Detecting Buildings in Historical Photographs Using Bag-of-Keypoints

Natália C. Batista*, Ana Paula B. Lopes*† and Arnaldo de A. Araújo*

*Computer Science Department, Federal University of Minas Gerais – UFMG
31270–010, Belo Horizonte, MG, Brazil
†Exact and Technological Sciences Department, State University of Santa Cruz – UESC
45662–000, Ilhéus, BA, Brazil
{natalia, paula, arnaldo}@dcc.ufmg.br

Abstract—The strategies for the preservation of historical documents can include their digitization, which is an effective way to make them publicly available while preventing degradation of the original sources. The Arquivo Público Mineiro (APM), the Archives of the State of Minas Gerais, has a collection of historical photographs from Brazil, and some of them have been digitized. The availability of digital copies of historical photographs makes it possible to apply Content-Based Image Retrieval (CBIR) techniques to alleviate the huge manual effort that is put nowadays into their description and indexing. On the other side, such images are usually more challenging than modern photographs, because of the poor quality of the originals and several degradation effects.

In this work, it is proposed a technique based on a bag-of-keypoints representation to identify images containing buildings in the APM photographic collection. The bag-of-keypoints is an efficient image representation technique, which has been proved robust to occlusion and variations due to pose, scale, illumination and several transformations. Experiments were performed on the images from the APM collection, to classify them between building and non-building, using bag-of-keypoints representations of those images. Results show that, despite of the poor quality of the images, the bag-of-keypoints representation is able to provide good detection rates, indicating the suitability of the proposed method for the task.

Keywords-Historical photographs; Image classification; Buildings recognition; Bag-of-keypoints;

I. INTRODUCTION

The strategies for the preservation of historical documents can include their digitization, which is an effective way to make them publicly available while preventing degradation of the original sources.

Frequently, historical documents present a high level of degradation, caused by aging or damage due to manipulation, resulting in digital images of poor visual quality. In extreme cases, such images can even be completely unreadable [1]. The digitization of such material can reduce the need for direct manipulation of the original copies and, as a consequence, it can help in reducing the speed of the degradation process.

Due to the spreading of projects for digitizing historical collections for preservation purposes, a vast number of digital images of historical documents has been generated. But digitization alone can be useless if there are not efficient ways of organizing and indexing the huge amount of visual data produced by such projects. Ideally, the digitization project ends up by making a set of computational tools available, so that the final users – historians, students, researchers and even the general lay public – can easily and quickly query and retrieve desired documents.

The Arquivo Público Mineiro [1] (APM), the Archives of the State of Minas Gerais, is an example of institution dealing with the preservation of documents from the cultural heritage of Brazil. The APM has an important photographic collection, and some of them have been already digitized, indexed and made publicly available by means of an information system [2]. Figure 1 presents some examples of the APM collection, from which it is possible to have an idea of the mentioned problems that can be found in this collection.

Currently, in the APM information system, image retrieval is based on textual information, obtained by manual indexing. The task of manually describe and index historical collections can last several people/years to be completed, depending on the collection size and the number of annotated concepts [3].

In such context, tools that are able to alleviate the workload in the annotation process can become very valuable [2]. Some examples of such tools are described in [4], [5], [6].

In [2], a detailed analysis of the image database from APM is performed, based on interviews with users of the previously mentioned information system. This analysis resulted in a broad list of different needs toward the photographic collection. Such needs include: separation of colored from black and white photographs; extract pieces of information written by hand, with typewriters or rubber stamps; identification of photographs where there are only busts or faces or groups of people; distinction of images by emulsion used in the revelation process; identification of images in external and internal environments; finding photographs depicting rural, urban and mountain areas; and identification of specific objects in the scene (like bridges, churches, buildings, vegetation, façades, plates, transmission

---

Figure 1: Examples of digitized documents from the APM photographic collection.

lines, railroads, etc.).

Several of such tasks require Content-Based Information Retrieval (CBIR) techniques. In this research area, one of the greatest challenges consists of finding descriptors which are able to express, from low level features, like color and texture, the semantic content of the image. The distance between the low level visual information and its semantic is commonly referred to in the literature as the *semantic gap* [7].

In this work, in the spirit of creating CBIR tools to alleviate the task of indexing such material, it is proposed a methodology to aid in the identification of images containing buildings in the APM historical photographic collection.

The text proceeds as follows: next section (II) describes some related work. Section III details the construction of bag-of-keypoints representations, explaining some of the main implementation decisions and the classification method. The methodology for testing the proposed approach is detailed in Section IV and the experimental results are presented and discussed in Section V. Finally, the work is concluded in Section VI.

II. RELATED WORK

Generally speaking, it is possible to classify objects into two broad types: man-made objects and natural objects. Man-made objects are mainly characterized by the regularity in appearance and in structural and spatial relationship among the objects and their components. By the other side, natural objects, like trees, vegetation, rivers, stones and clouds, can co-exist with man-made objects, what makes the task of detecting their presence a non-trivial one.

Buildings are a particular class of man-made objects, and their presence in the image typically generates a greater number of edges, joints, lines and parallel groups of lines, when compared to images where the presence of buildings does not predominate. Those structures correspond to the presence of corners, windows, doors, walls, etc.

Most approaches found in the literature to retrieve building images use lines and edges as features. In [8], rules of perceptual grouping are applied to the classification of scenes which contain man-made objects and natural objects, by extracting lines, coterminations, junctions, parallel groups and polygons. The feature space is partitioned in three classes: structured images (i.e., those ones exhibiting significant structure due to man-made objects), non-structured images (images with no man-made object at all) and intermediate ones (images exhibiting a mixture of man-made and natural objects).

Other proposals for recognition and location of man-made objects are [9] and [10]. In [9], consistent line clustering is applied, exploring color, orientation and spatial features of line segments. In [10], orientation-based features of image blocks are extracted, and each block is individually classified. Then, spatial dependencies are captured from this data by using a model based in a causal multiscale random field.

The proposals described in [11] and [12] deal more specifically with retrieval of buildings images. In [11], the similarity between two images is evaluated by the distance of the centroids of the peaks in each band, with every image being split in the Hough space. In [12], the recognition is performed in two hierarchical stages. The first stage uses an indexing method based in localized color histograms computed in structures with dominant orientation to select a subset of models. In the second stage, matching of scale-invariant keypoints descriptors is used to identify the best

model from the subset.

The work in [13] proposes an approach for detecting and tracking of apartment buildings for the development of a video-based navigation system. In their proposal, boosting techniques are used to create strong classifiers to detect some primitives. Then, a cascade of those strong classifiers is used to detect patterns representing buildings by using Haar features. Based on the spatial locality of the detected patterns, a agglomerative hierarchical clustering is adopted to combine the parts together.

Another approach for detecting buildings aimed at robot navigation is proposed in [14], in which it is possible also to perform the recognition of buildings by means of the detection of line segments to create a mesh of parallelograms which represents the building façade.

One of the advantages of the approaches that make use of lines and edges features is the fact that neither the segmentation or the detailed representation of the object are not needed. In other words, the decision relative to the presence of certain man-made objects can be done without the need to locate and recognize specific instances of the objects, which would require a greater knowledge about their properties, leading to less generalizable methods.

To the best of our knowledge, all proposals related to ours make use of image databases comprised of recent photographs, mostly containing modern buildings. It is worth mentioning that those bases have rather distinct characteristics when compared to the APM database, not only because of the already mentioned degradation process of the original photographs, but also because both the photographs themselves and the content in them are not contemporary ones. Then, it is very likely that these specific characteristics make the detection of buildings in those images even more challenging. In this work, it is proposed a classification of a subset of the APM photographs based on a bag-of-keypoints representation for them.

III. IMAGE CLASSIFICATION

The state of the art approach in CBIR is based in the recognition of several object categories [15]. In that case, a specialized classifier is built for every class. Then, all detectors are applied to the images, each one determining whether there is an object of that class in a image or not. The final feature vector representing the images for the CBIR system is therefore comprised of a vector in which every component indicates the presence or absence of one of the considered concepts[2].

Typically, these base classifiers for CBIR systems are built on a large set of examples and counter-examples of the target class, from which some features are extracted for training a classifier. The evaluation of the classifier can be performed

---

[2]Sometimes, those components are the probabilities for the presence of those concepts, instead of a simple binary presence/absence indicator.

by splitting some ground-truth set into training and testing subsets.

Distinct objects that belong to the same category frequently present great variations in their appearance due to changes in pose, scale, illumination conditions and imaging techniques applied. Occlusions and several transformations can also contribute to a large intra-class variation.

Ideally, the representation must be flexible enough to cope with a large variety of visually different instances of the same class, yet keeping the discriminative power between classes [16].

The work described in [17] presents a computationally efficient approach which has shown good results for objects and scenes categorization. The method, called bag-of-keypoints, is an analogy to the bag-of-words representation from the textual information retrieval area, in which each document is represented as a feature vector counting words occurrences in its text [16].

The basic idea of bag-of-keypoints is to describe the image as an unordered collection of local features taken around interest points. For a compact representation, a visual vocabulary is built, typically by clustering such features. Each cluster of interest point descriptors is treated as a visual word in the vocabulary. By mapping the interest points extracted from one image to the words in the visual vocabulary, it is possible to count their occurrences much like it is done in the textual case [18]. The resulting histogram is called the bag-of-keypoints for that image, and given a set of images represented by their bag-of-keypoints, any classifier can be applied to these feature vectors.

The bag-of-keypoints approach and its extensions (like, for example, [19], [20] and [21]), have recently achieved good results, due to their computational efficiency and robustness to intra-class variations from several sources. Besides, being a very generic method, it is able to deal with a great variety of objects and scenes. Because of these advantages, in this work, an approach based in interest points to classify the APM images is chosen.

The next subsections detail the the bag-of-keypoints approach, discussing some implementation decisions made along the way.

A. Detection and Description of Interest Points

Interest points are those points in the image that possess a great amount of information in terms of local signal changes [22]. Among the several existing methods to select interest points in images and to provide a local descriptor for them, the Scale-Invariant Feature Transform (SIFT) [23] algorithm is one which has been presenting the best results for several different tasks [16], [18].

The SIFT algorithm transforms the image information in a set of scale-invariant coordinates, related to the local features. Each interest point is associated to a location in the image, to one specific scale and to one orientation. The

Figure 2: Some images from the APM collection marked with the points detected by the SIFT algorithm (red crosses).

descriptors of those interest points are created by computing the gradient magnitude and the orientation around its location in the selected scale.

Since the SIFT descriptor has 128 dimensions, in order to make the next processing step feasible, the Principal Component Analysis (PCA) [24] method is applied for dimensionality reduction. The PCA is a statistical technique which is able to identify patterns in the data and express them in a new vector space where their differences and similarities are enhanced. Besides, it makes it possible to compress the original data by sorting the dimensions in that new vector space so that the most important components appear in the first positions. This way, it is possible to prune out the final components of the projected vectors without loosing a significant amount of information.

Figure 2 shows the SIFT points detected in some images from the APM collection.

B. Building the visual vocabulary

The visual vocabulary is generated by clustering the interest points detected in a subset of images, and each discovered cluster represents a visual word of the vocabulary. In this work, the clustering is performed by applying the k-means algorithm [25], which is the most commonly applied method for this task, mainly because of its simplicity and relatively low computational cost.

The k-means algorithm partitions a set of points among k disjoint subsets, such as the intra-class distance is minimized and the inter-class distance is maximized. Differently from the vocabulary considered in text retrieval, whose values comes naturally from the training *corpus*, the vocabulary size in the visual case is determined by the number of clusters predefined for the k-means algorithm. Typical sizes for visual vocabularies found in the literature go from hundreds to thousands of words, and the issue of how to optimize the vocabulary size is an open research topic. In the experiments performed for this work, the vocabulary size is fixed in 100 words.

Figure 3 shows one example of regions around interest points collected from a subset of the APM photographs and clustered together by k-means. From this figure, it is possible

Figure 3: Some patches of a specific cluster of the visual vocabulary built from the APM images, containing 128 regions of 13x13 pixels. The central point of each patch is an interest point detected by SIFT.

verify that there is a noticeable visual resemblance among such patches.

C. Building the bag-of-keypoints histogram

In the typical bag-of-keypoints approach, an image is characterized by a histogram of visual words, i.e., by counting the number of interest points that belong to every cluster formed in the previous step.

This is done by comparing the point descriptors to the cluster centroids, associating every descriptor to that cluster to which the distance to its centroid is the smallest one. In our proposal, the words in the vocabulary have equal weights, and the histogram is normalized to one. The resulting histogram is the bag-of-keypoints representation for the image, which is going to be used as its feature vector for classification purposes.

D. Classification

After building the feature vectors for all images in the target collection, the next step is to categorize the image set. Each visual category is mapped onto a class, which is associated with a label.

In a supervised learning method, the classification occurs in two steps: training and testing. In the training step, a set of images to which the label is known is used to create the model. A statistical decision procedure is used to enable the classifier to distinguish among the previously established classes. In order to verify the quality of the classifier, the classification provided by the model is compared to the

labels of testing images, kept separated from the training ones. The model created can then be used to classify new, unlabeled images.

The classifier selected in this work is the Support Vector Machines (SVM) classifier [26], commonly used together with bag-of-keypoints approaches, due to its good performance in a huge variety of scenarios [17], [18]. The basic idea behind SVM is finding a hyperplane in the feature space which is able to separate two distinct classes within a certain margin. Multi-class SVMs are built on the combination of binary ones.

When using SVM, the initial feature space can be mapped onto a higher dimensional feature space, in which a linear separating hyperplane can be found more easily. This is done by using a suitable kernel function, and this technique is known as the kernel trick.

Ideally, the kernel must give emphasis only to regions containing the target concept, while at the same time tolerating the background variations without amplifying their effects. Although there are several general purpose kernels available, it is an open issue which one is the most effective to be applied with bag-of-keypoints to visual classification. Since it is one of the most common choice among several existing proposals [18], and also because it provided the best results in preliminary tests, the Radial Basis Function (RBF) kernel is selected for this work.

IV. METHODOLOGY

A subset of 600 images were randomly selected from the APM images. Those photos were originally digitized using the RGB color system, with a 600 dpi resolution and stored in TIFF format. For this work, such images were reduced to a resolution of 72 dpi, transformed to monochromatic and stored in PGM format, in order to reduce the computational complexity involved.

To build the ground-truth used in the experiments, 18 volunteers were selected for a manual annotation procedure. The volunteers selected one out of three categories for each evaluated image. The categories presented for them were *building*, *non-building* and *intermediary* and each volunteer evaluated a set of 100 images.

The building category should be selected if the person considered that the photograph contained some building façade in the majority of its area. The non-building class should be selected for those photographs in which no building appeared at all or it appeared only in a very small area, and the intermediary class should be the choice for the case the volunteer could not decide between building and non-building.

By this process, every image received the categorization of three different people, and the final class was decided by voting. This resulted in 145 images classified as buildings, 88 classified as intermediary and 367 images classified as non-buildings.

Table I: Classification results, using a 95% confidence interval.

Exp.	Building	Non-Building
1	$50.2\% - 53.4\%$	$57.5\% - 60.1\%$
2	$71.6\% - 74.6\%$	$69.0\% - 70.5\%$

To verify the ability of this algorithm to correctly classify the APM photographs, the annotated image set was randomly split into five subsets so to perform a classification using 5-fold cross validation scheme. For vocabulary building, a random sampling of sixty images was selected (and separated from the other images used in the experiments), and all of the interest point descriptors obtained from them were used (about 23,000).

An overview of the method is shown in Figure 4.

V. EXPERIMENTAL RESULTS

The experiments for this work were performed in two groups. The first group with the default configuration described in Section III, i.e., training the classifier with three different classes, which yielded recognition rates of 51.8% for the building class and 58.8% for the non-building class.

In the second group of experiments, the classifier was trained only with the building and non-building classes. Then, the training set was balanced to keep the number of building and non-building images similar. This was done by randomly selecting a number of non-buidings images to remove from the folds, since the number of non-buildings images was larger. This new experimental setting resulted in an average recognition rate of 73.1% for the building class and 69.8% for the non-building one.

In each case, 30 runs were executed randomly varying the folds composition. The results from both experiments, using a confidence interval of 95% are summarized in Table I. These values suggest that the intermediary class confuses the classifier. This is better understood by looking at Figure 5, where some examples of images labeled as intermediary by the users are shown.

Table II and Table III show the confusion matrices for one of the 30 runs from each group of experiments. The results for the images in the intermediary class are also presented. In both cases, the numbers show clearly the difficulty for the classifier to discriminate those images from the intermediary class. Even in the case depicted in Table II, in which the classifier is explicitly trained taking into account three classes, there is a high level of confusion in the classification of images from the intermediary class.

As already mentioned, we were not able to find similar proposals in the literature dealing with historical photographs. The proposals presented in [27] and [8] are the most similar to ours, in a sense that they also tried to
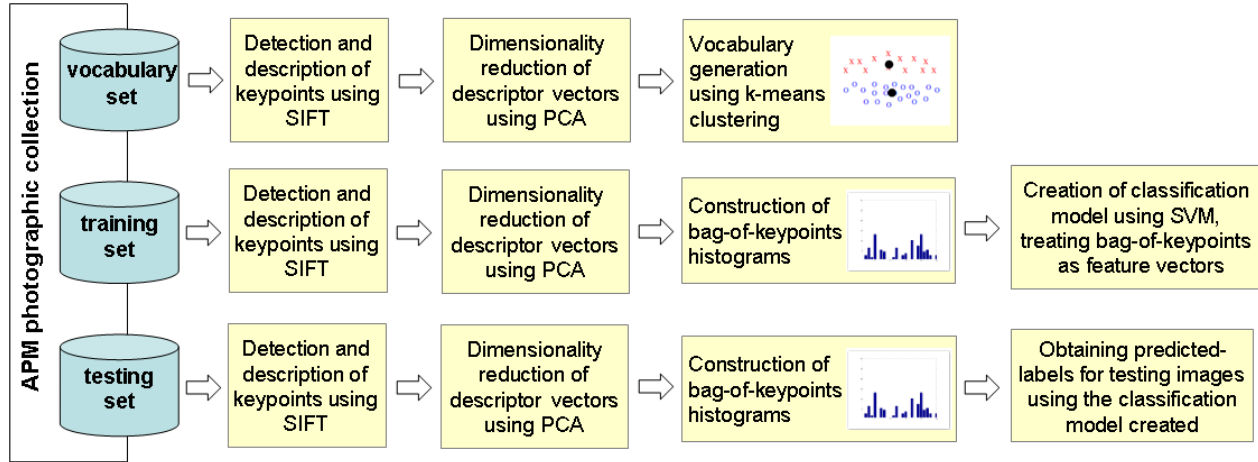
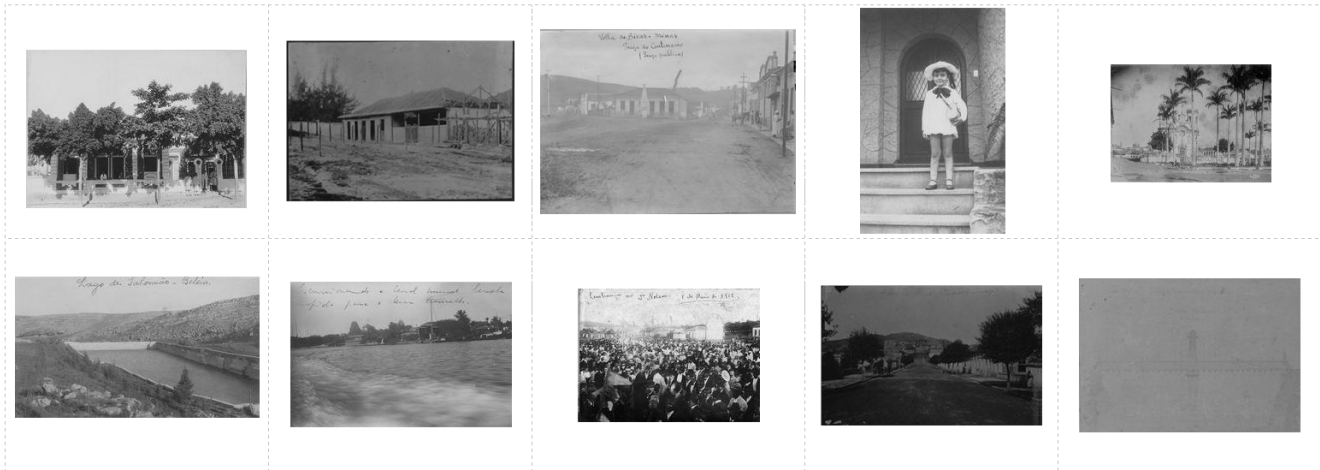Figure 4: Overview of the bag-of-keypoints approach applied to APM images.

Figure 5: Some examples of images from APM collection labeled as intermediary by the users. First row: images classified as buildings. Second row: images classified as non-buildings.

Table II: Confusion matrix for one run from the fist group of experiments.

	Building	Non-Building	Intermediary
Building	52.2%	21.7%	26.1%
Non-Building	17.9%	59.7%	22.4%
Intermediary	23.5%	41.2%	35.3%

Table III: Confusion matrix for one run from the second group of experiments.

	Building	Non-Building
Building	73.9%	26.1%
Non-Building	28.4%	71.6%
Intermediary	58.8%	41.2%

identify buildings from low level visual cues. Nevertheless, those authors worked on modern photographs presenting very good visual quality. Their reported recognition rates for the building class were 92% and 77.6%, respectively. A coarse comparison of those values with the ones achieved by our approach indicates that the proposed method is indeed able to achieve reasonably good recognition rates, despite of all the challenging problems found in the target images.

VI. CONCLUSION

This work presented the application of a bag-of-keypoints approach for the identification of buildings in a a set of digitized historical photographs.

By their own nature, the photographs of this collection presented varied degradation levels, which make them particularly challenging for computer vision algorithms.

Since the bag-of-keypoints representation approach has proven its robustness in several recognition and classification

scenarios in the literature, it was selected for this problem, and a particular implementation based on SIFT descriptors was provided. The experimental results indicate the suitability of the proposed method for the task, despite of the poor visual quality of the images from the target database.

REFERENCES

[1] F. A. R. Bertholdo, "Técnicas de limiarização para melhorar a qualidade visual de documentos históricos," Master's thesis, September 2007.

[2] C. J. S. Oliveira and A. de Albuquerque Arajo, "Levantamento dos projetos que envolvem bases de imagens digitais do arquivo público mineiro," Digital Image Processing Group, Computer Science Department, Federal University of Minas Gerais, UFMG, Belo Horizonte, Brazil, Tech. Rep. RT.DCC.001/2004, 2004.

[3] L. S. Kennedy, S.-F. Chang, and I. V. Kozintsev, "To search or to label?: predicting the performance of search-based automatic image classifiers," in *Proceedings of the ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM, 2006, pp. 249–258.

[4] A. Lopes, C. Oliveira, and A. Albuquerque Araujo, "Face recognition aiding historical photographs indexing using a two-stage training scheme and an enhanced distance measure," in *Proceedings of the Brazilian Symposium on Computer Graphics and Image Processing*, October 2008, pp. 11–18.

[5] C. J. S. Oliveira, N. C. Batista, A. de Albuquerque Arajo, and J. L. de Souza Pio, "Photographic prints identification," in *Proceddings of the International Conference on Systems, Signals and Image Processings and Semantic Multimodal Analysis of Digital Media*, B. Enyedi and A. Reichardt, Eds. Budapest, Hungary: Harlequin, s.r.o. Kosice, September 2006, pp. 231–234.

[6] E. Valle, M. Cord, and S. Philipp-Foliguet, "Content-based retrieval of images for cultural institutions using local descriptors," in *Geometric Modeling and Imaging–New Trends*, August 2006, pp. 177–182.

[7] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[8] Q. Iqbal and J. Aggarwal, "Retrieval by classification of images containing large manmade objects using perceptual grouping," *Pattern Recognition Journal*, vol. 35, no. 7, pp. 1463–1479, July 2002.

[9] Y. Li and L. G. Shapiro, "Consistent line clusters for building recognition in cbir," in *Proceedings of the International Conference on Pattern Recognition*, vol. 3. Washington, DC, USA: IEEE Computer Society, August 2002, pp. 952–956.

[10] S. Kumar and M. Hebert, "Man-made structure detection in natural images using a causal multiscale random field," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2003, pp. 119–126.

[11] X. Yuan and C.-T. Li, "Cbir approach to building image retrieval based on linear edge distribution," in *Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance*. Los Alamitos, CA, USA: IEEE Computer Society, 2006, p. 95.

[12] W. Zhang and J. Košecká, "Hierarchical building recognition," *Image Vision Computing*, vol. 25, no. 5, pp. 704–716, 2007.

[13] Y. Hu, D. S. Jang, J. H. Park, S. I. Cho, and C. W. Lee, "Real-time apartment building detection and tracking with adaboost procedure and motion-adjusted tracker," *ETRI JOURNAL*, vol. 30, no. 2, pp. 338–340, 2008.

[14] H.-H. Trinh and K.-H. Jo, "Line segment-based facial appearance analysis for building image," in *Proceedings of the International Forum on Strategic Technology*, October 2006, pp. 332–335.

[15] S.-F. Chang, W.-Y. Ma, and A. Smeulders, "Recent advances and challenges of semantic image/video search," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, April 2007, pp. 1205–1208.

[16] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proceedings of the European Conference on Computer Vision*. Springer, 2006, pp. 490–503.

[17] G. Csurka, C. Dance, J. Willamowski, L. Fan, and C. Bray, "Visual categorization with bags of keypoints," in *Proceedings of the Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1–22.

[18] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. New York, NY, USA: ACM, 2007, pp. 494–501.

[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 2169–2178.

[20] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1243–1256, 2008.

[21] M. Marszalek and C. Schmid, "Spatial weighting for bag-of-features," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2118–2125, 2006.

[22] S. Agarwal, A. Awan, and D. Roth, "Learning to detect objects in images via a sparse, part-based representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1475–1490, November 2004.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[24] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*.   Prentice Hall of India, 2006.

[25] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[26] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, June 1998.

[27] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, October 2005, pp. 604–610.