IET Research Journals

IET Journals
The Institution of Engineering and Technology

A Novel Bag of Words KAZE (BoWK) with Two-Step Classification for High Resolution Remote Sensing Images

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

Usman Muhammad[1,2] *Weiqiang Wang*[1] *Abdenour Hadid*[2] *Shahbaz Pervez*[3]

[1] *School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing, China*
[2] *Center for Machine Vision and Signal Analysis (CMVS), University of Oulu, Finland*
[3] *Yanbu University College, Saudi Arabia*
** E-mail: usman@mail.bnu.edu.cn*

Abstract: The bag-of-words(BoW) model has been widely used for scene classification in recent state-of-the-art methods. However, inter-class similarity among scene categories and very high spatial resolution imagery makes its performance limited in the remote-sensing domain. Therefore, this research presents a new KAZE based image descriptor that makes use of the BoW approach to substantially increase classification performance. Specifically, a novel multi-neighborhood KAZE is proposed for small image patches. Secondly, the spatial pyramid matching (SPM) and bag-of-words representation can be adopted to use the extracted features and make an innovative Bag of Words KAZE (BoWK) descriptor. Third, two bags of multi-neighborhood KAZE features are selected in which each bag is regarded as separated feature descriptors. Next, canonocial correlation analysis (CCA) is introduced as feature fusion strategy to further refine the BOWK features, which allows a more effective and robust fusion approach than the traditional feature fusion strategies. Experiments on three challenging remote-sensing data sets show that the proposed BoWK descriptor not only surpasses the conventional KAZE descriptor, but also yields significantly higher classification performance than the state-of-the-art methods used now. Moreover, the proposed BoWK appraoch produces rich informative features to describe the scene images with low computational cost and much lower dimension.

1 Introduction

The currently available instruments (e.g., multi/hyper spectral, synthetic aperture radar, etc.,) for earth observation not only provide high spatial resolution remote-sensing images but also allow us to study the ground surface in greater detail. However, the large-scale nature, intra-class variability or inter-class similarity makes the classification task very challenging. Different scene categories may share some identical thematic classes. For instance, images from harbor and beach, which are two different scene categories, may both consist of ships, water, and trees at the same time but differ in the density and spatial distribution of these three thematic classes [1]. Under this situation, the spatial information plays a key role in the analysis and understanding of remotely sensed datasets. Infact, it is more semantically meaningful to collect the information from different parts of the images. For example, 'sky' in scene category exists on the upper side of images, while 'water' lies on the lower part of the images. In order to encode local features expeditiously, we argue this spatial information is important and should be combined efficiently to get better performance. In this regard, selecting suitable features for scene classification is a crucial task for researchers and the existing methods can be generalized into three main classes, namely: low-level visual features, mid-level visual features and methods based on high-level visual information. Low-level feature approaches are based on shape, color or textual information, and the most popular descriptors are scale invariant feature transform (SIFT), color histogram (CH), local binary pattern (LBP) and GIST [2]. Although these approaches have been applied successfully in different applications, but the pixelwise information or only object cannot accomplish the entire scene understanding due to high-diversity and non-homogenous spatial distributions.

Mid-level feature-based approaches attempt to develop a global scene representation based on the distribution of the features, and bring the necessary flexibility to cope with deformations. A well-known approach is the bag-of-words model. It was first introduced for text analysis and then extended to represent images by the frequency of "visual words" using a clustering scheme (k-means) [3]. In order to improve the k-means clustering codebook, several variant coding methods such as multi-task joint sparse coding [4], fisher kernel (FK) framework [5], kernel collaborative representation [2], and multi-scale CLBP (MS-CLBP) descriptor [6] have been introduced in the BOW model for improving the reconstruction accuracy of local features. On the other hand, local descriptor such as KAZE [7], which is handcrafted and designed to describe 2D features in a nonlinear scale space by means of nonlinear diffusion filtering, is generally utilized in many applications, but how to merge with BoW effectively and efficiently still remains a challenge. The work reports in [42] that KAZE features are invariant to rotation, scale, and have more distinctiveness at varying scales with the cost of moderate increase in computational time. Therefore, this research attempts to explore KAZE with three basic steps in the BoW pipeline for scene classification: extracting KAZE features based on the proposed multi-neighborhood strategy, constructing the codebook and encoding KAZE features on the codebook.

High-level methods are usually based on deep learning models, which have gained great popularity due to stack of learned convolutional filters. Since the introduction of ImageNet Large-Scale Visual Recognition Challenge (ILSVRC), the pre-trained convolutional neural networks (CNNs) such as AlexNet [27], VGG-Net [25], GoogLeNet [31], etc., provide the scene classification as an end-to-end problem. However, the backpropagation process, the stochastic gradient descent (SGD) strategy or training from scratch is a highly time consuming process, and limited training samples often cause an overfitting issue in deep learning model.

Recently, the BoW model has been investigated by using the spatial information based on the local features [8]. However, the traditional SIFT-based BoW model uses the Gaussian scale space framework, and Gaussian derivatives for deriving the Gaussian kernel for scale-space smoothing, but the Gaussian blurring does not repect the distinctive boundaries of objects, smoothes information and noise to the same extent at all scale levels, which causes decrease

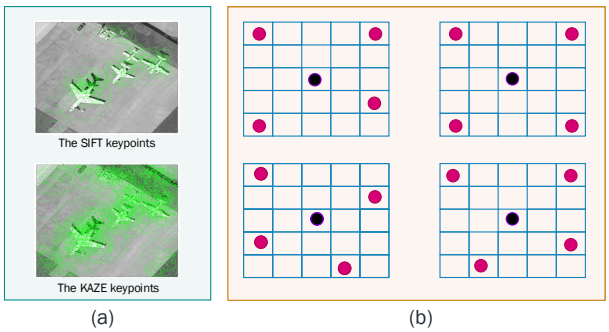

Fig. 1: (a) Displays the KAZE keypoints which are densely distributed while the SIFT keypoints around the regions. (b) Displays the four 4-neighborhood masks used for computing KAZE descriptor.

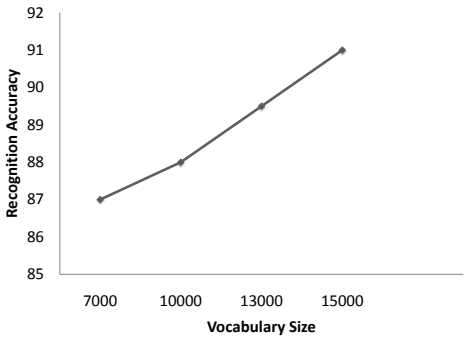

Fig. 2: Effect of changing the vocabulary size.

in localization accuracy and distinctiveness [9]. To alleviate this limitation, KAZE features are proposed which blur the noise and keep the details or edges at the same time. Our main contributions include four aspects.

(1) The multi-neighborhood strategy is proposed to compute KAZE features from each image.

(2) An iterative keypoint selection algorithm [2] is selected to ignore unhelpful keypoints, which may have a negative effect on computational efficiency and image representation.

(3) Two bags of KAZE features are selected in which each bag is regarded as separated feature descriptors.

(4) We introduce the canonical correlation analysis as a feature fusion strategy to fuse KAZE features in a BoW model, which helps us to achieve an improved classification performance.

In the rest of the work, we explain the proposed framework in Section II, and present experimental results in Section III. Finally, in Section IV, we draw conclusions for the proposed framework.

2 The Proposed Methodology

First, we give a brief introduction about KAZE descriptor, and then explain the process of computing the proposed Bag of Words KAZE (BoWK) descriptor with canonical correlation analysis.

2.1 KAZE Features

The main difference between traditional SIFT-base BoW descriptor and the proposed KAZE is the construction of the scale space. KAZE features are extracted on non-linear scale space while SIFT is based on Gaussian scale space (GSS). Equation 1 shows the classic nonlinear diffusion formulations [7]:

$$\frac{\partial L}{\partial t} = div(c(x,y,t).\nabla L) \tag{1}$$

where div and ∇ are respectively the divergence and gradient operators. c is the conductivity function which depends on the local image differential structure while t is scale parameter. c is dependent on the gradient magnitude as shown in (Eq.(2)), helping in the reduction of diffusion at edges, and encouraging smoothing within a region instead of smoothing across boundaries, thus resulting in more smoothening of regions as compared to edges.

$$c(x,y,t) = g(\mid \nabla L_\sigma(x,y,t) \mid) \tag{2}$$

where σ is the amount of blur, and the luminance function ∇L_σ is the gradient of a Gaussian smoothed version of the original image L. This property of the conductivity function preserves the boundary and reduces image noises in BoW model. The KAZE approach is more sensitive to the original image resolution, without applying any downsampling at each new octave as done in SIFT. From Fig.1(a), it can be seen that KAZE not only focus on objects keypoints but also concentrate at the boundary.

2.2 Dense Sampling: A modified KAZE for Small Image Patches

The first step while computing the proposed BoWK descriptor is to choose sampling strategy (e.g. densely, randomly, using a keypoint detector). The sampler is a vital element of any bag-of-feature methods. Researchers are using multiscale keypoints detectors ((Laplacian of Gaussian, SIFT, Harris-affine, etc.) as samplers to select regions of interest within the image but surprisingly the randomly sampled or dense patches are often more discriminant than keypoint based sampling methods [10]. In the proposed framework, the image is partitioned into several equal sized blocks using a uniform gird and each block is treated as a separate region for feature extraction. In order to improve classification performance, overlapping image blocks are proposed. Suppose an image S be represented by a set of x_i (KAZE) at P locations placed with their indices $i = 1, \cdots, P$. L the regions of interest with P_m defining the set of locations/indices inside the region m. Let q and g represent some coding (vector quantization) and sampling operators, respectively. The vector v expressing the whole image, which is extracted by sequentially coding, sampling over all regions, and concatenating [11]:

$$\alpha_i = q(x_i), i = 1, \cdots, P \tag{3}$$

$$h_m = g(\{\alpha_i\}_{i_\in P_m}), m = 1, \cdots, L \tag{4}$$

$$v^T = [h_1^T \cdots h_L^T] \tag{5}$$

The target is to decide which operators q and g provide adequate classification performance using v as input to either a non-linear classifier (SVM), or a liner classifier. q minimizes the distance to a codebook, formed by an unsupervised algorithm (K-means), and g computes the average over the sampling region:

$$\alpha_i \in \{0,1\}^K, \alpha_{i,j} = 1, iff j = argmin_{k \leq K} \parallel x_i - d_k \parallel_2^2 \tag{6}$$

$$h_m = \frac{1}{\mid P_m \mid} \sum_{i \in P_m} \alpha_i, \tag{7}$$

where d_k defines the k-th codeword. Different strategies have been proposed to select discriminant neighborhoods using local binary patterns (LBP) [12], [13]. In our work, we propose a four 4-pixel neighborhoods strategy as shown in Fig.1(b) to generate the multi-neighborhood KAZE descriptor. If the traditional KAZE is applied to small image patches, it will become more sensitive to distortions and consumes more time. To alleviate this problem, four smaller

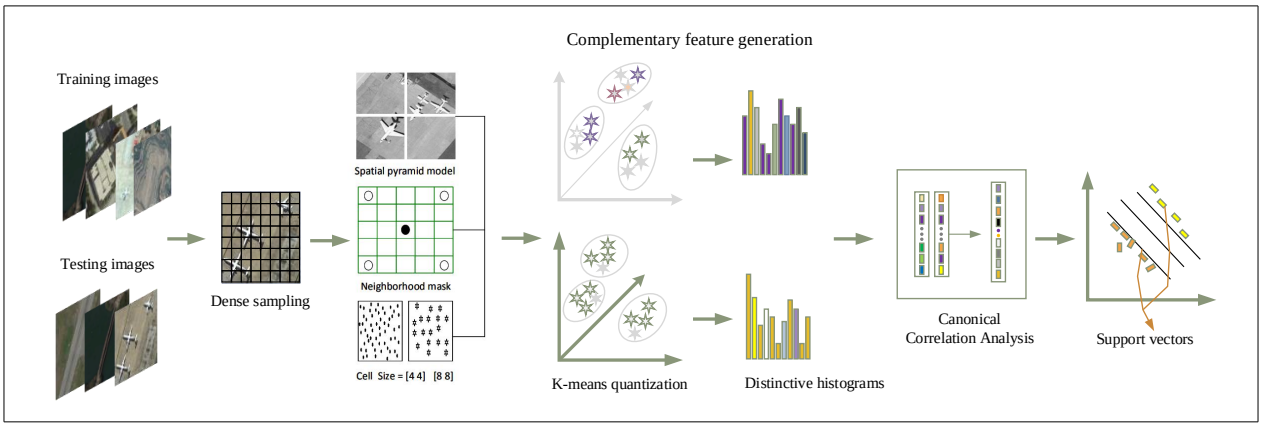

Fig. 3: Overall architecture of the proposed method.

neighborhoods of four pixels each are utilized. These four neighborhoods, resulting in 128-dimensional feature vector, which is used to describe each image patch. The 128-dimension of improved descriptor may seem high, but it performs better than the lower-dimensional ones based on the results of the experiments.

2.3 Histogram of Visual Words

To obtain a compact representation of the scene images, the popular k-means clustering method is proposed to code the extracted features into a visual vocabulary without losing too many details. The vocabulary size is very crucial because an inappropriate choice of k may yield poor results. It varies from a few hundreds to several thousands. If the codebook is too small, different dissimilar local regions could be merged to the same visual word, which can limit the discrimination of local features. On the other hand, if we keep the codebook size too high, many similar local regions could be mapped to different visual words. In order to find the optimum vocabulary size, we have compared with different sizes to get the final vocabulary size . The results for UC Merced dataset are shown in fig.2. We used a relatively larger codebook size for all three datasets (10000 for WHU-RS, 150000 for UC Merced, 180000 for NWPU-RESISC45), because we want to increase the amount of information by using more dimensions for canonical correlation analysis space which may result in increasing the accuracy. Moreover, the variance of classification performance over different sizes of vocabulary is 1% to 1.5% which proves that the proposed approach is not so sensitive to size of vocabulary.

During the construction of the visual vocabulary, each patch in an image is mapped to a specific codeword through the k-means clustering process and the image, thus, can be represented by a histogram of visual words. The histogram becomes a feature vector for the image..

In order to introduce spatial information, the scheme proposed by Lazebnik $et\ al.$ [14] which is based on spatial pyramid matching is proposed. An image is tiled into a number of smaller rectangular blocks and the proposed KAZE is computed for each block and concatenated. In our work, we follow only the second level of this pyramid to keep the computational complexity low.

2.4 Iterative Keypoint Selection

To remove unhelpful keypoints, an iterative keypoint selection method is proposed in [2]. Authors use response value with neighboring keypoints to reflect the saliency of keypoints. Thus, keypoints will remove according to Equation (8).

$$D(\hat{X}) = D + \frac{1}{2}\frac{\partial D^T}{\partial X}\frac{\partial^2 D^{-1}}{\partial X^2}\frac{\partial D}{\partial X} < \theta \tag{8}$$

where D is the function value [15] in the location of keypoints and $X = (x, y, \sigma)$ is the offset of keypoints. This strategy helps us to choose discriminative keypoints which are not different from neighboring keypoints. The representative keypoints are those which are closest to the cluster center. Keypoints whose Euclidean distance in KAZE feature space are within a threshold T of those representative keypoints will be removed. This is how the first iteration is performed. On the basis of the first iteration, the initial keypoints are selected and will be used in the second iteration and the process will be the same as in the first iteration. The iteration performs continually until no keypoints will be filtered out or rest of the keypoints are insufficient to be clustered.

2.5 Canonical Correlation Analysis

Feature fusion is the process of combining two or more feature vectors to obtain a single feature vector, where the fused features contain rich information to describe the image scene well. How to properly integrate feature vectors is always a challenge. In order to extract more discriminant descriptors to represent the structure of scene images, different methods have been introduced in the literature for feature fusion. Among them, two famous approaches are serial feature fusion and parallel strategy . Serial fusion strategy [39] simply fuses two feature vectors into a single one. Let's consider that a and b are two features obtained from an input image with c,d vector dimension, respectively, and then the concatenated feature is v with size equal to $(c + d)$.

Parallel strategy [40], [41] that concatenates the two features vector into a complex vector. Each input image I generated two sets of features, that is, A_1 and A_2 comprising two sets of features. The final fused feature representation is formulated as

$$A_a(I) = A_1(I) + iA_2(I) \tag{9}$$

where i is the imaginary unit.

Sun $et\ al$ [16] introduces a robust feature fusion method based on canonical correlation analysis (CCA). This method establishes the correlation criterion function between the two groups of feature vectors, to extract their canonical correlation features. The dimension of the fused vector will be equal to or less than the dimension of the two vectors. It has some resemblance to principal component analysis (PCA) and linear discriminant analysis (LDA), but PCA searches for patterns only within single multivariate data. In order to perform CCA fusion successfully, dimension of two features set should be same. Suppose that $X \in R^{p\times n}$ and $Y \in R^{q\times n}$ represent two matrices, each consists of n training feature vectors from two different sets. p and q are the dimensions of each vector.

Let's assume that $S_{xx} \in R^{p\times p}$ and $S_{yy} \in R^{q\times q}$ contain the within-sets covariance matrices of X and Y and $S_{xy} \in R^{p\times q}$ contains the between-set covariance matrix (note that $S_{yx} = S_{xy}^T$). The overall covariance matrix $(p + q) \times (p + q)$ is then computed:

$$S = \begin{pmatrix} cov(x) & cov(x,y) \\ cov(y,x) & cov(y) \end{pmatrix} = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} \tag{10}$$

Table 1 Number of average remaining keypoints in our method and other methods

Method	UC MERCED Dataset	WHU-RS Dataset
BoWK without keypoint selection	6,879,392±10,255	8,039,120±12,665
IKS [21]	356,952±10,403	575,168±29,110
Class-Specific Codebook [2]	403,857±11,259	635,497±14,227
The proposed Method	**260,934±6,500**	**320,000±9,130**

It is complicated [17] to follow the relationship between these two sets of vector from matrix S because these feature vectors may not follow a consistent pattern. The objective of CCA is to find the linear combinations, $\overset{*}{X} = W_x^T X$ and $\overset{*}{Y} = W_y^T X$, which maximizes the pair-wise correlations across the two feature sets:

$$corr(\overset{*}{X}, \overset{*}{Y}) = \frac{cov(\overset{*}{X}, \overset{*}{Y})}{var(\overset{*}{X}).var(\overset{*}{Y})} \tag{11}$$

Where $cov(\overset{*}{X}, \overset{*}{Y}) = W_x^T S_{xy} W_y, var(\overset{*}{X}) = W_x^T S_{xx} W_x$ and $var(\overset{*}{Y}) = W_y^T S_{yy} W_y$. Maximization is conducted by maximizing the covariance between $\overset{*}{X}$ and $\overset{*}{Y}$ using Lagrange multipliers subject to satisfy the following constraints $var(\overset{*}{X}) = var(\overset{*}{Y}) = 1$. Both transformation matrices, W_x and W_y, are then computed by using the eigenvalue equations:

$$\begin{cases} S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \hat{W}_x = R^2 \hat{W}_x \\ S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} \hat{W}_y = R^2 \hat{W}_y \end{cases} \tag{12}$$

where $\hat{W}_x$ and $\hat{W}_y$ are the eigenvectors and R^2 is the diagonal matrix of eignvalues or it could be defined as squares of the canonical correlations. The number of non-zero eignvalues can be find in each equation, that is $d = rank(S_{xy} \leq min(n, p, q)$, which will be fixed in descending order, $r_1 \geq r_1 \geq \cdots \geq r_d$. As mentioned earlier, both the transformation matrices, W_x and W_y, composed of the sorted eigenvectors corresponding to the non-zero eigenvalues. $\overset{*}{X}$,$\overset{*}{Y} \in R^{d \times n}$ are consider as canonical variates. It could be observed that the sample covariance matrix denoted in Eq.(9) will be of the form:

$$\overset{*}{S} = \begin{pmatrix} 1 & 0 & \cdots & 0 & r_1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 & 0 & r_2 & \cdots & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & \\ 0 & 0 & \cdots & 1 & 0 & 0 & \cdots & r_d \\ - & - & - & - & - & - & - & - \\ r_1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & r_2 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & & \vdots & & \ddots & \\ 0 & 0 & \cdots & r_d & 0 & 0 & \cdots & 1 \end{pmatrix} \tag{13}$$

The matrix explains that the canonical variates have nonzero correlation only on their corresponding indices. It also express that the canonical variates are uncorrelated within each other because of identity matrices in the upper left and lower right corners. Hence, it is possible to perform feature-level fusion either by concatenation or summation of the transformed feature vectors:

$$Z_1 = \begin{pmatrix} \overset{*}{X} \\ \overset{*}{Y} \end{pmatrix} = \begin{pmatrix} W_x^T X \\ W_y^T Y \end{pmatrix} = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{14}$$

or

$$Z_2 = \overset{*}{X} + \overset{*}{Y} = W_x^T X + W_y^T Y = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{15}$$

where Z_1 and Z_2 are called the Canonical Correlation Discriminant Features (CCDFs). The overall proposed framework is given in Algorithm 1.

Table 2 Comparison in average computational time for vector quantization

Dataset	The proposed Method	Class-Specific Codebook [2]	IKS [21]
UC Merced	**60.14±1.2 min**	135±2.5 min	478±3.3 min
WHU-RS	**90.57±2.5 min**	221±2.9 min	630±5.1 min

Table 3 Overall classification accuracy (%) of each feature set

Method	UC MERCED Dataset	WHU-RS Dataset	NWPU dataset
KAZE with four neighborhoods	91.85 ±1.20	92.60±0.50	62.80±0.80
KAZE with three neighborhoods	89.80 ±0.90	91.27±0.80	60.40±1.20
The proposed BOWK (Fusion by addition)	97.52±0.80	99.47±0.60	66.87±0.90

Table 4 Description of the fused features used for final classification

Dataset	Size
WHU-RS	802
UC Merced	1680
NWPU	6299

2.6 Two-Step classification description

In order to employ canonical correlation analysis, two distinctive visual words vocabularies based on four neighborhoods using the size of 13×13 with 4 pixels spacing from each block, and three neighborhoods using the size of 11×11 with 8 pixels spacing from each block are used. Fig.3. describes the feature fusion procedure, where features are extracted from the input scene image. Then, two distinctive visual words vocabularies are selected and their transformations were calculated based on CCA. After that, we combine (concatenation/addition) the transformed features to represent the input images by single informative features. The overall BoW scene classification framework is given in Algorithm I.

Algorithm I : BOWK Algorithm.
Input :
1: The training category set $S_{train} = \{S_{tr1} \cdots S_{trP}\}$;
2: The testing category set $S_{test} = \{S_{ts1} \cdots S_{tsP}\}$;
visual dictionary D;
3: **for** $i = 1$ to P **do**
4: Extract dense features from S_{train} and S_{test} using multi-neighborhoods KAZE features;
X_{htrain}=$\{x_{htrain_1}, x_{htrain_2}, .., x_{htrain_n}\}$,
X_{htest}=$\{x_{htest_1}, x_{htest_2}, .., x_{htest_m}\}$;
5: Keypoints selection, T;
6: Generation of codebook using k-means clustering;
7: Select C_{train1} and C_{test1} using four 4-pixel neighborhoods $(X_{htrain}, gridsize = 4)$;
8: Select C_{train2} and C_{test2} using three 8-pixel neighborhoods $(X_{htrain}, gridsize = 8)$;
9: Compute their transformation matrix W_x and W_y;
10: Project C_{train1} , C_{train2} into the CCA subspace. $C_{train1} = W_x * C_{train1}, C_{train2} = W_y * C_{train2}$;
11: Project C_{test1} , C_{test2} into the CCA subspace. $C_{test1} = W_x * C_{test1}, C_{test2} = W_y * C_{test2}$;
12: Fuse C_{train1}, C_{train2} and C_{test1} , C_{test2} by concatenation.
13: **end for**
14: Perform SVM classification
15: **Output**: Overall Accuracy

3 Datasets And Experimental Setup

In this section, we first describe three datasets, which are used to evaluate our approach. Then, the parameters settings of all experiments are defined. Finally, the results are compared for each dataset and discussed.

3.1 Datasets

The proposed 'NWPU-RESISC45' dataset, represents 31,500 high resolution images, classified into 45 scene classes, covering more than 100 countries and regions all over the world. This benchmark

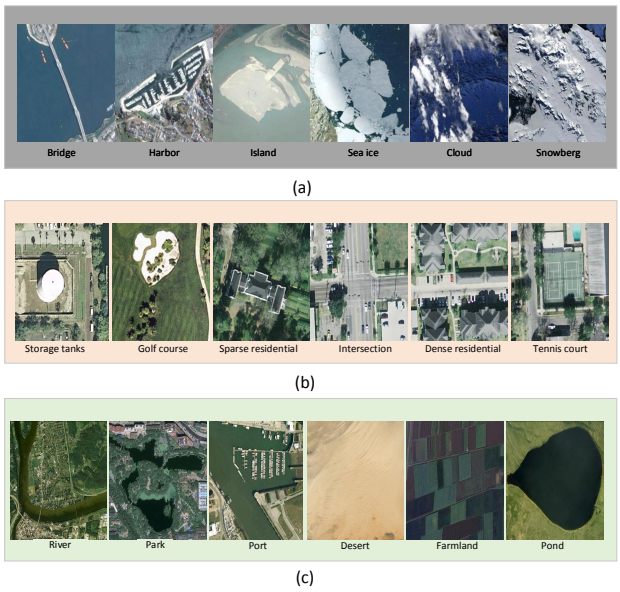

Fig. 4: The images in the first row(a), second row(b) and thirds row(c) belong to six different classes of the 'NWPU-RESISC45 dataset, the UC Merced (UCM) dataset, and the WHU-RS dataset, respectively.

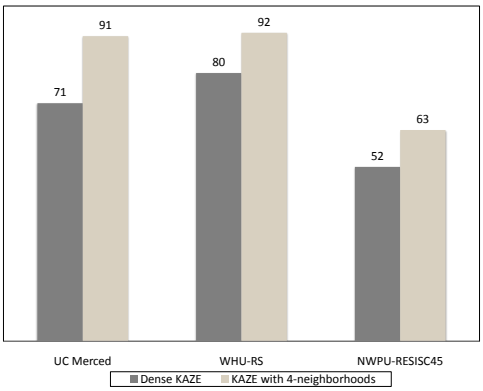

Fig. 5: The proposed KAZE with 4-neighborhoods, the dense KAZE and their mean average classification performance on three datasets.

is recently introduced [18], and designed to alleviate the overfitting issue in deep learning models, where each class consists of 700 images with the size of 256×256 pixels. The images are obtained from Google Earth (Google Inc.), where the spatial resolution ranges from 30 to 0.2 m per pixel. This dataset is consider as the largest satellite scenes images dataset so far, and 15 times larger than the most common and widely-used UC Merced dataset. Hence, the rich image variations, highly overlapping classes, and the large scale make the dataset rather challenging. Fig.4(a), displays some images in the first row.

The second dataset is the UC Merced (UCM) dataset [19], which is manually extracted from the USGS National Map Urban Area Imagery collection for various urban areas around the country with a pixel resolution of one foot. It is classified into 21 classes, and each class is composed of 100 images with size of $256 \times 256 \times 3$. Inter-class similarity among categories, for instance, images from beach and river can be easily mixed with other each, which make this dataset a challenging one. Some example images are shown in the second row of Fig.4 (b).

The third proposed dataset, WHU-RS dataset [20] , is comparatively smaller, and formed from satellite images of Google Earth. It contains 950 images with 19 scene classes, and a size of 600×600. Images are vary greatly in high depth of field, scale and orientation, which makes it more complicated than the above datasets. Fig.4 (c) displays some scenes images in the third row.

By observing the sample images in Fig. 4(b), we find the great similarity between the dense residential, intersection, and the tennis court categories. Fig. 4(a) shows some of the sea images, which causes the classification to become very challenging even for a human. Except for the desert and farmland categories in Fig.4(c), other categories such as river, park, port, and pond also share similar thematic classes. In summary, these data sets are challenging for the BoW model.

3.2 Experimental Setup

To analyze the scene classification performance on the mentioned datasets, the color images are converted to grayscale images. For the UC Merced and WHU-RS datasets, the linear SVM classifier is trained on a set of 80% images per category for training and remaining 20% images for testing. The SVM classifier is trained on Matlab by using statistics and machine learning toolbox with one-vs-all. As described earlier, two bags of KAZE features are selected as two different feature sets and the fusion is performed into two principal phases: in the first phase the transformation matrices W_x and W_y are calculated and projected the training feature sets into the CCA subspace, and then fusion is performed by addition on two transformed features sets. In the second phase, the testing feature sets are projected into CCA subspace and fusion is performed on the two transformed testing feature sets by addition. All the experiments are performed in MATLAB 2017b by using an Intel Core i7-4370 (3.80 GHz) computer with a 8 GB of RAM memory, and experiments are repeated five times to obtain convincing classification performance.

For the keypoint selection algorithm, we choose a threshold of response value 0.080, a bit higher than [2], and keeps the same for all datasets. Table 1 represents the average number of remaining keypoints which are considered in our approach and also compared with baseline methods [21] [2] with their standard deviation. It can be observed, multi-neighborhood strategy gives us large number of keypoints from the training set , 6,879,392 and 8,039,120 keypoints over UC Merced and WHU-RS dataset, respectively. Although, the proposed method aims a slightly lower selection rate, these keypoints are more discriminant and achieve more than 90% accuracy on two proposed datasets.

However, as can be seen in table 2, a large number of keypoints were selected by [21], [2], but these approaches take a lot more computational time to complete the vector quantization with a little increase in classification accuracy. Hence, the proposed approach is low cost, since the algorithm removes indiscriminative keypoints.

3.3 Comparative Assessment of the KAZE, the BOWK and state-of-the-art approaches

To make a fair comparison with traditional KAZE, first, we compare the performance without using feature fusion strategy. The classification results in different visual vocabularies summarized in Table 3. It could be observed that KAZE with four neighborhoods performs better in comparison with most of the recent works such as [5], [22], [34], [35], [36], [37], [24], [18], [6] and [38]. Moreover, using feature fusion strategy based on CCA achieves improvement beyond the state-of-the-art (e.g., CNN-based methods) with a small size of feature descriptor, which equals 802 for WHU-RS dataset, 1680 for UC Merced dataset, and 6299 for NWPU dataset. The features size based on CCA is reported in Table 4. Further detail is provided in Fig.5, where the highest classification rate for the UC Merced, WHU-RS, and NWPU dataset is 91%, 92% and 63%, respectively, which is 14%, 12% and 11% higher than traditional dense KAZE based BoW model. The proposed KAZE with four neighborhoods produces decent classification performance on all the datasets and outperforms the KAZE by a fair margin. In Fig.6, the category wise

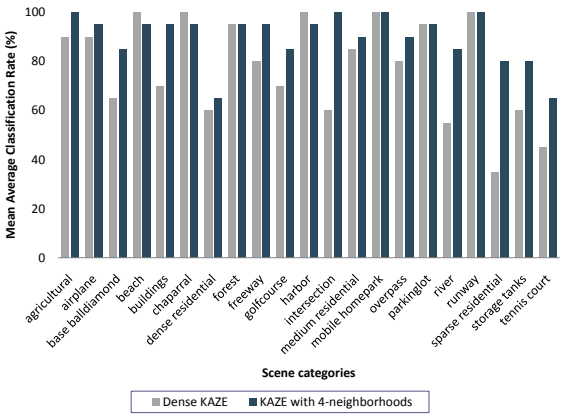

Fig. 6: The comparative mean average classification performance of the proposed KAZE with 4-neighborhoods and the dense KAZE descriptor on the 21 categories of the UC Merced dataset.

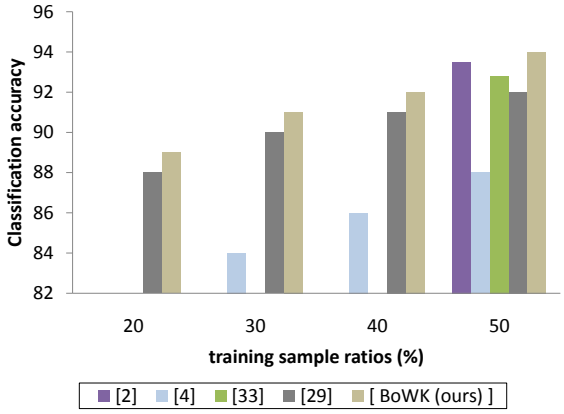

Fig. 7: The influence of the training sample ratios with BoWK and previous methods for the UCM dataset.

classification rates for the UC Merced dataset are expressed. It could be observed that the most confusion occurs in the "dense residential", "intersection", "river", "sparse residential", "storage tanks", "golf course" and "tennis court" categories. Due to high similarity of tennis court and dense residential images, it makes it challenging for BoW model to classify efficiently. Except for the dense residential category, the proposed KAZE with four neighborhoods out performs the dense KAZE based BoW model by a fair margin.

A comparative assessment with the baseline BoW approaches and deep learning methods is made after using feature fusion strategy described earlier to evaluate the classification performance. The number of training samples are taken as a key factor to demonstrate our approach. A competent classification approach can achieve a fine performance even with fewer training samples. Based on this logic, the size of the training samples is decreased. To make a comparison with previous methods on UC Merced dataset, we randomly select 20%, 30%, 40% and 50% as training samples, and remaining for test. It could be seen from Fig.7, that proposed framework on UC Merced dataset is superior in comparison with state-of-the arts approaches right from the start (20% sample ratios). Infact, a substantial difference can be observed even with fewer training samples, and the accuracy is increased with the addition of training samples, exceeding 90% with just 30% training ratios.

In Table 5, results from some state-of-the-art methods are shown, and summarized here for a comparison analysis. The work reported

Table 5 Overall classification accuracy (%) of reference and the proposed BoWK on the UC-Merced dataset and WHU-RS dataset with 80% ratios.

Methods	UC-Merced	WHU-RS
LPCNN [34]	89.90	-
CCNN [36]	91.56	-
GoogLeNet + fine-tuning [31]	97.10	96.14
GoogLeNet [33]	92.80±0.61	93.00
D-CNN with AlexNet [27]	96.67±0.10	-
D-CNN with GoogLeNet [27]	97.07±0.12	-
UFL [30]	95.71±0.13	-
FBC [1]	85.53±1.24	-
FK-S [5]	91.63±1.49	-
FV+HCV [22]	91.80±1.30	-
SDSAE [44]	93.57±1.02	-
S-UFL [35]	82.72±1.18	-
Fusion strategy 1 (GoogLeNet) [45]	96.17±0.90	-
Fusion strategy 2 (GoogLeNet) [45]	97.12±0.96	-
OverFeat [37]	90.91±1.19	-
Partlets-based method [43]	91.33±1.11	-
SPP-net+MKL [24]	96.38±0.92	95.07±0.79
Fusion by addition [23]	97.42±1.79	98.70±0.22
KCRC [2]	93.80±0.58	93.70±0.57
MTJSLRC [4]	91.07±0.67	91.74±1.14
MS-CLBP1 [6]	90.60±1.40	93.30±0.80
CaffeNet [25]	95.02±0.81	96.24±0.56
VGG-VD-16 [25]	95.21±1.20	96.05±0.91
D-DSML-CaffeNet [26]	96.76±0.36	96.64±0.68
MLF [28]	89.62±1.67	88.16±2.76
AlexNet-SPP-SS [29]	96.67±0.94	95.00±1.12
salM^3LBP-CLM [32]	95.75±0.80	96.38±0.76
BoWK (ours)	**97.52±0.80**	**99.47±0.60**

Table 6 Overall classification accuracy (%)of reference and the proposed BoWK on the NWPU-RESISC45 dataset with 20% ratios.

Methods	NWPU-RESISC45
BoVW [18]	44.97±0.21
BoVW+SPM [18]	32.96±0.47
LLC [18]	40.03±0.34
BoVW with dense SIFT [38]	44.97±0.28
AlexNet [38]	59.22±0.18
BoWK (ours)	**66.87±0.90**

in [1], presents a fast binary coding scheme (FBC) for global feature representations using randomly-sampled image patches. To generate a class-specific codebook, an improved class-specific codebook using kernel collaborative representation based classification (KCRC) is proposed [2]. Multiple features, e.g., shape, color and textual features, are used in [4]. Then, a multi-task joint sparse and low-rank representation is adopted to combine the features. The fisher kernel (FK) coding framework is introduced to extend the BOVW model in [5], by characterizing the low-level features with a gradient vector. Authors report in [6], introducing the completed local binary patterns (CLBP) operator for the first time on remote sensing land-use scene classification. A new method based on hierarchically coding structures is introduced in [22], where multiple bags of visual words (BoVW) coding layers and one fisher coding layer is used to develop the coding structure. Then, semi-local features were encoded with fisher vectors and aggregated to form a final global representation. A large patch convolutional neural network (LPCNN) is introduced in [34] , where authors replace the fully-connected layer with global average pooling layer to decrease the kernels parameters. An unsupervised feature learning approach to extract patches based on a saliency detection algorithm is proposed in [35]. A new convolutional neural network for dealing with the scale variation of the patterns in the scenes is developed [36]. An effective partlets based method is proposed in [43], and further training is performed to enhance the VHR image land-use classification. The solution for automatic semantic annotation is proposed by [44], based on a unified annotation framework by combining discriminative high-level feature (sparse autoencoder) learning and weakly supervised feature transferring.

Basically, feature fusion methods are showing great popularity in deep learning models. The most relevant work is performed by [23], [24] and [45]. In [23], discriminant correlation analysis (DCA) is introduced showing that a feature fusion with few features can be performed efficiently by fusing two fully connected layers of VGG-Net architecture. However, the authors claimed that the main drawback of CCA is to ignore the class information. Our proposed approach uses manually labeled datasets. We aim to use CCA to combine two bags of KAZE features into a single one,

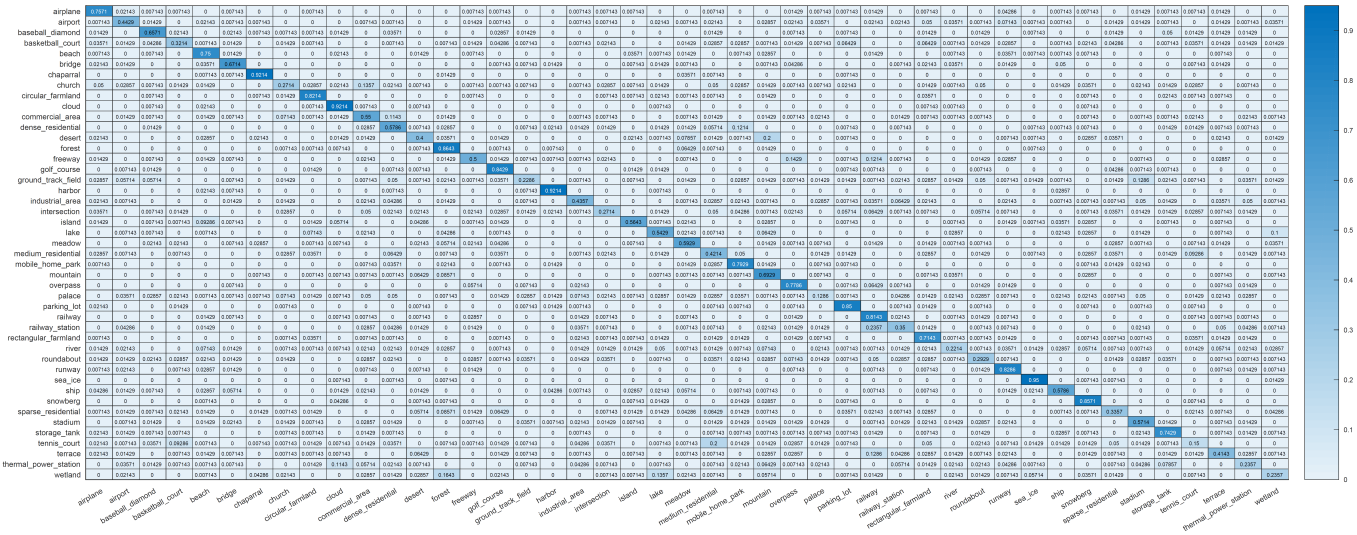

Fig. 8: Confusion matrix for the BoWK with NWPU-RESISC45 dataset.

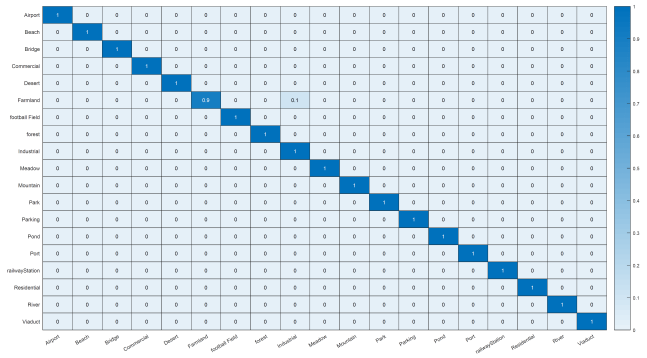

Fig. 9: Confusion matrix for the BoWK with WHU-RS dataset.

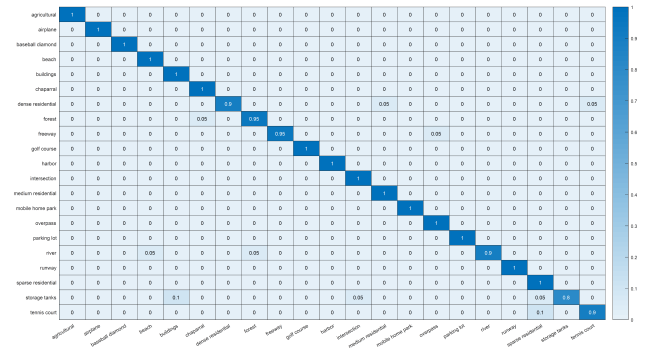

Fig. 10: Confusion matrix for the BoWK with UC Merced dataset.

and to maximize the correlation. Therefore, the class structures are preserved and fusion can be performed either by addition or concatenation. CCA increases the performance by more than four percent on each dataset. We also compared our proposed approach against serial and parallel feature fusion strategies. With regard to the UC-Merced dataset, the classification results in Table 5 once again show that the proposed feature fusion strategy is better than the serial and parallel feature fusion strategies using the GoogLeNet architecture. The authors in [45] also used other deep learning models such as CaffeNet, VGG-Net-16 to perform fusion.

VGG-Net-16 and CaffeNet seem to give a better performance than our proposed approach by using parallel feature fusion strategy. However, we should note that all these VGG-Net-16 and CaffeNet are pre-trained models trained on ImageNet whose images are all natural images. Thus, the pre-trained neural networks seem to be more suitable for handling natural images. This means that CCA could be further explored by fusing convolutional or fully-connected layer features in deep learning models. The Alex-Net is explored with spatial pyramid pooling (SPP-net), and transfer learning is performed to ensures the effectiveness of each layer. In order to fuse the multi features, the multi-kernel learning is used [24], where authors extract features from different layers of a pre-trained model and then fuse them to get the final classification.

The work presented in [25], attempts to tune the weights of CaffeNet using fine-tuning approach based on VGG-VD-16 architecture. Metric learning (ML) [26], [27], has been utilized frequently into the convolutional neural models to further increase the discrimination of deep representations. The mid-level features were extracted via using sparse autoencoder in [28]. To increase the depth of convolutional layers, the side supervision strategy (SS) is proposed for AlexNet model [29]. Other approaches including, an unsupervised representation for deconvolutional networks [30], fine-tuned GoogLeNet [31], fine-tuned CaffeNet with VGG-VD-16 [25], fusion of local and global features [32], OverFeat [37] and six fine-tuned ConvNets [33]. Making a comparison with above methods, the proposed approach achieves the best accuracy (99.47%) for WHU-RS dataset using 80% samples as training data and obtained an impressive accuracy (97.52%) for UC Merced dataset. For NWPU-RESISC45 dataset, we select 20% training sample ratios for training and rest for testing as same in [18]. Table 6, shows the performance comparison of the BoWK with baselines models such as BoVW [18], BoVW+SPM [18], LLC [18] and AlexNet [38]. The proposed framework competes low-level or mid-level based approaches with fair margin of over 20%. This is a significant improvement for NWPU-RESISC45 dataset. Hence, the proposed BoWK framework achieves very competitive accuracy in the literature of scene classification when compared with low-level based approaches, high-level methods, and deep learning frameworks.

For further analysis, a confusion matrix of NWPU-RESISC45 dataset, UC Merced dataset and WHU-RS dataset is shown in Fig.8, Fig.9, and Fig.10 respectively. We use heat map function in Matlab to visualize the confusion matrix. The rows and columns of the matrix represent the actual and predicted classes. The class labels range is 1:21 for UC Merced dataset, 1:19 for WHU-RS dataset and 1:45 for NWPU dataset. The vertical color bar indicates the proportion of samples over the actually total class samples. The storage tanks category in UC Merced dataset, which is hard to be classified because of inter-class similarity with sparse residential, and achieving lower accuracy. From the confusion matrix of NWPU-RESISC45 dataset, It can be observe that other classes such as wetland, thermal power station, tennis court, river, roundabout,

medium residential, industrial area, commercial area, church, or airport are easily confused due to similar structures and background color. In summary, these datasets are challenging, even though we have achieved a comparable performance.

4 Conclusion

Since the introduction of pre-trained CNN, the focus is shifting to high-level semantic features rather than acquiring low-level and mid-level features. However, limited training samples, the stochastic gradient boosting or training a convolutional neural network (CNN) from scratch is a highly time consuming process. Therefore, this paper focuses on a BoW approach for remote-sensing scene classification, and introduces a novel neighborhood strategy to compute two distinct KAZE based BoW features, with the objective that the learned encodings are maximumly correlated. The original BoW model discards the spatial information. In this regard, the proposed approach not only overcomes the spatial information problem but also takes advantage of canonical correlation analysis (CCA) to maximize the correlation, and to combine them into a final feature set. In comparison with related fusion methods, the proposed fusion strategy(CCA) proved to be more robust than other fusion methods. Moreover, the proposed framework is low cost and has more discriminative features with low dimension. This unsupervised learning approach is well suited for off-line classification, where the classification accuracy is the prime goal. The proposed framework could be used in various applications where deep learning models are not easy to train such as biometrics (face recognition), Chinese character recognition, etc. It is also computational efficient as it could be trained without GPU. However, this approach falls behind real-time requirements and feature fusion process reduce the diversity of feature representations. In future work, these challenges should be further investigated.

Acknowledgment

The authors would like to acknowledge the world academy of sciences (TWAS) for giving an opportunity to conduct this research at University of Chinese Academy of Sciences.

5 References

1 Hu, Fan, et al. "Fast binary coding for the scene classification of high-resolution remote sensing imagery." Remote Sensing 8.7 (2016): 555.
2 Yan, Li, et al. "Improved Class-Specific Codebook with Two-Step Classification for Scene-Level Classification of High Resolution Remote Sensing Images." Remote Sensing 9.3 (2017): 223.
3 Jiang, Yu-Gang, Chong-Wah Ngo, and Jun Yang. "Towards optimal bag-of-features for object categorization and semantic video retrieval." Proceedings of the 6th ACM international conference on Image and video retrieval. ACM, 2007, pp. 494-501
4 Qi, Kunlun, et al. "Multi-Task Joint Sparse and Low-Rank Representation for the Scene Classification of High-Resolution Remote Sensing Image." Remote Sensing 9.1 (2016): 10.
5 Zhao, Bei, et al. "The Fisher kernel coding framework for high spatial resolution scene classification." Remote Sensing 8.2 (2016): 157.
6 Chen, Chen, et al. "Land-use scene classification using multi-scale completed local binary patterns." Signal, image and video processing 10.4 (2016): 745-752.
7 Alcantarilla, Pablo FernÃąndez, Adrien Bartoli, and Andrew J. Davison. "KAZE features." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2012, pp. 214-227.
8 Li, Hongguang, et al. "Superpixel-Based Feature for Aerial Image Scene Recognition." Sensors 18.1 (2018): 156.
9 Liu, Yuxuan, et al. "S-AKAZE: An effective point-based method for image matching." Optik-International Journal for Light and Electron Optics 127.14 (2016): 5670-5681.
10 Nowak, Eric, FrÃŀdÃŀric Jurie, and Bill Triggs. "Sampling strategies for bag-of-features image classification." Computer VisionâĂŞECCV 2006 (2006): 490-503.
11 Boureau, Y-Lan, et al. "Learning mid-level features for recognition." Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, 2559-2566.
12 Banerji, Sugata, Abhishek Verma, and Chengjun Liu. "Novel color LBP descriptors for scene and image texture classification." 15th International Conference on Image Processing, Computer Vision, and Pattern Recognition, Las Vegas, Nevada. 2011, p. 1.
13 Gu, Jiayu, and Chengjun Liu. "Feature local binary patterns with application to eye detection." Neurocomputing 113 (2013): 138-152.
14 Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." Computer vision and pattern recognition, 2006 IEEE computer society conference on. Vol. 2. IEEE, 2006, pp. 2169-2178.
15 Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.
16 Sun, Quan-Sen, et al. "A new method of feature fusion and its application in image recognition."Pattern Recognition 38.12 (2005): 2437-2448.
17 Haghighat, Mohammad, Mohamed Abdel-Mottaleb, and Wadee Alhalabi. "Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition." IEEE Transactions on Information Forensics and Security 11.9 (2016): 1984-1996.
18 Cheng, Gong, Junwei Han, and Xiaoqiang Lu. "Remote sensing image scene classification: benchmark and state of the art." Proceedings of the IEEE 105.10 (2017): 1865-1883.
19 Yang, Yi and Newsam, Shawn. "UC MERCED DATASET." , http://vision.ucmerced.edu/datasets/landuse.html (accessed on November 20 2017)
20 Xia, Gui-Song and Yang, Wen and Delon, Julie and Gousseau, Yann and Sun, Hong and Maître, Henri. "SIRI-WHU Dataset.", http://www.lmars.whu.edu.cn/xia/AID-project.html (accessed on November 20 2017)
21 Lin, Wei-Chao, et al. "Keypoint selection for efficient bag-of-words feature generation and effective image classification." Information Sciences 329 (2016): 33-51.
22 Wu, Hang, et al. "Hierarchical coding vectors for scene level land-use classification." Remote Sensing 8.5 (2016): 436.
23 Chaib, Souleyman, et al. "Deep feature fusion for VHR remote sensing scene classification." IEEE Trans. Geosci. Remote Sens 55.8 (2017): 4775-4784.
24 Liu, Qingshan, et al. "Learning multiscale deep features for high-resolution satellite image scene classification." IEEE Transactions on Geoscience and Remote Sensing 56.1 (2018): 117-126.
25 Xia, Gui-Song, et al. "AID: A benchmark data set for performance evaluation of aerial scene classification." IEEE Transactions on Geoscience and Remote Sensing 55.7 (2017): 3965-3981.
26 Gong, Zhiqiang, et al. "Diversity-Promoting Deep Structural Metric Learning for Remote Sensing Scene Classification." IEEE Transactions on Geoscience and Remote Sensing 56.1 (2018): 371-390.
27 Cheng, Gong, et al. "When deep learning meets metric learning: remote sensing image scene classification via learning discriminative CNNs." IEEE transactions on geoscience and remote sensing 56.5 (2018): 2811-2821.
28 Li, Erzhu, et al. "Mid-level feature representation via sparse autoencoder for remotely sensed scene classification."IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10.3 (2017): 1068-1081.
29 Han, Xiaobing, et al. "Pre-Trained AlexNet Architecture with Pyramid Pooling and Supervision for High Spatial Resolution Remote Sensing Image Scene Classification." Remote Sensing 9.8 (2017): 848.
30 Lu, Xiaoqiang, Xiangtao Zheng, and Yuan Yuan. "Remote sensing scene classification by unsupervised representation learning." IEEE Transactions on Geoscience and Remote Sensing 55.9 (2017): 5148-5157.
31 Castelluccio, Marco, et al. "Training convolutional neural networks for semantic classification of remote sensing imagery." Urban Remote Sensing Event (JURSE), 2017 Joint. IEEE, 2017, pp. 1-4.
32 Bian, Xiaoyong, et al. "Fusing local and global features for high-resolution scene classification." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 10.6 (2017): 2889-2901.
33 Nogueira, Keiller, OtÃąvio AB Penatti, and Jefersson A. dos Santos. "Towards better exploiting convolutional neural networks for remote sensing scene classification." Pattern Recognition 61 (2017): 539-556.
34 Zhong, Yanfei, Feng Fei, and Liangpei Zhang. "Large patch convolutional neural networks for the scene classification of high spatial resolution imagery." Journal of Applied Remote Sensing 10.2 (2016): 025006.
35 Zhang, Fan, Bo Du, and Liangpei Zhang. "Saliency-guided unsupervised feature learning for scene classification." IEEE Transactions on Geoscience and Remote Sensing 53.4 (2015): 2175-2184.
36 Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.
37 Penatti, OtÃąvio AB, Keiller Nogueira, and Jefersson A. dos Santos. "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?." Computer Vision and Pattern Recognition Workshops (CVPRW), 2015 IEEE Conference on. IEEE, 2015,pp. 44-51.
38 Cheng, Gong, et al. "Remote sensing image scene classification using bag of convolutional features." IEEE Geoscience and Remote Sensing Letters 14.10 (2017): 1735-1739.
39 Liu, Chengjun, and Harry Wechsler. "A shape-and texture-based enhanced Fisher classifier for face recognition." IEEE transactions on image processing 10.4 (2001): 598-608.
40 Yang, Jian, and Jing-yu Yang. "Generalized KâĂŞL transform based combined feature extraction." Pattern Recognition 35.1 (2002): 295-297.
41 Yang, Jian, et al. "Feature fusion: parallel strategy vs. serial strategy." Pattern recognition 36.6 (2003): 1369-1381.
42 Tareen, Shaharyar Ahmed Khan, and Zahra Saleem. "A comparative analysis of SIFT, SURF, KAZE, AKAZE, ORB, and BRISK." Computing, Mathematics and Engineering Technologies (iCoMET), 2018 International Conference on. IEEE, 2018, pp. 1-10.
43 Cheng, Gong, et al. "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images." IEEE Transactions on Geoscience and Remote Sensing 53.8 (2015): 4238-4249.

44 Yao, Xiwen, et al. "Semantic annotation of high-resolution satellite images via weakly supervised learning." IEEE Transactions on Geoscience and Remote Sensing 54.6 (2016): 3660-3671.
45 Yu, Yunlong, and Fuxian Liu. "A Two-Stream Deep Fusion Framework for High-Resolution Aerial Scene Classification." Computational intelligence and neuroscience 2018 (2018).