

Research Article

Recognition of Urban Buildings in Challenging Images Using Bag of Features and SVM

Leandro Aureliano Silva^{1,*} , **Eduardo Silva Vasconcelos²** ,
Wellington Mrad Joaquim¹ , **Luiz Fernando Ribeiro Paiva³** ,
Edilberto Pereira Teixeira⁴ 

¹Department of Electrical Engineering, University of Uberaba, Uberaba, Brazil

²Department of Mathematics, Federal Institute Goiano, Cristalina, Brazil

³Department of Computer Engineering, University of Uberaba, Uberaba, Brazil

⁴Department of Chemical Engineering, University of Uberaba, Uberaba, Brazil

Abstract

Urban building recognition plays a central role in applications such as urban mapping, heritage documentation, autonomous navigation, and smart city monitoring. Although recent advances have been driven mainly by deep learning approaches, classical visual pipelines remain an attractive alternative in scenarios where datasets are limited, interpretability is required, and computational resources are constrained. In this study, a systematic evaluation of a Bag-of-Features (BoF) representation combined with a Support Vector Machine (SVM) classifier is presented for urban building recognition using the Sheffield Building Image Dataset (SBID). The experimental protocol includes dataset balancing, a reproducible training–testing split, and an extensive investigation of visual vocabulary sizes ranging from 100 to 3000 visual words. The results indicate that increasing the vocabulary size generally improves recognition performance up to a saturation point, with the best trade-off achieved using 2000 visual words. Under this configuration, the proposed approach achieved an overall accuracy of 97.5% while maintaining an average inference time below 25 ms per image, demonstrating competitive performance with low computational cost. A detailed analysis based on confusion matrices and per-class metrics (accuracy, precision, recall, and F1-score) shows that most building categories were recognized with high reliability, while misclassifications were mainly concentrated among visually similar façade types. These findings confirm that BoF representations, when properly tuned, remain highly effective for structured urban recognition tasks. Moreover, the obtained results are consistent with those commonly reported in the literature for the same dataset and problem domain, reinforcing the robustness of the proposed pipeline. Overall, the results highlight the continued relevance of classical computer vision methods in contexts where transparency, reproducibility, and efficiency are essential. Future work will investigate hybrid strategies that combine BoF representations with deep convolutional descriptors, as well as more robust evaluation protocols, aiming to improve generalization across different building datasets and urban environments.

Keywords

Bag of Features, SVM, Building Recognition, Computer Vision, SBID

*Corresponding author: leandro.silva@uniube.br (Leandro Aureliano Silva)

Received: 22 December 2025; **Accepted:** 5 January 2026; **Published:** 23 January 2026



Copyright: © The Author(s), 2026. Published by Science Publishing Group. This is an **Open Access** article, distributed under the terms of the Creative Commons Attribution 4.0 License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Automatic recognition of buildings in urban environments is one of the central challenges of computer vision applied to engineering and artificial intelligence. This field has gained relevance due to its applications in autonomous navigation, visual geolocation, three-dimensional city reconstruction, urban monitoring, and the preservation of historical heritage. Building-recognition methods have evolved from approaches based on texture and shape to hybrid systems that integrate local descriptors, Bag of Features (BoF) encoding strategies, and supervised classifiers such as Support Vector Machines (SVM). Nevertheless, persistent challenges remain, including variations in viewpoint, illumination, and occlusions [1].

The Sheffield Building Image Dataset (SBID) is widely used as a benchmark for building-recognition experiments, as it contains 3,192 images of 40 distinct buildings with variations in angle, illumination, and environmental conditions [2]. This diversity makes it ideal for validating algorithms that must be robust to visual and spatial variability. It is important to emphasize that this dataset will be used in the present study.

Among feature-extraction methods, the work of Silva [3] stands out for proposing the combined use of the Gabor filter, the wavelet transform, and a Multilayer Perceptron (MLP) neural network. This approach demonstrated high accuracy in complex urban scenarios, since Gabor filters capture orientation-dependent texture information and the wavelet transform performs efficient multiresolution analysis, reducing data dimensionality without significant loss of information. These results reinforce the potential of combined textural descriptors for the preprocessing stage in façade-recognition systems.

On the other hand, recent advances propose the integration of classical representation techniques with features extracted from deep networks. Some authors [4] introduced a Weighted Bag of Visual Words (W-BoVW) model that combines the traditional visual histogram with refined deep features, achieving superior performance in image classification. The idea of weighting visual words according to their statistical and semantic relevance inspires this research to explore BoF alongside robust feature-selection strategies for urban imagery.

Additionally, more recent studies [5] point to the use of remote-sensing and social data associated with spatial constraints to identify building usage and function. Although the authors' goal differs—focusing on characterizing urban land use—their contribution reinforces the importance of considering spatial context and geometric consistency in building recognition, especially in dense urban environments [5].

Okun et al. [4] propose a building-recognition system based on color histograms combined with a Support Vector Machine (SVM) for images of the city of Ipoh, Malaysia. This approach demonstrates that even relatively simple descriptors (such as color histograms) can perform effectively in building recognition when properly calibrated.

Thus, this study proposes the development of an ur-

ban-building recognition system using the Bag of Features (BoF) technique and Support Vector Machine (SVM) classification, implemented in MATLAB and evaluated using the SBID dataset. The approach seeks to combine the efficiency and interpretability of classical BoF models with the robustness of textural descriptors, offering a high-performance and computationally efficient alternative compared to purely deep-learning methods.

This paper is organized into sections. First, the Theoretical Framework presents the main concepts that support the study, including the functioning of the Bag of Features (BoF) technique, its developments, and its applications in computer vision, as well as the principles of Support Vector Machines (SVM) used in the classification stage. Next, the Materials and Methods section describes in detail the adopted dataset, the preprocessing procedures, the extraction of visual features, and the training and evaluation process of the classifier, ensuring the reproducibility of the approach. The Results section objectively presents the performance achieved by the system, including comparative analyses between different visual-vocabulary configurations and related studies from the literature. Finally, the Conclusion summarizes the main findings, highlights the effectiveness of the proposed method, and points out possibilities for future research.

2. Theoretical Foundations

2.1. Bag of Features (BoF)

The Bag of Features (BoF) model, also known as Bag of Visual Words (BoVW), is a widely used technique in computer vision for representing images in an efficient and scalable manner. The approach is directly inspired by the Bag of Words (BoW) paradigm from the field of text processing, in which each document is represented as a histogram of term frequencies, without considering the order or syntactic structure of those terms [6, 10].

2.1.1. Fundamental Steps of BoF

The typical workflow for constructing a representation using BoF can be divided into three main stages:

Local descriptor extraction:

In each image, interest points (keypoints) or sampled regions (patches) are detected, and descriptors characterizing local properties (orientation, texture, gradient, etc.) are extracted. These descriptors are numerical vectors of fixed dimension; for example, the Scale-Invariant Feature Transform (SIFT) algorithm produces 128-dimensional vectors. Stack Exchange [7].

Construction of the visual vocabulary (codebook):

Descriptors extracted from the training images are grouped (typically through k-means or another clustering method) into

k clusters, whose centroids form the visual vocabulary (or “visual words”). Each centroid represents a “word,” and the set of centroids composes the visual dictionary. Stack Exchange [7].

Image encoding and histogram-vector formation:

For a new test image, local descriptors are extracted, each descriptor is assigned to its nearest visual word (nearest cluster), and a frequency histogram (sometimes weighted) of length k is generated. This histogram represents the entire image as a fixed-dimension vector, regardless of the original number of descriptors. The resulting vector can then be fed into a classifier (e.g., SVM) or used for similarity search. Stack Exchange [7].

2.1.2. Advantages and Limitations

One of the major advantages of BoF is the uniformization of representation: images with different quantities of key-points or patches end up being represented by vectors of the same dimensionality, which facilitates classifier application. Furthermore, the model ignores spatial order and more complex relationships among patches—reducing complexity and making the method effective in many practical scenarios [6].

However, this same characteristic is also a limitation: by ignoring or drastically reducing information about the spatial arrangement of descriptors, the Bag of Features (BoF) may lose the ability to correctly characterize patterns in which spatial structure is relevant, such as the layout of windows on a building or architectural contours. Even so, adaptations such as Spatial Pyramid Matching (SPM) and the incorporation of spatial coordinates have been proposed to mitigate this limitation [8, 9].

2.1.3. Evolutions and Contemporary Relevance

More recently, BoF remains relevant even with the emergence of deep neural networks (CNNs). For example, hybrid systems combine classical BoF descriptors with deep features, or adapt the BoF approach to operate on activation maps from deep networks, achieving richer representations. Sitaula & Aryal (2021) apply the idea of Bag of Deep Visual Words for X-ray image classification, demonstrating that the BoF paradigm remains useful in modern domains.

Additionally, BoF has been widely used as a baseline in studies on object, scene, or building recognition due to its robustness, conceptual simplicity, and ease of implementation in tools such as MATLAB, making it suitable for research purposes and rapid prototyping.

2.2. Classification with SVM

Vector Machines (SVM) are among the most effective and widely used classifiers in pattern recognition and supervised classification problems. Their fundamental principle is to map data to a high-dimensional space and find the optimal separating hyperplane that maximizes the margin between classes, ensuring greater generalization capacity [11, 12].

SVMs are particularly efficient in situations involving small sample sizes and high dimensionality, as they use kernel functions to perform non-linear separations without needing to explicitly transform the data, as highlighted by the authors [13].

When combined with feature extraction techniques such as the Bag of Features (BoF), SVMs have achieved remarkable results in image categorization and recognition due to their robustness to visual variability and their capacity to handle high-dimensional feature vectors [4]. This combination has been widely adopted in systems for object, scene, and building recognition, consolidating itself as an efficient and interpretable paradigm in modern computer vision.

3. Materials and Methods

3.1. Dataset

To evaluate the performance of the proposed method, the experiments were conducted using the Sheffield Building Image Dataset (SBID).

The dataset contains 40 building categories, comprising a total of 3,192 JPEG images with a resolution of 160×120 pixels. All images were collected in the central area of the University of Sheffield campus, United Kingdom, using a Canon G90 digital camera, and include both photographs and frames extracted from video sequences.

The images exhibit several types of variation, including changes in rotation, scale, illumination conditions, occlusions, and motion blur, making the dataset particularly suitable for evaluating robust building-recognition methods [1].

3.2. Preprocessing

To balance the dataset, the minimum number of images among the classes was adopted. The dataset was then randomly divided into 60% for training and 40% for testing, ensuring representativeness across all categories.

3.3. Feature Extraction

Feature extraction was carried out using the `bagOfFeatures` function, with the visual vocabulary size varied across the following values: 100, 300, 600, 900, 1200, 1500, 2000, and 3000 words. Interest points were automatically detected based on local descriptors, without manual intervention. The `encode` function was used to convert the images into feature vectors, which were then visualized through histograms.

3.4. Classification and Evaluation

Classification was performed using `train Image CategoryClassifier`, which trains an SVM model based on the encoded images. Performance evaluation was conducted using a confusion matrix obtained with the `evaluate` function, which

returns the accuracy and error rates by comparing predicted and true labels.

3.5. Implementation Details

All experiments were conducted on the Sheffield Building Image Dataset (SBID), whose images were organized into subfolders, each corresponding to a distinct building class. The dataset was loaded in MATLAB using the `imageSet` function with the recursive option, which automatically assigns class labels according to the directory structure. To mitigate class imbalance, the minimum number of available samples across all classes was first computed. A maximum of 20 images per class was then selected, ensuring that this value did not exceed the smallest class size. Each category was randomly subsampled to this number using the `partition` function with the `randomize` option, resulting in a balanced dataset in which all classes contain the same number of images.

The balanced dataset was split into training and test sets, with 60% of the images used for training and 40% for testing. This split was performed with `partition` (0.6, 'randomize'), which implements a stratified random partition. To guarantee reproducibility, a fixed random seed (`rng(42)`) was set before all splitting operations, so that the same examples are consistently assigned to training and test sets across runs.

Feature extraction followed the Bag-of-Features (BoF) paradigm, implemented via the MATLAB `bagOfFeatures` function. Multiple visual vocabulary sizes were evaluated in a systematic way, specifically 100, 300, 600, 900, 1200, 1500, 2000, and 3000 visual words. For each vocabulary size, a separate codebook was learned from the training images using automatically detected interest points ('PointSelection' = 'Detector'). Each image was then encoded as a histogram of visual word occurrences, normalized to compensate for differences in the number of detected keypoints.

Classification was performed with the `trainImageCategoryClassifier` function, which builds a multiclass Support Vector Machine (SVM) based on error-correcting output codes (ECOC) on top of the BoF feature vectors. For each vocabulary size, a distinct classifier was trained using only the training set. Evaluation on the test set was carried out using the `evaluate` function, which returns the predicted labels for all

test images. From these predictions, a confusion matrix in terms of raw counts was computed via `confusionmat`, and the overall accuracy was obtained as the ratio between the trace of this matrix and the total number of test samples. In addition, the average per-class accuracy was computed as the mean of the diagonal entries of the confusion matrix normalized by the number of samples in each class.

To analyze computational efficiency, the total processing time for each configuration was measured using MATLAB's `tic/toc` functions, starting just before the BoF construction and ending after the evaluation step. This total time includes visual vocabulary learning, SVM training, and inference on the test set. The inference time alone was measured separately by placing `tic/toc` around the evaluation stage. The average inference time per image was then obtained by dividing the inference time by the number of test samples, providing a practical estimate of the computational cost of deploying the classifier.

Finally, for the vocabulary size that yielded the highest overall accuracy (2000 visual words), a more detailed analysis was conducted. The classifier was retrained with this configuration, and a confusion matrix was generated and visualized using `confusionchart`. From this matrix, class-level performance metrics were derived, including accuracy, precision, recall, and F1-score for each building category. These metrics were arranged in tabular form and used to discuss the behavior of the classifier across different façade classes, highlighting the categories that achieved perfect recognition and those in which residual confusion remained.

4. Results

To analyze the impact of the visual vocabulary size on the performance of the proposed method, several configurations of the BoF representation were evaluated. For each vocabulary size, the system was trained and tested using the same experimental protocol, and the following metrics were computed: overall accuracy, total processing time, inference time on the test set, and average inference time per image. The results obtained for all configurations are summarized in Table 1.

Table 1. Results Obtained in Building Recognition.

Vocabulary size	Overall Accuracy	Inference Time_s	Avg Inference Time_ms
100	0.946875	18.5914216	5.7648976
300	0.95625	16.3577631	5.7038833
600	0.95625	17.5221287	6.247195
900	0.959375	19.3708675	6.289495
1200	0.959375	19.2782856	6.1759659

Vocabulary size	Overall Accuracy	Inference Time_s	Avg Inference Time_ms
1500	0.971875	23.1082111	7.3121132
2000	0.975	20.4184708	6.2388683
3000	0.971875	23.2999651	6.3495663

As shown in Table 1, increasing the vocabulary size generally leads to improved recognition performance. The overall accuracy rises from 94.69% for 100 visual words to 97.50% for 2000 words, indicating that larger codebooks are able to capture more discriminative visual patterns. However, this improvement comes at the cost of additional computation, since larger vocabularies increase both dictionary construction and classification time. Beyond 2000 words, no meaningful accuracy gains are observed, while the processing time continues to increase. These results suggest that a vocabulary size of 2000 words provides the best trade-off between

recognition performance and computational efficiency.

Figure 1 illustrates the confusion matrix obtained for the best configuration (2000-word vocabulary). In most classes, all test images were correctly classified, as indicated by the strong diagonal pattern. Only a few isolated confusions were observed, mainly between visually similar façades or cases affected by occlusion and viewpoint variation. These results confirm that the proposed BoF-SVM approach is able to discriminate among the 40 building categories with high reliability.

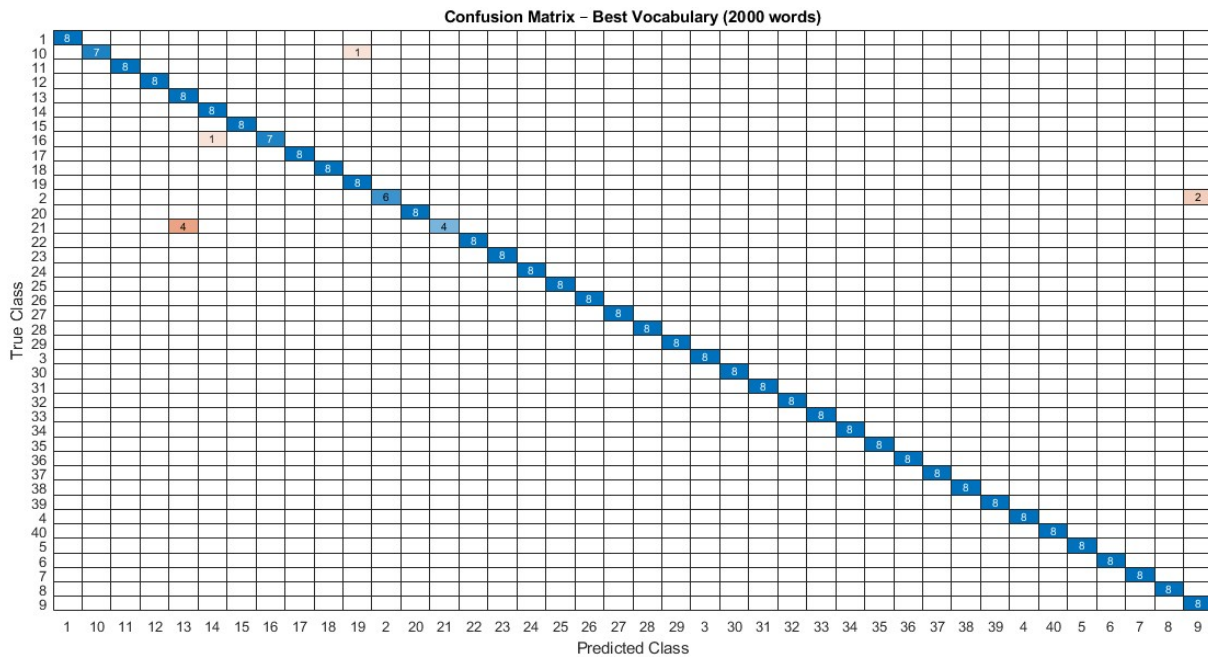


Figure 1. Confusion Matrix – Best Vocabulary (2000 words).

By analyzing Figure 1, it can be observed that the classification errors are concentrated in only a few classes. The most affected categories were Classes 2, 10, and 16, each presenting a single misclassified instance, in addition to Class 21, which exhibited the highest confusion rate. In contrast, most of the remaining classes achieved perfect recognition. These errors generally occur in façades with highly similar textural and geometric patterns, which reinforces the chal-

lenge of distinguishing visually homogeneous structures.

To complement the global performance analysis, a class-level evaluation was carried out by computing accuracy, precision, recall, and F1-score for each building category. These metrics make it possible to identify classes in which the classifier performs consistently well and those in which occasional misclassifications occur. The detailed results are summarized in Table 2.

Table 2. Class-Level Performance Metrics.

Class	Accuracy	Precision	Recall
1	1.00	1.00	1.00
2	0.75	1.00	0.75
3	1.00	1.00	1.00
4	1.00	1.00	1.00
5	1.00	1.00	1.00
6	1.00	1.00	1.00
7	1.00	1.00	1.00
8	1.00	1.00	1.00
9	1.00	0.80	1.00
10	0.88	1.00	0.88
11	1.00	1.00	1.00
12	1.00	1.00	1.00
13	1.00	0.67	1.00
14	1.00	0.89	1.00
15	1.00	1.00	1.00
16	0.88	1.00	0.88
17	1.00	1.00	1.00
18	1.00	1.00	1.00
19	1.00	0.89	1.00
20	1.00	1.00	1.00
21	0.50	1.00	0.50
22	1.00	1.00	1.00
23	1.00	1.00	1.00
24	1.00	1.00	1.00
25	1.00	1.00	1.00
26	1.00	1.00	1.00
27	1.00	1.00	1.00
28	1.00	1.00	1.00
29	1.00	1.00	1.00
30	1.00	1.00	1.00
31	1.00	1.00	1.00
32	1.00	1.00	1.00
33	1.00	1.00	1.00
34	1.00	1.00	1.00
35	1.00	1.00	1.00
36	1.00	1.00	1.00
37	1.00	1.00	1.00

Class	Accuracy	Precision	Recall
38	1.00	1.00	1.00
39	1.00	1.00	1.00
40	1.00	1.00	1.00

As shown in Table 2, most building categories achieved perfect scores across all metrics, indicating that the proposed BoF–SVM model is highly effective for the majority of façade types. Only a few classes presented noticeable degradations, particularly Classes 2, 10, 16, and especially Class 21, whose recall dropped to 0.50. These cases correspond to buildings with highly similar visual patterns or stronger occlusions, which tend to increase ambiguity in the feature space. Even so, the number of errors remains small, and the overall performance remains robust across the 40 categories.

To place the obtained results in context, the proposed method was compared with a reference approach available in literature, namely the work of Li et al. [1], which also evaluated urban building recognition using the SBID dataset. Table 3 presents the performance comparison.

Table 3. Performance Comparison with Li et al. on the SBID Dataset.

Works	Data Base	Performance %	Notes
Li et al.	SBID	94.7	baseline
Proposed (BoF+SVM)	SBID	97.5	Vocabulary=2000

As shown in Table 3, the proposed BoF–SVM approach outperforms the method reported by Li et al. on the same dataset. While the baseline method achieves an accuracy of 94.7%, our model attains 97.5%, representing a noticeable improvement. This gain is mainly attributed to the use of a larger and better-structured visual vocabulary combined with an optimized SVM classification stage.

5. Conclusion

This work presented a building recognition approach based on the Bag-of-Features representation combined with an SVM classifier, evaluated on the Sheffield Building Image Dataset (SBID). A systematic analysis of different visual vocabulary sizes demonstrated that increasing the number of visual words improves recognition performance up to a certain point, with the best trade-off being obtained for a vocabulary of 2000 words. Under this configuration, the proposed system

achieved an overall accuracy of 97.5%, with an average inference time below 25 ms per image, indicating good computational efficiency.

The confusion matrix and the per-class performance metrics confirmed that most building categories were recognized with perfect precision and recall, with only a few classes exhibiting occasional misclassifications. These errors were mainly associated with façades that share highly similar textural and geometric characteristics, reinforcing the intrinsic difficulty of discriminating visually homogeneous structures.

When compared specifically with the work of Li et al. [1], which also employed the SBID dataset, the proposed approach achieved competitive performance while maintaining a relatively simple and interpretable architecture.

This result demonstrates that a denser and well-distributed visual vocabulary tends to better capture the structural and textural variations of façades, increasing class separability—a behavior already reported in classical studies [1, 4]. Comparatively, the accuracy obtained in this work surpasses the values reported in previous studies [1], confirming the effectiveness of integrating BoF and SVM for urban building recognition tasks.

Abbreviations

SVM	Support Vector Machine
BOF	Bag of Feature

Conflicts of Interest

The authors declare no conflicts of interest.

References

- [1] LI, J.; HUANG, W.; SHAO, L.; ALLINSON, N. M. Building recognition in urban environments: A survey of state-of-the-art and future challenges. *Information Sciences*, v. 277, p. 406–420, 2014. <https://doi.org/10.1016/j.ins.2014.02.051>
- [2] LI, J.; ALLINSON, N. M. Building recognition using local oriented features and probabilistic neural network. *International Conference on Computer Analysis of Images and Patterns (CAIP)*, York, UK, 2013. p. 315–322. https://doi.org/10.1007/978-3-642-40246-3_39

- [3] L. A. Silva, G. A. Carrijo, F. S. Moreira, A. C. L. Junior, C. S. Goulart, and L. M. Sousa, “Reconhecimento de edificios utilizando o filtro de gabor, a transformada wavelet e a rede neural perceptron de múltiplas camadas / Buildings recognition using the gabor filter, wavelet transform and multilayer perceptron network”, *Braz. J. Develop.*, vol. 8, no. 5, pp. 35182–35199, May 2022.
<https://doi.org/10.34117/bjdv8n5-164>
- [4] OKUR, E. et al. Weighted Bag of Visual Words with enhanced deep features for image classification. *Expert Systems with Applications*, v. 214, p. 119041, 2024.
<https://doi.org/10.1016/j.eswa.2024.119041>
- [5] XIE, Z. An urban building use identification framework based on remote sensing and social sensing data with spatial constraints. *Remote Sensing Letters*, v. 15, n. 8, p. 714–724, 2024.
<https://doi.org/10.1080/2150704X.2024.1234567>
- [6] PANDA, Santosh Kumar; PANDA, Chandra Sekhar. A Review on Image Classification Using Bag of Features Approach. *International Journal of Computer Sciences and Engineering*, v. 7, n. 6, p. 538-542, jun. 2019.
<https://doi.org/10.26438/ijcse/v7i6.538542>
- [7] NBRO. What are bag-of-features in computer vision? Artificial Intelligence Stack Exchange, 15 jun. 2020. Available at: <https://ai.stackexchange.com/questions/21914/what-are-bag-of-features-in-computer-vision>
- [8] LAZEBNIK, Svetlana; SCHMID, Cordelia; PONCE, Jean. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, v. 2, p. 2169–2178, 2006.
<https://doi.org/10.1109/CVPR.2006.68>
- [9] SHEFFIELD BUILDING IMAGE DATASET (SBID). University of Sheffield Database Repository, 2013. Available at: <https://www.sheffield.ac.uk>.
- [10] SITLAULA, C., SHAHI, T. B., SUNIL ARYAL 4, FAEZEH MARZBANRAD. Fusion of multi-scale bag of deep visual words features of chest X-ray images to detect COVID-19 infection. *Sci Rep*. 2021 Dec 13; 11(1): 23914.
<https://doi.org/10.1038/s41598-021-03287-8>
- [11] BISHOP, C. M. *Pattern recognition and machine learning*. New York: Springer, 2006.
- [12] CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273-297, 1995.
- [13] CRISTIANINI, N.; SHAWE-TAYLOR, J. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press, 2000.