Expert Systems with Applications 39 (2012) 1679–1686

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A Bayesian approach for object classification based on clusters of SIFT local features

Leonardo Chang [a,b,1], Miriam M. Duarte [b,2], L. Enrique Sucar [b,*], Eduardo F. Morales [b,2]

[a] Advanced Technologies Application Center, 7th Avenue, Number 21812, Siboney, Playa, Havana, Cuba
[b] National Institute for Astrophysics, Optics and Electronics, Luis Enrique Erro No. 1, Sta. María Tonanzintla, Puebla, México C.P. 72840, Mexico

ARTICLE INFO

Keywords:
Object class recognition
Local features
SIFT
Clustering
Bayesian networks

ABSTRACT

Several methods have been presented in the literature that successfully used SIFT features for object identification, as they are reasonably invariant to translation, rotation, scale, illumination and partial occlusion. However, they have poor performance for classification tasks. In this work, SIFT features are used to solve object class recognition problems in images using a two-step process. In its first step, the proposed method performs clustering on the extracted features in order to characterize the appearance of the different classes. Then, in the classification step, it uses a three layer Bayesian network for object class recognition. Experiments show quantitatively that clusters of SIFT features are suitable to represent classes of objects. The main contributions of this paper are the introduction of a Bayesian network approach in the classification step to improve performance in an object class recognition task, and a detailed experimentation that shows robustness to changes in illumination, scale, rotation and partial occlusion.

1. Introduction

Recognizing object classes is one of the oldest problems in Computer Vision. However, it remains as one of the most challenging problems, due to the undefined nature of object classes similarity. Sometimes, objects in different classes can have more similar characteristics compared to others in the same class. Several researchers have addressed this problem in many ways. In Ullman (1996), object recognition is divided into three subcategories, based on the major principles they employ: the approach that uses invariant properties and feature spaces, those that use parts and structural descriptions, and finally alignment approaches.

The first subcategory assumes that there are certain invariant characteristics that are common to an entire class of objects. Most current research is based on this assumption, looking for different and increasingly robust invariant features. The second subcategory used parts and structural decomposition. In this case, it is assumed that all the objects are composed of a set of generic components. The third subcategory is the alignment approach. The main idea of these methods is to compare the new object to be recognized with a stored model and estimate the changes that separate them. Generally, stored models are 3D models, which make these methods very expensive in computational terms.

Also, most objects class recognition methods characterize objects by their global appearance, usually of the entire image. These methods are not robust to occlusion or variations such as rotation or scale. Moreover, these methods are only applicable to rigid objects. Local invariant features have become very popular to give solution to the limitations of these methods in object detection and recognition.

In the last few years, local features (e.g. Harris-Affine (Mikolajczyk & Schmid, 2004), SIFT (Lowe, 2004), SURF (Bay, Ess, Tuytelaars, & Van Gool, 2008)) have proven to be very effective in finding distinctive features between different views of a scene. The traditional idea of these methods is to first identify structures or significant points in the image and to obtain a discriminant description of these structures from its surroundings, which is then used for comparison using a similarity measure between these descriptors.

A keypoint detector is designed to find the same point in different images even if the point is in different locations and scales. Different methods have been proposed in the literature. A study and comparison of these approaches is presented in Tuytelaars and Mikolajczyk (2007).

Local features have been mainly used for the identification of particular objects within a scene. For instance, a particular book is given to a system, which extracts its SIFT features and uses them to recognize that particular book. However, such features cannot be used to recognize another book or books in general on the scene.

One of the most popular and widely used local approaches is the SIFT (Scale Invariant Features Transform) method, proposed

* Corresponding author. Tel.: +52 222 266 3100; fax: +52 222 247 2580.
E-mail addresses: lchang@cenatav.co.cu (L. Chang), mduarte@ccc.inaoep.mx (M.M. Duarte), esucar@inaoep.mx (L. Enrique Sucar), emorales@inaoep.mx (E.F. Morales).
[1] Tel.: +53 7 272 1670; fax: +53 7 273 0045.
[2] Tel.: +52 222 266 3100; fax: +52 222 247 2580.

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.06.059

by Lowe (2004). The features extracted by SIFT are largely invariant to scale, rotation, illumination changes, noise and small changes in the viewing direction. The SIFT descriptors have shown better results than other local descriptors (Mikolajczyk & Schmid, 2005).

For object class recognition, many methods use clustering as an intermediate level of representation (Agarwal, Awan, & Roth, 2004; Leibe, Seemann, & Schiele, 2005). Due to the robustness of local features and the good results of clustering in object classification, several authors have recently been investigating the use of clustering for object class recognition using local features based approaches. In Dorkó and Schmid (2005), for invariant region detection, the authors use the Harris-Laplace (Mikolajczyk & Schmid, 2001) and the Kadir and Brady detector (Kadir & Brady, 2001). These regions are described using the SIFT descriptor (Lowe, 2004). In their work, Dorkó and Schmid perform clustering of descriptors to characterize class appearance. Then, they build classifiers of smaller parts of objects from the clusters formed. By discarding several of these clusters they kept only the most discriminative ones.

Mikolajczyk, Leibe, and Schiele (2005) evaluate the performance of various methods based on local features in the object class recognition task. The invariant region detectors evaluated were Harris-Laplace, SIFT, Hessian–Laplace, and MSER. The evaluated features descriptors were SIFT, GLOH, SIFT-PCA, Moments, and Cross-Correlation. In their paper the authors evaluate several detector–descriptor combinations. Clustering is also performed on the descriptors to characterize the appearance of classes. To classify a new sample, the extracted descriptors are matched with the clusters obtained and a threshold determines the class membership.

Zhang, Lazebnik, and Schmid (2007), proposed a method that represents images as distributions of features extracted from a sparse set of keypoint locations and learns a Support Vector Machine classifier with kernels based on two effective measures for comparing distributions, the Earth Mover's Distance and the χ^2 distance. They evaluated the performance and classifiers (i.e., SIFT and SPIN detectors, and the EMD kernel and SVM for classification).

In Wang et al. (2010) to recognize object classes, the authors proposed a coding scheme called Locality-constrained Linear Coding (LLC) instead of the vector quantization in spatial pyramid matching (Lazebnik, Schmid, & Ponce, 2005). LLC project each descriptor into its local-coordinate system by using the locality constraints. In order to generate a final representation the projected coordinates are integrated by max pooling. To describe feature points they used SIFT descriptor and a SVM linear classifier is utilized for classification.

In these works it is mentioned that their proposed methods have invariance to occlusion, changes in illumination, rotation and scale. However, there is no experimentation for the above; neither do they express how robust these methods are. It is also assumed that their proposed methods outperform a straightforward classification method using local features, but no evidence of this is given. In this paper we analyze these facts through a set of detailed experiments over our proposed method.

In this work we use SIFT features to recognize object classes (e.g., books, cameras) in order to provide robustness to changes in scale, rotation, illumination and partial occlusion. The proposed method, in the training phase, also performs clustering on the features extracted from the training set. Each feature in each cluster is labeled with its corresponding class in order to characterize the appearance of object classes. In the classification step, for a new image, the SIFT features are extracted, and for each feature the cluster from the learned model to which it belongs is identified. Information from the identified clusters is then used to find the most probable class. To represent this idea, we introduce the use of a three layer Bayesian network. Three experiments were conducted to test the performance of the proposed method. These experiments showed quantitatively that the use of SIFT local features, clustering and Bayesian networks are suitable to represent and recognize object classes, and that the proposed method significantly outperforms the direct use of SIFT features for object classification. They also showed the invariance of the method in the presence of changes in illumination, scale, rotation and partial occlusion.

The main contributions of this paper are the following. Firstly, we introduce a Bayesian network approach in the classification step to improve performance on this stage. Secondly, we show that clustering over local features provides robustness to changes in illumination, scale, rotation and partial occlusion. We also show that this kind of approach outperforms a straightforward classification method using SIFT features. These last two issues are mentioned in the literature but there is no detailed experimental evidence to support them.

The rest of the paper is organized as follows. Section 2 briefly presents some existing approaches for recognizing object classes using local features. Section 3 describes SIFT local features that are used by our proposed method. The proposed method for learning classes appearance based on clustering, and, a description of the clustering method used are presented in Section 4. In Section 5, the proposed method for recognizing the class of an unseen sample image is explained.. Results and discussion for the three conducted experiments are given in Section 6. Section 7 concludes the paper with a summary of our proposed method, main contributions, and future work.

2. Learning and recognition methods

2.1. Learning object classes

As mentioned in the previous section, matching of SIFT features has shown good results in finding a particular object in different views of a scene. However, the aim and specifications for object class recognition are not the same. In order to recognize object classes, a model that is able to generalize beyond each object in the training set and that allows us to learn a general structure of each class is desired. Moreover, learning should be possible from a small number of samples. With this aim and in accordance with several studies reported in the literature (mentioned in Section 1), clustering is performed on feature descriptors extracted from the training images.

Clusters are expected to have high accuracy i.e., each cluster is representative of only one class. In practice, this does not always occur so there could be clusters that are shared by several classes. Additional methods will be needed in the classification stage to solve these ambiguities.

Fig. 1 shows a high level diagram of the class learning method proposed, which is summarized as follows:

1. For each training image, SIFT local features are extracted.
2. Then, clustering is performed over the features descriptors.
3. Finally, each descriptor in each cluster is labeled with its corresponding class.

2.2. Recognizing object classes

Given a new sample image, classification is performed by first extracting the SIFT features from the input image. Then, for each of these features, a cluster is associated from the learned model and finally, from this instantiation of the model, the class of the input object is determined. Fig. 2 shows a layout of the proposed method.

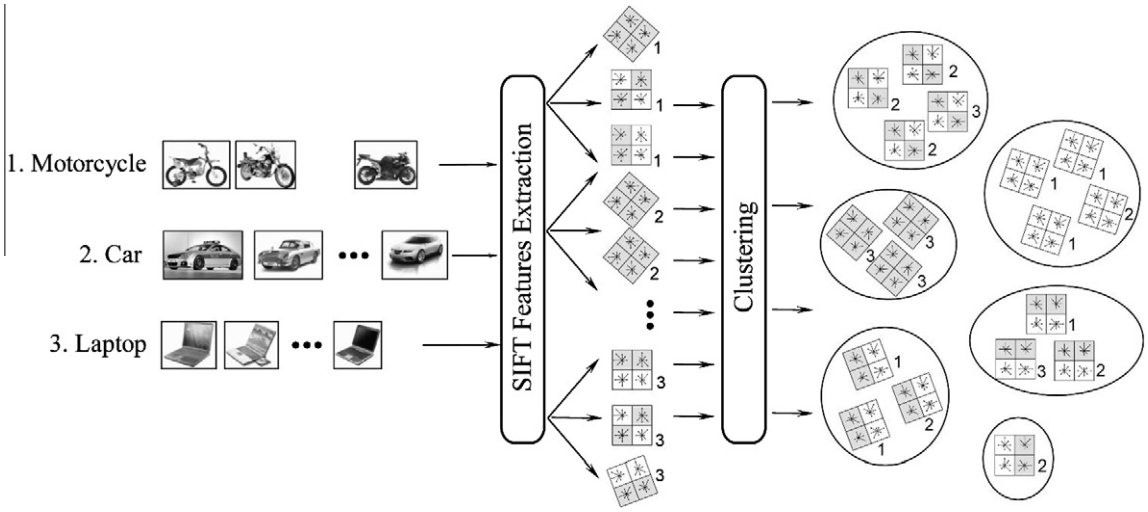

Fig. 1. SIFT local features are extracted from the training set formed by several sample images per class. Later, features descriptors are clustered (the elements inside each circle represents a cluster). Each feature in each cluster is labeled with its corresponding class (represented in the figure by the subscripted number on every element of a cluster). Note that features from different classes could be in the same cluster.

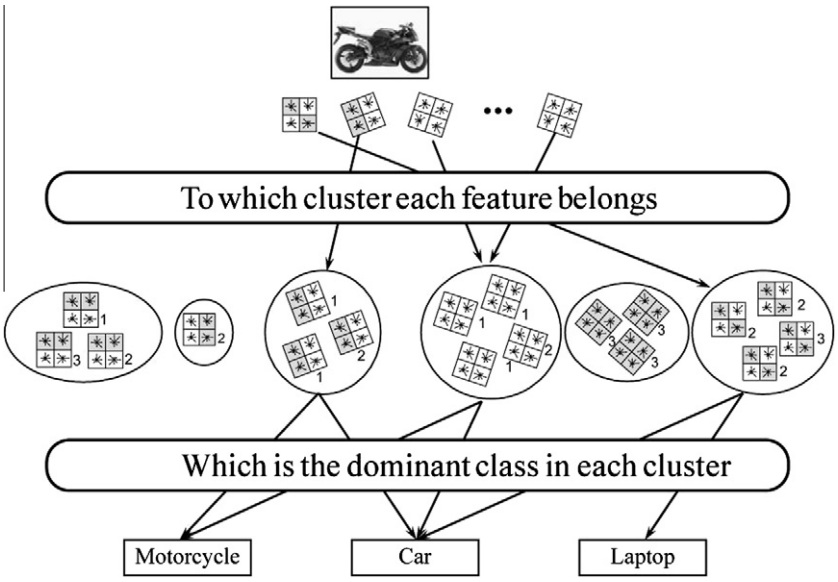

Fig. 2. Classification scheme for a new image. SIFT features are extracted from this image and for each feature the cluster from the learned model to which it belongs is identified. The object class is the majority class in these clusters.

This idea can be represented as a three layer Bayesian network (BN). Bayesian networks provide a means of expressing joint probability distributions over many interrelated hypotheses (Jensen, 1996). A Bayesian network consists of a directed acyclic graph and a set of local distributions. Each node in the graph represents a random variable. A random variable denotes an attribute, feature, or hypothesis about which we may be uncertain. Each random variable has a set of mutually exclusive and collectively exhaustive possible values. That is, exactly one of the possible values is or will be the actual value, and we are uncertain about which one it is. The graph represents direct qualitative dependence relationships; the local distributions represent quantitative information about the strength of those dependencies. The graph and the local distributions together represent a joint distribution over the random variables denoted by the nodes of the graph.

The graphical representation of this BN is shown in Fig. 3. In the first layer we have the trained object classes represented by $c_1, c_2, \ldots, c_C$, where C is the number of classes. In the second layer, clusters obtained in the training phase are represented by $k_1, k_2, \ldots, k_K$, where K is the number of obtained clusters. Finally, the third layer represents the features extracted from the new object, and are represented by the nodes $f_1, f_2, \ldots, f_F$, where F is the number of features extracted from the image.

Usually, the parameters of a Bayesian network are previously estimated. In this work, only the probabilities for a class given each cluster are previously estimated (this probability is calculated from the clusters obtained in the training phase). The probability for each feature in a cluster is calculated online, as described below in steps 1 and 2 of the classification process. Therefore, the parameters of this Bayesian network change for each test image.

Using this model, the classification of a new image I is performed as follows:

1. SIFT features are extracted from the input image I.
2. For each feature f extracted from I, cluster k_f to which it belongs is obtained. The cluster with the highest membership probability of the feature f is selected. This probability is a function of

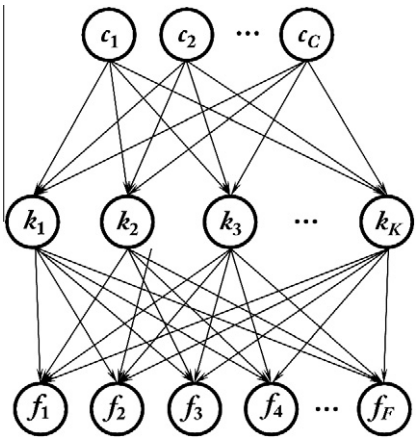

Fig. 3. Graphical representation of the three layer Bayesian network used to classify a new object. In the first layer we have the classes for which the model was trained. The nodes in the second layer represent the clusters obtained in the training phase. Finally, the third layer represents the features extracted from the new object.

Fig. 4. Examples of local features found on a graffiti image. The center of the grids indicates the keypoint location. Size and orientation of the grids represents scale and orientation for each keypoint, respectively.

the distance between the cluster and the feature, which is normalized by the distance between the two most distant clusters. The same distance D defined in Eq. (4) is used:

$$k_f = \arg\max_i P(f|k_i)P(k_i), \quad \text{where} \tag{1}$$

$$P(f|k_i) = 1 - \frac{D(f,k_i)}{\max_{k,l} D(k_k,k_l)}. \tag{2}$$

3. For each cluster $k_{f_1}, k_{f_2}, \ldots, k_{f_F}$ selected in the previous step (note that more than one feature could be in the same cluster), the probability of each class given this evidence is obtained, this probability is extracted from the trained model, propagating further the probability obtained in step 2.
4. Finally, the object class is the one whose sum of occurrence probabilities given each cluster selected in step 2 is the highest:

$$c^* = \arg\max_i \sum_f P(C_i|k_f)P(k_f). \tag{3}$$

3. Obtaining and clustering local features

3.1. SIFT local features

Methods based on comparisons of entire images or windows within them, are suitable for learning and describing the global structure of objects, but cannot deal with partial occlusion problems, large viewpoint changes, or with non-rigid objects.

In the past decade, there were significant advances in solving these problems with the development of local invariant features. The use of these features allows us to find local structures that are present in different views of the image. It also allows a description of these structures invariant to image transformations such as translation, rotation, scale and viewpoint.

SIFT is one the most widely used local approaches. It finds local structures (grids in Fig. 4) that are present in different views of the image. It also provides a description of these structures reasonably invariant to image variations such as translation, rotation, scale, illumination and affine transformations. Moreover, several studies have shown that the SIFT descriptor performs better than others (Mikolajczyk & Schmid, 2005).

The first stages of the SIFT algorithm find the scale-invariant keypoints in a certain scale and assign an orientation to each one. For each keypoint, scale and orientation are represented in Fig. 4 by the size and orientation of the grid, respectively. The results of these steps guarantee invariance to image location, scale and rotation. The keypoints detected by SIFT have a high repeatability, i.e., given two images of the same object or scene, taken under different viewing conditions, a high percentage of the features detected on the scene should be found in both images. Then, a descriptor is computed for each keypoint. This descriptor must be highly distinctive and partially robust to other variations such as illumination and 3D viewpoint.

To create the descriptor, Lowe proposed an array of 4×4 histograms of 8 bins (Lowe, 2004). These histograms are calculated from the values of orientation and magnitude of the gradient in a region of 16×16 pixels around the point so that each histogram is formed from a subregion of 4×4. Fig. 5 shows on its left side the gradient magnitude and orientation in a region around a keypoint, these values are weighted by a Gaussian window, indicated by the overlaid circle. The orientation histograms for each 4×4 region are shown on the right. Arrows longitude represent the accumulated gradient magnitude in each orientation. The descriptor vector is the result of the concatenation of these histograms. Since there are $4 \times 4 = 16$ histograms of 8 bins each, the resulting vector is of size 128. This vector is normalized in order to achieve invariance to illumination changes.

The distinctiveness of these descriptors allows us to use a simple algorithm to compare the collected set of feature vectors from one image to another in order to find correspondences between feature points in each image. These correspondences are adequate to identify particular objects in the image, but not to identify object classes. With this purpose in mind, in this paper SIFT feature descriptors are clustered to characterize object classes and are incorporated in a Bayesian network classifier.

3.2. SIFT features clustering

To build clusters of SIFT descriptors, the agglomerative hierarchical clustering method proposed by Johnson (1967) is used. Unlike K-means or EM-clustering, this algorithm does not depend on initialization. Furthermore, it has been reported to be superior to K-means (Jain & Dubes, 1998).

Given F features descriptors extracted from all the images in the training set, the clustering is initialized with F clusters, each one containing one descriptor only. In each iteration, the two clusters with the highest cohesion are merged.

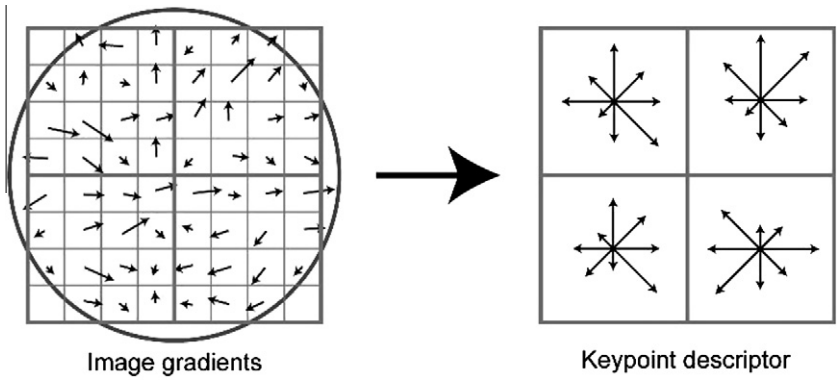

Fig. 5. This figure, for purposes of explanation, shows a descriptor of 2×2 regions calculated from a region of 8×8 pixels, but Lowe proposes to use descriptors of 4×4 regions in neighborhood of 16×16 pixels around the point.

The similarity between any two clusters can be measured in several ways; the most common are single linkage, complete linkage and average linkage. In this paper, average linkage is used, which is defined as the average distance of every element in a cluster to every other element in the other cluster:

$$D(k,l) = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N} d(k_m, l_n), \tag{4}$$

where M and N are the number of descriptors in the clusters k and l, respectively.

Agglomerative clustering produces a hierarchy of associations of clusters until the cut off criterion halts the process. Therefore, after each iteration, a new cluster is obtained from the pair of clusters with the highest similarity above a given value. This value is used as the cut off criterion and, ideally, should be the one that ensures that the resulting clusters are those that best characterize objects classes. In this work the cut off value is selected in the training phase. The cut off value that maximizes the accuracy in the classification for a given training set is selected. Therefore, there will be two training sets, one for obtaining the classes appearance models and another to obtain the cut off parameter.

The time complexity of the agglomerative hierarchical clustering algorithm is $O(n^4)$ (Olson, 1995), where n is the number of objects (i.e. $n = F$). In order to optimize the process of updating the similarities matrix, the Lance–Williams formulas (Lance & Williams, 1967), allow us to calculate the similarity between groups based on similarities among clusters of the previous level. Using this technique a complexity of $O(n^3)$ is achieved.

Day–Edelsbrunner algorithm (Day & Edelsbrunner, 1984) is a variation of a hierarchical clustering algorithm that uses the Lance–Williams formulas. This algorithm uses one priority queue per cluster to find the more similar pairs of clusters. Each queue has the remaining clusters ordered by similarity. At each iteration, update of the priority queue of each cluster is required. Using this structure, finding the more similar pairs of clusters is done in $O(n)$, therefore, the algorithm achieves a time complexity of $O(n^2 \log n)$.

Moreover, since the average linkage satisfies the principle of reducibility (Murtagh, 1983), the reciprocal nearest neighbors (RNNs) can be joined without having to find the more similar pairs of clusters. Two clusters are RNNs if one of them is most similar to the other and vice versa. In order to find the RNNs, a queue where each cluster is most similar to the one preceding it is created. The obtained clusters are ordered by the same similarity value that joined them in order to obtain the same dendrogram than the original hierarchical clustering algorithm. This mechanism achieved a time complexity of $O(n^2)$.

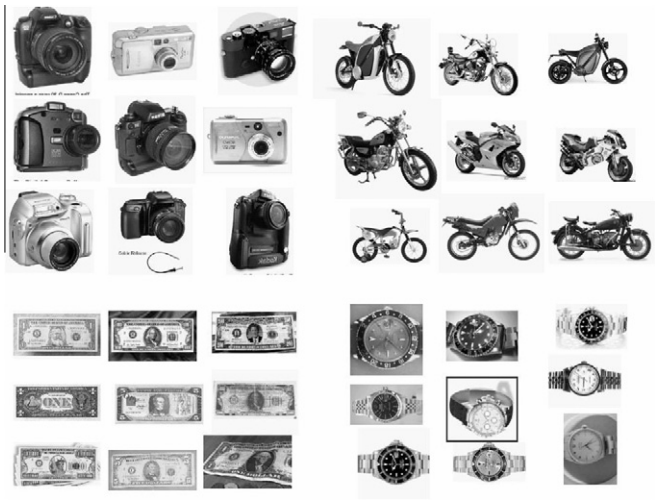

Fig. 6. Example images from the training set. The training set is composed of 20 images for each of the four classes. These images were randomly selected from the database. At top left: cameras, top right: motorcycles, bottom left: dollar bills, and bottom right: wristwatches.

4. Evaluation

In this section the evaluations and tests performed to the proposed method are presented. The data collection used and the measurement of variables are also described.

For the conducted experiments, images from the CalTech101 collection[3] (Li & Perona, 2005) were used. This database contains 101 different classes of objects and different numbers of images per class, the compression format is JPG and the average size is 300×300 pixels. Each image contains only one object centered in the image.

In order to test the performance of the proposed method, a system was trained to recognize four classes of objects (i.e., camera, dollar bill, motorcycle, and wristwatch), which were randomly selected. For training the classes appearance models, 20 images per class were used. Another 20 images per class were selected in order to train the clustering cutoff parameter, the resulting value of this parameter was 0.91. All the images in both training sets were also randomly selected. Example images for camera, dollar bills,

[3] Available online at: "http://www.vision.caltech.edu/Image_Datasets/Caltech101/\#Download".

motorcycle, and wristwatch from Caltech101 database are shown in Fig. 6.

Three experiments were conducted to evaluate the proposed method. The goal of the first experiment is to measure the performance of the proposed method in normal conditions (i.e., illumination, occlusion, rotation and scale problems-free images). The second experiment compares the method proposed in this paper with a straightforward classification method also using SIFT features. Finally, the third experiment measures how the performance of the proposed method behaves in the presence of partial occlusion and variation in illumination, scale and rotation in the test set.

The performance indicators used were recall, precision, true negative rate and accuracy. The recall rate measures the proportion of actual positives which are correctly identified as such:

$$recall = \frac{tp}{tp + fn}. \quad (5)$$

Precision is defined as the proportion of the true positives against all the positive results:

$$precision = \frac{tp}{tp + fp}. \quad (6)$$

The true negative rate (TNR) measures the proportion of negatives which are correctly identified:

$$TNR = \frac{tn}{fp + tn}. \quad (7)$$

The accuracy is the proportion of true results, both true positives and true negative, in the population:

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn}, \quad (8)$$

where tp, tn, fp, fn refer to the number of true positives, true negatives, false positives and false negatives, respectively.

Table 1
Performance indicators for Experiment 1.

Measures	Camera	Dollar bill	Motorcycle	Wristwatch	Average
Recall (%)	84.0	100	99.0	89.0	90.7
Precision (%)	94.6	89.2	90.5	98.9	93.3
True negative rate (%)	98.3	96.0	96.7	99.7	97.6
Accuracy (%)	93.5	95.0	95.0	94.5	94.5

5. Results and discussion

This section presents a quantitative evaluation of the proposed method and discusses the main results obtained.

5.1. Experiment 1

In Experiment 1, results were obtained for 100 test images per class. The goal of this experiment is to measure the performance of the proposed method in normal conditions. These images have small variations in occlusion, scale, illumination and rotation. Images from the training set were not in the test set.

Table 1 shows the results obtained in Experiment 1.

As could be seen in Table 1, all the measures averages were over 90%, which indicates the high performance of the proposed method.

5.2. Experiment 2

In order to evaluate the improvement introduced by the clustering of SIFT descriptors on the representation of object classes and the use of a Bayesian network in the classification phase, in this section we compare the method proposed in this paper with a straightforward classification method also using SIFT features,

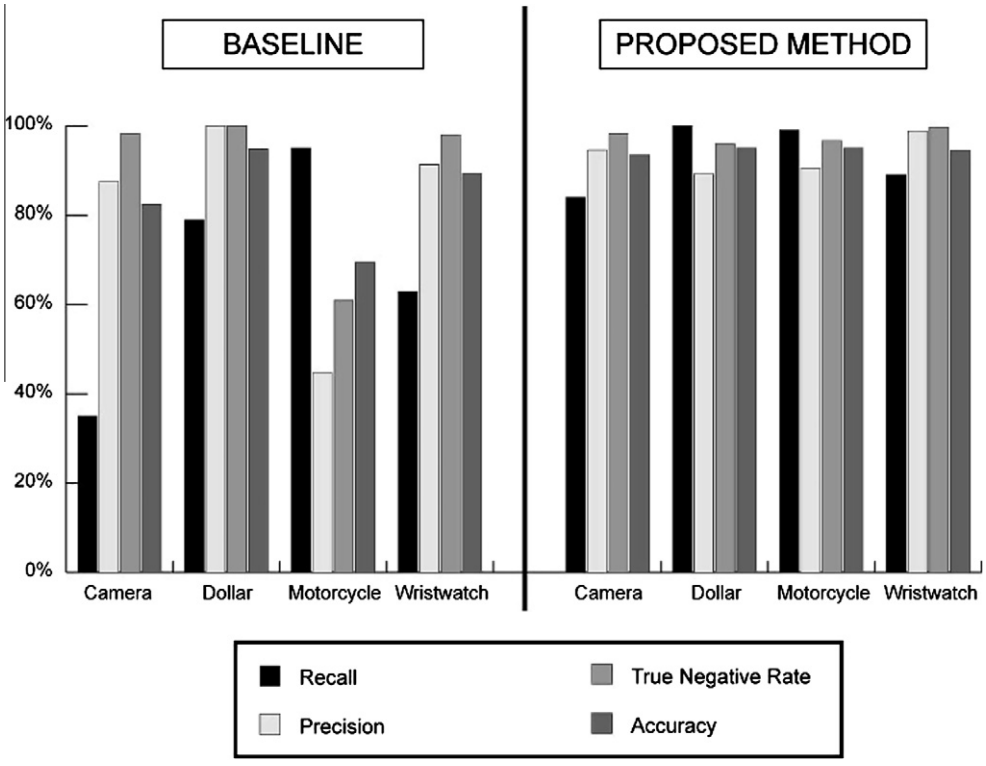

Fig. 7. Comparison between the Baseline method and the results obtained by our proposed method in Experiment 1.

which is taken as the baseline. This method is summarized as follows:

1. Extract SIFT features of each image from the training set.
2. For a new image *I* extract its SIFT features.
3. This image is matched with each of the images of the training set. The matching method used is the one proposed by Lowe (2004).
4. The class of the input image will be the one that receives the highest number of correspondences with image *I*.

To perform this experiment, the same training and test sets that were used in Experiment 1 were used. A comparison between the Baseline method and the results obtained by our proposed method in Experiment 1 are shown in Fig. 7. Each group of bars represents the values of recall, precision, true negative rate, and accuracy obtained for each class. On the left are the values for the baseline method and those of Experiment 1 on the right. We can see that the proposed method outperforms the baseline method. Our method performance measures are above 90%, while for the baseline method values are around 80%.

Table 2 summarizes the above comparison between the Baseline method and our method. The values in this table correspond to the average of each of these measures for all trained classes.

As could be noticed in Fig. 7 and Table 2, the proposed method outperforms the baseline method by a wide margin. This result gives evidence of the improvement introduced by the clustering of SIFT descriptors on the representation of object classes and the use of a Bayesian network in the classification phase.

Table 2
Comparison of baseline and Experiment 1.

Measures	Baseline	Proposed method
Recall (%)	68.0	90.7
Precision (%)	80.9	93.3
True negative rate (%)	89.3	97.6
Accuracy (%)	84.0	94.5

5.3. Experiment 3

The aim of Experiment 3 is to test the robustness of the proposed method to changes in illumination, occlusion, scale and rotation. For Experiment 3, the same model obtained in Experiment 1 is used to classify. To build the test set, 10 images that were correctly classified in Experiment 1 were randomly selected for each class. Variations in occlusion, scale, illumination and rotation were artificially introduced to each of these images, resulting in 40 images

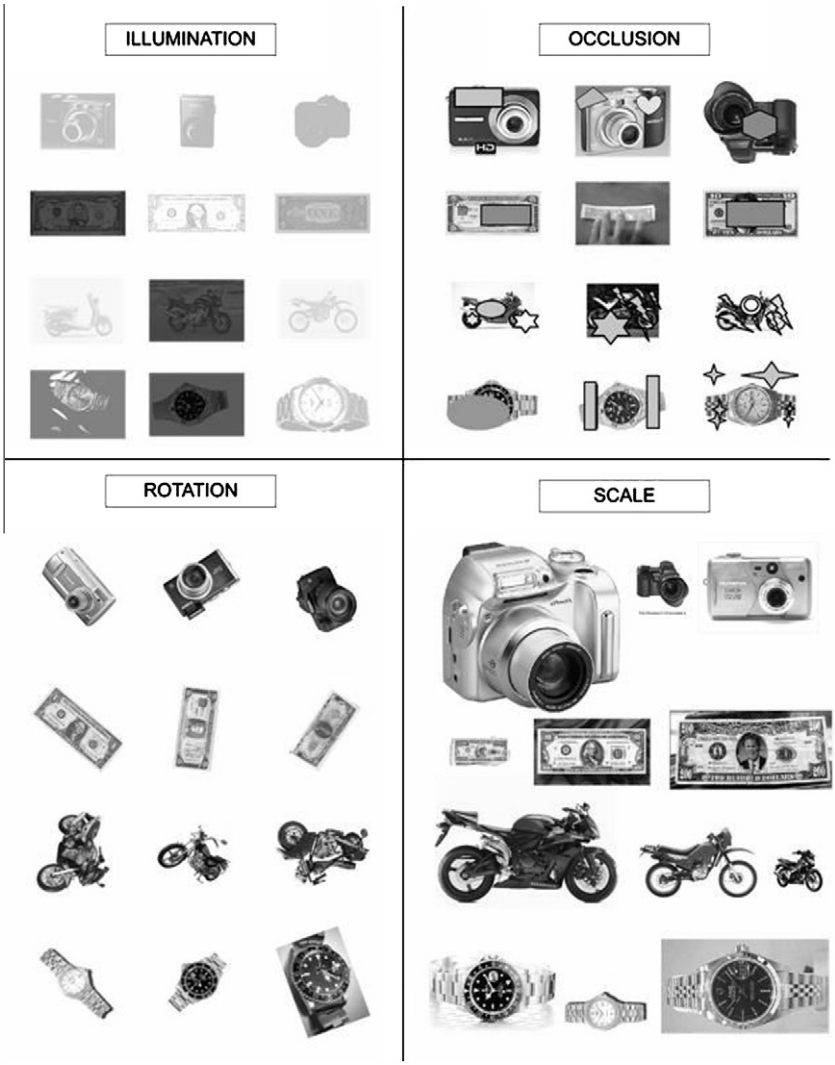

Fig. 8. Example images from the test set used for the Experiment 3. These images present partial occlusion and changes in illumination, rotation and scale.

Table 3
Performance indicators for Experiment 3.

Measure	Camera	Dollar bill	Motorcycle	Wristwatch
Recall (%)	94.8	95.3	92.5	100
Precision (%)	92.0	98.0	94.0	96.0
True negative rate (%)	97.5	99.3	97.9	98.7
Accuracy (%)	96.5	98.0	96.5	99.0

Table 4
Recall and precision measures (%) for each type of image alteration in Experiment 3.

Variation	Measure	Camera	Dollar bill	Motorcycle	Wristwatch
Occlusion	Recall	100	100	90.9	100
	Precision	90.0	100	100	100
Illumination	Recall	100	76.9	81.8	100
	Precision	70.0	100	90.0	90.0
Scale 2×	Recall	90.9	100	100	100
	Precision	100	100	90.0	100
Scale 0.5×	Recall	100	100	100	100
	Precision	100	100	100	100
Rotation	Recall	83.3	100	90.0	100
	Precision	100	90.0	90.0	90.0

per class. Example images from the test set used in this experiment are shown in Fig. 8.

Table 3 shows the performance results obtained in Experiment 3. As it could be seen, the average values of performance are maintained above 95%, showing the robustness of the proposed method to variations in illumination, occlusion, scale and rotation.

The recall and precision measures obtained for each kind of variation introduced to the test set are shown in Table 4. It could be noticed that there were no major falls in recall and precision rates, showing the largest variations (30%) in the precision on the camera class with illumination changes.

6. Conclusions

As a result of this work, a method for recognizing object classes using SIFT features have been developed. The proposed method performs clustering on the descriptors of the detected points to characterize the appearance of object classes. It also introduces the use of a three layer Bayesian network in the classification stage to improve classification rates. Three experiments were conducted to evaluate the proposed method. They showed that SIFT features are suitable to represent object classes, and evidenced the improvement achieved by clustering SIFT descriptors and using a Bayesian network for classification. These experiments also showed quantitatively the invariance of the method to illumination changes, scale, rotation and occlusion. It also provided experimental evidence that supports that a method based on clustering of SIFT features outperforms a straightforward object recognition method based on SIFT features to identify object classes.

As future work, the localization of objects in the image will be investigated, trying to learn the spatial relationships between the local features and clusters that describe an object class.

Acknowledgments

We want to acknowledge the distinction and support provided by Sociedad Mexicana de Inteligencia Artificial (SMIA) and the 9th Mexican International Conference on Artificial Intelligence (MICAI-2010) in order to enhance, improve, and publish this work. This project was supported in part by CONACYT grant No. 103876. L. Chang was supported in part by CONACYT scholarship No. 240251.

References

Agarwal, S., Awan, A., & Roth, D. (2004). Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(11), 1475–1490<http://www.dx.doi.org/10.1109/TPAMI.2004.108>.

Bay, H., Ess, A., Tuytelaars, T., & Van Gool, L. (2008). Speeded-up robust features (SURF). *Computer Vision and Image Understanding, 110*(3), 346–359<http://www.dx.doi.org/10.1016/j.cviu.2007.09.014>.

Day, W. H., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification, 1*, 7–24.

Dorkó, G., & Schmid, C. (2005). Object class recognition using discriminative local features, Tech. rep.. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jain, A. K., & Dubes, R. C. (1998). *Algorithms for clustering data*. Upper Saddle River, NJ, USA.: Prentice-Hall, Inc..

Jensen, F. V. (1996). *Introduction to Bayesian networks* (1st ed.). Secaucus, NJ, USA: Springer-Verlag New York, Inc..

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 2*, 241–254 <http://www.garfield.library.upenn.edu/classics1985/A1985AQU6100001.pdf>.

Kadir, T., & Brady, M. (2001). Scale, saliency and image description. *International Journal of Computer Vision, 45*(2), 83–105.

Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *The Computer Journal, 4*, 373–380.

Lazebnik, S., Schmid, C. & Ponce, J. (2005). A maximum entropy framework for part-based texture and object recognition. In *Proceedings of international conference of computer vision'05*.

Leibe, B., Seemann, E., & Schiele, B. (2005). Pedestrian detection in crowded scenes. *CVPR'05: Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 878–885). Washington, DC, USA: IEEE Computer Societyhttp://www.dx.doi.org/10.1109/CVPR.2005.272.

Li, F.-F., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. *Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Volume 2, pp. 524–531). Washington, DC, USA: IEEE Computer Society<http://www.dx.doi.org/10.1109/CVPR.2005.16>.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110http://www.dx.doi.org/10.1023/B:VISI.0000029664.99615.94.

Mikolajczyk, K., & Schmid, C. (2001). Indexing based on scale invariant interest points. *Proceedings of the 8th international conference on computer vision* (pp. 525–531). Vancouver, Canada <http://www.perception.inrialpes.fr/Publications/2001/MS01a>.

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International Journal of Computer Vision, 60*, 63–86. doi:10.1023/B:VISI.0000027790.02288.f2<http://www.portal.acm.org/citation.cfm?id=990376. 990402>.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(10), 1615–1630<http://www.dx.doi.org/10.1109/TPAMI.2005.188>.

Mikolajczyk, K., Leibe, B., & Schiele, B. (2005). Local features for object class recognition. In *ICCV'05: Proceedings of the tenth IEEE international conference on computer vision* (pp. 1792–1799). Washington, DC, USA: IEEE Computer Society<http://www.dx.doi.org/10.1109/ICCV.2005.146>.

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal, 26*, 354–359.

Olson, C. (1995). Parallel algorithms for hierarchical clustering. *Parallel Computing, 21*, 1313–1325.

Tuytelaars, T., & Mikolajczyk, K. (2007). Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision, 3*(3), 177–280<http://ww.dblp.uni-trier.de/db/journals/ftcgv/ftcgv3.html>.

Ullman, S. (1996). *High-level vision: Object recognition and visual cognition* (illustrated ed.). The MIT Press<http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0262210134>.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., & Gongin Y. (2010). Locality-constrained linear coding for image classification. In *2010 IEEE conference on computer vision and pattern recognition (CVPR)*.

Zhang, J., Lazebnik, S., & Schmid, C. (2007). Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer*, 73.