Satellite image classification using HOG and DAISY Feature Descriptors

Themba Ngobeni
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
1334236@students.wits.ac.za

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

***Abstract*—Advancements in remote sensing techniques provide crucial information for a variety of applications, including landscape changes, land cover categorization, enhanced weather forecasting, and climate observation. With the recent breakthroughs that have made satellites both more powerful and cheaper to deploy, the volume of high-resolution satellite images collected has increased exponentially. Manual classification techniques of these high-resolution satellite images are too inaccurate and inefficient to handle the challenge. Hence, automation of satellite image classification has become a crucial part of the field of remote sensing. The major purpose of this work is to evaluate the ability of a feature extraction model to identify satellite pictures utilizing different approaches. This study presents actual data demonstrating that the Bag of Feature (BoF) approach outperforms conventional feature extraction algorithms for multilabel scene classification. We employ local feature descriptors such as DAISY features for local feature extraction and classification, and global feature descriptors such as the histogram of oriented gradients (HOG) for global feature extraction and classification, using the Bag of Features technique. The bag of feature encoding is augmented by the Mini-Batch K-Means method to decrease the complications of the feature encoding technique. A method for pooling is applied to combine DAISY and HOG data relevant to each image. Finally, we perform a 10-fold cross-validation experiment with a support vector machine (SVM) classifier and classify our results on a dataset of 21 scene categories. The hybrid model attained an accuracy of 81.76% and the KNN model attained an accuracy of 67.14%. The study demonstrates how this categorization method may be used to identify changes in land use and land cover, as well as aid in the improvement of satellite classification.**

***Index Terms*—Bag of Feature(BoF)/Bag of Visual Words(BoVW),Feature extraction,Convolutional Neural Network (CNN), histogram of oriented gradients (HOG), Support vector machine(SVM)**

I. Introduction

Computer vision enables computers to extract important data from digital images, videos, and other visual inputs. Its purpose is to allow computer systems to accurately recognize objects and their properties such as dimension, textures, colours, size, and spatial configuration in order to offer the most comprehensive description of the visual data. This is done by using feature extraction which is a critical component of the computer vision pipeline.

With the recent advancements in space travel and technology in the past several decades, remote sensing has become one of the most efficient techniques for surveying the Earth at global spatial scales. These satellite observations enable constant monitoring of the the atmosphere, the land, and the ocean. Furthermore, satellite devices can detect hazardous or dangerous conditions without endangering people or equipment. Large-scale continuous satellite observations reinforce comprehensive field observations by providing an unrivaled quantity of information for scientific data analytic and integration. At regular intervals, satellite remote sensing devices acquire high spatial resolution land images. The amount of data received at data centers is enormous, and it is expanding exponentially as technology advances at a quick pace, and data volumes have grown at an exponential rate [1]. Effective and efficient procedures for extracting and interpreting relevant information from high spatial resolution land images are desperately needed. Satellite image classification is a useful approach for extracting information from massive amounts of satellite imagery. The bag of features technique, which is developed from text mining algorithms, has shown excellent efficiency on image classification.

The approach in this work comprises primarily of processes for the extraction of features and classification. We enhance this conventional approach by developing an algorithm that combines both local and global descriptors at distinct levels of resolution with the purpose of boosting classification performance. As the extracted feature descriptors involved would be extremely considerable in application, we developed the bag of features technique employing the Mini-Batch K-Means method rather than the usual K-Means approach to considerably minimize the computational time complexity of our learning phase. We assess the methods by using ground truth UC Merced land-use image data comprising 21 land-use categories. To merge DAISY and HOG features from single image, we use an L2 pooling technique. In a cross-validation setting, we train a multi-class Support Vector Machine (SVM) classifier using several kernel parameters and track the classification outcomes.

The main objective of the research paper is to explore

and enhance image identification techniques using the bag of features method,and improve the accuracy and overall effectiveness of the algorithm proposed by [2] mainly for object recognition in satellite imagery. The necessity for efficient, accurate, and reliable autonomous classification models of enormous quantities of images into diverse categories in massive databases is the main drive for this research. This goal will be attained by using just the retrieved features and disregarding any meta-data.

The following is a breakdown of how this document is structured: Section III examines the contributions of machine learning in the realm of scene categorization. Section IV demonstrates the different design processes and their impacts on boosting the current effectiveness of our scene recognition approach. In Section V, we study the effect of vocabulary size on the encoding stage as well as the implications of alternative SVM kernel configurations along with variables, and we describe the experimental results. In Section V, we review the important observations in light of our results. Section VI summarizes the research.

II. Background

There have been several experiments where the researchers studied how supervised algorithms behaved in object recognition. When there is a large volume of input, neural networks produce good models [3]. The present research is being conducted with a comparatively limited data collection. As a result, CNNs are not thought to be the best approach for this analysis. Various supervised algorithms were tested using a variety of efficiency metrics in [4]. According to the findings of this study, logistic regression performed worse than other models [5] compared k Nearest Neighbors and Support Vector Machines for image classification. Support Vector Machines was found to outperform k Nearest Neighbors(kNN) by a small margin. Based on these results, the supervised learning algorithm to be applied was chosen as Support Vector Machine. The modified Seeded Region Growing method for ground cover segmentation was established by [6] For the classification of land cover, the Seeded Region Growing scheme used the details of the the MS edge and spectral bands of the sharpened satellite image. The adjoining solution does not provide positive outcomes on all input pictures due to the dependence on the threshold.

[7] proposed a multi functional support vector machine set model that combines several spatial, spectral and pixel level functions. For the combination of multiple properties three algorithms were created, including the semantic object-based solution, the vote for certainty and probabilistic fusion. The performance characteristics demonstrated that the Support vector machine Ensemble model built outperforms the existing voting and probabilistic classification accuracy models.

A. Extraction of features

The initial goal is to extract raw features from each picture in the collection using data descriptors. Clusters may be represented as numerical vectors using feature representation algorithms. These vectors are known as feature descriptors. This step of the overall framework minimizes data dimensionality by eliminating redundant data.

B. Codebook generation

The descriptors generated in the feature extraction are subsequently converted to codewords. Codewords are simple vector representations of clusters that are identical. That is, a K-means clustering technique or a similar approach is utilized to assign features to the closest terms in the vocabulary. The codewords create a codebook (a codebook is a collection of codewords).

Method	Ref.	Year	Accuracy
BOVW+SCK	[8]	2010	77.71%
Dirichlet	[9]	2013	92.80%
VLAD	[10]	2014	94.30%
DCNN+SIFT BOVW	[11]	2018	95%
Inception-v3-CapsNet	[12]	2020	80%
Proposed		2021	81.76%

TABLE I
CLASSIFICATION ACCURACY (%) OF REFERENCE AND PROPOSED METHODS ON THE UC-MERCED DATASET.

C. Development of feature vectors

In this last phase, we represent each image to build a normalized histogram of codewords, which means the codewords are represented as a frequency. The resulting codeword histograms will then be employed to develop a classification method. Implementing the Bag of Features model entails a number of potential considerations. Some of these problems are addressed in Section IV.

III. Research Methodology

In our experiments, we construct an object recognition method that includes extraction of features, encoding, pooling, and classification of images. The kernel used by the Linear Support Vector Machine (SVM) classifier is given as K, and SVM hyperparameters such as the C parameter and the Gamma parameter ($gamma$) are all configurable parameters in our model.To scientifically choose the appropriate hyperparameters, we employ 10-fold cross-validation, which facilitates the determination of a more precise estimate of model prediction performance. For a linear SVM, we first determine the optimal K by experimentation. We take this K and then use a grid search to find the best C that corresponds to an SVM model with a popular kernel function called the Radial Basis Function. The following sections details each phase of our process.

A. Satellite Image Acquisition

The UC Merced land-use data, which is a collection of aerial pictures with a high pixel density obtained from the United States Geological Survey (USGS) National Map Urban Area Imagery collection, is explored in this study. Many research

papers have utilized the UC Merced dataset as a baseline for land-use classification assessment. The UC Merced land-use dataset contains 21 classes of aerial photography with a range of spatial patterns, some with color presentation. A hand selection of 100 photographs was generated to build the dataset, with each of the photos having around 256 pixels for each RGB class. Each shot is 256x256 pixels in size, with each photo having a 30 cm dynamic range.

B. Feature Extraction

DAISY descriptor is a novel image local feature descriptor recently proposed by [13]. The main principle behind DAISY involves combining numerous scale Gaussian filter algorithms to combine multiple gradient maps of the source picture. This descriptor demonstrates the efficiency and efficacy of Gaussian filter functions in dense matching tasks since they are separable. We then present a more explicit specification of our DAISY descriptor. For each input image I, we initially compute H the number of orientation maps, the number of orientation maps generated using pixel difference is given by

$$G_o = (\frac{\partial I}{\partial o})^+, 1 \leq o \leq H \tag{1}$$

After that, the orientation maps are combined with numerous Gaussian filter functions at various Σ values to produce combined orientation maps for various areas. These combined orientation maps are indicated by the following equation

$$G_o{}^{\Sigma} = G_{\Sigma} * (\frac{\partial I}{\partial o})^+ \tag{2}$$

Since the Gaussian filters are separable, the combined orientation maps could be computed in a short period of time to lessen the computation cost. In other words, a larger Gaussian filter Σ_2 is generated by convolution with a smaller one Σ_2 using the following procedure.

$$G_o{}^{\Sigma_2} = G_{\Sigma_2} * (\frac{\partial I}{\partial o})^+ = G_{\Sigma} * G_{\Sigma_1} * (\frac{\partial I}{\partial o})^+ = G_{\Sigma} * G_o{}^{\Sigma_1} \tag{3}$$

HOG descriptor for the whole picture at several granularity's (using the parameters pixels per cell, cells per block, and orientations), essentially enabling us to select features from various dimensions. We utilize Mini-Batch KMeans to cluster the DAISY descriptors taken from all training photos to create a bag-of-visual-words. Cross-validation is used to identify the optimal visual vocabulary size (K).

C. Support Vector Machine

The SVM is a supervised learning algorithm that distinguishes between data points and is formally characterized as a hyperplane in an N-dimensional space (N = the number of features). The SVM finds an ideal hyperplane that categorizes subsequent samples given a set of training data. In linear SVM, the hyperplane is learned by converting the problem using some linear algebra . For a linear kernel, the following equation is used to predict a new input using the dot product of the input (x) and each support vector (x_i), where$B(0)$ and a_i are coefficients estimated from the training data:

$$f(x) = B(0) + sum(a_i * (x * x_i)) \tag{4}$$

The Polynomial kernel is denoted by:

$$K(x, x_i) = 1 + sum(x * x_i)^d \tag{5}$$

At which polynomial degree should be manually provided to the learning method. This is the same as the linear kernel if d is set to be d=1. The Radial kernel is denoted by:

$$K(x, x_i) = exp(-\gamma * sum(x - {x_i}^2)) \tag{6}$$

γ is a learning parameter that must be specified. A reasonable default value for γ is 0.1, where γ is frequently $0 < \gamma < 1$. The radial kernel is relatively local and may construct complex areas inside the feature space, such closed polygons in 2-D space

D. K - nearest neighbors algorithms

The euclidean distance was selected as the distance measure for this technique; given a query data point q, the distance to any other point in the training set X may be computed as

$$D(q, X) = \sqrt{(X - q)^t (X - q)} \tag{7}$$

The ideal value for k may be found by doing a parameter sweep in a suitable search space.

E. Encoding

To encode each image as a histogram of visual words, we use the local DAISY features corresponding to each keypoint. In this case, we'll employ the normal "bag-of-visual-words" technique. To put it another way, we use the K-means algorithm which finds key trends by grouping comparable DAISY features together into 'K' clusters to create "visual words" in a vocabulary. The K-means algorithm finds k centroids and then assigns each data point to the vocabulary closest to it, keeping the centroids as compact as possible. The size of the vocabulary is represented by the letter K. Using this vocabulary, we create a histogram corresponding to each image, with 'K' as the dimensionality. The "DAISY histogram" is an encoded representation of the image that makes up part of our hybrid feature descriptor.

F. Pooling

The arrangement of the features in the input is sensitive to the location of the features in the output feature maps, which is an issue. Down-sampling the feature maps is one method of dealing with this sensitivity. This has the impact of generating the down sampled feature maps more robust towards variations of the feature in the image, which is referred to as "local translation invariance" in technical terms. By summarizing the existence of features in portions of the feature map, pooling layers provide a method for down sampling feature maps. In our scene identification pipeline, we adopt a two-level pooling method. By encoding the frequency of every visual word within each picture, we generate a histogram from the DAISY features. This is accomplished by selecting each key point in the picture and retrieving the cluster id for that DAISY descriptor in our "DAISY histogram," then incrementing the count for that bin. This is essentially a "sum

pooling" approach. The resulting histogram is L2 normalized to generate a DAISY histogram feature, which we refer to as "DAISY histogram." To build our hybrid feature descriptor, we extract the HOG global descriptor for each picture and perform level 2 pooling normalization before appending it with the accompanying DAISY histogram feature.

G. Classification

For classification, we employ the traditional SVM classifier with various kernels and K - nearest neighbors algorithms. For both KNN and SVM implementations, we use the sklearn framework . We use K fold cross validation to partition the dataset into a training and validation set at random with the value of K=10. From the training split, we create the "visual vocabulary" as well as training feature vectors. The total accuracy, confusion matrix, and typical information retrieval statistics such as precision, recall, and fmeasure for our trials are reported.

Where appropriate, the experiments were carried out to establish parameters of the model. The training datasets were used in all of the experiments. For the value of k on the KNN model, the specified search space is given by 3, 5, 7, 13. In the search space, we conduct a parameter sweep and analyze the training error and accuracy for each unique value of k. Table II below shows the training errors and accuracy.

K	3	5	7	13
Training Error	0.273	0.303	0.323	0.379
Accuracy	72.6%	69.6%	67.6%	62.0%

TABLE II
SHOWS THAT IN THE GIVEN SEARCH SPACE, THE OPTIMAL VALUE OF K IS 3.

IV. Results

In this section, we report experimental results from image classification on 21 land-use categories. We ran a series of tests to analyze the effectiveness of the classification technique as well as compare it to existing findings. We employ the UC Merced Land Use dataset, [8] , comprising of a collection of high-pixel-density aerial photographs. Many studies have utilized this dataset in recent years, allowing for broad comparison of observations with the research Table I show some of the studies. All experiments in this paper were carried out on Google Colaboratory, a GPU-accelerated Python environment that may be used for up to 12 hours in a single session for research and teaching. Google Colaboratory makes use of a Tesla K80 GPU with 25 GB of memory, a single-core 2.3 GHz Intel Xeon CPU, and 12 GB of RAM.

Following the implementation of the scene classification algorithm proposed above, the optimal value of visual vocabulary is experimentally determined by running the algorithm three times for each K and comparing their performances. For classification, a Support Vector Machine (SVM) architecture was implemented. An ideal k has been determined to be 600. We performed a search to select the most appropriate value of the cost of misclassification, known as (C), and the parameter of a Gaussian kernel known as γ (gamma) by altering the logarithm of both C and γ within the domain of $-3 \leq \log(C) \leq 3$ and $-3 \leq log(\gamma) \leq 2$ respectively, with the k value of 600 after reaching the optimal K using a linear SVM. The optimum values for $\log(C)$ and $log(\gamma)$, are 1.679 and -0.163, respectively. For optimal K, C, and γ, the Hybrid algorithm is ran using a radial basis function kernel , and the accuracy was 81.42%. The algorithm managed to achieve an average precision of 85% , with both the average recall and f1-score accuracy being 85%. The kNN approach and optimal K of 3 obtained an accuracy of 68.02% ,with an average precision of 73% , and both the average recall and f1-score of 71%. The linear kernel approach and optimal K obtained an efficiency of 76.19% ,with an average precision of 84% , and an average recall and f1-score of 83% and 83% respectively. This shows that non-linear support vector machines enhance the performance of the algorithm. We observed that using just HOG descriptors in the linear support vector machine algorithm results in an accuracy of 36.73%, which is the worst of all.

Classifier	Accuracy
Hybrid Classifier	81.42%
KNN	68.02%
Linear Classifier	76.19%
HOG only	36.73%
DAISY only	73.40%

TABLE III
SHOWS ACTUARIES OF THE CLASSIFIERS

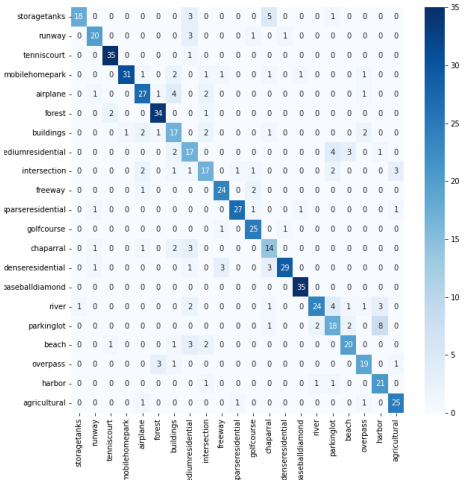

Fig. 1. Confusion Matrix: Support vector machine (SVM) with DAISY features and a linear kernel

Different algorithms for satellite imagery classification have recently been proposed, and the majority of these algorithms have been validated on the UC-Merced dataset utilizing the

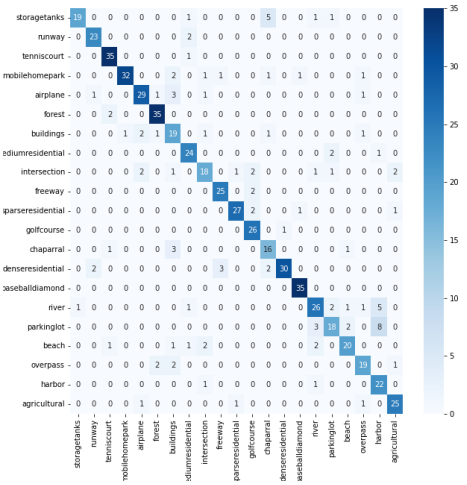

Fig. 2. Confusion Matrix: Support vector machine (SVM) with hybrid features and linear Kernel

identical experimental protocol and 10-fold cross-validation technique. As a result, there has been sufficient information gathered to establish a meaningful reference to the present state of the art algorithms and approaches. Table I illustrates the overall accuracy results of all related techniques, as reported in the original research.

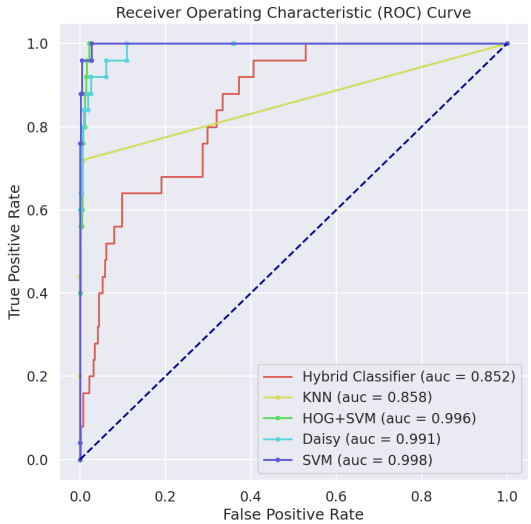

Fig. 3. ROC Curve for all classifiers used

The hybrid Support Vector Machine Model produces the best classification score of 81.42 % and 76.19%, as shown in Table III. The performance of the model using DAISY descriptors and the SVM with linear kernel (Linear Classifier) is equivalent. The kNN model outperforms the HOG features with satisfactory results. The confusion matrices of the top two best performing classifiers are shown in Figure 1 and Figure 2. Figure **??** shows the performance evaluation of classification techniques with different threshold settings.

V. Conclusion

Environmental monitoring, disaster forecasting, geological surveying, and other activities depend significantly on data from remotely sensed pictures. Many satellites have been launched in response to the rising demand for remotely sensed pictures, and hundreds of photos are captured on a daily basis. As a result, the number of remotely detected photographs in the database skyrockets.As a consequence, rapidly and correctly obtaining useful photos from a broad, raw image collection becomes a task. One of the solutions to the problems listed above is a bag of features. A bag of features is basically a simplified representation of a picture. Rather than simple feature matching, we construct a global representation of an object. Consequently, we take a set of attributes, build a representation of the picture in a simplified fashion, and classify it.

We integrated both global and local feature descriptors from the photos and built a hybrid feature descriptor for each individual image in the proposed scene recognition algorithm [2]. As a local feature descriptor, DAISY descriptors associated with important regions in images were utilized, while a HOG feature linked to an image was used as a global feature descriptor. To substantially reduce the complexity of our encoding strategy, we used a bag of features encoding using the Mini-Batch K-Means method. A two-level pooling strategy was applied to integrate the DAISY and HOG features for each photo. Finally, to assess the performance of our model, a multi-class SVM classifier was trained using multiple kernel parameters in a cross-validation setting. The hybrid model, which combines SVM with an RBF kernel, demonstrated to be among the most effective, with an accuracy of 81.42%, while the linear classifiercame in second, with an accuracy of 76.19%. The experimental results in this research are comparable to those obtained by [2].

Prominent constraints that may be affecting model performance in feature extraction image categorization are the insufficient of credible ground truth data. While the UC Merced land-use dataset is one of the most popular for satellite image classification, it is limited to merely 21 classes, which is a pretty small number compared to the number of existing satellite classes. This makes developing models that generalize to a broad variety of classes challenging. For future approach a convolutional neural network may be employed and a dataset having a considerable number of classes can be utilized. This additional datasets might boost prediction performance as a number of distinct scene classes will be covered and data sensitive algorithms will increase performance owing to existence of more data. In addition, a

new dataset with a huge number of image samples will make it easier for models to adapt to complex networks.

ACKNOWLEDGMENT

I would like to acknowledge my supervisor, Dr. Ritesh Ajoodha, for his assistance and direction in the completion of this project.

REFERENCES

[1] X. Zheng, X. Sun, K. Fu, and H. Wang, "Automatic annotation of satellite images via multifeature joint sparse coding with spatial relation constraint," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, no. 4, pp. 652–656, 2012.

[2] J. Wilson and M. Arif, "Scene recognition by combining local and global image descriptors," *arXiv preprint arXiv:1702.06850*, 2017.

[3] C. M. Bishop, "Pattern recognition and machine learning (information science and statistics)," 2007.

[4] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 161–168.

[5] J. Kim[1], B. Kim, and S. Savarese, "Comparing image classification methods: K-nearest-neighbor and support-vector-machines," in *Proceedings of the 6th WSEAS international conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics*, vol. 1001, 2012, pp. 48 109–2122.

[6] Y. G. Byun, Y. K. Han, and T. B. Chae, "A multispectral image segmentation approach for object-based image classification of high resolution satellite imagery," *KSCE Journal of Civil Engineering*, vol. 17, no. 2, pp. 486–497, 2013.

[7] X. Huang and L. Zhang, "An svm ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 1, pp. 257–272, 2012.

[8] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, 2010, pp. 270–279.

[9] T. Kobayashi, "Dirichlet-based histogram feature transform for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3278–3285.

[10] R. Negrel, D. Picard, and P.-H. Gosselin, "Evaluation of second-order visual features for land-use classification," in *2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2014, pp. 1–5.

[11] X. Gong, L. Yuanyuan, and Z. Xie, "An improved bag-of-visual-word based classification method for high-resolution remote sensing scene," in *2018 26th International Conference on Geoinformatics*. IEEE, 2018, pp. 1–5.

[12] B. Petrovska, E. Zdravevski, P. Lameski, R. Corizzo, I. Štajduhar, and J. Lerga, "Deep learning for feature extraction in remote sensing: A case-study of aerial scene classification," *Sensors*, vol. 20, no. 14, p. 3906, 2020.

[13] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 5, pp. 815–830, 2009.

[14] M. Shahbaz, A. Guergachi, A. Noreen, and M. Shaheen, "Classification by object recognition in satellite images by using data mining," in *Proceedings of the World Congress on Engineering*, vol. 1, 2012, pp. 4–6.

[15] C. Vaiphasa, S. Piamduaytham, T. Vaiphasa, and A. K. Skidmore, "A normalized difference vegetation index (ndvi) time-series of idle agriculture lands: A preliminary study," *Engineering Journal*, vol. 15, no. 1, pp. 9–16, 2011.

[16] A. M. Cheriyadat, "Unsupervised feature learning for aerial scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 1, pp. 439–451, 2013.