is first found to be a good sparse sampling method in [27, 28], where it has been reported that random sampling outperforms the keypoint-based sampling in natural image classification. However, one should note that this is not a big surprise, by observing the fact in compressive sensing/sampling that using random matrices we are able to exactly reconstruct sparse vectors and closely approximate compressible vectors stably with high probability using just random measurements [45]. Here, we test random sampling for scene classification of HRRS imagery. For fast implementation, we generate a uniform random noise image to guide the sampling procedure.

Saliency-based sampling  In this case, the informative part of image is defined as regions which are salient to the other parts of the image. The saliency is inspired from the visual attention mechanism, which is unique to primates that have a remarkable ability to interpret complex scenes in real time. In the process of human visual attention mechanism, the information is selected in a way to reduce the complexity of scene analysis. In the past decade, saliency detection has attracted much attentions [32, 34, 41–44], in order to analyze the salient parts of image scenes quickly and accurately. Some work in scene classification of HRRS images has tried to use saliency-based sampling [2], but it is still not clear how it can help the scene classification in a general setting compared with other strategies.

In the following, we briefly review some state-of-the-art saliency detection methods for guiding our sampling.

- Itti’s model [40]: This model is among the early investigations on saliency computation. It is based on the “feature integration theory”, explaining human visual search strategies [46]. The model first extracts different low-level features, e.g. colors, intensity and orientations, independently on several spatial scales. Then the center-surround differences, normalization and across-scale combination are sequently operated on each feature to generate a conspicuity map, and the final saliency map is the sum of the three normalized conspicuity maps on the hypothesis that different features contribute independently to the saliency map.

- Graph-Based Visual Saliency (GBVS) [41]: It is based on Itti’s model but uses a novel idea from graph theory. More precisely, after low-level features being extracted, a fully-connected directed graph is constructed from the feature map and modeled with a Markov chain. The
activation map can be derived from the equilibrium distribution and pairwise contrast. From the activation map, a new graph is constructed and treated as another Markov chain, aiming at concentrating activation into a few key locations. The saliency values are obtained from an equilibrium distribution over map locations.

- **Spectral Residual (SR) model** [42]: This model defines the saliency measure as the spectral residual in frequency domain. Specifically, given an image, the spectral residual is first computed as the ratio between the amplitude of the Fourier coefficients and its local geometric average. The saliency map is then constructed from the spectral residual as its inverse Fourier transformation. This model is global and simple which is independent of feature extractions or prior knowledge of the objects.

- **Frequency-tuned (FT) model** [32]: It is also computed in the frequency domain. Given an image $f$, the saliency measure uses low-level features, e.g. color and luminance, and is defined as,
  \[ Sa(p) = \| f_{\mu} - f_{\omega_{hc}}(p) \|_2, \]
  where $f_{\mu}$ is the mean feature vector of $f$, $f_{\omega_{hc}}$ is an average image blurred with a Gaussian filter with a cut-off frequency as $\omega_{hc}$. Experimentally, $\omega_{hc}$ is set to be $\frac{2\pi}{11}$ so as to eliminate fine texture details as well as noise and coding artifacts. This model can output full resolution saliency maps with well-defined boundaries and efficiently eliminate false detections arising from texture, noise and blocking artifacts.

- **Attention by Information Maximization (AIM)** [43]: This model is rooted in information theory, where the saliency is determined by quantifying the Shannon’s self-information of each local image patch. It consists of two stages: independent feature extraction and the estimation of the Shannon’s measure of self-information. In the first step, in order to analyze features independently, the independent coefficients corresponding to the contribution of different features are computed by using independent component analysis (ICA). In the second step, the distribution of each basis coefficient across the entire image is estimated, and the Shannon’s measure of self-information is finally computed from the joint distribution.

- **Saliency Using Natural statistics (SUN)** [44]: This method is based on an observation from the aspect of information theory, which implies that the rarer a feature is the more informative it is. The measure of saliency is derived from the statistics of a collection of natural images instead of the particular image being processed. In our work, we use the bottom-up saliency. Let $F$ be a used features, e.g. color or intensity, and $p_F$ be the probability density function of $F$ estimated from a collection of natural images, then the saliency value of a location $p \in \Omega$ is computed as $1/p_F(F_p)$.
- **Salient Segmentation (SEG) [33]** method: It uses a statistical framework and local feature contrast to define the saliency measure. Considering a rectangular window $W$ made up of two disjoint parts, an inner window $K$ (the kernel) and the border $B$, see Fig. 2, one can assume that the points in $K$ are salient while $B$ belongs to the background. Define a random variable $Z$ describing the distribution of pixels in $W$, the saliency measure of a point $p \in K$ is thus defined to be the conditional probability,

$$SA_{seg}(p) = p_Z(Z \in K | F(Z) \in Q_{F(p)}),$$

where $F$ denotes the feature map, which maps every point $p$ to a certain feature $F(p)$, and $Q_{F(p)}$ denotes the bin which contains $F(p)$.

- **Random Center-Surround (RCS) saliency** [34]: This model is based on computing local saliency over random rectangular regions of interest. Given an image $f : \Omega \mapsto R^D$ with $D$ channels, for each channel, $n$ sub-windows are randomly generated with uniform distribution. In the $d$-th channel, the local saliency of a point $p$ is defined as the sum of the absolute differences between the pixel intensity and the mean intensity of the random sub-windows by which it is contained. The global saliency map is then computed by fusing the channel specific saliency maps by a pixel-wise Euclidean norm. Furthermore, normalization, median filtering and histogram equalization are applied to the global saliency map so as to preserve edges, eliminate noise and enhance the contrast.

- **Pixel-wise Image Saliency by Aggregating (PISA) [35]**: It computes image saliency by aggregating pixel-wise complementary appearance contrast measures with spatial priors. The model uses two complementary measures and takes advantage of an efficient edge-aware image representation and filtering technique [47]. Given an image $f$, the initial saliency value $\tilde{S}(p)$ for each pixel $p \in \Omega$ is defined as follow:

$$\tilde{S}(p) = U^c(p) \cdot D^c(p) + U^g(p) \cdot D^g(p),$$

where $U^c(p)$ and $U^g(p)$ are the color and structure contrast respectively, $D^c(p)$ and $D^g(p)$ are their spatial priors. In order to take the spatial coherence into account and remove spurious noises, the cross-based local multipoint filtering (CLMF) technique [47] is used to smooth out $\tilde{S}$ and produce a spatially coherent yet edge-preserving saliency map.

### 3.3 Evaluating the sampling performances

The main goal of this work is to compare different sampling strategies in the context of scene classification of HRRS images, thus we need to quantitatively evaluate their performances. For doing so, we fix the other procedures of the pipeline shown in Fig. 1 but vary the sampling method, and measure the sampling performances by using the classification overall accuracy (OA). It is defined as the number of right predict labels divided by the total number of test images.

In order to achieve the same sampling rate for fair comparisons, we threshold the response maps of different sampling methods. The response maps of different sampling methods on two example images are demonstrated in Fig. 3. SIFT indicates the normalized DoG response map for keypoint-based sampling. For saliency-based sampling, the saliency maps are linearly normalized to be in the ranges of $[0, 255]$ as the response maps. The response map of random sampling is produced from an independent uniform distribution between $[0, 255]$. These maps are of the same size as the treated images, and can be used as masks to guide the sampling procedure: the brighter the region, the more likely it is to be sampled from the image. The response map of dense sampling is not shown here, as it is uniformly all white. For fair comparisons, we set a sampling rate $r$ between $[0, 1]$ as the threshold. For every image, we sort the values in its response map in descending order and select the first $r$ proportion as the sampled points. Therefore, the patches centered on the sampled points are chosen as the representative set. With $r$ changes from 0 to 1, we get an OA curve with respect to $r$ for each sampling method.
Figure 3: The response maps of different sampling methods on two example images. SIFT indicates the normalized DoG response map for keypoint-based sampling. For saliency-based sampling, the saliency maps are linearly normalized to be in the ranges of \([0, 255]\) as the response maps. The response map of random sampling is produced from an independent uniform distribution between \([0, 255]\). These maps are of the same size as the treated images, and can be used as masks to guide the sampling procedure: the brighter the region, the more likely it is to be sampled from the image. The response map of dense sampling is not shown here, as it is uniformly all white.

4 Experimental Results

In this section, we describe the comparative results of different sampling strategies on two main datasets used for scene classification in HRRS images.

4.1 Datasets

Two different high-resolution remote sensing image datasets are used for our experiments:

- **UC-Merced dataset** [12]: It contains 21 scene categories, with 100 samples per class. Fig. 4 shows a few example images representing different scenes that are included in this dataset. The images are of size 256 × 256 pixels with the pixel resolution of 1 foot. This dataset is widely used and has been reported to be challenging as it contains highly overlapping classes (e.g. the dense residential, medium residential and sparse residential), which mainly differ in the density of structures and is difficult to distinguish even for humans.

- **RS19 dataset** [48]: It contains 19 classes of scenes in high-resolution satellite imagery which are exported from Google Earth with various resolutions. Fig. 5 shows some example images of each class. For each class, there are 50 samples and the image size are 600 × 600 pixels. This dataset is very challenging due to the changes in resolution, scale, orientation, and illuminations of the images.
These two datasets are widely used for testing the scene classification of HRRS images, see e.g. [1–9].

Figure 4: UC-Merced dataset: some example images are displayed. It contains 21 scene categories, with 100 samples per class. The images are of size $256 \times 256$ pixels with the pixel resolution of 1 foot.

Figure 5: RS19 dataset: It contains 19 classes of scenes in high-resolution satellite imagery which are exported from Google Earth with various resolutions. For each class, there are 50 samples and the image size are $600 \times 600$ pixels.
4.2 Experimental settings

A representative set of patches are sampled from the dataset using different strategies with different proportions \( r \) from 0.01 to 1 with the step to be 0.01. As we use SIFT descriptor to describe each patch, the patch size is set to be \( 16 \times 16 \) pixels, and the spacing is set to be 4 pixels which can save the memory space for the descriptors without affecting the performance. Then, 10000 patches per class are randomly chosen from the representative set to generate the dictionary with 1000 words for both datasets using k-means clustering algorithm. Thus every image is described as a histogram of 1000 bins representing the frequency of each word. For classification, we employ the libSVM classifier [49]. We randomly select a subset of images from the dataset for training to build the classification model, and the rest is used for testing to measure the performance. This process is repeated 100 times, and the average classification accuracy are demonstrated with \( r \).

With the UC-Merced dataset, we randomly select 80% of the samples from each class to initialize the training set, while for the RS19 dataset, we randomly select 60% for training.

4.3 Results and analysis

4.3.1 Overall testing

Fig. 6 shows the overall classification accuracy using different sampling methods on the UC-Merced dataset and the RS19 dataset. The horizontal axis indicates the sampling ratio \( r \), the proportion of the sampled patches with respect to dense sampling. The vertical axis is the classification accuracy corresponding to different ratios. Thus, the classification accuracy using dense sampling is the point with the ratio to be 1, where the sparse sampling strategies converge. Observe that because of the random choosing of the training set in the SVM classifier, there are slight variations at the point \( r = 1 \), at which the classification curves of different sampling strategies should converge.

As shown in the Fig. 6 (a), all the curves share the same properties: they are increasing with slight shakes within the standard deviation, which implying that the classification accuracy of the dense sampling is higher than all the sparse sampling methods. We can also learn from the results that as the ratio becomes bigger, the slope of the curves become smaller, and some of the curves almost become flat when the ratio is larger than 0.7. This implies that some sparse sampling methods can obtain the similar results as dense sampling when the information contained in the sampled patches approaches to a certain degree. Comparing the sampling strategies on the RS19 dataset in Fig. 6.(b), we can get the same conclusion with the UC-Merced dataset that dense sampling outperforms sparse sampling and random sampling has an evident advantage over other sparse sampling methods from the gradually increasing curves.

From the overall accuracy on the two datasets, we can come to the conclusion that:

- Dense sampling performs best among all, while random sampling is obviously better than the other sparse sampling methods, like keypoint-based sampling and saliency-based sampling when the ratio is low, mainly due to the fact that it can extract balanced land-cover information which can help to improve the performance in scene classification. However, one should note that dense sampling has the high spatial complexity.

- As the sampling ratio become larger, the differences among the sparse sampling methods become smaller, since the information extracted become richer.

However, intuitively, sparse sampling methods may be helpful for some structural scenes. Thus, here we roughly divide the scene in HRRS images into 4 types: single texture scenes, multiple texture scenes, object-based scenes and structural scenes, and we analyze the classification performance on each kind of scenes in what follows.

4.3.2 Testing on single texture scenes

Fig. 7 shows two examples of the single texture scenes, which mainly contains only one kind of textures, like agricultural, meadow, chaparral, forest, mountain, etc. From the curves of agricultural, we can see that most of the sparse sampling methods become flat when the ratio becomes larger...
than 0.2, and the meadow curves keep stable when the ratio is larger than 0.4. The reason is that this type of scene mainly consists of a single texture with large information redundancy, and thus can be easily distinguished using a small part of the land cover features. We can also find that the curve using PISA is a little different from others. This is because PISA highlights the whole salient regions with clear edges, thus when the ratio is low, it first samples the most salient region which maybe distracters like the farm lanes as shown in Fig. 8 that will affect the classification performance. For single texture scenes, the performances of most sparse sampling methods are similar and can easily approach the performance of dense sampling.

Therefore, for the single texture scenes with high information redundancy, we can use sparse sampling, especially random sampling, to replace dense sampling, since it can reach the same classification accuracy while with less spatial complexity, which can save time and memory for processing.

![Comparison of classification accuracy for different sampling methods on UC-Merced dataset](image)

**Figure 6:** Comparisons of the overall classification accuracy using different sampling methods on UC-Merced dataset (a) and RS19 dataset (b)

4.3.3 Testing on multiple texture scenes

This type of scenes is composed of several textures, such as the beach, river, etc. As this type need to be classified by different texture features like water, grass, trees and sands, most of the curves become stable after 60% of patches are extracted and share the similar trend. We can also

![Comparison of classification accuracy for different sampling methods on various scenes](image)

**Figure 7:** Comparisons of the classification accuracy using different sampling methods on the scenes of agricultural (a) and meadow (b)
learn that random sampling performs better when the ratio is low and sparse sampling can get similar results with dense sampling when the ratio becomes higher. Note that the accuracy of PISA is much lower than others when the ratio is low. The reason is that it may firstly samples the most salient region containing only one kind of texture which is not helpful for the classification performance.

Figure 8: An example of the agricultural scene image (a) and its saliency map using PISA (b)

Figure 9: Comparisons of the classification accuracy using different sampling methods on river (a) and beach (b)

4.3.4 Testing on object-based scenes

Object-based scenes mainly consist of a clear object and its backgrounds, such as airplane and storage tanks. The result is very interesting that most curves drop a little bit after increasing rapidly at the beginning, and then become stable in the end. This phenomenon is due to the fact that the classification of this type of scenes is similar to the object recognition problem in essence. Thus, most saliency-based method can firstly extract the features of the objects which can help to improve the performance. However, as the ratio becomes higher, the features of the objects become less dominant for more background parts are extracted. Thus, the trend of the curves can
be explained. Therefore, for this type of scenes, saliency-based sampling with low ratio is preferred.

![Graphs showing classification accuracy for different sampling methods on airplane and storage tanks.](a) airplane (b) storage tanks

Figure 10: Comparisons of the classification accuracy using different sampling methods on airplane (a) and storage tanks (b)

### 4.3.5 Testing on structural scenes

The last is composed of more complex scenes which contain densely distributed structures as well as textures, such as buildings, commercial, residential areas, etc. As is shown in Fig.11, the curves of this type are increasing all the way, which implies that the more information is extracted, the better the performance. Because this type of scene classification not only depends on the different land cover features they contain, but also depends on their different proportion. Therefore, without considering the spatial complexity, dense sampling is the best for this kind of complex scenes, while random sampling outperforms the others when the sampling rate is less than 1.

![Graphs showing classification accuracy for different sampling methods on buildings and residential.](a) buildings (b) residential

Figure 11: The classification accuracy using different sampling methods on buildings (a) and residential (b)

### 4.4 Discussion

From the experiment results on the two datasets we choose, we now can come to the conclusion that although some sparse sampling methods may be better than others for a certain type of scenes, the performance of dense sampling is the best on the whole dataset and random sampling has an
advantage over other sparse sampling methods, which is consistent with the conclusions for natural image classification [15].

As scene classification is obviously composed of various scenes, both simple and complex, in the case that there is no constrain on the spatial complexity, dense sampling is the first choice so as to improve the classification performance. However, if the dataset is too large under the limits of hardware conditions, random sampling strategy is recommended for its robustness, good performance and lower spatial complexity. If there is no prior information about the types of scenes, both keypoint-based and saliency-based sampling strategies are not recommended for the following two reasons. On the one hand, compared with random sampling, the samples they extracted are more dependent on their saliency measures, which is disadvantageous to the classification performance. On the other hand, they need to compute various low-level features for generating the response maps or saliency maps, which is thus time-consuming.

5 Conclusion

This paper studies the performance of various sampling strategies in the scene classification of high-resolution remote sensing imagery. To compare different sampling strategies, we adopt the classic bag-of-visual-words model to construct a unified scheme embedded with different sampling methods, e.g. dense sampling, random sampling, keypoint-based sampling and many recent saliency-based sampling methods. However, the conclusion on this experimental setting can be also extended to other scene classification approaches, such as those based on topic models and deep learning. We conduct the experiments on two commonly used datasets in the literature: UC-Merced dataset and the RS19 dataset. From the experiment results, we come to the consistent and meaningful conclusions that: although keypoint-based and saliency-based methods may show a little improvement in some classes of scene classification, they cannot compete with random sampling on the whole dataset. Moreover, the simplest dense sampling strategy provides the best results among all, without considering the spatial complexity. Therefore, if there is no constrain on the computational and spatial complexity, dense sampling is the first choice to improve the classification performance in scene classification of high-resolution remote sensing imagery. However, if the hardware source is limited and no prior information on the types of the scene is available, random sampling should be the first choice. The results in this paper can be applied to many scene classification approaches for HRRS images, such as many recently proposed deep learning-based ones [2, 5, 9].

Acknowledgment

The authors would like to thank all the researchers who kindly sharing the codes used in our studies. This research is supported by the National Natural Science Foundation of China under the contract No.91338113.

References

[1] S. Chen and Y. Tian, “Pyramid of spatial relations for scene-level land use classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 4, pp. 1947–1957, 2015.

[2] F. Zhang, B. Du, and L. Zhang, “Saliency-guided unsupervised feature learning for scene classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 53, no. 4, pp. 2175–2184, 2015.

[3] L. Zhao, P. Tang, and L. Huo, “A 2-d wavelet decomposition-based bag-of-visual-words model for land-use scene classification,” International Journal of Remote Sensing, vol. 35, no. 6, pp. 2296–2310, 2014.
[4] R. Kusumaningrum, H. Wei, R. Manurung, and A. Murni, “Integrated visual vocabulary in latent dirichlet allocation–based scene classification for ikonos image,” Journal of Applied Remote Sensing, vol. 8, no. 1, pp. 083 690–083 690, 2014.

[5] A. M. Cheriyadat, “Unsupervised feature learning for aerial scene classification,” IEEE Transactions on Geoscience and Remote Sensing, vol. 52, no. 1, pp. 439–451, 2014.

[6] H. Sridharan and A. Cheriyadat, “Bag of lines (bol) for improved aerial scene representation,” IEEE Transactions on Geoscience and Remote Sensing Letters, vol. 12, no. 3, pp. 676–680, 2014.

[7] W. Shao, W. Yang, and G.-S. Xia, “Extreme value theory-based calibration for the fusion of multiple features in high-resolution satellite scene classification,” International Journal of Remote Sensing, vol. 33, no. 8, pp. 2395–2412, 2012.

[8] G. Sheng, W. Yang, T. Xu, and H. Sun, “High-resolution satellite scene classification using a sparse coding based multiple feature combination,” International Journal of Remote Sensing, vol. 34, no. 23, pp. 8588–8602, 2013.

[9] F. Hu, G.-S. Xia, Z. Wang, L. Zhang, and H. Sun, “Unsupervised feature coding on local patch manifold for satellite image scene classification,” in IGARSS, 2014, pp. 1273–1276.

[10] J. Sivic and A. Zisserman, “Video google: A text retrieval approach to object matching in videos,” in Proc. IEEE International Conference on Computer Vision, 2003, pp. 1470–1477.

[11] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1794–1801.

[12] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2010, pp. 270–279.

[13] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis,” Machine learning, vol. 42, no. 1-2, pp. 177–196, 2001.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” the Journal of Machine Learning research, vol. 3, pp. 993–1022, 2003.

[15] L. Fei-Fei and P. Perona, “A bayesian hierarchical model for learning natural scene categories,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 524–531.

[16] E. Nowak, F. Jurie, and B. Triggs, “Sampling strategies for bag-of-features image classification,” in Proc. European Conference on Computer Vision, 2006, pp. 490–503.

[17] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.

[18] A. Bosch, A. Zisserman, and X. Muñoz, “Scene classification via plsa,” in Proc. European Conference on Computer Vision, 2006, pp. 517–530.

[19] ———, “Scene classification using a hybrid generative/discriminative approach,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 4, pp. 712–727, 2008.

[20] K. Mikolajczyk and C. Schmid, “Scale & affine invariant interest point detectors,” International Journal of Computer Vision, vol. 60, no. 1, pp. 63–86, 2004.

[21] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004.
[22] G.-S. Xia, J. Delon, and Y. Gousseau, “Accurate junction detection and characterization in natural images,” International Journal of Computer Vision, vol. 106, no. 1, pp. 31–56, 2014.

[23] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, “Evaluating bag-of-visual-words representations in scene classification,” in Proceedings of the International Workshop on Multimedia Information Retrieval, 2007, pp. 197–206.

[24] G. Csurska, C. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in Proc. European Conference on Computer Vision, vol. 1, no. 1-22, 2004, pp. 1–2.

[25] S. Lazebnik, C. Schmid, and J. Ponce, “A sparse texture representation using affine-invariant regions,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2003, pp. II–319–II–324.

[26] D. Gokalp and S. Aksoy, “Scene classification using bag-of-regions representations,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[27] F. Jurie and B. Triggs, “Creating efficient codebooks for visual recognition,” in IEEE International Conference on Computer Vision, vol. 1, 2005, pp. 604–610.

[28] J. Winn, A. Criminisi, and T. Minka, “Object categorization by learned universal visual dictionary,” in IEEE International Conference on Computer Vision, vol. 2, 2005, pp. 1800–1807.

[29] C. Siagian and L. Itti, “Rapid biologically-inspired scene classification using features shared with visual attention,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 2, pp. 300–312, 2007.

[30] G. Sharma, F. Jurie, and C. Schmid, “Discriminative spatial saliency for image classification,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 3506–3513.

[31] A. Borji and L. Itti, “Scene classification with a sparse set of salient regions,” in IEEE International Conference on Robotics and Automation, 2011, pp. 1902–1908.

[32] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 1597–1604.

[33] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, “Segmenting salient objects from images and videos,” in Proc. European Conference on Computer Vision, 2010, pp. 366–379.

[34] T. N. Vikram, M. Tscherepanow, and B. Wrede, “A saliency map based on sampling an image into random rectangular regions of interest,” Pattern Recognition, vol. 45, no. 9, pp. 3114–3124, 2012.

[35] K. Shi, K. Wang, J. Lu, and L. Lin, “Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2115–2122.

[36] C. E. Shannon, “Communication in the presence of noise,” Proc. Institute of Radio Engineers, vol. 37, no. 1, p. 10C21, 1949.

[37] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in Proc. Computer Vision and Pattern Recognition, Washington, DC, USA, 2005, pp. 886–893.

[38] G.-S. Xia, J. Delon, and Y. Gousseau, “Shape-based invariant texture indexing,” International Journal of Computer Vision, vol. 88, no. 3, pp. 382–403, 2010.

[39] G. Liu, G.-S. Xia, W. Yang, and L. Zhang, “Texture analysis with shape co-occurrence patterns,” in Proc. International Conference on Pattern Recognition, 2014, pp. 1627–1632.
[40] L. Itti, C. Koch, and E. Niebur, “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, 1998.

[41] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Advances in Neural Information Processing Systems*, 2006, pp. 545–552.

[42] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[43] N. Bruce and J. Tsotsos, “Saliency based on information maximization,” in *Advances in Neural Information Processing Systems*, 2005, pp. 155–162.

[44] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, “Sun: A bayesian framework for saliency using natural statistics,” *Journal of Vision*, vol. 8, no. 7, p. 32, 2008.

[45] E. J. Candès and M. B. Wakin, “An Introduction To Compressive Sampling,” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[46] A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[47] J. Lu, K. Shi, D. Min, L. Lin, and M. N. Do, “Cross-based local multipoint filtering,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 430–437.

[48] G.-S. Xia, W. Yang, J. Delon, Y. Gousseau, H. Sun, H. Maître et al., “Structural high-resolution satellite image indexing,” in *ISPRS TC VII Symposium-100 Years ISPRS*, vol. 38, 2010, pp. 298–303.

[49] C.-C. Chang and C.-J. Lin, “Libsvm: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.