sensors

MDPI

Article

Superpixel-Based Feature for Aerial Image Scene Recognition

Hongguang Li [1,2], Yang Shi [3,*], Baochang Zhang [4] and Yufeng Wang [3]

[1] Institute of Unmanned Systems, Beihang University, Beijing 100191, China; lihongguang@buaa.edu.cn
[2] Key Laboratory of Advanced Technology of Intelligent Unmanned Flight System of Ministry of Industry and Information Technology, Beijing 100191, China
[3] School of Electronic and Information Engineering, Beihang University, Beijing 100191, China; wyfeng@buaa.edu.cn
[4] School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China; bczhang@buaa.edu.cn
[*] Correspondence: newsy@buaa.edu.cn; Tel.: +86-10-8231-7391

Received: 24 October 2017; Accepted: 4 January 2018; Published: 8 January 2018

Abstract: Image scene recognition is a core technology for many aerial remote sensing applications. Different landforms are inputted as different scenes in aerial imaging, and all landform information is regarded as valuable for aerial image scene recognition. However, the conventional features of the Bag-of-Words model are designed using local points or other related information and thus are unable to fully describe landform areas. This limitation cannot be ignored when the aim is to ensure accurate aerial scene recognition. A novel superpixel-based feature is proposed in this study to characterize aerial image scenes. Then, based on the proposed feature, a scene recognition method of the Bag-of-Words model for aerial imaging is designed. The proposed superpixel-based feature that utilizes landform information establishes top-task superpixel extraction of landforms to bottom-task expression of feature vectors. This characterization technique comprises the following steps: simple linear iterative clustering based superpixel segmentation, adaptive filter bank construction, Lie group-based feature quantification, and visual saliency model-based feature weighting. Experiments of image scene recognition are carried out using real image data captured by an unmanned aerial vehicle (UAV). The recognition accuracy of the proposed superpixel-based feature is 95.1%, which is higher than those of scene recognition algorithms based on other local features.

Keywords: superpixel-based feature; image scene recognition; aerial remote sensing

1. Introduction

1.1. Background

Aerial remote sensing significantly complements satellite remote sensing. Recent developments in unmanned aerial vehicles (UAVs), platforms, positional and attitudinal measurement sensors, imaging sensors, and processing approaches have opened up considerable opportunities for applying remote sensing in national environmental protection [1], land use survey [2], marine environmental monitoring [3], water resource development, crop growth monitoring and assessment [4], wildlife multispecies remote sensing [5], forest protection and monitoring [6], natural disaster monitoring and evaluation [7], target surveillance [8], and Digital Earth. Subsequently, these applications have greatly promoted the development of aerial remote sensing.

Scene recognition has long been a popular and significant research field [9]. Image scene recognition, the most common task of aerial remote sensing applications, is the process of marking images according to semantic categories, such as seashore, forest, field, mountain, and city scenes.

Scene recognition has also been the research focus in machine learning, computer vision, and image processing, among others.

1.2. Related Work

Despite the rapid development of convolutional neural networks in recent years, traditional methods for feature-based machine learning offer important application value to image scene recognition. Image feature description and extraction are core technologies for many aerial remote sensing applications. Image scene features can be described accurately by their hierarchical levels using the following techniques: scale-invariant feature transform (SIFT) [10,11] and histogram of oriented gradient (HOG) [12] for low-level features; Bag-of-Words [13,14] modeling, sparse coding [15], and deformable parts modeling [16] for middle-level features; and topic modeling [17,18] and spatial pyramid matching [19,20] for high-level features.

Low-level features are closest to the digital storage forms of images (i.e., color, texture, and shape). Regarded the most direct source of image information, low-level features can be used to obtain higher-level information. Thus, low-level features provide the basis for image scene cognition. Middle-level features are extracted statistically or by reasonable judgment through image mining techniques. Thus, middle-level features can describe semantic content, which implies higher-level accuracy for image scene recognition. High-level features can be obtained by modeling the features of the middle layer to subsequently describe the content closest to the human perception of an image.

Among the above feature levels and image recognition techniques, the middle-level semantic feature of the Bag-of-Words model effectively solves the "semantic gap" between low-level image information and high-level semantics. The multi-level characterization of the Bag-of-Words model is more advanced than low-level imaging techniques and is therefore widely used in scene recognition. In Bag-of-Words modeling, local features are regarded as major factors affecting scene recognition performance. Scene feature descriptions are generally divided into two kinds. The first set of descriptors is based on the detection of points of interest, such as Harris corner detector [21], features from accelerated segment test (FAST) [22], Gaussian laplacian, and Gaussian difference, among others. The other set of descriptors is based on dense extraction, such as SIFT, HOG, and extraction of local binary patterns (LBP) [23], among others. However, these methods employ bottom-to-top descriptions and consider image characteristics (point, color, texture, and shape) from a certain aspect only, thereby leading to incomplete image information. These conditions imply that even the advanced feature modeling processes (e.g., Bag-of-Words model) employ bottom-led mathematical deduction and statistics while neglecting top-led semantic expression.

Scenes from current aerial imaging processes are often characterized as follows: (1) depth of information is lost and entire scenes are approximated as a plane given the far-off distance between the aerial equipment and captured images; (2) contents of image planes are categorized as classical types of landforms, such as grasslands, deserts, buildings, and rivers, among others; and (3) the recognized image scenes are constrained to the classical types of landforms only. Based on these delimited aerial image scene characteristics, the bottom-to-top method for low-level feature description cannot acquire local surface content. Thus, the Bag-of-Words model is not as effective when dealing with scene recognition.

1.3. Contribution of This Paper

This study first considers aerial image landforms for scene identification. Then, the study proposes a top-down feature-based image scene recognition method for aerial application.

The main work and contributions of this study are as follows: (1) A novel superpixel-based feature description method is proposed. To utilize landform information, the study establishes the features acquired by top-task landform superpixel extraction to bottom-task expression of feature vectors. (2) Using the proposed superpixel-based features, a type of Bag-of-Words model is constructed and an image scene recognition algorithm is designed. (3) Experiments are carried out using real image data captured by UAVs.

2. Methodology

2.1. Aerial Image Semantic Hierarchical Structure

To better demonstrate the motivation and basis of the proposed method, this section introduces a hierarchical structure model of aerial image semantics.

Image semantics is the description of the image content itself. According to the abstraction degree of image semantics, Eakin [24] divides it into 3 layers: low-level feature layer, object layer and semantic concept layer. From bottom to top, it includes feature semantics, object semantics, spatial relations semantics, scene semantics, behavior semantics and emotion semantics, as shown in Figure 1a. High level semantics are often more abstract than low-level semantics and are quantified by a lower level of semantics.

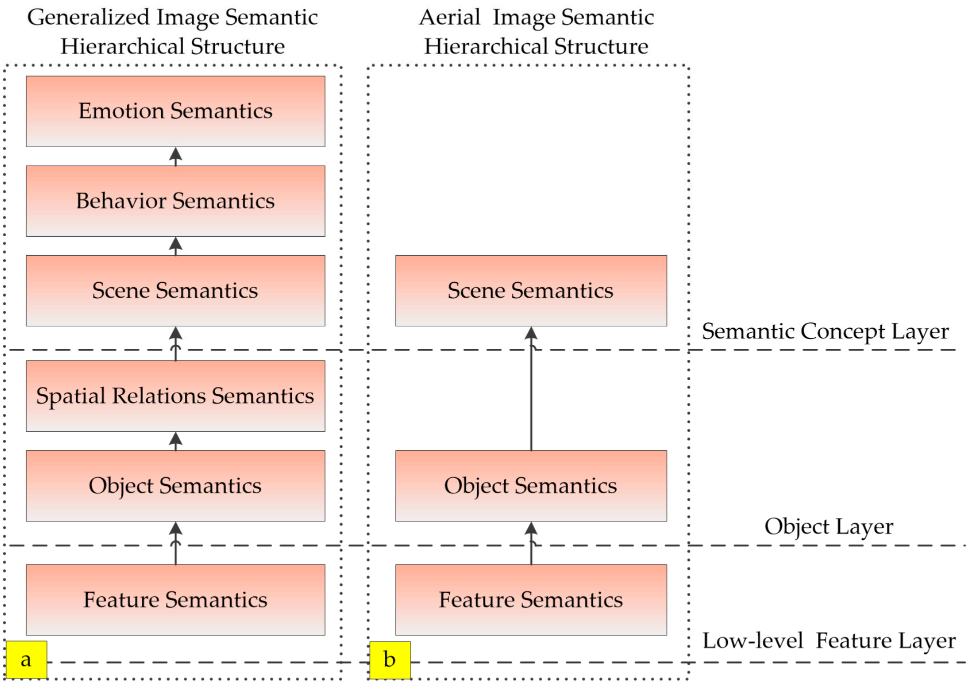


Figure 1. Image semantic hierarchical structure: (**a**) generalized image semantic hierarchical structure; and (**b**) aerial image semantic hierarchical structure.

Because of the characteristics of long-distance imaging, there are few information about spatial relations semantics, behavior semantics and emotional semantics in aerial images. Its main semantic features are feature semantics, object semantics and scene semantics, as shown in Figure 1b. Feature semantics include texture, color, shape, structure and so on, which correspond to the basic visual perception. The object semantics is embodied in the landform area or target area of the aerial image, which can be used to model the semantic features for scene and target recognition. Scene semantics refers to the image scene label, which involves a higher level of abstract attributes, and is derived from the object semantics.

Based on the above aerial image semantic hierarchical structure, a novel superpixel-based feature is proposed in this paper, which corresponds to the object semantics in the semantic structure. In the process of proposed feature based image scene recognition, landform is a useful and basic clue for method design. In the object semantic layer, superpixel is used to express landform information in Section 2.2.1. Then, Sections 2.2.2–2.2.4 describe how to extract the low-level landform feature of superpixel region, which corresponds to the feature semantics in the semantic structure. Finally, Section 2.3 gives the overall flow of the scene recognition of aerial images.

2.2. Superpixel-Based Feature Description

A novel process for superpixel-based feature description is proposed for capturing landforms aerially during scene identification. The process mainly consists of the following steps: simple linear iterative clustering (SLIC) [25] based superpixel segmentation, adaptive filter bank construction, Lie group-based feature quantification, and feature weighting based on the visual saliency model (Figure 2). SLIC is used to extract superpixels, whereas filter banking, feature quantification, and feature weighting are conducted to transform the superpixels into feature vectors.

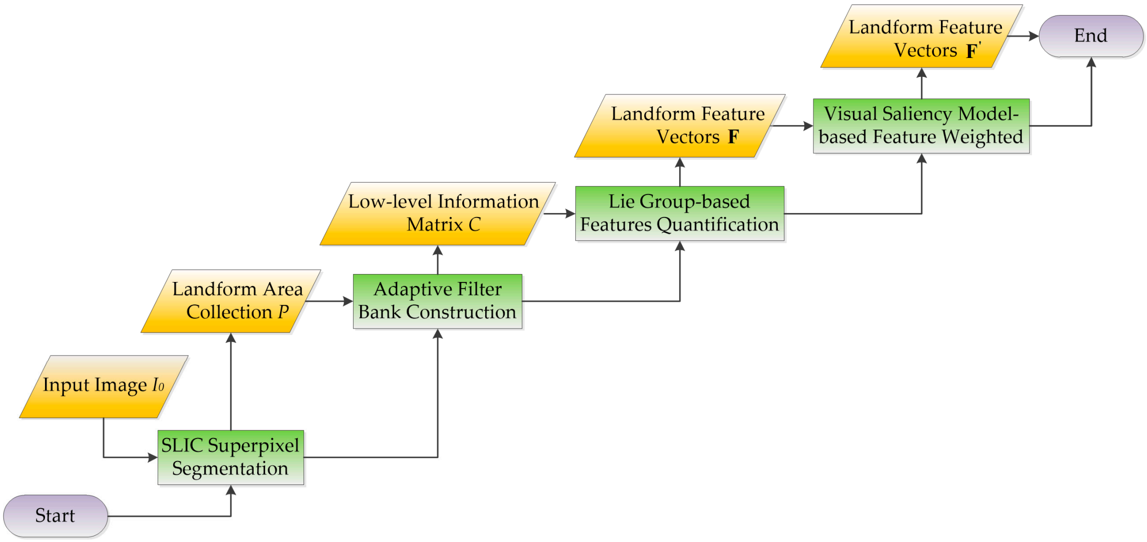

Landform Feature Vectors F'
End
Landform Feature Vectors F
Visual Saliency Model-based Feature Weighted
Low-level Information Matrix C
Lie Group-based Features Quantification
Landform Area Collection P
Adaptive Filter Bank Construction
Input Image I_0
SLIC Superpixel Segmentation
Start

Figure 2. Flowchart of superpixel-based feature extraction (green: process; yellow: data type).

In Figure 2, input image I_0 is initially segmented by SLIC superpixel algorithms to collect landform data of the image. Based on on prior landform information, 2D linear discriminant analysis [26,27] is conducted to adaptively construct the filter bank. Second, the filter bank is convolved with the image to obtain matrix C with low-level information (color, texture, context, etc.). Each pixel corresponds to a filter response vector, and the image content of the irregular surface is described by a feature matrix that comprises the corresponding feature vectors of the pixels. Third, the Lie group theory based on Riemann manifold geometry [28–31] is used to analyze the topological relations of surface pixels, and then the feature matrix is mapped into vector space to generate feature vector $\mathbf{F}$. Finally, local feature vector $\mathbf{F}'$ is obtained by weighting the feature vectors according to the visual saliency model.

2.2.1. SLIC Superpixel Segmentation

Aerial image scenes differ from low-altitude outdoor images. Scenes in low-altitude outdoor images are mostly composed of a background and a target, which are defined differently. Backgrounds may include skies, roads, grasslands, and buildings; by contrast, targets may include pedestrians and vehicles. Objects such as pedestrians and vehicles are proportionally small compared with the image because of the far-off imaging distance of aerial platforms; thus, the depths of field of image planes are nearly the same. Image scenes are generally composed of different landforms, such as grasslands, deserts, buildings, rivers, and so on. Therefore, landforms comprise the basic components of scene semantics of aerial images.

Statistical analyses of aerial images generally depict landforms as irregularly shaped images with different colors and textures. In particular, the superpixels of images comprise irregular pixel blocks with a certain visual sense, i.e., adjacent pixels with similar textures, colors, brightness and other features. Thus, the superpixel area of aerial images can represent actual landform surfaces.

This study uses color and space distance based on the SLIC algorithm [25] to segment aerial images into landform areas. The specific steps of SLIC superpixel segmentation are as follows:

(1) Cluster center initialization is conducted by assuming N pixels in the image. The size of each superpixel is approximately N/K after setting the number of superpixels to K. To avoid clustering at the edge of an object, the initial cluster is "centered" to the position wherein the gradient value is smallest, for instance, at the center of a 3 × 3 neighborhood window. The image gradient calculation formula is

$$G(x,y) = \|v(x+1,y) - v(x-1,y)\|^2 + \|v(x,y+1) - v(x,y-1)\|^2 \tag{1}$$

where $v(x,y)$ is the pixel value at point (x, y). The cluster center is defined by $(O_i, i = 1, \ldots, K)$.

(2) Similarity measurement in SLIC is expressed by the following:

$$d(i,k) = d_{lab} + \frac{m}{S} d_{xy} \tag{2}$$

$$d_{lab} = \sqrt{(l_k - l_i)^2 + (a_k - a_i)^2 + (b_k - b_i)^2} \tag{3}$$

$$d_{xy} = \sqrt{(x_k - x_i)^2 + (y_k - y_i)^2} \tag{4}$$

where d_{lab} is the color difference between two pixels; l, a, and b are the three components of Lab color space; d_{xy} is the spatial distance between two pixels; (x, y) is pixel position; $d(i,k)$ is the similarity between ith pixel and kth cluster center (i.e., the smaller the value, the higher the similarity); and $m \in [1,20]$ is the parameter that balances color and spatial information in the similarity measure, which is set to 10 in this study. The desired superpixel size is $S \times S$, where $S = \sqrt{N/K}$.

(3) The K-means clustering method involves an iteration of the clustering center. Based on the measured similarity, the K-means clustering method is performed on the $2S \times 2S$ region in the X – Y image plane. The process is repeated until convergence (i.e., the maximum number of times) is reached. The initial cluster centers are uniformly initialized in the image with all pixels situated near their cluster center.

(4) The process of generating the superpixel regions is conducted given that some small areas may be present in the regions. Each generated small area is labeled as a superpixel even if these areas are not connected to the superpixel. The small areas are thus connected to the largest superpixel to ensure the integrity of every superpixel. For instance, Figure 3b is the result of the superpixel region segmentation of Figure 3a.

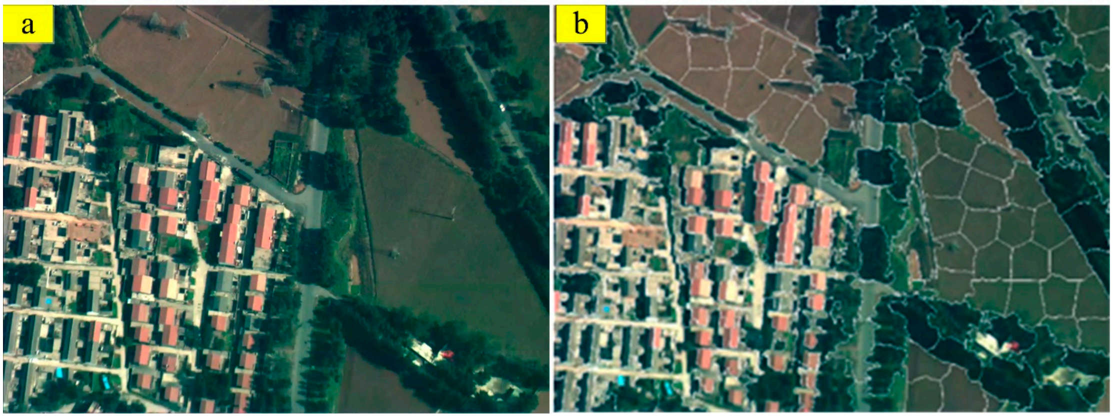


Figure 3. Superpixel segmentation of an aerial image: (**a**) aerial image; and (**b**) result of superpixel region segmentation of (**a**).

2.2.2. Adaptive Filter Bank Construction

Different types of landforms in aerial images vary greatly in color, texture, and shape. To fully describe landform features, the aerial images are filtered and low-level feature modeling is performed using a filter bank. Filter banks are an array of band-pass filters that separate input signal from multiple components. Filter types, scale parameters, and direction parameters are determined in accordance with the actual situation on the basis of different task requirements.

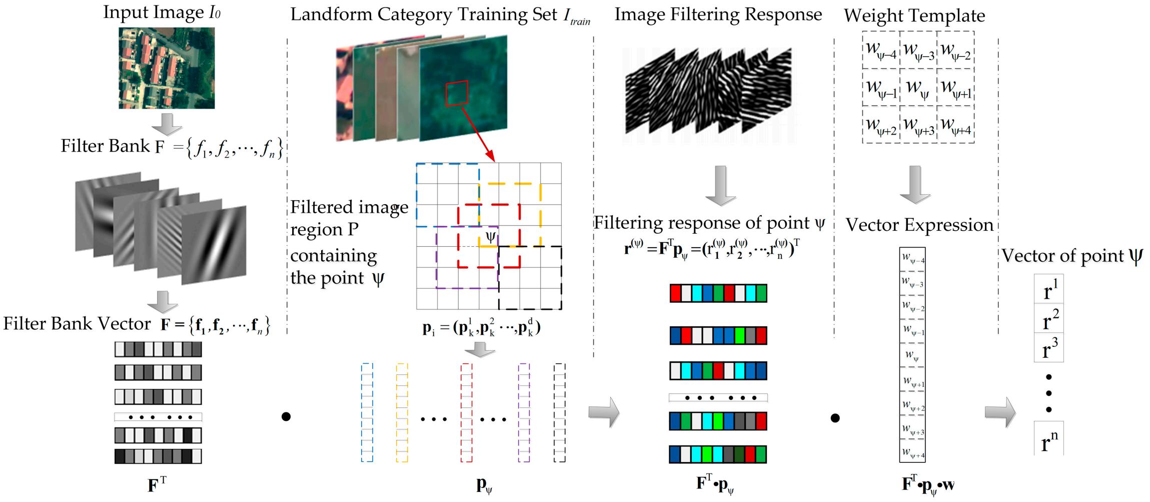
In the construction of the filter bank, any prior information about the surface is assumed to have five types of image landforms, namely, buildings, trees, grasses, bare lands, and roads. Then, 2D linear discriminant analysis [27] is conducted on the basis of the surface prior information to adaptively construct image filter banks and their weight templates. Consequently, the filter banks are convolved with the input image to generate low-level features.

If image I_0 is filtered by filter $\mathbf{F}_0$, then the response image is I_0'. Then, filtering response $\mathbf{r}^{(\psi)}$ corresponding to point ψ in response image I_0' is regarded a convolution value of filter $\mathbf{F}_0$ and the image area in which point ψ is contained. The process can be expressed as Equation (5).

$$\mathbf{r}^{(\psi)} = \mathbf{F}_0{}^{\mathrm{T}} \mathbf{p}_{\psi} \tag{5}$$

where $\mathbf{F}_0{}^{\mathrm{T}} \in \mathbb{R}^{\mathrm{L_n} \times 1}$ represents the vector form of filter $\mathbf{F}_0$ and $\mathbf{p}_{\psi}$ represents the vector form of the image region in which point ψ is the center.

The constructed adaptive filter bank and corresponding processes are shown in Figure 4.



Input Image I_0
Landform Category Training Set I_{train}
Image Filtering Response
Weight Template
Filter Bank F $=\{f_1, f_2, \cdots, f_n\}$
Filtered image region P containing the point ψ
ψ
Filtering response of point ψ
$\mathbf{r}^{(\psi)} = \mathbf{F}^{\mathrm{T}}\mathbf{p}_{\psi} = (\mathbf{r}_1^{(\psi)}, \mathbf{r}_2^{(\psi)}, \cdots, \mathbf{r}_n^{(\psi)})^{\mathrm{T}}$
Vector Expression
Vector of point ψ
Filter Bank Vector $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \cdots, \mathbf{f}_n\}$
$\mathbf{p}_i = (\mathbf{p}_k^1, \mathbf{p}_k^2 \cdots, \mathbf{p}_k^d)$
$w_{\psi-4}$ $w_{\psi-3}$ $w_{\psi-2}$
$w_{\psi-1}$ w_{ψ} $w_{\psi+1}$
$w_{\psi+2}$ $w_{\psi+3}$ $w_{\psi+4}$
r^1 r^2 r^3 r^n
$\mathbf{F}^{\mathrm{T}}$
$\mathbf{p}_{\psi}$
$\mathbf{F}^{\mathrm{T}} \bullet \mathbf{p}_{\psi}$
$\mathbf{F}^{\mathrm{T}} \bullet \mathbf{p}_{\psi} \bullet \mathbf{w}$

Figure 4. Adaptive filter bank construction. Filter banks are convolved with input image I_0, to generate low-level feature ψ.

In Figure 4, $\mathbf{F} = \{\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f}_n\}$ is the vector expression of filter bank $F = \{\mathrm{f}_1, \mathrm{f}_2, \ldots, \mathrm{f}_n\}$. After the convolution of $\mathbf{F} = \{\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f_F}\}$ and image I_0, the corresponding filtering response of point ψ is defined as Equation (6).

$$\mathbf{r}^{(\psi)} = \mathbf{F}^{\mathrm{T}} \mathbf{p}_{\psi} = \left(r_1^{(\psi)}, r_2^{(\psi)}, \ldots, r_n^{(\psi)}\right)^{\mathrm{T}} \forall 1 \leq k \leq n r_k^{(\psi)} = \mathbf{f}_k{}^{\mathrm{T}} \mathbf{p}_{\psi} \tag{6}$$

After weighted calculation, the corresponding feature vectors of point ψ are obtained by Equation (7).

$$\mathbf{r}_{\psi} = \frac{1}{d} \sum_{i \in \mathbb{R}^{(\psi)}} \left(r_1^{(\psi)}, r_2^{(\psi)}, \ldots, r_n^{(\psi)}\right)^{\mathrm{T}} = \frac{1}{d} \sum_{i \in \mathbb{R}^{(\psi)}} \mathbf{F}^{\mathrm{T}} \mathbf{p}_i \tag{7}$$

where d is filter template size, while $\mathbf{p}_i = (\mathbf{p}_k^1, \mathbf{p}_k^2 \ldots, \mathbf{p}_k^d)$ is the vector form of the corresponding region of an image containing point ψ. The pixels in different positions affect the feature vectors of ψ points differently. Subsequently, weight template $\mathbf{w}$ is defined for the obtained feature vectors, and the corresponding vector form is $\mathbf{w} = (w_1, w_2, \ldots, w_d)$. The feature vector value of point ψ is expressed as Equation (8).

$$\mathbf{r}_\psi = \mathbf{F}^\mathrm{T}\mathbf{p}_i\mathbf{w} \tag{8}$$

To ensure that the extracted feature vectors contain abundant image information and can strongly express the surface, all prior information about the aerial landform image is used for low-level feature extraction. Based on the Fisher criterion [32], the 2D-LDA method is used to adaptively learn the filter bank and extract the low-level features of the landforms.

Variable $\mathbf{r}_m^n$ is defined as the vector form of the nth image region extracted from the mth landform category. Intra-class and inter-class differences are quantified by Equation (9).

$$\begin{cases} \mathrm{L}_{\text{intra}} = \sum\limits_{m=1}^{M}\sum\limits_{n=1}^{N}(\mathbf{r}_m^n - \overline{\mathbf{r}_m})(\mathbf{r}_m^n - \overline{\mathbf{r}_m})^\mathrm{T} \\ \mathrm{L}_{\text{inter}} = \sum\limits_{m=1}^{M} N_m(\overline{\mathbf{r}_m} - \bar{\mathbf{r}})(\overline{\mathbf{r}_m} - \bar{\mathbf{r}})^\mathrm{T} \end{cases} \tag{9}$$

where M represents surface type; N_m represents the number of features extracted from the mth landform category; $\overline{\mathbf{r}_m}$ is the mean of all feature vectors in the mth landform category; and $\bar{\mathbf{r}}$ represents the mean of all feature vectors. After the derived feature vector formula is added to the model, intra-class and inter-class differences are determined as Equation (10).

$$\begin{cases} \mathrm{L}_{\text{intra}} = \sum\limits_{m=1}^{M}\sum\limits_{n=1}^{N}\mathbf{F}^\mathrm{T}(\mathbf{P}_m^n - \overline{\mathbf{P}_m})\mathbf{w}\mathbf{w}^\mathrm{T}(\mathbf{P}_m^n - \overline{\mathbf{P}_m})^\mathrm{T}\mathbf{F} \\ \mathrm{L}_{\text{inter}} = \sum\limits_{m=1}^{M} N_m\mathbf{F}^\mathrm{T}(\overline{\mathbf{P}_m} - \overline{\mathbf{P}})\mathbf{w}\mathbf{w}^\mathrm{T}(\overline{\mathbf{P}_m} - \overline{\mathbf{P}})^\mathrm{T}\mathbf{F} \end{cases} \tag{10}$$

where $\mathbf{P}_m^n$ represents the vector form of the corresponding region of the nth image in the mth landform category training set I_{train}; $\overline{\mathbf{P}_m}$ represents the mean value of the mth category; and $\overline{\mathbf{P}}$ is the mean value of the entire dataset.

Then, filter bank $\mathbf{F}$ and its weighting template $\mathbf{w}$ are solved by the objective functions of intra-class difference and inter-class difference, as shown in Equation (11).

$$J(\mathbf{F}, \mathbf{w}) = \underset{(\mathbf{F},\mathbf{w})}{\operatorname{argmax}} \frac{\sum\limits_{m=1}^{M} N_m\mathbf{F}^\mathrm{T}(\overline{\mathbf{P}_m} - \overline{\mathbf{P}})\mathbf{w}\mathbf{w}^\mathrm{T}(\overline{\mathbf{P}_m} - \overline{\mathbf{P}})^\mathrm{T}\mathbf{F}}{\sum\limits_{m=1}^{M}\sum\limits_{n=1}^{N}\mathbf{F}^\mathrm{T}(\mathbf{P}_m^n - \overline{\mathbf{P}_m})\mathbf{w}\mathbf{w}^\mathrm{T}(\mathbf{P}_m^n - \overline{\mathbf{P}_m})^\mathrm{T}\mathbf{F}} \tag{11}$$

Equation (11) can be solved by the 2D-LDA algorithm. Equations (12) and (13) are used as optimization functions to solve $\mathbf{F}$ and $\mathbf{w}$.

$$\begin{cases} J(\mathbf{F}) = \text{maxtrace}\left(\mathbf{F}^\mathrm{T}\mathrm{L}_{\text{intra}}^{\mathbf{w}}\mathbf{F}\right)^{-1}\left(\mathbf{F}\mathrm{L}_{\text{inter}}^{\mathbf{w}}\mathbf{F}^\mathrm{T}\right) \\ \mathrm{L}_{\text{intra}}^{\mathbf{w}} = \sum\limits_{m=1}^{M}\sum\limits_{n=1}^{N}(\mathbf{P}_m^n - \overline{\mathbf{P}_m})\mathbf{w}\mathbf{w}^\mathrm{T}(\mathbf{P}_m^n - \overline{\mathbf{P}_m})^\mathrm{T} \\ \mathrm{L}_{\text{inter}}^{\mathbf{w}} = \sum\limits_{m=1}^{M} N_m(\overline{\mathbf{P}_m} - \overline{\mathbf{P}})\mathbf{w}\mathbf{w}^\mathrm{T}(\overline{\mathbf{P}_m} - \overline{\mathbf{P}})^\mathrm{T} \end{cases} \tag{12}$$

$$\begin{cases} J(\mathbf{w}) = \text{maxtrace}\left(\mathbf{w}^\mathrm{T}\mathrm{L}_{\text{intra}}^{\mathbf{F}}\mathbf{w}\right)^{-1}\left(\mathbf{w}\mathrm{L}_{\text{inter}}^{\mathbf{F}}\mathbf{w}^\mathrm{T}\right) \\ \mathrm{L}_{\text{intra}}^{\mathbf{F}} = \sum\limits_{m=1}^{M}\sum\limits_{n=1}^{N}(\mathbf{P}_m^n - \overline{\mathbf{P}_m})^\mathbf{T}\mathbf{F}\mathbf{F}^\mathrm{T}(\mathbf{P}_m^n - \overline{\mathbf{P}_m}) \\ \mathrm{L}_{\text{inter}}^{\mathbf{F}} = \sum\limits_{m=1}^{M} N_m(\overline{\mathbf{P}_m} - \overline{\mathbf{P}})^\mathbf{T}\mathbf{F}\mathbf{F}^\mathrm{T}(\overline{\mathbf{P}_m} - \overline{\mathbf{P}}) \end{cases} \tag{13}$$

The final filter bank $\mathbf{F}$ and its weight template $\mathbf{w}$ can be obtained after multiple iterations of the optimization function. Then, filter bank $\mathbf{F}$ and weight template $\mathbf{w}$ are convolved with image I_0 to obtain the filter response image set $\mathbf{I}_{\text{res}} = \{I_1, I_2, \ldots, I_{\text{F}}\}$. The filter bank is usually convolved with a gray image of the input image for low-level feature extraction. However, low-level features can only describe the texture, context, and other information of an image, but cannot reflect color characteristics. To solve this problem, this study combines nine channel images of the input image corresponding to RGB, HSV, and Lab color space with the filter response image to form the image feature set $\mathbf{I}_{\text{fea}} = \{I_1, I_2, \ldots, I_{\text{F}}, I_{\text{F}+1}, \ldots, I_{\text{F}+9}\}$.

2.2.3. Lie Group-Based Feature Quantification

The aerial images are divided into a number of superpixel regions in which every pixel point corresponds to a multidimensional feature vector. Inputs to the Bag-of-Words model usually take the form of a feature expression vector of a local region. Thus, to effectively integrate low-level features to the Bag-of-Words model, the feature matrix describing the superpixel region is quantized as a feature vector.

This study therefore introduces the Lie group structure [28] to the Riemann manifold space [30]. The Lie group can reflect the correlation between two feature vectors of each pixel and describe the spatial structure information of all the pixel feature vectors in the entire superpixel region.

A total of N superpixels pixels in an input image can be extracted after superpixel segmentation. An arbitrary superpixel $\{\mathrm{P}_i \in \mathrm{I}, i = 1, \ldots, N\}$ in an image contains C pixels $\{x_j \in \mathbb{R}^{\mathrm{k}\times 1}, j = 1, \ldots, \mathrm{C}\}$, and the superpixel region is expressed by the Gaussian model in accordance with the maximum likelihood method (Equation (14)).

$$\phi(x_i|\mu,\Sigma) = \frac{\exp(-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu))}{\sqrt{(2\pi)^k \det(\Sigma)}} \tag{14}$$

where $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\Sigma = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$ are the mean value and covariance of the vector while $\det(\cdot)$ is the matrix determinant. The Gaussian model with K dimensions is a Riemann manifold, and $\phi(\mu,\Sigma)$ is defined as the Gaussian model with μ mean value and Σ covariance. For any random vector of the Gaussian distributions, a unique transformation can be used to meet the requirements of Equation (15).

$$\begin{bmatrix} x \\ 1 \end{bmatrix} = \begin{bmatrix} H & \mu \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_0 \\ 1 \end{bmatrix} \tag{15}$$

where H is an upper triangular matrix that satisfies $\Sigma = HH^T$. Covariance matrix Σ and H are both positive definite. Thus, the upper triangular positive definite affine matrix can be expressed as Equation (16).

$$\mathbf{M} = \begin{bmatrix} H & \mu \\ 0 & 1 \end{bmatrix} \tag{16}$$

A double shooting may occur between matrix $\mathbf{M}$ and Gaussian function. According to Lie group theory, an invertible affine scaling matrix can form a Lie group. Thus, the upper triangular positive definite affine matrix is assumed equivalent to the Gaussian function. After affine transformation, $\phi(\mu,\Sigma)$ can be uniquely transformed into $(k+1)\times(k+1)$ dimensional positive definite symmetric matrix, as expressed in Equation (17).

$$\phi(\mu,\Sigma) \sim S = \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix} \tag{17}$$

The vector space S^+_{k+1} of positive definite symmetric matrices with $(k+1) \times (k+1)$ dimensions is a Lie group and a Riemann manifold with local topology properties equivalent to vector space. For any element in the Lie group, a corresponding Lie algebra exists in which the tangent space is a vector space. The elements of the Gaussian Lie group can be mapped between tangent space (Gaussian Lie algebras) and manifolds (Gaussian Lie group) by means of log operations and exponential matrix operations. Finally, by adopting upper triangular positive definite affine matrix transformation, the $(k+1)(k+2)/2$ (i.e., 190) dimensional feature vector is obtained for describing superpixel region P_i. Consequently, the final input image is expressed by N 190-dimensional manifold-based superpixel feature f_i.

2.2.4. Visual Saliency Model-Based Feature Weighting

For two images belonging to different categories but with similar contents, the contents of the salient regions often serve an important basis for determining image categories. To enhance the expression capability of low-level features and image semantic features, a feature weighting method based on saliency is proposed in this study for aerial scene recognition.

The steps for the algorithm are as follows: A visual saliency map S_{map} with the same size as the input image is obtained according to the visual saliency model. For the feature f_i of superpixel region P_i, the feature weight w_i is defined as the average value of the saliency map corresponding to all pixels in the superpixel region, as shown in Equation (18).

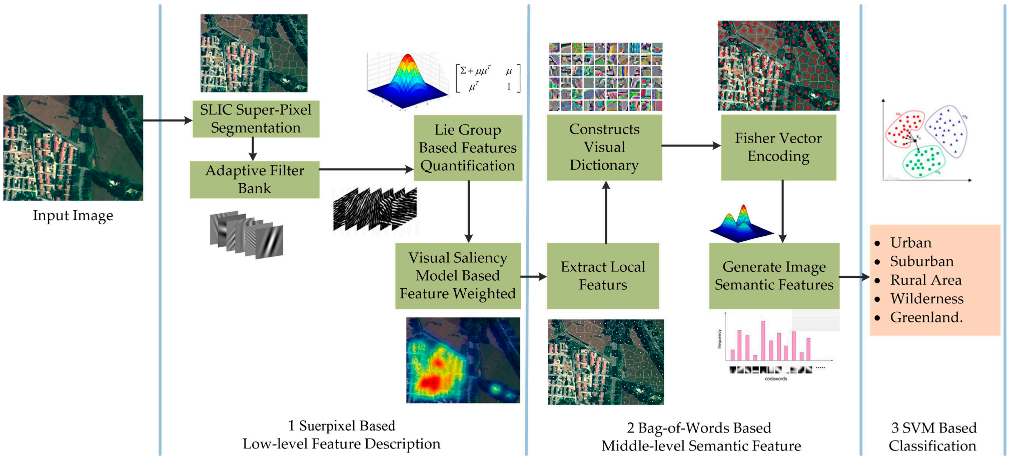
$$w_i = mean(S_{map}\{P_i\}) \tag{18}$$

The feature corresponding to each area P_i of the landform is $w_i * f_i$.

2.3. *Scene Recognition of Aerial Images*

The image scene recognition algorithm based on the Bag-of-Words model is determined using the proposed superpixel-based feature for aerial remote sensing applications.

The Bag-of-Words model constructs a visual dictionary using the low-level features of an image; performs feature coding to obtain middle-level semantic features; and integrates a classifier to realize image scene recognition. As shown in Figure 5, the three main steps of Bag-of-Words modeling are as follows: first, the superpixel-based low-level features are extracted; second, the dictionary is generated; and, finally, image scene classification is conducted.



Input Image
SLIC Super-Pixel Segmentation
Adaptive Filter Bank
Lie Group Based Features Quantification
Visual Saliency Model Based Feature Weighted
Constructs Visual Dictionary
Extract Local Featurs
Fisher Vector Encoding
Generate Image Semantic Features
codewords
Urban
Suburban
Rural Area
Wilderness
Greenland.
1 Suerpixel Based Low-level Feature Description
2 Bag-of-Words Based Middle-level Semantic Feature
3 SVM Based Classification

Figure 5. Scene recognition of aerial images. The method includes three steps: superpixel-based low-level feature extraction, Bag-of-Words model-based middle-level semantic feature extraction, and SVM-based classification.

In the first step, superpixel-based low-level features are extracted using the proposed method. In the second step, the main feature encoding methods (vector quantization, sparse, local linear constraint, Fisher vector [32], etc.) are conducted for Bag-of-Words modeling. Fisher vector encoding, which simultaneously conducts generative and discriminative modeling to record first- and second-order differences between a local feature and its nearest visual words, is generally considered for its strong capability to express features. Thus, Fisher vector encoding is selected to further quantify the local features of the extracted aerial image and generate semantic features to complete the aerial image scene recognition. In the third and final step, the SVM classifier is used to recognize five image scenes (i.e., urban, suburban, rural, wilderness, and green land).

3. Result and Discussion

3.1. Experimental Data

The experimental data used in this study are collected from UAV images. The dataset for scene recognition is divided into five categories, each containing 100 images (30 images for training and 70 images for testing). The designated scene categories are urban, suburban, rural, wilderness, and green land, as shown in Figure 6. The landform images used to train adaptive filters are collected from the training set of scene recognition data. Each landform image has the size of 100 × 100 (resolution), and landforms are categorized as buildings, trees, grasses, bare lands, and roads, as shown in Figure 7.

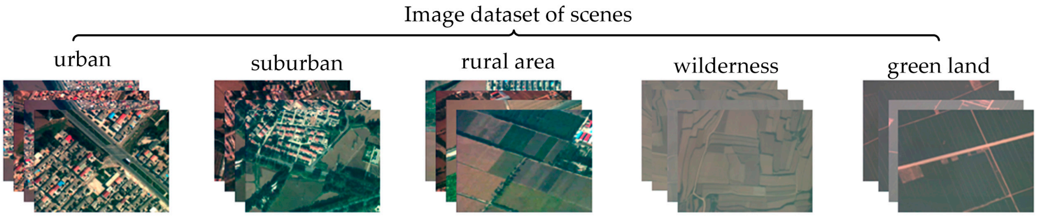


Figure 6. Image dataset of scenes. The designated categories are urban, suburban, rural, wilderness, and green land.

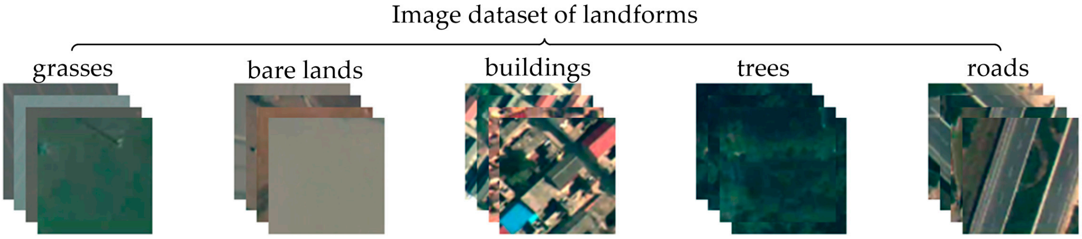


Figure 7. Image dataset of landforms. The landforms are categorized as buildings, trees, grasses, bare lands, and roads.

3.2. Experimental Results Analysis

This study applies Bag-of-Words model-based scene recognition and compares this technique with typical feature-based Bag-of-Words modeling to verify the expression capability of the proposed superpixel-based features. Model performance is evaluated in terms of scene recognition accuracy and extraction time. Filter template size and saliency model are considered during feature extraction, as both are regarded as important factors that can affect feature expression.

3.2.1. Influence of Filter Template Size on Feature Expression

The 2D linear discriminant analysis is conducted based on prior image surface information, which is also used to construct the image filter bank. Then, the filter bank is convolved with the image to obtain the low-level features of the image. This method does necessarily consider filter type, scale factor, and direction factor; however, filter template size needs to be artificially determined. This requirement suggests that filter template size can influence the capability of the proposed method to express the local features and semantic features of the image; moreover, the level of influence can be reflected by the results of the final scene recognition. The experimental results are shown in Table 1. Experiments for the low-level features are carried out under non-overlapping saliency weight conditions.

Table 1. Influence of filter template size on recognition results.

Template Size	3 × 3	5 × 5	7 × 7	9 × 9
Recognition Accuracy	88.3%	88%	88.3%	88.9%

Experimental results show that filter template size has little influence on low-level and semantic features. Differences in scene recognition accuracy for the different templates are below 1%. The situation can be explained by large single superpixel areas (i.e., segmentation result), whereas filter template size is small. Image filtering is usually conducted in one-pixel steps; therefore, pixel filter response is related only to neighboring image regions and the filter's own parameters. In addition, different filter templates are trained using the same training dataset, and this approach leads to similar filtering responses despite the different size templates in the same location.

3.2.2. Influence of Saliency Model on Scene Recognition

To test the influence of the saliency model on scene recognition, this study introduces four kinds of saliency models for comparison, namely, the Itti model [33], Erdem model [34], Achanta model, [35] and SIM model [36].

The Itti visual saliency model [33] is based on the three characteristics of brightness, color, and direction, and then the salient region of an image is determined after normalization. The color difference of different landforms is large, and artificial building parts (buildings and roads) have relatively high reflection coefficients than natural surfaces; thus, the brightness varies greatly for different regions in aerial images. Erdem [34] introduced region covariance into visual saliency to improve model robustness by nonlinear feature fusion. Subsequently, the capability of local features to describe image content is strengthened. The visual saliency model proposed by Achanta [35] theoretically starts with a frequency domain. Then, the saliency value of each pixel is set as the Euclidean distance between two pixel values after Gaussian low-pass filtering and pixel value averaging of the entire image in Lab space. The method uses fixed value filtering. When the salient object is relatively large, the calculation of the Lab space mean is affected; consequently, the saliency value of the salient region is less than that of the background. No obvious differences exist between target and background in aerial images, and salient areas mostly comprise artificial buildings and roads. When the proportion of the building or road to the image is extremely large, the effect of local features weighting using the salient model is unsatisfactory. To a certain extent, the category differentiation of image scenes is reduced, thereby weakening the final expressive capability of the image semantic features. SIM saliency maps [36], which are based on mathematical and statistical methods, adopt biological bottom visioning using multi-scale weight optimization for feature extraction. The method effectively captures landform information in aerial images, and the multi-scale weight optimization method can well adapt to landform regions with different scales. Therefore, the capability to express image features is enhanced to a great extent.

The influence of feature weights (i.e., extracted from different saliency maps) on the scene recognition of aerial images is shown in Table 2. Filter template size is specified as 9 × 9. According to the four typical saliency models, feature weighting can be designed to complete local feature modeling. In this study, SIM is used for the proposed algorithm.

Table 2. Influence of saliency models on recognition accuracy.

Saliency Model	Null	Itti	Erdem	Achanta	SIM
Recognition Accuracy	88.9%	92.1%	94.2%	85.4%	95.1%

3.2.3. Comparison of Feature Performances

This study evaluates and compares the expression capability of the proposed superpixel-based features with the six commonly used features of Bag-of-Words modeling (i.e., dense SIFT, dense HOG, dense LBP, dense Gabor, Harris interest points, and FAST interest points). The final result is measured in terms of scene recognition accuracy and time consumption to extract the local features of each image, as shown in Table 3.

Table 3. Comparison and analysis of local feature description methods.

Local Feature Type	Recognition Accuracy	Time Consumption
Dense SIFT	78.3%	16.86 s
Dense HOG	79.1%	15.23 s
Dense LBP	82.6%	15.96 s
Dense Gabor	72.8%	45.75 s
Harris Interest Points	73.5%	0.73 s
FAST Interest Points	78.2%	0.53 s
Proposed Feature	**95.1%**	21.57 s

Dense local features were compared with interest point local features. All of the dense features, except that of dense Gabor, showed stronger expression capability than those of interest points. The condition can be explained by the sparsely distributed local features of the interest points, which resulted in limited image information. The information loss effect is further expanded after quantifying the Bag-of-Words model, the analysis of which showed a decline in semantic feature expression capability. Gabor features can comprehensively extract low-level image information using time and frequency domains. The Gabor filter sifts dense image blocks in several directions and scales in the dense modeling of local image regions; however, redundant information is generated, and this phenomenon weakens feature expression.

The capability of dense local features to describe image content is weaker than that of the proposed feature. The phenomenon can be explained by features (i.e., dense local features) that are mainly extracted by traversing image blocks, and this approach cannot accurately capture local surface areas. This dense local feature method also often uses a single feature descriptor to model the local area of images. However, the local content of the image cannot be fully covered; thus, the capability to generate image semantic features is weaker than the method proposed in this study.

The number of extracted dense local features is much larger than the number of interest point features because aerial image sizes are frequently large. Therefore, in terms of time consumption, the extraction of dense local features consumes much more time than those of interest point features. With regard to local feature modeling, image segmentation and saliency weighting extraction are required by the proposed method, and these additional processes may lead to relatively high time complexity. However, final results verify the high recognition accuracy of the proposed method. Moreover, the time consumption of the proposed method has not increased significantly unlike the dense SIFT, dense HOG, and dense LBP methods. The time consumption of the proposed method is even lower than that of dense Gabor features.

4. Conclusions

The superpixel-based feature description method proposed in this study can be applied to image scene recognition for aerial remote sensing applications. The image scene recognition algorithm based on the Bag-of-Words model is designed using the proposed feature. The following conclusions can be drawn from the experiments: (1) The proposed superpixel feature can be extracted adaptively by using the landform characteristics of aerial images. Image saliency, which is highly adaptable to aerial images, is introduced into the algorithm to complete the local area modeling. (2) The scene recognition accuracy of aerial images based on the proposed feature is 95.1%, and this result is higher than those of scene recognition algorithms based on other local features.

In the future, the proposed method can be improved and applied to other types of remote images, such as satellite images. Finally, the method may also be transplanted to embedded systems.

Acknowledgments: This work was supported by the National Key Research and Development Project (No. 2016YFB1200100), the Open Projects Program of National Laboratory of Pattern Recognition, and Shenzhen peacock plan.

Author Contributions: Hongguang Li wrote the program and the manuscript. Baochang Zhang revised the paper. Yang Shi and Yufeng Wang performed the experiments and analyzed the data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Teng, H.; Rossel, R.A.V.; Shi, Z.; Behrens, T.; Chappell, A.; Bui, E. Assimilating satellite imagery and visible-near infrared spectroscopy to model and map soil loss by water erosion in australia. *Environ. Model. Softw.* **2016**, *77*, 156–167. [CrossRef]
2. Chen, C.; Zhang, B.; Su, H.; Li, W.; Wang, L. Land-use scene classification using multi-scale completed local binary patterns. *Signal Image Video Process.* **2016**, *10*, 745–752. [CrossRef]
3. Tong, X.; Liu, X.; Chen, P.; Liu, S.; Luan, K.; Li, L.; Liu, S.; Liu, X.; Xie, H.; Jin, Y. Integration of UAV-based photogrammetry and terrestrial laser scanning for the three-dimensional mapping and monitoring of open-pit mine areas. *Remote Sens.* **2015**, *7*, 6635–6662. [CrossRef]
4. Peña, J.M.; Torressánchez, J.; Serranopérez, A.; de Castro, A.I.; Lópezgranados, F. Quantifying efficacy and limits of unmanned aerial vehicle (UAV) technology for weed seedling detection as affected by sensor resolution. *Sensors* **2015**, *15*, 5609–5626. [CrossRef] [PubMed]
5. Chrétien, L.P.; Théau, J.; Ménard, P. Visible and thermal infrared remote sensing for the detection of white-tailed deer using an unmanned aerial system. *Wildl. Soc. Bull.* **2016**, *40*, 181–191. [CrossRef]
6. Getzin, S.; Nuske, R.S.; Wiegand, K. Using unmanned aerial vehicles (UAV) to quantify spatial gap patterns in forests. *Remote Sens.* **2014**, *6*, 6988–7004. [CrossRef]
7. Kakooei, M.; Baleghi, Y. Fusion of satellite, aircraft, and UAV data for automatic disaster damage assessment. *Int. J. Remote Sens.* **2017**, *38*, 1–24. [CrossRef]
8. Moranduzzo, T.; Melgani, F.; Bazi, Y.; Alajlan, N. A fast object detector based on high-order gradients and gaussian process regression for UAV images. *Int. J. Remote Sens.* **2015**, *36*, 2713–2733.
9. Huang, Y.; Cao, X.; Zhang, B.; Zheng, J.; Kong, X. Batch loss regularization in deep learning method for aerial scene classification. In Proceedings of the Integrated Communications, Navigation and Surveillance Conference, Herndon, VA, USA, 18–20 April 2017.
10. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 20–27 September 1999.
11. Lowe, D.G.; Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
12. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; pp. 886–893.
13. Csurka, G.; Dance, C.R.; Fan, L.; Willamowski, J.; Bray, C. Visual categorization with bags of keypoints. *Workshop Stat. Learn. Comput. Vis. ECCV* **2004**, *44*, 1–22.
14. Willamowski, J.; Arregui, D.; Csurka, G.; Dance, C.R.; Fan, L. Categorizing nine visual classes using local appearance descriptors. *ICPR Workshop Learn. Adaptable Vis. Syst.* **2004**, *17*, 21.
15. Olshausen, B.A.; Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **1996**, *381*, 607. [CrossRef] [PubMed]

16. Felzenszwalb, P.; Mcallester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
17. Gao, X.; Hua, G.; Niu, Z.; Tian, Q. Context aware topic model for scene recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2743–2750.
18. Zang, M.; Wen, D.; Wang, K.; Liu, T.; Song, W. A novel topic feature for image scene classification. *Neurocomputing* **2015**, *148*, 467–476. [CrossRef]
19. Grauman, K.; Darrell, T. The pyramid match kernel: Discriminative classification with sets of image features. In Proceedings of the Tenth IEEE International Conference on Computer Vision, Beijing, China, 17–21 October 2005; Volume 1452, pp. 1458–1465.
20. Kim, J.; Liu, C.; Sha, F.; Grauman, K. Deformable spatial pyramid matching for fast dense correspondences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; Volume 9, pp. 2307–2314.
21. Harris, C. A combined corner and edge detector. *Alvey Vis. Conf.* **1988**, *15*, 147–151.
22. Rosten, E.; Porter, R.; Drummond, T. Faster and better: A machine learning approach to corner detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2009**, *32*, 105–119. [CrossRef] [PubMed]
23. Ahonen, T.; Hadid, A.; Pietikainen, M. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef] [PubMed]
24. Content-Based Image Retrieval: A Report to the JISC Technology Application Program. Available online: http://www.leeds.ac.uk/educol/documents/00001240.htm (accessed on 24 October 2017).
25. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274. [CrossRef] [PubMed]
26. Zhao, Z.; Jiao, L.; Hou, B.; Wang, S.; Zhao, J.; Chen, P. Locality-constraint discriminant feature learning for high-resolution SAR image classification. *Neurocomputing* **2016**, *207*, 772–784. [CrossRef]
27. Liang, Z.; Li, Y.; Shi, P. A note on two-dimensional linear discriminant analysis. *Pattern Recognit. Lett.* **2008**, *29*, 2122–2128. [CrossRef]
28. Liyu, G. *Research on Lie Group Based Image Generalized Gaussian Feature Structure Analysis*; Huazhong University of Science & Technology: Wuhan, China, 2012.
29. Wei, J.; Jianfang, L.; Bingru, Y. Semi-supervised discriminant analysis on Riemannian manifold framework. *J. Comput. Aided Des. Comput. Graph.* **2014**, *26*, 1099–1108.
30. Wang, Q.; Li, P.; Zhang, L.; Zuo, W. Towards effective codebookless model for image classification. *Pattern Recognit.* **2016**, *59*, 63–71. [CrossRef]
31. Li, P.; Wang, Q.; Hui, Z.; Lei, Z. Local log-Euclidean multivariate Gaussian descriptor and its application to image classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 803–817. [CrossRef] [PubMed]
32. Belhumeur, P.N.; Hespanha, J.P.; Kriegman, D.J. Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. In Proceedings of the European Conference on Computer Vision, Cambridge, UK, 14–18 April 1996.
33. Itti, L.; Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2001**, *2*, 194–203. [CrossRef] [PubMed]
34. Erdem, E.; Erdem, A. Visual saliency estimation by nonlinearly integrating features using region covariances. *J. Vis.* **2013**, *13*, 11. [CrossRef] [PubMed]
35. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009.
36. Murray, N.; Vanrell, M.; Otazu, X.; Parraga, C.A. Low-level spatiochromatic grouping for saliency estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 2810–2816. [CrossRef] [PubMed]