Bag of Visual Words for Land-Use Land-Cover Classification

Ai-Linh Alten
Computer Science
San José State University
San José, CA, USA
ai-linh.alten@sjsu.edu

***Abstract*—Remote sensing image scene classification plays an important role in a variety of applications. With the advance of remote sensing technologies that provide Very High Resolution (VHR) imagery, there are many opportunities to extend scene classification to more semantic classes. This paper explores the traditional bag of visual words (BOVW) approach for land-use land-cover (LULC) classification. This will be evaluated using an image dataset of 21 land-use classes. In addition, different sizes of the visual codebook will be evaluated to determine the stability of the traditional BOVW model. This model will cluster a non-spatial representation of the feature set to retrieve the visual codewords in codebook generation. With the BOVW approach, a more robust method of providing semantic scene classification can be used in alternative to other standard LULC methods.**

Keywords—land-use classification, bag of visual words, remote sensing, Very High Resolution satellites, visual codewords, visual codebook

I. INTRODUCTION

With the advent of newer remote sensing platforms that can acquire Very High Resolution (VHR) Earth Observation (EO) data that is high spatial-, temporal-, and spectral-resolution, there are now opportunities to provide better image scene classification methods. VHR remote sensing data can now be offered in a wide range of spectral resolutions between 0.6 μm to 14.4 μm. Spatial resolutions are now as low as 31 cm to 3 m. VHR data dates back to 1999 when the first high resolution satellite IKONOS was launched under the Lockheed Corporation. IKONOS can provide 1 m spatial resolution with one panchromatic band and four multispectral bands in 4 m resolution [2]. As of recent, DigitalGlobe's WorldView-3 which is launched by United Launch Alliance and Lockheed Martin in 2014 now provides 31 cm in panchromatic and 1.24 m in its eight multispectral bands [2]. Fig. 1(a)-(d) shows images of spatial resolutions 250m, 10m, 2.5m, and 31cm. Spatial resolution is the representation of 1 pixel in m^2. For example, Fig. 1(a) MODIS image has pixels that are an area of $250m^2$. Fig. 1(a) compared with Fig. 1(b) shows that San Francisco's coastline can now be determined with the decreased spatial resolution. Fig. 1(c) and Fig. 1(d) show finer spatial resolutions in which objects can be seen. However, to pick out even smaller objects like cars or people, 31cm spatial resolution is required as seen in Fig. 1(d). With newer and finer VHR imagery, a greater range of objects and spatial patterns can be observed. This can provide opportunity for more semantic classes in LULC.

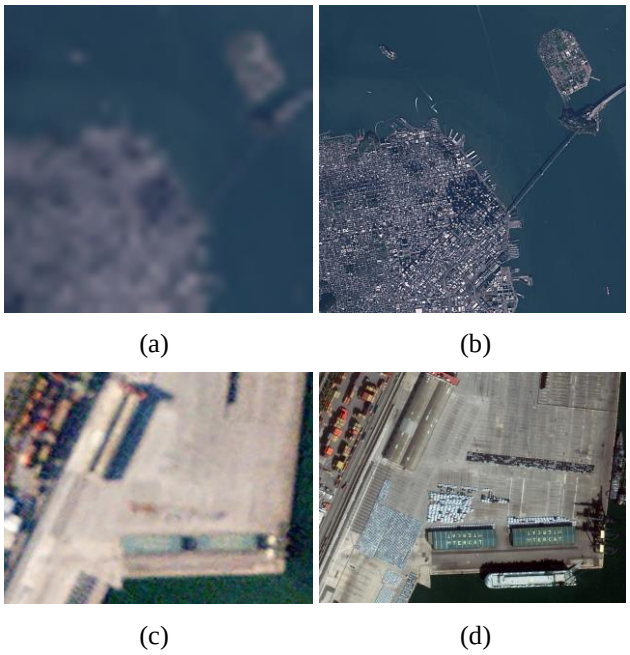
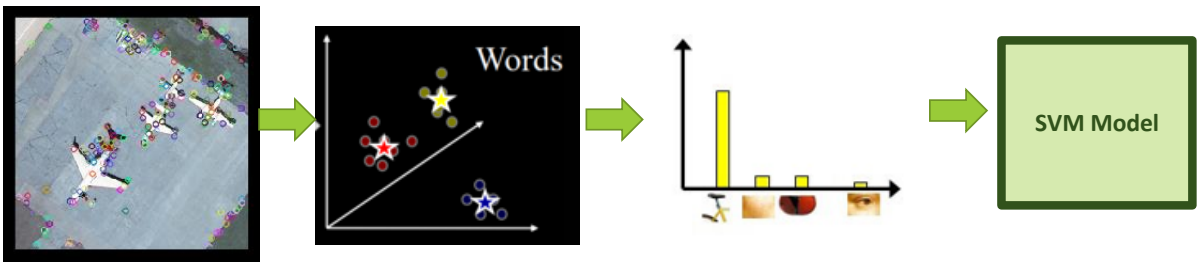


Fig. 1. Images with differing spatial resolution. (a) San Francisco (8km x 8km) captured by MODIS bands 1 and 2 at 250m resolution. (b) Same image captured by Sentinel-2 bands red, green, and blue at 10m resolution. (c) Port de Barcelona captured by Planet Labs Dove at 2.5m resolution. (d) Same image captured by WorldView-3 at 31cm resolution. Increased spatial resolution opens doors for more spatial analysis methods.

Land-Use Land-Cover (LULC) classification provides information on the purpose the land serves and the type of surface cover on the ground respectively. Land use usually involve baseline mapping and subsequent monitoring since it is important to observe the change in the land over time. Some examples of land use can be recreation, wildlife habitat, or agriculture [1]. Land cover identifies and delineates land cover and can be important for global monitoring studies, resource management, and planning activities. Some examples of land cover can be vegetation, urban infrastructure, water, bare soil, or others [1]. Usually classes are in the following main categories: agricultural, forest, urban (built-up), vegetation, and water. However, with VHR imagery, more sub-categories or semantic classes can be used to represent these main categories. For



(a) (b) (c) (d)

Fig. 2. Bag of Visual Words methodology. (a) First step is to compute features from images in dataset (SIFT). (b) Visual codewords are computed by clustering the feature descriptors (k-means). (c) In codebook generation, frequencies of visual codewords are quantized. (d) Image classification is performed.

instance, built-up can further be broken down into residential, commercial, and industrial classes. LULC can be applied to many applications of remote sensing: natural resource management, wildlife habitat protection, baseline mapping for Geographic Information Systems (GIS), urban expansion or encroachment, routing and logistics for seismic or exploration or resource extraction activities, damage delineation caused by natural disasters for damage assessment, legal boundaries for tax and property evaluation, and target detection [1].

This paper explores the concept of using the Concept-based Image Retrieval (CBIR) method called Bag of Visual Words (BOVW) for scene classification. BOVW is commonly used as a robust method to perform image classification on complex semantic categories [3]. Although there are state-of-the-art deep learning computer vision methods that are useful in learning the image representations, it is hard to say if the performance of such models will perform well on other complex datasets [3]. BOVW has been used in the last decade and proves to be a good leading strategy for image classification [3]. BOVW is a three-step method that first uses feature extraction methods such as Scale Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG). Second, quantizing the feature descriptors is performed using clustering algorithms such as K-Means to generate visual codewords. These visual codewords are then transformed into a normalized histogram of visual word frequencies called the visual codebook. The visual codebooks can be used to train powerful statistical learning models such as Support Vector Machine (SVM) in order to perform the scene classification [3].

The BOVW method for LULC described in this paper will allow for a robust capability to classify the variety of semantic classes that comes with land use and land cover. This method is evaluated using a ground truth dataset of 21 land-use classes. The traditional BOVW will be compared to other approaches in BOVW LULC. As the simple BOVW model in this paper shows to not perform as well as the state-of-the-art extensions for BOVW LULC, this paper will show the stability of the traditional model by increasing the visual codebook size.

II. Existing Systems

This paper is motivated by one of the first BOVW LULC methods suggested in "Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification." Yang and Newsam compare several extensions to the BOVW method. In their experiments they used non-linear SVM to perform the multi-class classification. The histogram kernels they tested were the traditional BOVW, Spatial Pyramid Match Kernel (SPMK), and BOVW combined with Spatial Co-occurrence Kernel (SCK). Their results showed that BOVW+SCK method outperforms the traditional BOVW and the SPMK [4].

In "Remote Sensing Image Scene Classification Using Bag of Convolutional Features," Cheng, Li, Yao, Guo, and Wei suggest using features using convolutional neural networks (CNN). In their proposed method they extract convolutional features from AlexNet, GoogleNet, and VGGNet-16 to remove the need to use human-engineered features such as color and shape information. They compare their Bag of Convolutional Features model to the traditional BOVW model with dense SIFT and shows that classification accuracy is improved and proved that taking advantage of the state-of-the-art CNN models can provide more semantic properties for scene classification [5].

In "Color-Boosted Saliency-Guided Rotation Invariant Bag of Visual Words Representation with Parameter Transfer for Cross-Domain Scene-Level Classification," Yan, Zhu, Liu, and Mo describe a several-step Knowledge Discovery in Databases (KDD) approach for BOVW. Their method helps remove influence of background information rather than extracting features on whole images, effect of rotation transformation, and provide the optimal parameters for SVM in image classification. In their first step, the color saliency method (CBGCSRD) is the pre-processing step that helps remove the background noise that a feature detection algorithm could pick up. They convert the RGB space to their own predefined color space that is uncorrelated to photometric events. The second step is the Rotation Invariant BOVW that is the feature augmentation and extraction step. They partition the images and get features of several orientations of the images. For feature extraction Dense-SIFT is used. K-means is used to create the visual codewords. Lastly, parameter tuning is used to find the best parameters for SVM to help the learning process and reduce bias. Their method is then compared to variations of this method and the traditional BOVW across two public datasets. Their method shows to have roughly 91%-95% accuracy and shows a good semantic annotation result compared to the ground-truth image [6].

III. Bag of Visual Words

This section describes the bag of visual words (BOVW) method for image classification. BOVW is based off the original

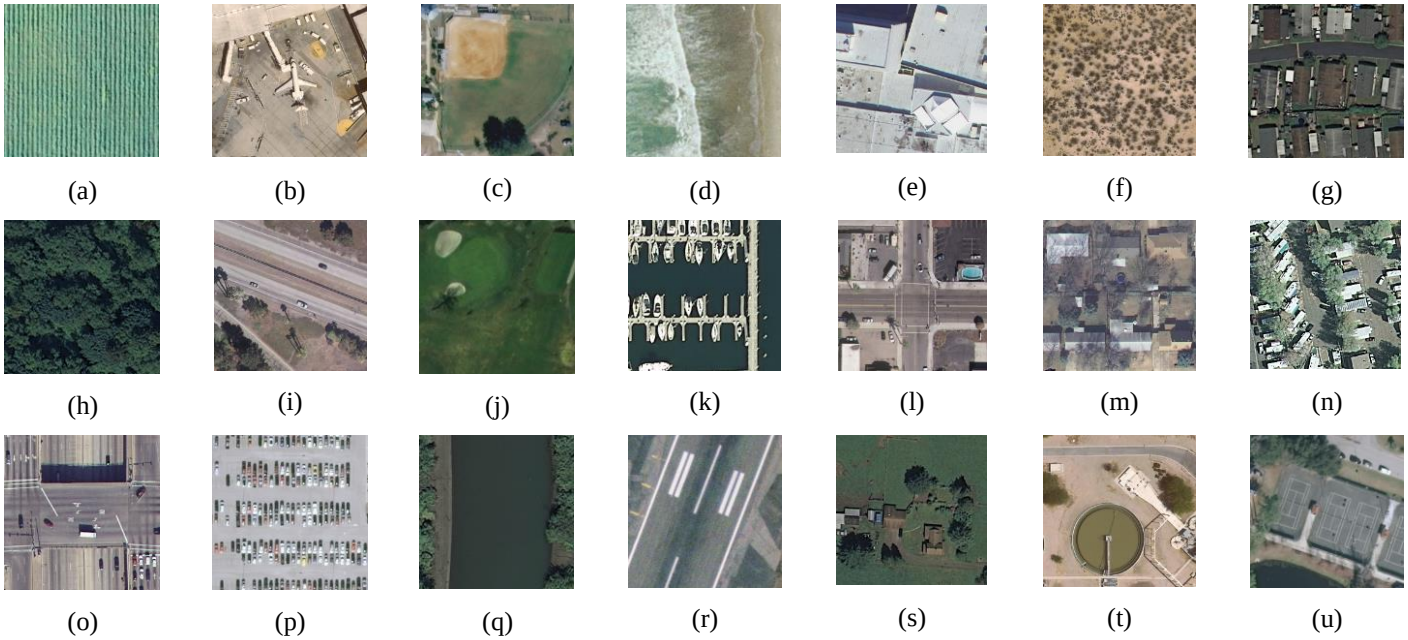


Fig. 3. UC Merced Land Use Dataset[1]. (a) agricultural. (b) airplane. (c) baseballdiamond. (d) beach. (e) buildings. (f) chaparral. (g) denseresidential. (h) forest. (i) freeway. (j) golfcourse. (k) harbor. (l) intersection. (m) mediumresidential. (n) mobilehomepark. (o) overpass. (p) parkinglot. (q) river. (r) runway. (s) sparseresidential. (t) storagetanks. (u) tenniscourt.

bag of words model that is well known in text analysis and information retrieval. In this model, a text document is the "bag" of words which is represented by word frequencies. The frequencies can be used for document classification. The similar concept is used in BOVW for images in which the words are represented by image descriptors. Fig. 2 shows an example of the BOVW methodology.

A. Scale Invariant Feature Transform

In the traditional BOVW method, David Lowe's Scale Invariant Feature Transform (SIFT) is the most commonly used feature extraction algorithm. SIFT is well known for computing features when images or particular objects or other things are at different scales, rotations, illuminations, and viewpoints. SIFT is particularly good for image matching. There are several steps in SIFT:

1) Interest Point Detection: the SIFT descriptor is computed from image intensities from interest points in the image key points. These key points are created from scale space extrema of differences-of-Gaussians (DoG) in a differences-of-Gaussians pyramid. Key points where there are edges and low contrast regions are eliminated [7]. DoG is derived from subtracting two Laplacian of Gaussians (LoG):

$$DOG(x,y;s) = L(x,y;s+\Delta s) - L(x,y;s) \quad (1)$$

$L(\bullet)$ is the Laplacian operator in which the image is smoothed by a convolution of Gaussian kernels. s is the variance of the Gaussian kernel [7].

2) Image Descriptor: After each key point is obtained, image descriptors are computed by using histogram of local gradient in several derections around the interest points. To get scale invariance, the local neighbourhood is normalized. To get rotational invariance, the dominant orientation in the interest point's neighborhood is used for orienting the 4x4 grid over the point [7].

B. Build a Bag of Visual Words

First, SIFT descriptors are extracted from interest points in each image. Each feature vector is 128-dimension and are considered as the visual word. K-means clustering is applied to create the visual codebook or dictionary. This visual codebook is used to quantize extracted features during testing phase. In this implementation, SIFT features are used for feature extraction, k-means for generating visual codewords, these are then quantized as frequency histograms which are normalized. Codebook generation is known as vector quantization in which the set of interest points computed by SIFT are grouped by other points closer to them [8]. The normalized histograms are represented as *images* x *codewords*. The normalized histograms are then used in a linear SVM for image classification.

IV. Experiment

A. Dataset

The UC Merced dataset[1] is a publicly-available dataset that has been manually extracted from large images of the US Geological Survey (USGS) National Map Urban Area Imagery collection. The spatial resolution is 30cm and each image is 256 x 256 pixels in RGB. The dataset consists of 21 classes and 100 images of each like in Fig. 3 making this 2,100 images in total.

B. Experimental Setup

The traditional BOVW will be tested by performing the multi-class scene classification. OpenCV-Python package is

[1] http://weegee.vision.ucmerced.edu/datasets/landuse.html

TABLE I. ACCURACY SCORES OF CODEBOOK SIZES

# of codewords	*μ (+/-) σ*	*Standard Error*	*95% CI*
10	0.7634 (+/-) 0.0926	0.0414	[0.6822, 0.8445]
25	0.8248 (+/-) 0.1527	0.0683	[0.6909, 0.9587]
50	0.8463 (+/-) 0.1630	0.0729	[0.7034, 0.9892]
75	0.8463 (+/-) 0.1640	0.0734	[0.7025, 0.9900]
100	0.8439 (+/-) 0.1766	0.0790	[0.6891, 0.9987]
125	0.8396 (+/-) 0.1820	0.0814	[0.6801, 0.9992]
150	0.8558 (+/-) 0.1602	0.0716	[0.7154, 0.9963]

Fig. 4. Comparison of codebook sizes. Mean accuracy and standard deviation of 5-fold cross validation, standard error, and 95% confidence interval.

first used to extract the SIFT features. K-means++ for clustering the codewords, StandardScaler for normalizing the frequency histograms, and SVM from the Scikit-learn library was used to build the BOVW model. The StandardScaler normalizes the histograms around a mean of 0 and standard deviation of 1. The linear kernel was used for SVM. To test the performance of the BOVW model the following experiments were done:

- Stratified 5-fold cross validation is performed on increasing codebook sizes (n = number of codewords): 10, 25, 50, 75, 100, 125, 150. This cross validation tests the performance of SVM by partitioning the codebooks in five equal parts.
- Stratified train/test split of 80/20 was used for performance metrics. To see performance of the SVM, 80/20 split was used only on the normalized frequency histograms or codebooks. In this experiment, the codebook size is 25 codewords.
- To validate performance of entire BOVW model on unseen instances, 80/20 split was used on the UC Merced dataset. In this experiment, image features are created on unseen instances and are clustered by the pre-trained k-means model of 25 clusters. The frequency histogram of codewords is then normalized by the pre-trained StandardScaler. And finally, the histogram is classified by the pre-trained SVM.
- Stratified 5-fold cross validation is performed on the BOVW model by partitioning the images by five. This is tested with 10 clusters and 25 max iterations for k-means.

V. RESULTS

A. 5-Fold Cross Validation

Table 1 and Fig. 5 shows the average accuracy score across entire 2100 image dataset. It can be seen here that the average

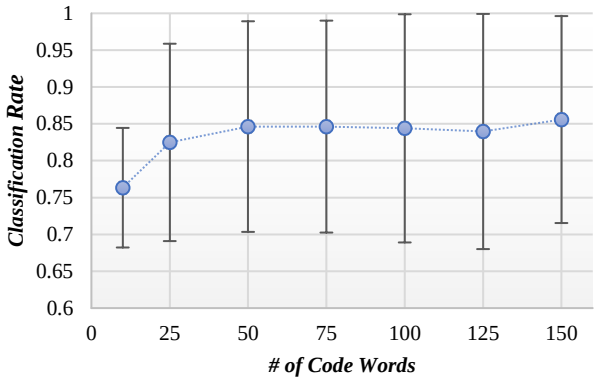


Fig. 5. Classification rate with 95% confidence interval.

accuracy improves from 76% to about 85% when the codebook size increases from 10 to 50 codewords. Codebook sizes from 50 to 150 codewords stabilizes at 85% average accuracy. In Table 1 and Fig. 5, it can be seen that the standard deviation and the 95% confidence interval increases as codebook size increases. From 50 to 150 codewords the 95% confidence interval ranges from about 70% to 99% accuracy. This means the population mean accuracy of the 5-fold can be within that range with 95% confidence.

B. Results of SVM on Codebooks

Fig. 7 provides the confusion matrix of the 21 land use classes. This figure shows that the model performs well on the 20% unseen cases of the normalized frequency histograms. The classes that weren't accurate by 100% were golfcourse, storagetanks, agricultural, chaparral, sparseresidential, intersection, tenniscourt, parkinglot, harbor, and river. The class with the lowest accuracy score, storagetanks, had 80% accuracy with 10% identified as denseresidential and 10% identified as golfcourse. The unweighted mean accuracy score is 95.24%, precision is 95.60%, recall is 95.24%, and f-score is 95.21%.

C. Results of BOVW

Fig. 8 provides the confusion matrix for the 21 land use classes. A stratified shuffle of the 2,100 images into 80/20 train/test set is used. Results shown in figure 8 show a very poor classification performance on unseen cases. It can be seen that most of the images were predicted as mediumresidential, harbor, overpass, and buildings. The unweighted mean accuracy is 5.71%, precision is 2.32%, recall is 5.71%, and f-score is 3.04%. In Table 2, the 5-fold cross validation results shows an average accuracy of 8.97% with standard deviation of 0.0381. The 95% confidence interval falls between 7.5% to 10.5%.

TABLE II. ACCURACY SCORES OF BOVW

# of codewords	*μ (+/-) σ*	*Standard Error*	*95% CI*
10	0.0897 (+/-) 0.0381	0.0076	[0.0747, 0.1046]

Fig. 6. Classification accuracy of BOVW.

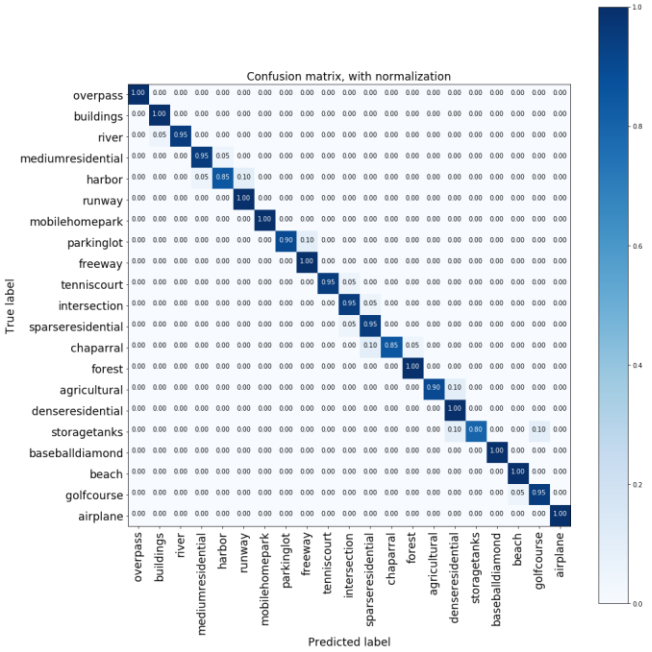


Fig. 7. Confusion matrix of result from SVM using test set from codebooks.

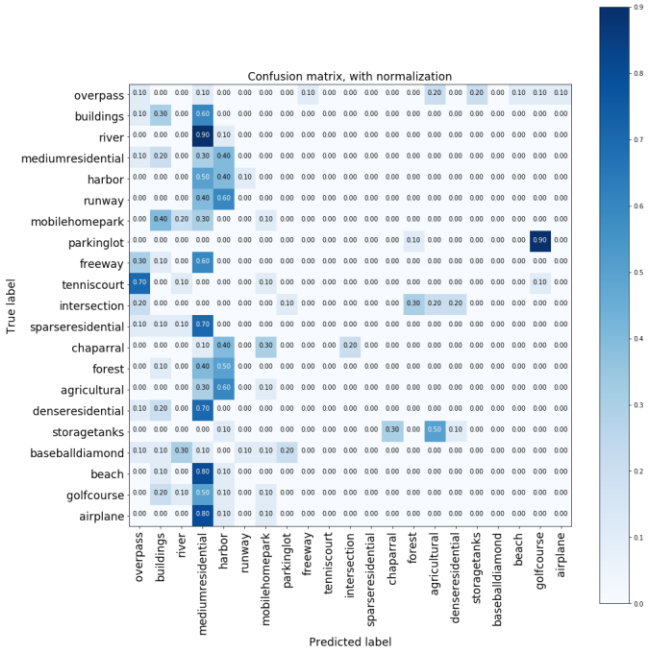


Fig. 8. Confusion matrix of result from BOVW using test set from images.

VI. Discussion of Results

Table 1 and figure 5 shows that the best average accuracy rate for the traditional BOVW is approximately 84.5% as codebook size increases. Accuracy is able to improve with the increase in codebook size. Variance and the spread of the 95% confidence interval increases with increased codebook size. This could mean the linear SVM model is unstable. Since 95% confidence interval ranges from about 70% to 99% mean accuracy, this means the population mean accuracy score can be within this range with 95% confidence. This spread is 29%-wide so it is very high. It is hard to say if the 95% confidence interval is valid since central limit theorem specifies that a sample size needs to be sufficiently large in order for data to be normally distributed. In this case, 5-fold may not be enough to know if the spread of the confidence interval is truly around 29%. However, it is a good indicator of the validity of our model so this could mean the model will require better features or pre-processing of the dataset. Comparatively to the results provided in [4], the model performs better than their traditional BOVW, Spatial Pyramid Match Kernel, and BOVW with Spatial Co-occurrence Kernel.

With accuracy, precision, recall, and f-score remaining around 95% for the multi-class classification, this means the SVM was able to identify an equal amount of false positives. The model performs well with unseen cases of the normalized histograms. It can be verified that the linear SVM classifier is a good model to use in the third step of the BOVW method for scene classification.

The BOVW model falls apart somewhere between 1st and 2nd step of the codebook generation where feature extraction and clustering takes place. With accuracy of 5.71% and low precision, recall, and f-score, the this shows that there are almost no true positives that were identified in the BOVW model. With such a low accuracy, this could mean the codebook generation with k-means was unable to group the unseen image SIFT features to its respective codeword. A lot of features are extracted in a scene which may become a problem in codebook generation. For instance, in the airplane imagery a lot of SIFT features extracted also included runways and buildings, hence it can be difficult to correctly classify airplane. It is interesting to note in confusion matrix a lot of the images were classified as mediumresidential. A lot of the classes have similarity to mediumresidential as buildings and residential homes tend are of similar shape and orientation. SIFT will require texture descriptors or color descriptors as well to help with the scene classification.

VII. Conclusion and Future Work

In this study, the traditional BOVW has been evaluated using Very High Resolution imagery at 30cm resolution. The traditional BOVW may not perform as well as current land-use land-cover classification methods, it does show that it can be a good and robust alternative for many semantic classes. Codebook generation needs special care especially in scene classification. Methods proposed by [5] for color saliency and feature extraction using Dense-SIFT can remove the unwanted features for each scene and can improve classification performance. SVM is unstable with higher codebook sizes so alternatives may be necessary. In future work, pre-processing with color saliency and using more that provide color and homogeneous texture will be used to see if BOVW model can be improved. If commercial satellite dataset is obtainable, spectral features can also be used in codebook generation.

REFERENCES

[1] Nrcan.gc.ca. (2018). *Land Cover & Land Use | Natural Resources Canada*. [online] Available at: https://www.nrcan.gc.ca/node/9373 [Accessed 10 Dec. 2018].http://www.gisat.cz/content/en/satellite-data/supplied-data/very-high-resolution

[2] Law M.T., Thome N., Cord M. (2014) Bag-of-Words Image Representation: Key Ideas and Further Insight. In: Ionescu B., Benois-Pineau J., Piatrik T., Quénot G. (eds) Fusion in Computer Vision. Advances in Computer Vision and Pattern Recognition. Springer, Cham

[3] Y. Yang and S. Newsam. Bag-Of-Visual-Words and Spatial Extensions for Land-Use Classification. In *ACM GIS*, 2010.

[4] G. Cheng, Z. Li, X. Yao, L. Guo, and Z. Wei. Remote Sensing Image Scene Classification Using Bag of Convolutional Features. *Geoscience and Remote Sensing*, 2017.

[5] L. Yan, R. Zhu, Y. Liu, and N. Mo. Color-Boosted Saliency-Guided Rotation Invariant Bag of Visual Words Representation with Parameter Transfer for Cross-Domain Scene-Level Classification. *MDPI Remote Sensing*, 2018.

[6] Lindeberg, T. (2018). *Scale Invariant Feature Transform*.

[7] Sinha, U. (2018). *SIFT: Theory and Practice: Introduction - AI Shack*. [online] Aishack.in. Available at: http://aishack.in/tutorials/sift-scale-invariant-feature-transform-introduction/ [Accessed 10 Dec. 2018].

[8] Y. Zhang, J. Chen, X. Huang, and Y. Wang. A Probabilistic Analysis of Sparse Coded Feature Pooling and Its Application for Image Retrieval. *NCBI*, 2015.