Pooling-Based Bag of Visual Words for Geographic Image Classification

Huang, Yunjia
Graduate School of Design, Kyushu University

Inoue, Kohei
Faculty of Design, Kyushu University

Li, Dongqi
Graduate School of Design, Kyushu University

出版情報：Journal of the Institute of Industrial Applications Engineers. 13 (1), pp.1-10, 2025-01-25. Institute of Industrial Applications Engineers
バージョン：

Journal of the Institute of Industrial Applications Engineers Vol.13, No.1, pp.1–10, (2025.1.25)
DOI: 10.12792/JIIAE.13.1 Online edition: ISSN 2187-8811 Print edition: ISSN 2188-1758

Paper

Pooling-Based Bag of Visual Words for Geographic Image Classification

YUNJIA HUANG* Member, KOHEI INOUE†‡ Member
DONGQI LI* Non-member

(Received July 12, 2024, revised November 30, 2024)

Abstract: This paper proposes a pooling-based Bag-of-Visual-Words for geographic image classification. Traditional image classification methods need a higher consumption of computational resources and slow convergence speed. In contrast, deep learning-based methods are highly demanded for large amounts of labeled data and computational resources. To address these problems, this paper introduces pooling operations, combined with the Bag-of-Visual-Words, to propose an efficient and resource-saving geographic image classification method. In the feature extraction stage, this paper combines SIFT features with maximum pooling and average pooling to generate a new feature space to capture the local and global information of the image more comprehensively. Experimental results show that the pooling-based Bag-of-Visual-Words not only improves the classification accuracy but also significantly reduces the consumption of computational resources and accelerates the convergence speed when dealing with geographic image classification. The results show that the method in this paper outperforms the original Bag-of-Visual-Words model and Spatial Pyramid Matching-based Bag-of-Visual-Words in terms of accuracy and computational efficiency, demonstrating its potential in geographic image classification with the Multi-temporal Scene Dataset of Wuhan University.

Keywords: Pooling Operation, Bag of Visual Words, Geographic image classification, SIFT features, K-means clustering

1. Introduction

Image scene classification is one of the most critical components of computer vision. Over the years, it has evolved significantly and found applications in various fields, including automatic surveillance systems [1], self-driving vehicles [2], medical imaging [3], and satellite image analysis [4]. The ability to classify scenes accurately is crucial for extracting meaningful information from visual data. For instance, surveillance systems help in detecting suspicious activities, while self-driving cars aid in understanding the environment and making navigation decisions.

With the rapid growth of image data on the Internet and the advent of sophisticated deep learning techniques, scene classification has become even more important. The vast amount of visual content available online presents both opportunities and challenges for image scene classification [5] [6]. The traditional methods, which rely on hand-crafted features and classical image processing techniques, have been effective to some extent. These methods include pixel-level differences and alignment [7]∼[9]. However, they often struggle with complex scene changes and require extensive domain knowledge for feature engineering.

‡ Corresponding: k-inoue@design.kyushu-u.ac.jp
* Graduate School of Design, Kyushu University
† Faculty of Design, Kyushu University

Deep learning-based methods, particularly those utilizing Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown great promise in automatically learning discriminative features from raw image data [10] [11]. These methods have outperformed traditional approaches in many cases, but they come with their own set of challenges, such as the need for large, labeled datasets, substantial computational resources, and robustness to environmental factors [12] [13].

The Bag of Visual Words (BoVW) [14], inspired by the bag-of-words model in Natural Language Processing [15], represents an image as a fixed-length vector where each element corresponds to a visual word. These visual words are obtained by clustering feature points extracted from training images. During image classification, the BoVW model constructs feature vectors by counting the occurrences of each visual word in the image and then classifies these feature vectors using a classifier.

To address some limitations of the BoVW model, the Spatial Pyramid Matching (SPM) technique was introduced by Lazebnik et al. in 2006 [16]. SPM captures spatial layout information of an image through multi-scale segmentation, enhancing the BoVW model's ability to handle images with complex spatial structures. Despite its advantages, SPM-based methods can be computationally intensive and may still struggle with

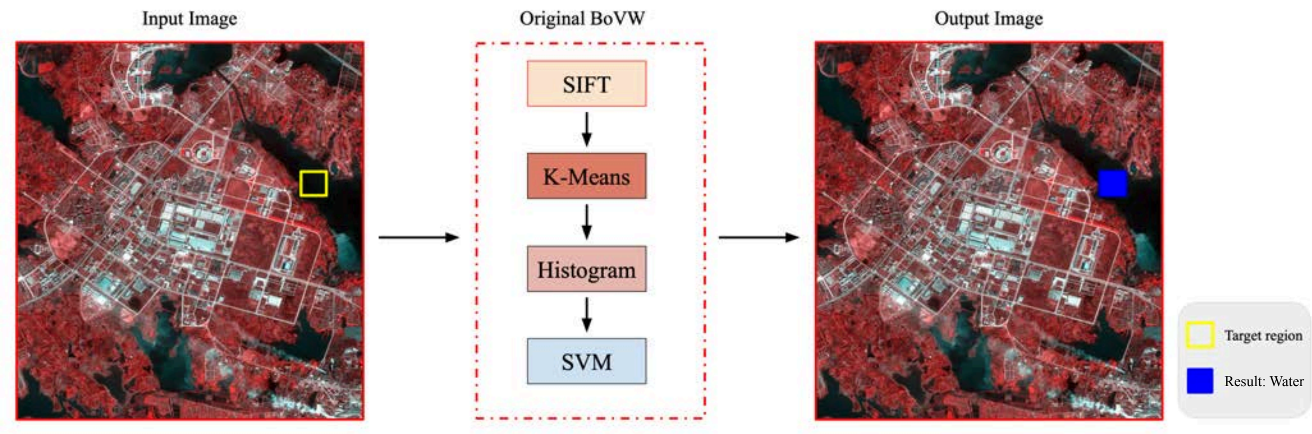


Figure 1: The flowchart of Bag of Visual Words.

certain types of variability in image data.

Pooling operations, such as maximum pooling and average pooling [17], are widely used in deep learning, particularly in CNNs, to reduce the size of the feature map and extract the most salient features. Maximum pooling selects the maximum value in each pooling region, preserving the most prominent features, while average pooling computes the average value in each region. These operations help in reducing the dimensionality of the feature map, improving computational efficiency, and enhancing the model's translational and local invariance.

Given the strengths and weaknesses of existing methods, this paper proposes a novel approach called pooling-based BoVW for geographic image classification. Our method aims to bridge the gap between traditional image processing techniques and modern deep learning methods. By integrating innovative feature extraction algorithms with the power of pooling operation, our approach addresses the complexities of multi-scene recognition.

Our main contributions are as follows:

Pooling-based BoVW Model: We propose a Bag-of-Visual-Words model that integrates maximum pooling and average pooling operations. This approach effectively addresses some drawbacks associated with the Spatial Pyramid Matching(SPM)-based BoVW method, providing a more robust feature representation.

Experimental Validation: We conduct extensive experiments on the Multi-temporal Scene Dataset of Wuhan University (MtS-WH) [18], demonstrating significant accuracy and computational efficiency improvements. Notably, in the 2002 test, our method achieved an Overall Accuracy of 93.45%.

The rest of the paper is organized as follows: Section 2 presents the research of materials and methods. Section 3 provides the experiments and visualisation results. Finally, Sections 4 and 5 provide a comprehensive discussion and summary of the paper, highlighting key points and suggesting potential avenues for future research.

2. Materials and Method

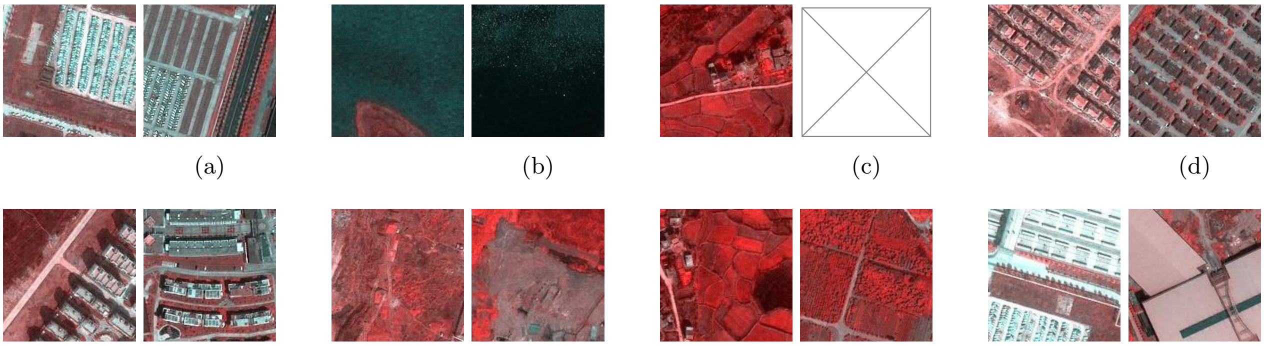
2.1 Materials Multi-temporal Scene Dataset (MtS-WH) [18] is a Wuhan University satellite remote sensing image dataset, which is shown in Figure 2. This dataset is mainly used to conduct theoretical research and validation of methods for scene change detection. The images were acquired in Hanyang District, Wuhan, in February 2002 and June 2009, respectively. 02 and 09 each contained 190 images for the training set and 1920 images for the test set. And they are divided into eight geographic attribute labels: 1-parking lot, 2-water, 3-sparse houses, 4-dense houses, 5-residential region, 6-idle region, 7-farmland, 8-industrial region. It should be noted that there are no 3-sparse houses in the 2009 dataset.

2.2 Methods In this section, we propose using a pooling-based BoVW model. This includes the process structure, design logic, and advantages of the approach.

2.2.1 Overall Structure The step of recognizing a geographical image by its geographical attributes can be simply understood as the image is fed into a classification model, and then a value will come out from the model, representing the predicted result of the input image. As shown in Figure 1 above, each image, after going through the BoVW block, will get the classification result of the image, such as water, houses, factories, etc.

Significantly, this thesis addresses two key limitations of the original BoVW algorithm: the absence of global feature extraction and the computational burden of the SPM-based BoVW algorithm. To overcome these, the original BoVW has been restructured, as depicted in Figure 3. These modifications are crucial as they pave the way for more efficient and accurate geographical image recognition.

As shown in Figure 3, this paper has a new design for feature extraction. The original feature extraction used



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

(e) (f) (g) (h)

Figure 2: The examples of Multi-temporal Scene Dataset pairs acquired in 2002 and 2009. (a) 1-parking lot, (b) 2-water, (c) 3-sparse houses, (d) 4-dense houses, (e) 5-residential region, (f) 6-idle region, (g) 7-farmland, (h) 8-industrial region.

SIFT algorithm to obtain (n, 128) dimensional features, but this paper, based on this, set up new feature extractions, respectively, maximum pooling and average pooling, and the above three features are stacked and combined to generate a new feature space different from the original features. And the subsequent parts of the algorithm are kept in sync with the original algorithm. However, it is worth stating that because the feature extraction session has been changed, and thus, in the K-means clustering session, the number of cluster centers K will also be changed accordingly to adapt the convergence of the whole algorithm. Meanwhile, The process overview of pooling-based BoVW is summarized in **Algorithm 1**.

Algorithm 1 Pooling-based BoVW Image Classification

Require: Non-empty training and test images, $K > 0$
Ensure: Output predicted labels for test images
1: Extract features from training images
2: **for** each Training images **do**
3: Compute local descriptors with Pooling+Sift
4: Collect descriptors into a single feature vector
5: **end for**
6: Perform K-Means on descriptors to get the visual vocabulary
7: Using K cluster centers
8: **repeat**
9: **for** each descriptor **do**
10: Assign descriptor to its nearest cluster center
11: **end for**
12: Update cluster centers to the mean of assigned descriptors
13: **until** cluster centers converge
14: Encode training images using the visual vocabulary
15: **for** each training image **do**
16: Initialize an empty histogram of length K
17: **for** each descriptor **do**
18: Assign descriptor to its nearest cluster center and update the histogram
19: **end for**
20: Normalize the histogram
21: **end for**
22: Train classifier using encoded training features and their labels
23: Extract features from test images
24: **for** each test image **do**
25: Compute local descriptors
26: Encode test image using the same visual vocabulary from step 6
27: **end for**
28: Predict labels for test images using the trained classifier (SVM)
29: Output predicted labels for test images

2.2.2 Pooling and SIFT

Pooling results are usually used to reduce the feature dimensions and extract the overall features of the image. In contrast, the SIFT algorithm is typically used to extract the local features of the image. Therefore, stacking pooling and SIFT results can combine global and local features to improve the algorithm's performance. This allows the BoVW algorithm to extract the feature information of the image more comprehensively.

Pooling is computed from deep learning models, the most common of which are maximum pooling and average pooling. As the word suggests, maximum pooling takes the maximum value in the sliding window as the output result. In contrast, average pooling takes the average value in the sliding window as the result. Different levels of information about the image can be obtained depending on the different calculations of the two pooling operations. They both are used to reduce the spatial dimension of the feature map, which reduces the number of parameters and computation of the model and helps in extracting the invariance and noise resistance of the features, which also provides the possibility of reducing the computation of the classification model and improving the computational efficiency. At the same time, the combination of salient and smooth features extracted to the image by pooling with the sift features can better increase the generalization performance of the model and make up for the defects brought about by the SIFT's sensitivity to image noise and the shortcomings of the SIFT's large

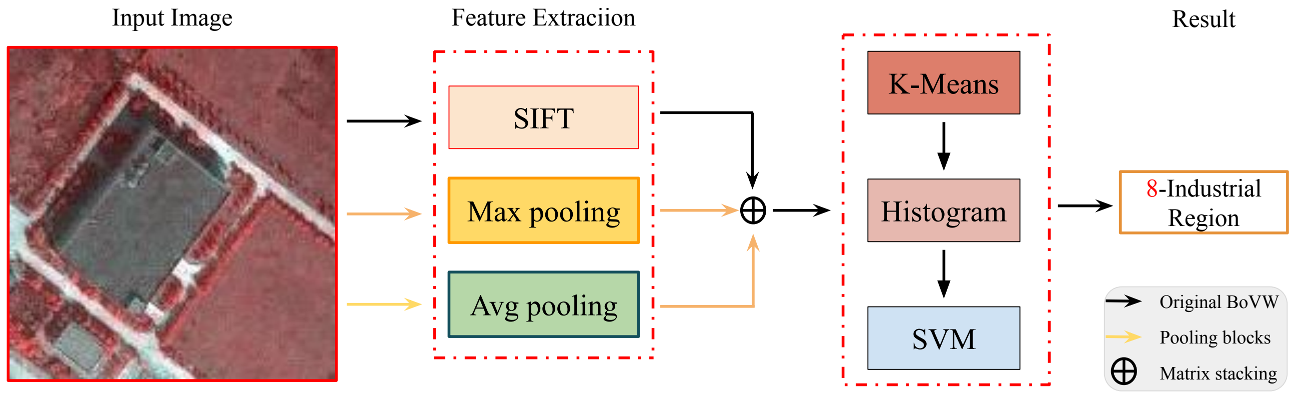


Figure 3: Pooling-based BoVW.

computational volume.

However, it should be noted that fusing pooling features with SIFT features also has specific problems. The features obtained after the SIFT processes the image are 128 dimensions, so certain processing is required when using pooling for feature extraction. Otherwise, access to the pooling features, which are also 128 dimensions, is impossible.

For this purpose, this paper borrows the idea of SPM to divide the image into sub-regions and then extract feature vectors from each sub-region. Each image is divided into 8x16, resulting in a total of 128 sub-regions. Next, pooling is computed for each sub-region; the formula is shown below.

$$region = (8, 16) \tag{1}$$
$$i = \{0, 1, 2, \ldots, 7\} \tag{2}$$
$$j = \{0, 1, 2, \ldots, 15\} \tag{3}$$

$$Feature\_Avg[:, i \times 16 + j] = average(region[i, j]), \quad Feature\_Avg \in \mathbb{R}^{1\times128} \tag{4}$$

$$Feature\_Max[:, i \times 16 + j] = maximum(region[i, j]), \quad Feature\_Max \in \mathbb{R}^{1\times128} \tag{5}$$

$$Feature\_sift = SIFT(image), Feature\_sift \in \mathbb{R}^{n\times128} \tag{6}$$

$$Features = \begin{bmatrix} Feature\_sift \\ Feature\_Max \\ Feature\_Avg \end{bmatrix} \in \mathbb{R}^{(1+1+n)\times128} \tag{7}$$

The image is divided equally into 8×16 sub-regions as described in Equation 1. Equations 4, 5 obtain the average and maximum eigenvalues of each sub-region by traversing each sub-region using average computation and maximum computation, respectively, to ultimately generate $Feature\_avg, Feature\_max \in \mathbb{R}^{1\times128}$. A further step is to merge the two pooling features and the SIFT features by vertically stacking them to form $Features \in \mathbb{R}^{(1+1+n)\times128}$, which serves as the final output features of the feature extraction phase provide data for the subsequent K-means clustering.

3. Results

This paper compares the algorithms, visualizes the results in this section, and analyzes the relevant tables to prove our algorithm's correctness and feasibility.

3.1 Parameters settings The experiments were conducted using the parameters of SIFT algorithm with a threshold (contrastThreshold) of *7e-3* for filtering the low contrast feature points and a threshold (edgeThreshold) of 10 for filtering the edge feature points. K-means was parameterized with a number of iterations of 10 where the value of K, the number of cluster centers was altered according to the respective algorithms and the classifiers were all classified using the Support Vector Machine (SVM) [19], and the kernel used, Radial Basis Function Kernel (RBF) [20], is a radial basis function, also known as a Gaussian kernel. All of the above parameters are in Python 3.9 Version.

3.2 Comparison experiment In this section, it is hoped to find the K-value for the best results of each of the six algorithms and compare the algorithms' running speed using their respective best K-values. By comparing the experiments, the pooling-based BoVW algorithm has a great lead in running speed and accuracy compared to the original BoVW and SPM-based BoVW.

3.2.1 The optimal K An optimal parameter search was performed for K in K-means clustering to find the optimal parameters of the algorithm. The value of K is set to range from 10 to 500, with experiments conducted at intervals of 10. The evaluation metric for the experiments is Overall Accuracy.

The experiments are divided into six groups: the original BoVW, SPM-based BoVW (Level 2 and Level 3), and pooling-based BoVW (Avg, Max, and Avg+Max). The experimental results by best K search are shown in Figure 4 and Table 1. The six algorithms achieved their highest accuracy at their respective optimal K values: original BoVW reached an accuracy of 93.20% with 400. SPM-based (Level 2) achieved its highest accuracy with a K value of 440, while SPM-based (Level 3) achieved its highest accuracy with a K of 470. Pooling-based (Avg), Pooling-based (Max), and Pooling-based (Avg+Max) achieved their highest accuracies with K values of 300, 330, and 220, respectively.

The comparison results show that the maximum achievable accuracy of these six algorithms is similar, but there is a noticeable difference in the K-value selection of the different algorithms. The K-value of the pooling-based BoVW (Avg+Max) is almost half of the K-value of the original BoVW when 93% accuracy is achieved, which also indicates that the algorithms that incorporate pooling computation are more potent than the original BoVW in terms of feature extraction.

3.2.2 Computational efficiency The computational efficiency of the six algorithms is compared by repeating the experiments ten times, with the mean distribution of the ten results representing the range of accuracy of the algorithms. The comparison experiments are based on the algorithms' training phase using the same dataset. The only difference is each algorithm's optimal K-value. The computation time statistics are performed for the four phases of feature extraction, K-means clustering, histogram building, and classifier classification.

As shown in Table 2 and Figure 5, the running time increases in the feature extraction phase by adding multiple multi-layer feature extraction on top of the original feature extraction with only the SIFT algorithm. Among them, SPM (Level 3) increases the time cost the most, reaching a time distribution of 12.22±2.12s. Whereas pooling-based BoVW has an increase in time, it is within acceptable limits.

The comparison shows that the difference in the K-means stage is the most obvious. The original BoVW has an operation time at this stage of 468.19±68.0s, whereas the SPM-based algorithm increases the computation time substantially, with 592.07±61.77s for Level 2 and 868.96±86.52s for Level 3. The SPM-based algorithm, relative to the original algorithm, almost doubles the computation time for Level 3. However, for pooling-based BoVW, the computation time for Avg, Max, and Avg+Max decreases rather than increases, with the Avg+Max-based algorithm reaching 175.45±9.06s. It is generally pleased that the pooling-based method can obtain the image's features and global and local information more efficiently. At the same time, the selection of the K-value also has a significant impact. The larger the K-value, the more computational content is required, and the computing time increases. There is no significant difference between the six algorithms when establishing histogram and classification sessions.

It can be seen that the pooling-based BoVW method, in terms of computing cost, is not just a step ahead of the original BoVW algorithm and the SPM-based algorithm but leaps and bounds ahead. Summing up the analysis of Section 3.2, the pooling-based BoVW method proposed in this paper has some accuracy improvement and a breakthrough in computing cost.

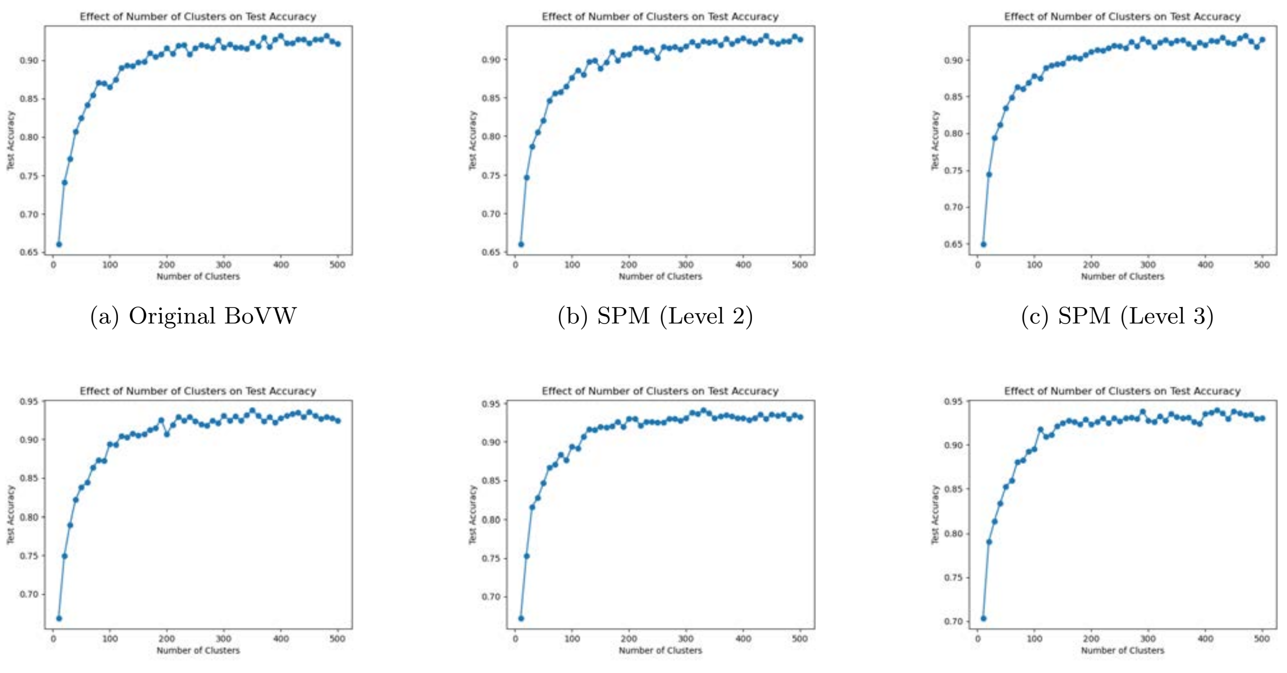
3.2.3 Visualisation and Analysis In this subsection, the pooling-based BoVW algorithm is used to visualize and analyze the algorithm results for 2002 and 2009. The algorithm has an Overall Accuracy of 93.45% with a kappa score of 0.92 and an Overall Accuracy of 93.34 % with a kappa score of 0.92, separately. As shown in Figure 6, the algorithm has better classification results for water, idle region, vegetation region, and industrial region, but is terrible for parking region, which is only in the upper and lower 50%.

Meanwhile, the misclassification images were reduced to Figure 7 based on the coordinates to demonstrate the robustness of the classification results better. Specifically, the prediction results are compared with the ground truth, and the misclassified images are marked with yellow squares on the graph to identify and analyze these errors easily. It is clear to see which categories of samples are prone to misclassification, and it also corresponds to Figure 6.

Through image classification, the image of the Wuhan Hanyang district is turned into different color blocks for representation, as shown in Figure 8. Since the experimental accuracy did not reach 100% for each category, it is impossible to measure the urban change from 2002 to 2009 directly by changing the labeled values (predicted classification results) of the exact coordinates. Only general conclusions can be drawn based on the visualization results of the images. For example, the red blocks (industrial) were not evident in 2002, but in 2009, the visualization result map consisted of a lot of red, indicating that Hanyang District has significantly been developing its industrial construction in these seven years. By analogy, the yellow blocks (idles) in 2002 were much less in 2009, and so on.

4. Discussion

Our experimental results showcase the novelty of the proposed pooling-based BoVW method, which not only outperforms the original BoVW and SPM-based BoVW methods in terms of accuracy and computational efficiency on the MtS-WH datasets but also offers a unique approach to image classification. This improvement results from the effective combination of global and local features through maximum and average pooling operations, which capture more comprehensive information from the images. The integration



(a) Original BoVW (b) SPM (Level 2) (c) SPM (Level 3)

(d) Pooling (Avg) (e) Pooling (Max) (f) Pooling (Avg+Max)

Figure 4: The different K and accuracy in the different BoVW methods.

Table 1: The optimal K and Overall Accuracy for methods.

	Optimal K	Overall Accuracy (%)
Original BoVW	400	93.20
SPM-based (Level 2)	440	93.04
SPM-based (Level 3)	470	93.28
Pooling-based (Avg)	300	93.04
Pooling-based (Max)	330	94.09
Pooling-based (Avg+Max)	220	93.45

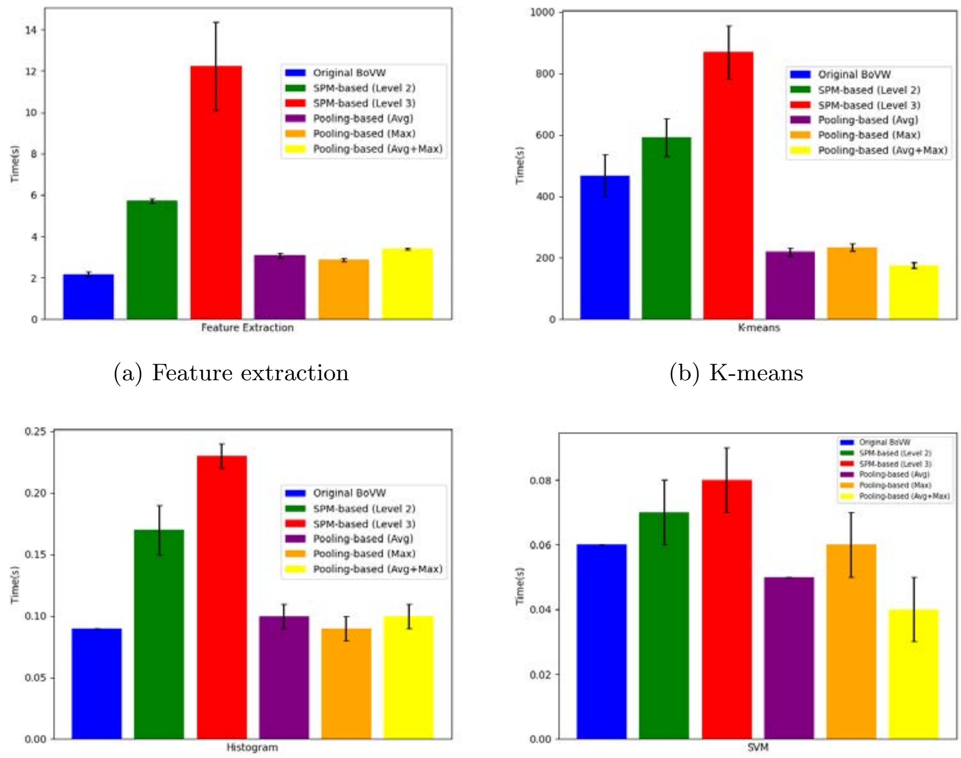
of SIFT features further enhances the feature space, leading to better clustering and classification performance. Maximum pooling effectively retains the most salient features, while average pooling provides a robust summary of the images, allowing our method to handle a variety of geographic image classification scenarios more effectively.

Our pooling-based BoVW method not only demonstrates superior performance in extracting and utilizing image features compared to the original BoVW, but also offers practical implications for real-world image classification tasks. While the original BoVW methods rely heavily on hand-crafted features and spatial pyramid matching, our method leverages deep learning techniques (pooling) to automatically extract more representative features, resulting in better performance. Despite requiring more initial computational resources for feature extraction and clustering, our method significantly improves the overall computational efficiency due to the reduced feature dimensions and enhanced feature representation. This practical implication of our method makes it a promising tool for real-world image classification tasks, where efficiency and accuracy are crucial.

While our study presents a significant advancement in image classification, there are areas that require further exploration. The reliance on the MtS-WH dataset, while comprehensive, may only encompass some possible geographic image variations. Future studies should validate our method on more extensive and diverse datasets to ensure its generalizability. Additionally, while our method reduces the demand for computational resources during classification, the initial feature extraction and clustering steps are computationally intensive. Further optimization of these steps is necessary to make the method more practical for real-time applications. Future research can focus on optimizing the feature fusion strategy, exploring different combinations of pooling operations, and investigating the impact of various clustering algorithms on visual vocabulary construction. Additionally, applying the proposed method to other types of image classification tasks and incorporating advanced techniques such as attention mechanisms or multi-scale feature extraction could enhance classification accuracy and efficiency, opening up new avenues for research and innovation in the field.

Table 2: The results of computational time.

	Optimal K	Feature Extraction (s)	K-Meann (s)	Histogram (s)	SVM (s)
Original BoVW	400	2.17±0.12	468.19±68.04	0.09±0.00	0.06±0.00
SPM-based (Level 2)	440	5.71±0.10	592.07±61.77	0.17±0.02	0.07±0.01
SPM-based (Level 3)	470	12.22±2.12	868.96±86.52	0.23±0.01	0.08±0.01
Pooling-based (Avg)	300	3.08±0.12	218.92±12.61	0.10±0.01	0.05±0.00
Pooling-based (Max)	330	2.87±0.08	234.58±12.34	0.09±0.01	0.06±0.01
Pooling-based (Avg+Max)	**220**	3.40±0.06	**175.45±9.06**	0.10±0.01	0.04±0.01



(a) Feature extraction (b) K-means

(c) Histogram building (d) SVM

Figure 5: The computational time results of the different methods.

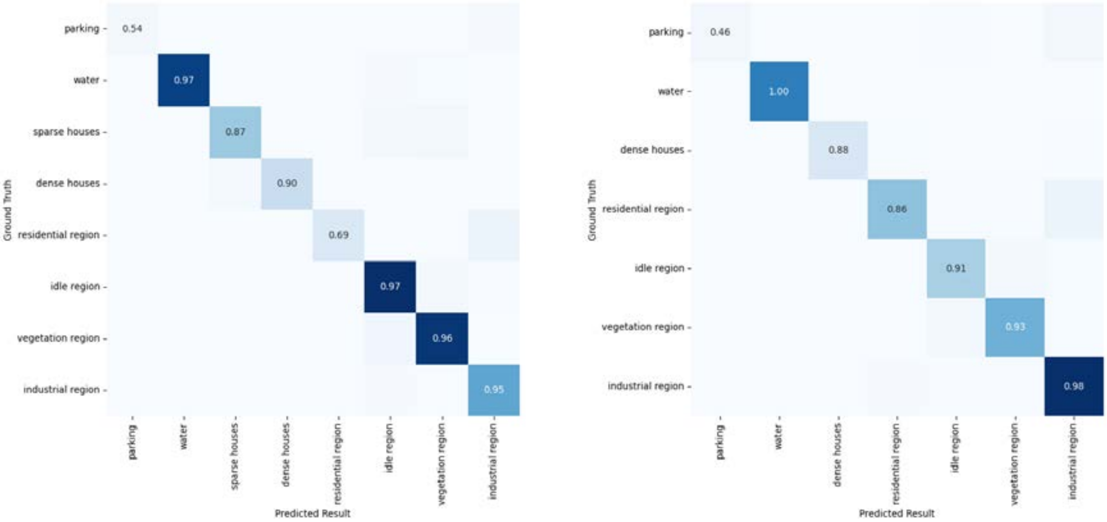
5. Conclusion

This paper proposes a pooling-based BoVW for geographic image classification. By combining pooling features and SIFT features, the method not only effectively solves some of the drawbacks of SPM-based BoVW but also significantly improves the algorithm's convergence speed and computational efficiency. Experimental results show that the pooling-based BoVW outperforms the original BoVW and SPM-based BoVW in the MtS-WH datasets in terms of accuracy and computational resources.

Firstly, through the maximum pooling and average pooling operations on the image, we successfully extracted the global and local features of the image and combined them with SIFT features to generate a more representative feature space. Then, based on feature extraction, we performed K-means clustering and encoded the images according to the visual vocabulary. The method's effectiveness in this paper is demonstrated by classifying the features in the training and testing.

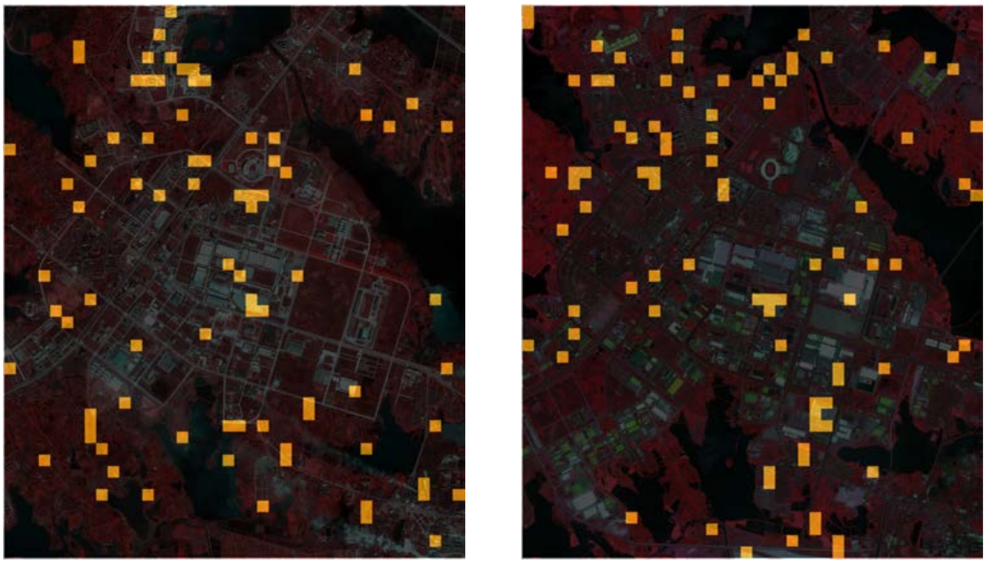
The experimental results show that the pooling-based BoVW method achieves significant improvement in accuracy and computational efficiency. This indicates that the combination of pooling features and SIFT features can extract image information more comprehensively, thus improving the performance of geographic image classification. At the same time, due to the reduced demand for computational resources, the method in this paper has higher operability and generalization in practical applications.

The research in this paper provides a new idea for geographic image classification and demonstrates the exciting potential of pooling operations in visual bag-of-words models through innovative feature extraction and clustering methods. Future research can further optimize the feature fusion strategy and explore more types of pooling operations to enhance classification performance further.



(a) The accuracy in 2002. (b) The accuracy in 2009.

Figure 6: The accuracy results of image classification.



(a) The misclassification in 2002. (b) The misclassification in 2009.

Figure 7: The yellow blocks to show misclassification.

References

[1] Maria Valera and Sergio A Velastin. Intelligent distributed surveillance systems: a review. *IEE Proceedings-Vision, Image and Signal Processing*, 152(2):192–204, 2005.

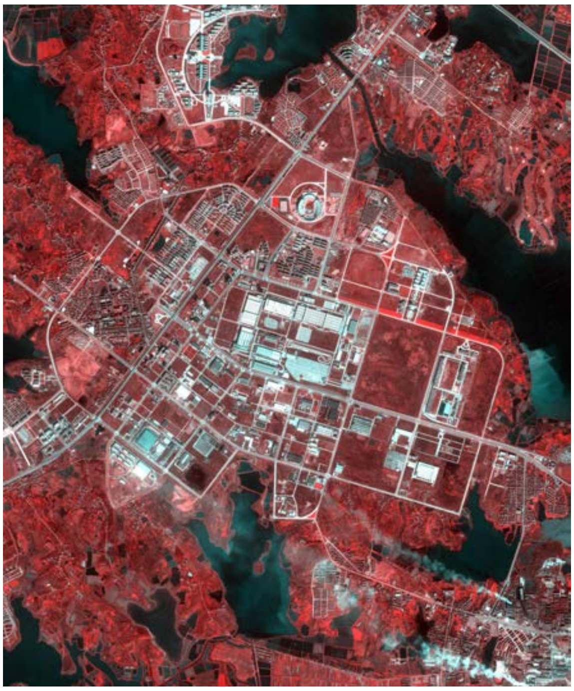
[2] Abhishek Gupta, Alagan Anpalagan, Ling Guan, and Ahmed Shaharyar Khwaja. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, 10:100057, 2021.

[3] Eka Miranda, Mediana Aryuni, and E Irwansyah. A survey of medical image classification techniques. In *2016 international conference on information management and technology (ICIMTech)*, pages 56–61. IEEE, 2016.

[4] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

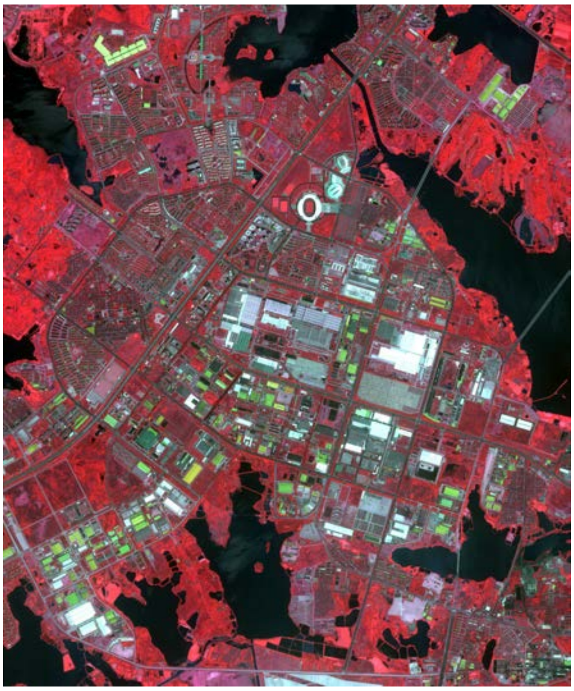
[5] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

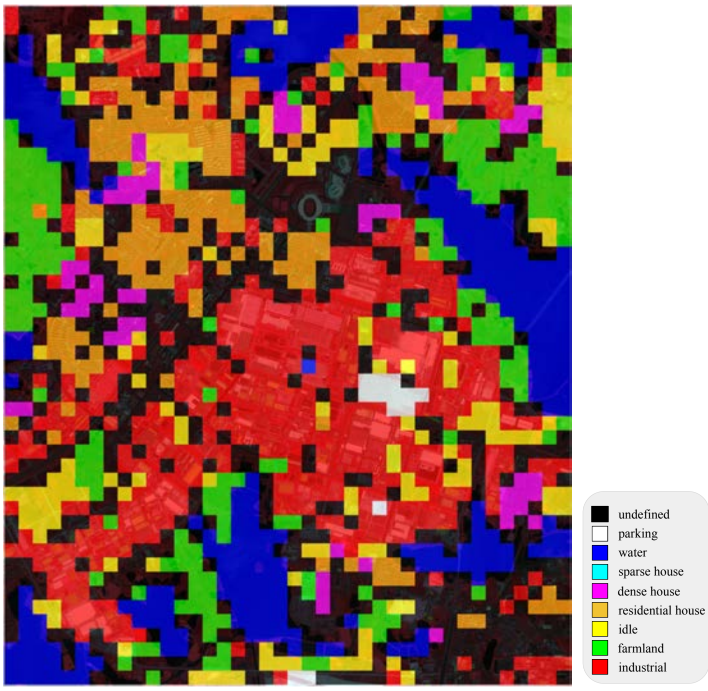


(a) The original image in 2002.

(b) The visualisation in 2002.



(c) The original image in 2009.



(d) The visualisation in 2009.

Figure 8: The visualisation results of image classification.

[7] Masroor Hussain, Dongmei Chen, Angela Cheng, Hui Wei, and David Stanley. Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of photogrammetry and remote sensing*, 80:91–106, 2013.

[8] Richard J Radke, Srinivas Andra, Omar Al-Kofahi, and Badrinath Roysam. Image change detection algorithms: a systematic survey. *IEEE transactions on image processing*, 14(3):294–307, 2005.

[9] Pol R Coppin and Marvin E Bauer. Digital change detection in forest ecosystems with remote sensing imagery. *Remote sensing reviews*, 13(3-4):207–234, 1996.

[10] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE geoscience*

and remote sensing magazine, 5(4):8–36, 2017.

[11] Jin Chen, Peng Gong, Chunyang He, Ruiliang Pu, and Peijun Shi. Land-use/land-cover change detection using improved change-vector analysis. *Photogrammetric Engineering & Remote Sensing*, 69(4):369–379, 2003.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep learning. *MIT press*, 2016.

[14] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[15] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Cambridge university press*, 2008.

[16] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE, 2006.

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[18] Chen Wu, Liangpei Zhang, and Bo Du. Kernel slow feature analysis for scene change detection. *IEEE Transactions on Geoscience and Remote Sensing*, 55(4):2367–2384, 2017.

[19] Bernhard Schölkopf and Alexander J Smola. Learning with kernels: support vector machines, regularization, optimization, and beyond. *MIT press*, 2002.

[20] Christopher M Bishop. Pattern recognition and machine learning. *Springer google schola*, 2:1122–1128, 2006.

Appendix

The following abbreviations are used in this manuscript:

BoVW	Bag of Visual Words
CNN	Convolutional Neural Network
MtS-WH	Multi-temporal Scene Dataset
RBF	Radial Basis Function Kernel
SIFT	Scale-Invariant Feature Transform
SPM	Spatial Pyramid Matching
SVM	Support Vector Machine

Acknowledgment

This work was supported by JST SPRING, Grant Number JPMJSP2136.

Yunjia Huang (Member) began his Ph.D. course at Kyushu University in 2022. His research focuses on image processing and neural networks.

Kohei Inoue (Member) received B.Des., M.Des., and D.Eng. degrees from Kyushu Institute of Design in 1996, 1998, and 2000, respectively. He is currently an Associate Professor at Kyushu University. His research interests include pattern recognition and image processing.

Dongqi Li (Non-member) began her Ph.D. course at Kyushu University in 2021. Her research focuses on using digital methods to study geographical environments and built heritage.