





Collaborative Clustering Approach Based on Dempster-Shafer Theory for Bag-of-Visual-Words Codebook Generation

Sabrine Hafdhellaoui, Yaakoub Boualleg^(✉) , and Mohamed Farah 

Univ. Manouba, ENSI, RIADI, LR99ES26 (SIIVT Team), Manouba, Tunisia
hafdellaouisabrine@gmail.com, yaakoub.boualleg@ensi-uma.tn,
mohamed.farah@riadi.rnu.tn

Abstract. Feature encoding methods play an important role in the performance of the recognition tasks. The Bag-of-Visual-Words (BoVW) paradigm aims to assign the feature vectors to the codebook visual words. However, in the codebook generation phase, different clustering algorithms can be used, each giving a different set of visual words. Thus, the choice of the discriminative visual words set is a challenging task. In this work, we propose an enhanced bag-of-visual-words codebook generation approach using a collaborative clustering method based on the Dempster-Shafer Theory (DST). First, we built three codebooks using the k-means, the Fuzzy C-Means (FCM), and the Gaussian Mixture Model (GMM) clustering algorithms. Then, we computed the Agreement Degrees Vector (ADV) between the clusters of the pairs (k-means, GMM) and (k-means, FCM). We merged the obtained ADVs using the DST in order to generate the clusters weights. We evaluated the proposed approach for Remote Sensing Image Scene Classification (RSISC). The results proved the effectiveness of our proposed approach and showed that it can be applied for different recognition tasks in various domains.

Keywords: Bag-of-Visual-Words · Codebook generation · Collaborative clustering · Dempster-Shafer Theory · Remote sensing image scene classification

1 Introduction

With the rapid advance of imaging technologies, a huge amount of visual information is becoming more and more available in different digital archives, whether they are publicly available on the Web or used for specialized applications such as in the Remote Sensing (RS) for which many freely high resolution images such as SPOT and SENTINEL are available. All these available satellite data are very useful for a wide range of critical applications in agriculture, deforestation, urban planning, etc.

The Remote Sensing Image Scene Classification (RSISC) aims to label and identify each image scene with its corresponding class. Since the classification performance is strongly affected by the effectiveness of the features vector, considerable efforts have been made to develop powerful feature representations. Early, image classification methods have been intensively using handcrafted global low-level visual features that are extracted from the whole image such as color, texture, and shapes [4]. Next, researches have focused more on local low-level features that are extracted from the interest points within the image such as Scale-Invariant Feature Transform (SIFT) [14], Speed Up Robust Feature (SURF) [2], Local Binary Pattern (LBP) [16], Histogram of Oriented Gradient (HOG) [6], and Pyramid Histogram of Oriented Gradient (PHOG) [3].

However, due to the high dimensionality of these features, as well as the time they need to be computed and processed, researches tend to map low-level image visual features into mid-level image representations through feature encoding methods such as Bag-of-Visual-Words (BoVW) [20], Vector of Locally Aggregated Descriptors (VLAD) [11], and Improved Fisher Kernel (IFK) [17].

Recently, Convolutional Neural Networks (CNNs), which are Deep-Learning (DL) architectures, showed significant progress in computer vision tasks. However, they have many limitations. First, learning CNN models from scratch requires a huge amount of labelled data. In addition, parameters tuning is an uninterpretable process and requires high computational power. Moreover, they are highly prone to overfitting. Transfer-Learning strategies have been proposed to alleviate the cited limits through fine tuning pretrained models or by using them as features extractors. More recently and motivated by the results of the use of CNNs for extracting deep features, new feature representations are proposed. In [5], the authors proposed to use the extracted deep convolutional feature maps from the pretrained CNN instead of using dense SIFT features. Also, in [13] multi-scale convolutional feature maps are aggregated using IFK to generate a better image representation.

Over the years, mid-level image representations, especially the BoVW method, have received increasing interest from the image classification community. This is because they have proven to be efficient for discriminating feature representations. The BoVW method was firstly proposed by Szelinski [22] based on the work of Sivic and Zisserman [20]. The main idea is to obtain image descriptions from a training set in order to generate a codebook or book of visual words by clustering the image features and using clustering centres as the words of the codebook. Then, the image is represented by the histogram of the visual words.

Most of the existing BoVW-based methods for RSISC have extensively explored various features as well as combinations of various strategies to generate the codebook of visual words. Sujatha et al. [21] proposed a multi-dictionaries model by combining the dictionaries resulting from the FCM clustering algorithm with different subsets of SIFT descriptors. In fact, during the feature-grouping step, n subsets of SIFT descriptors are randomly selected, and n dictionaries are generated using FCM. n histograms are therefore generated for each image and the final result is obtained from the concatenation of these n histograms. Jonathan et al. [15] investigated the use of a Dual BoVW model (Dual-BoVW) in

a relatively conventional framework to perform image classification. They showed the superiority of a BoVW with the combination of two local-feature descriptors by creating a dual codebook which contains both local features (Dual BoVW) compared to the conventional BoVW methods (BoVW and HOG-BoVW) with a single codebook. Zurita et al. [23] proposed a hybrid classification in the BoVW Model. Firstly, SIFT descriptor was used in the feature extraction phase. Then a dictionary of words was created through a clustering process using k-means, Expectation Maximisation algorithm (EM) and k-means in combination with EM. Finally, for the classification, they compared the algorithms of Support Vector Machine (SVM), Gaussian Naïve Bayes (GNB), k-Nearest Neighbours (kNN), Decision Tree (DT), Random Forest (RF), Neural Network (NN) and AdaBoost in order to determine the performance and accuracy of every method.

In the codebook generation phase, Sivic and Zisserman. [20] used the k-means classifier to construct the vocabulary of the BoVW model. Farquhar et al. [8] have proposed an extended grouping based on a generative model called the Gaussian Mixture Model (GMM). Avrithis and Kalantidis. [1] have proposed an approximate version of GMM, called Approximate Gaussian Mixture (AGM), Sujatha et al. [21] have proposed to use the Fuzzy C-Means classifier (FCM). Each of these classification models produces different visual words for the same sample of images, which poses the problem of choosing the best set and therefore the best classifier.

Besides, in order to improve the classification results, several authors have proposed techniques for fusing multiple unsupervised classifiers. Gañarski and Wemmert. [10] proposed a multiple-view voting method to combine unsupervised classifications. Forestier et al. [9] proposed the collaborative-clustering method to fuse data. This method allows the user to exploit different heterogeneous images in a global system. The process consists of three stages: initial and parallel execution of the classifiers, result refinement and result unification. In the second step, different classifications need to converge through an assessment and a resolution of the existing conflicts. The search is done two by two. For two results, the correspondence between classes is recommended via a similarity measure. Once the refinement is complete, the results are unified using a voting algorithm.

In order to overcome the problem caused by the conflict generated by different clustering algorithms in the codebook generation and to enhance the BoVW model performance, we propose a new codebook generation approach that uses at the same time the different results produced by several classifiers in a collaborative clustering method based on the Dempster-Shafer Theory (DST). Firstly, the extracted image features are clustered into k clusters using three classifiers that are k-means, GMM, and FCM. We then apply a collaborative clustering of these three clustering results, taking the results of k-means as reference clusters since k-means is usually used for this step. The collaboration is done between the results of k-means and GMM on the one side, then between those of k-means and FCM on the other side. Indeed, for each cluster resulting from k-means, we associate a mass function. These masses measure the degree of agreement between k-means results and those of GMM as well as those of FCM. Then we

use the DST to fuse the results. Finally, we obtain for each k-means cluster (visual word) a weight which indicates the cluster’s confidence. The Fusion-Agreement-Degree vector (visual words weights) is used for reweighting the final image representation.

The rest of this paper is structured as follows. In Sect. 2, we describe the DST for information fusion. In Sect. 3, we focus on the proposed approach. Section 4 describes the experimental implementation and evaluation of the obtained results. We will end with a conclusion summarizing our proposal.

2 The Dempster-Shafer Theory

Information fusion is a multilevel process which serves to combine information from multiple sources to improve decision-making. We report in this section necessary theoretical elements of the DST. This theory comes from the work of Dempster [7] which was resumed by Shafer [18]. It allows modelling information imperfections, particularly the conflicts. The formalism can be described as follows. Let θ be the framework of discernment, which describes all the possible hypotheses $\theta = \{H_1, H_2, H_3, \dots, H_k\}$. The set 2^θ of all the partitions of θ is given by:

$$2^\theta = \{A, A \subseteq \theta\} = \{\{H_1\}, \{H_2\}, \dots, \{H_k\}, \{H_1 \cup H_2\}, \dots, \theta\} \tag{1}$$

A first magnitude called ‘mass of belief’ can be constructed. This magnitude characterizes the veracity of a proposition A for an information source S . The mass m associated with this source is defined over all the partitions of the framework of discernment θ , i.e. 2^θ , as follows:

$$m : \begin{cases} 2^\theta \longrightarrow [0, 1] \\ A \longrightarrow m(A) \end{cases} \tag{2}$$

where $\sum_{A \in 2^\theta} m(A) = 1$ and $m(\emptyset) = 0$.

Each subset $A \subset \theta$ where $m(A) > 0$ is called a focal element of m . The union of the focal elements is called the nucleus. The complete ignorance of the hypotheses set corresponds to $m(\theta) = 1$.

The DST defines precisely a mass combination rule when there are different sources. Let $A, B \in 2^\theta$ and two sources S_1 and S_2 expressing the masses of belief $m_1(*)$ and $m_2(*)$ on the elements of 2^θ . The mass of the hypothesis A resulting from the fusion of masses $m_1(*)$ and $m_2(*)$ by the application of the Dempster rule is given by:

$$\begin{cases} (m_1 \oplus m_2)(A) = \frac{\sum_{B_1 \cap B_2 = A} m_1(B_1)m_2(B_2)}{1-K} \\ (m_1 \oplus m_2)(\emptyset) = 0 \end{cases} \tag{3}$$

where K is defined by:

$$K = \sum_{B_1 \cap B_2 = \emptyset} m_1(B_1)m_2(B_2) \tag{4}$$

As part of the Dempster combination rule, the mass on the empty set \emptyset must be zero and the sum of the masses on 2^θ must equal 1. Therefore, it is necessary to redistribute the mass assigned to the empty set on all the other masses. For this, the final mass distribution must be renormalized with the renormalization coefficient K . The rule of Dempster is a rule of a conjunctive consensus normalized by the conflict K . This global conflict is the sum of the partial conflicts resulting from the empty intersections of the focal elements of the different mass functions.

3 Proposed Approach

In this work, we propose an enhanced BoVW codebook generation based on a collaborative clustering approach. The proposed codebook uses the clustering results of several unsupervised classifiers at the same time. We use DST to reduce the conflict between the obtained clustering results and to generate the visual words' weights. The proposed weighted cookbook is tested within an image classification framework for RSISC.

As we can see from Fig. 1, the proposed approach takes as input the image descriptors and outputs a new feature representation based on a weighted codebook with a visual words weights vector, through four steps.

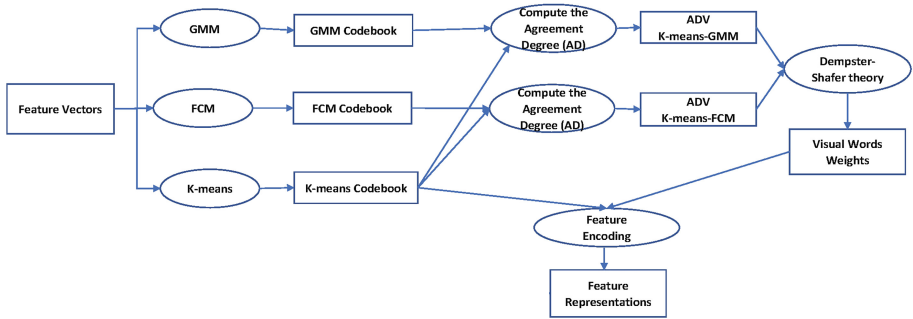


Fig. 1. The overall scheme of the proposed Bag-of-Visual-Words codebook generation.

3.1 Codebook Generation

We apply different unsupervised clustering algorithms separately on the input image feature set. We use three clustering algorithms namely k-means, FCM, and GMM.

Let R_1, R_2, R_3 denote the set of clustering results (codebooks) that are given by k-means, GMM, and FCM respectively, where $\|R_1\| = N_1$, $\|R_2\| = N_2$, and $\|R_3\| = N_3$. And let C denote the obtained clusters as follow: $C_i^1 \in R_1$, $C_j^2 \in R_2$, and $C_l^3 \in R_3$ where $i \in [1, N_1]$, $j \in [1, N_2]$, and $l \in [1, N_3]$.

3.2 Modelling

According to the traditional BoVW-based methods and inspired by [12], we consider that the clustering result provided by k-means presents the reference classes. To merge the obtained clustering results, we build the mass functions of the unsupervised classifiers GMM and FCM, by measuring the degree of agreement between the reference classes (k-means results) and the obtained clusters using GMM and then using FCM. In order to assign the masses, we look for the proportions of the reference classes in each cluster using the intersection matrix (*intrscM*) that is obtained using intercluster correspondence function [9] (see Eq. 5). Next, the Agreement Degree Vector (ADV) is generated using the Maximum function on the *intrscM* as described by Eq. 6 that represents the correspondence between the most similar clusters.

Define Mass Functions by Collaborating k-Means with GMM:

$$intrscM(R_1, R_c) = \begin{bmatrix} a_{1,1}^{1,c} & \cdots & a_{1,N_c}^{1,c} \\ \vdots & & \vdots \\ a_{N_1,1}^{1,c} & \cdots & a_{N_1,N_c}^{1,c} \end{bmatrix} \tag{5}$$

where $a_{i,j}^{1,c} = \frac{|C_i^1 \cap C_j^c|}{|C_i^1|}$.

$$m_1 = ADV(R_1, R_c) = \begin{pmatrix} \max(a_{1,1}^{1,c} \cdots a_{1,N_c}^{1,c}) \\ \vdots \\ \max(a_{N_1,1}^{1,c} \cdots a_{N_1,N_c}^{1,c}) \end{pmatrix} \tag{6}$$

where $c = 2$ to define mass functions by collaborating k-means with GMM and $c = 3$ to Define mass functions by collaborating k-means with FCM.

3.3 Fused Decision

In our method, there are two distinct and independent sources of information: the GMM and the FCM algorithms. Each source has its own mass vector (ADV) that is defined on the clustering result of k-means ($ADV(R_1, R_2) \neq ADV(R_1, R_3)$).

In order to benefit from both information sources, we fuse the normalized ADVs using the DST. We use the orthogonal Sum of DST (Eq. 3) where m_1 and m_2 are the mass functions corresponding to each cluster in R_1 with GMM clusters R_2 ($ADV(R_1, R_2)$) and with FCM clusters R_3 ($ADV(R_1, R_3)$), respectively. Finally, we get a new vector FADV (Fused-Agreement Degree Vector) where each value represents a weight associated with a k-means cluster (visual word).

3.4 Feature Encoding

Based on the clustering results of the k-means algorithm, with N_1 visual words (codebook size), we encode an input image using a global histogram representation which is determined by the frequency of each codebook visual word within

the image. Next, the obtained N_1 -dimensional image feature representation is reweighted using the visual words weights (FADV) based on a simple pairwise multiplication function.

4 Experiments and Results

4.1 Dataset

In order to evaluate our proposed approach, our experiments were conducted on the “NWPU-RESISC45” dataset¹ [5] which was proposed for REmote Sensing Image Scenes Classification (RESISC) by the Northwestern Polytechnical University (NWPU). The dataset is the largest publicly available aerial image dataset with 31500 remote sensing RGB images. It consists of 45 land-use classes, with 700 images per class. The aerial scene images of this dataset are acquired from Google Earth (Google Inc.) covering more than 100 countries and regions around the world. The image size is 256×256 pixel with a different special resolution that varies from 30 to 0.2 m. Figure 2 shows some sample images from this dataset.

4.2 Experimental Setup

To compare our proposed approach with the RSISC methods that are based on the BoVW feature encoding method, we selected two baseline methods; the traditional BoVW that uses the SIFT features, and BoCF which uses the convolutional feature maps from the VGG16 pretrained model [19]. In order to compare the classification performances of the proposed method with BoCF obtained results [5], we use the same training/test rate, the dataset was randomly split into 10% for the training and 90% for the test.

In the first experiment, we uses the $128D$ dense SIFT feature vector to describe the dataset’s images and we compared the results with the BoVW results. Secondly we used the $13 \times 13 \times 256$ deep-feature maps extracted from the pooling layer from the last convolutional block of the pretrained VGG16 CNN model in order to compare the obtained results with the BoCF results. Similarly to the BoVW and BoCF methods, in the codebook generation phase, for the k-means, GMM and FCM clustering algorithms, we set the number of clusters C to 500, 1000, 2000, 5000, and 10000 and we investigated the impact of the codebook size on the classification performance of the proposed approach.

The classification results were obtained using a linear SVM classifier. We evaluated the classification performance using the same evaluations metrics that used in [5] (i.e. Overall Accuracy (OA) and the confusion matrix metrics). All the experiments were performed with a t2.2xlarge machine that is available on the Amazon Web Service EC2 instance². The proposed approach was implemented in python 2.7.

¹ <http://www.escience.cn/people/JunweiHan/NWPU-RESISC45.html>.

² <https://aws.amazon.com/ec2/instance-types/>.



Fig. 2. Sample images from the NWPU-RESISC45 dataset [5]

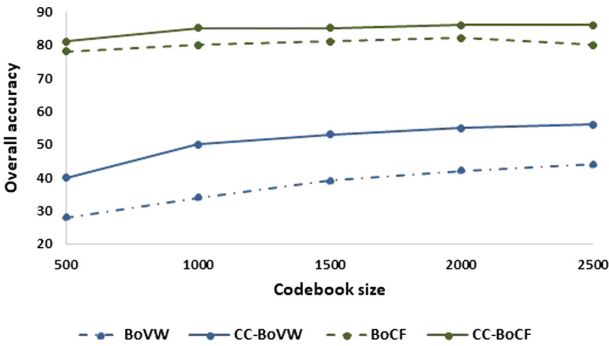


Fig. 3. Overall accuracies on the NWPU-RESISC45 dataset of BoVW with SIFT, BoCF with VGG16 deep features, and their enhanced versions using the proposed collaborative clustering method CC-BoVW and CC-BoCF respectively.

4.3 Results and Discussion

Figure 3 shows the Overall Accuracies (OA) that are obtained by using the baseline methods, the BoVW using SIFT features and the BoCF using the VGG16 deep features, and their enhanced versions using the proposed collaborative clustering method CC-BoVW and CC-BoCF in terms of the codebook size.

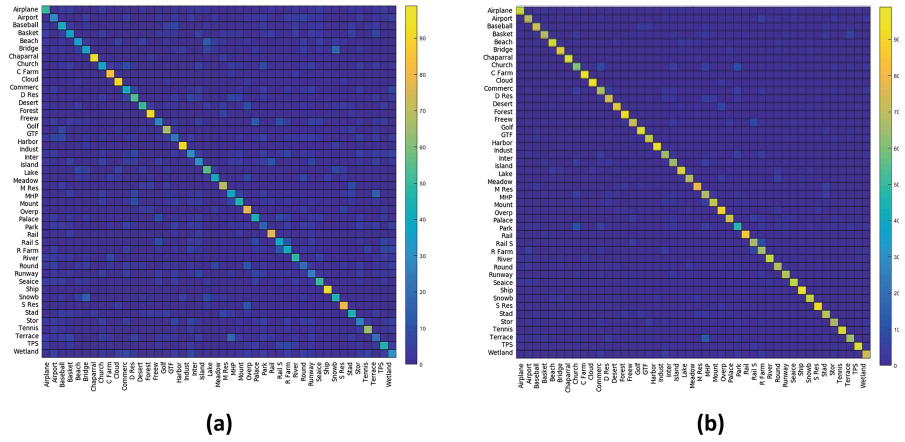


Fig. 4. Confusion matrices showing classification performance on the NWPU-RESISC45 dataset for BoVW with SIFT (a) and the enhanced version using the proposed collaborative clustering method CC-BoVW (b).

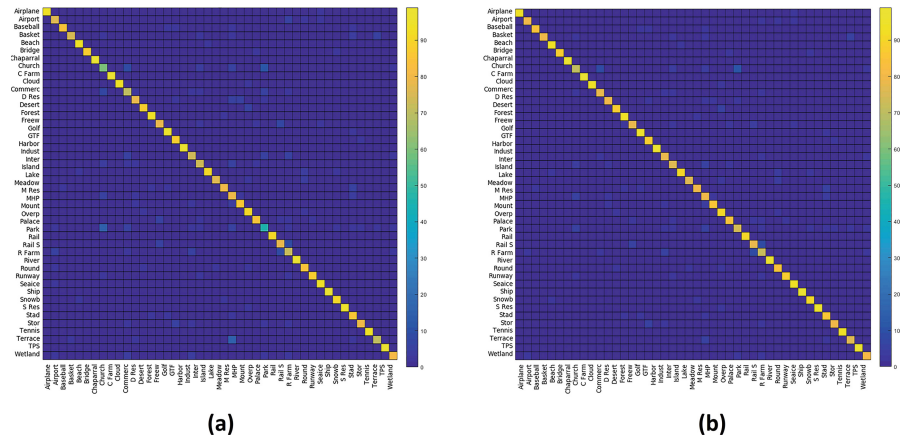


Fig. 5. Confusion matrices showing classification performance on the NWPU-RESISC45 dataset for BoCF with VGG16 deep features (a) and the enhanced version using the proposed collaborative clustering method CC-BoCF (b).

From Fig. 3, we can observe that the overall accuracy is influenced by the codebook size; for the BoCF, the codebook size 5000 gives the best OA. However, for the other methods, the largest codebook size gives the highest OA. The proposed approach achieves a better performance compared to the baseline methods on all the codebook size variation. For codebook size 10000, our proposed approach CC-BoVW achieves 54.2% OA, which is better than the traditional BoVW by 10.2%. Also, for the proposed CC-BoCV, the OA is boosted by 5.8% compared

to the traditional BoCF. For more details, the classification performance for each class is presented on the confusion matrices where the rows and columns of the matrix represent actual and predicted classes, respectively. In Fig. 4, the BoVW is compared to the proposed CC-BoCV, and in Fig. 5, the BoCF is compared with the proposed CC-BoCF.

5 Conclusion

In this paper, we propose a new Bag-of-Visual-Words codebook generation approach, based on a collaborative clustering method using the DST. In the codebook generation, each clustering algorithm gives a different codebook. In order to reduce the conflict between these sources of information, the collaborative clustering method was used to associate a weight value for each visual word of the k-means codebook. This weight represents the fused agreement degree of the GMM and the FCM codebooks with the k-means clusters. To fuse the obtained agreement degree vectors, we used the orthogonal sum of the DST. The proposed approach was evaluated on a remote sensing image scene classification framework, which achieves encouraging results compared to the RSISC baselines, and this showed that it can be applied for different recognition tasks in various domains.

References

1. Avrithis, Y., Kalantidis, Y.: Approximate gaussian mixtures for large scale vocabularies. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 15–28. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_2
2. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006). https://doi.org/10.1007/11744023_32
3. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, pp. 401–408. ACM (2007)
4. Cheng, G., Han, J.: A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **117**, 11–28 (2016)
5. Cheng, G., Li, Z., Yao, X., Guo, L., Wei, Z.: Remote sensing image scene classification using bag of convolutional features. *IEEE Geosci. Remote Sens. Lett.* **14**(10), 1735–1739 (2017)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
7. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. *Ann. Math. Statist.* **38**(2), 325–339 (1967). <https://doi.org/10.1214/aoms/1177698950>
8. Farquhar, J., Szedmak, S., Meng, H., Shawe-Taylor, J.: Improving ‘bag-of-keypoints’ image categorisation: Generative models and pdf-kernels (2005)

9. Forestier, G., Wemmert, C., Gañçarski, P.: Multisource images analysis using collaborative clustering. *EURASIP J. Adv. Sig. Process.* **2008**(1), 11 (2008)
10. Gañçarski, P., Wemmert, C.: Collaborative multi-strategy classification: application to per-pixel analysis of images. In: *Proceedings of the 6th International Workshop on Multimedia Data Mining: Mining Integrated Media and Complex Data*, pp. 15–22. ACM (2005)
11. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3304–3311. IEEE (2010)
12. Karem, F., Dhibi, M., Martin, A., Bouhleb, M.S.: Credal fusion of classifications for noisy and uncertain data. *Int. J. Electr. Comput. Eng. (IJECE)* **7**(2), 1071–1087 (2017)
13. Li, E., Xia, J., Du, P., Lin, C., Samat, A.: Integrating multilayer features of convolutional neural networks for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **55**(10), 5653–5665 (2017)
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **60**(2), 91–110 (2004)
15. Maas, J.L., Okafor, E., Wiering, M.A.: The dual codebook: combining bags of visual words in image classification. In: *Proceedings of the 28th Benelux Artificial Intelligence Conference (BNAIC)*, pp. 46–71 (2016)
16. Ojala, T., Pietikainen, M., Maenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 971–987 (2002)
17. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010. LNCS*, vol. 6314, pp. 143–156. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-15561-1_11
18. Shafer, G.: *A Mathematical Theory of Evidence*, vol. 42. Princeton University Press, Princeton (1976). ISBN 9780691100425
19. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014)
20. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 1470. IEEE (2003)
21. Sujatha, K., Keerthana, P., Priya, S.S., Kaavya, E., Vinod, B.: Fuzzy based multiple dictionary bag of words for image classification. *Procedia Eng.* **38**, 2196–2206 (2012)
22. Szeliski, R.: *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, Heidelberg (2010). <https://doi.org/10.1007/978-1-84882-935-0>
23. Zurita, B., Luna, L., Hernandez, J., Ramirez, J.: Hybrid classification in bag of visual words model. *Circ. Comput. Sci.* **3**(4), 10–15 (2018)