Object Class Recognition Using SIFT and Bayesian Networks

Leonardo Chang[1,2], Miriam Monica Duarte[2],
Luis Enrique Sucar[2], and Eduardo F. Morales[2]

[1] Advanced Technologies Application Center,
7a # 21812 b/ 218 y 222, Siboney, Playa, P.C. 12200, Havana, Cuba
lchang@cenatav.co.cu

[2] National Institute for Astrophysics, Optics and Electronics,
Luis Enrique Erro # 1, Tonantzintla, Puebla, Mexico
{lchang,mduarte,esucar,emorales}@ccc.inaoep.mx

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 problems of object class recognition 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 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.

Keywords: Object class recognition, local features, SIFT, clustering, Bayesian networks.

1 Introduction

In the last few years, local features 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 [11].

One of the most popular and widely used local approach is the SIFT (Scale Invariant Features Transform) method, proposed by Lowe [7]. The features extracted by SIFT are largely invariant to scale, rotation, illumination changes,

G. Sidorov et al. (Eds.): MICAI 2010, Part II, LNAI 6438, pp. 56–66, 2010.

noise and small changes in viewing direction. The SIFT descriptors have shown better results than other local descriptors [10].

The SIFT and 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.

In this work we use SIFT features to recognize object classes (e.g., books) in order to provide robustness to changes in scale, rotation, illumination and partial occlusion. The proposed method, in the training phase, 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. 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.

2 Related Work

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 features have become very popular to give solution to the limitations of these methods in object detection and recognition.

For object class recognition, many methods use clustering as an intermediate level of representation [1][6]. Due to the robustness of local features and the good results of clustering in objects classification, several authors have recently been investigating the use of clustering for object class recognition using local features based approaches. In [2], for invariant region detection, the authors use the Harris-Laplace [9] and the Kadir and Brady [5] detector. These regions are described using the SIFT descriptor [7]. 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.

In [8], Mikolajczyk *et al.* 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.

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.

The method proposed in this work also performs clustering on the descriptors of the features extracted from training images. The main difference with the previous mentioned methods is the use of a Bayesian network in the classification stage in order to improve performance on this stage. A deeper experimentation to measure the behavior against changes in illumination, scale, viewpoint and partial occlusion is presented as well. It is also shown how the use of clustering and Bayesian networks outperforms the traditional use of local features in object class recognition tasks.

3 SIFT Features Descriptors

SIFT is one the most widely used local approaches. It finds local structures 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.

The first stages of the SIFT algorithm find the coordinates of keypoints in a certain scale and assign an orientation to each one. The results of these steps guarantee invariance to image location, scale and rotation. 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 [7]. 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. The descriptor vector is a 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.

4 Learning Object Classes

A model 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 2), clustering is performed on feature descriptors extracted from the training images.

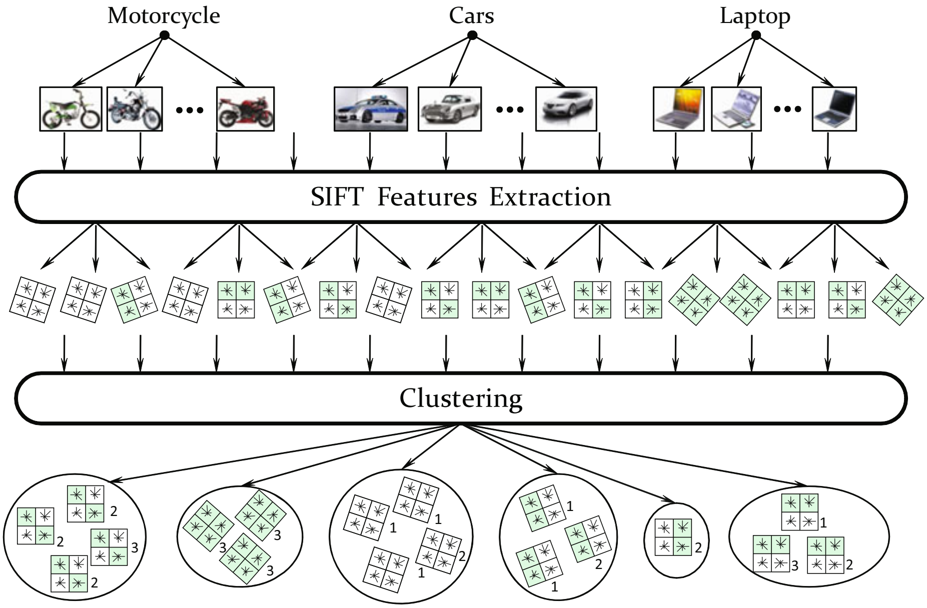

Fig. 1. SIFT local features are extracted from the training set formed by several sample images per class. Later, features descriptors are clustered and each feature in each cluster is labeled with its corresponding class.

Figure 1 shows a high level diagram of the class learning method used, 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.

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

4.1 SIFT Features Clustering

To build clusters of descriptors, the agglomerative hierarchical clustering method proposed by [4] is used. Unlike K-means or EM-clustering, this algorithm does not depend on initialization. Furthermore, it has been reported superior to K-means [3].

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.

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 any other element in other cluster:

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

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 cut off criterion.

5 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. Figure 2 shows a layout of the proposed method.

This idea can be represented as a three layer Bayesian network (BN). The graphical representation of this BN is shown in Figure 3. At the first layer we have the trained object classes represented by $c_1, c_2, ..., c_C$ where C is the number of classes. At the second layer, clusters obtained in the training phase are represented by $k_1, k_2, ..., k_K$ 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 $f_1, f_2, ..., f_F$ where F is the number of features extracted from the image.

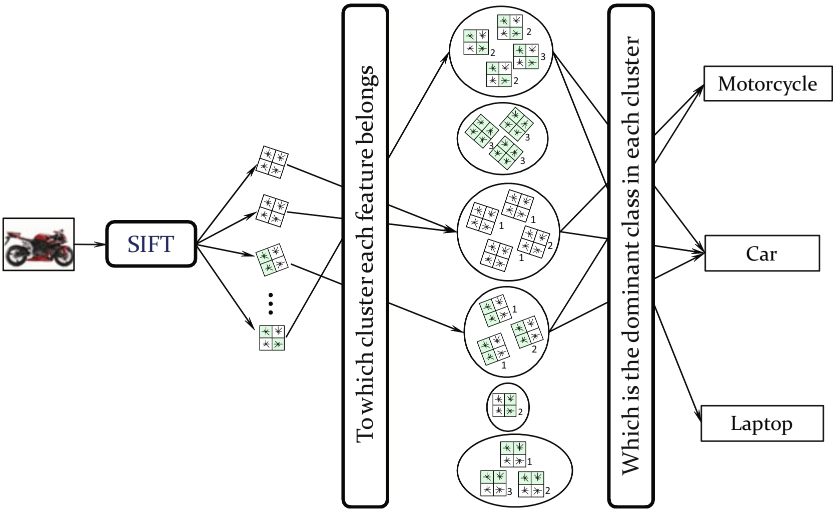
SIFT
To which cluster each feature belongs
Which is the dominant class in each cluster
Motorcycle
Car
Laptop
1
2
3

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.

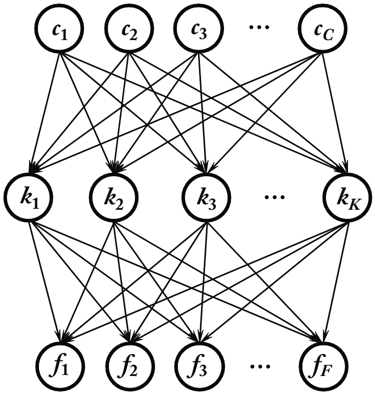

Fig. 3. Graphical representation of the three layer Bayesian network used to classify a new object

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 function of 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 Equation 1 is used:

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

$$P(f|k_i) = 1 - \frac{D(f, k_i)}{\max_{kl} D(k_k, k_l)}$$

3. For each cluster $k_{f_1}, k_{f_2}, ..., k_{f_F}$ 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:

$$c^* = \arg\max_i \sum_f P(C_i|k_f)P(k_f)$$

6 Experiments and Results

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

For the conducted experiments, images from the Pascal[1] collection were used. This database contains 101 different classes of objects and different numbers of images per class, the 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, 20 images per class were used, also randomly selected. Example images from the training set are shown in Figure 4.

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 ($recall = TP/(TP + FN)$). Precision is defined as the proportion of the true positives against all the positive results ($precision = TP/(TP + FP)$). The True Negative Rate (TNR) measures the proportion of negatives which are correctly identified ($TNR = TN/(FP+TN)$). The accuracy is the proportion of true results, both true positives and true negative, in the population ($accuracy = (TP + TN)/(P + N)$).

[1] Available online at:
"http://pascallin.ecs.soton.ac.uk/challenges/VOC/download/101objects.tar.gz"

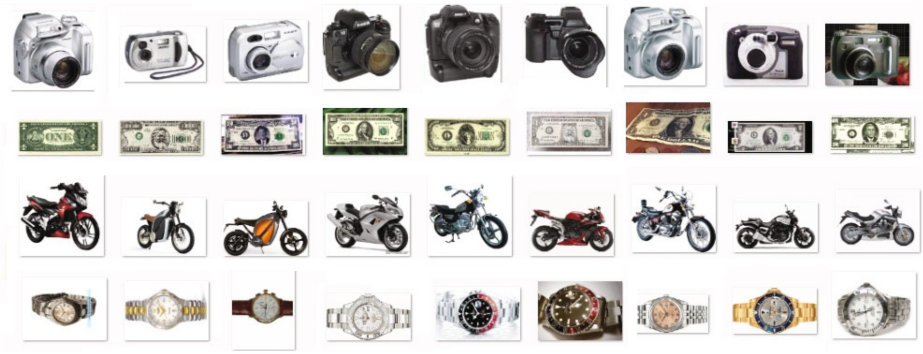

Fig. 4. Example images from the training set. The training set is composed of 20 images for each of the 4 classes. These images were randomly selected from the database.

6.1 Experiment 1

In Experiment 1, results were obtained for 100 test images per class. 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.

Table 1. Performance indicators for Experiment 1

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

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

6.2 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, which is taken as 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 in [7].
4. The class of the input image will be the one that receives the highest number of correspondences with image I.

Table 2 shows a comparison between the results obtained by the baseline method and the results obtained in Experiment 1. To perform this experiment, the same training and test sets that in Experiments 1 were used.

Table 2. Comparison of Baseline and Experiment 1

	Baseline	Experiment 1
Recall (%)	68.0	90.7
Precision (%)	80.9	93.3
TrueNegativeRate (%)	89.3	97.6
Accuracy (%)	84.0	94.5

As could be noticed in Table 2, the proposed method surpassed in each of the performance measures to 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.

6.3 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, 10

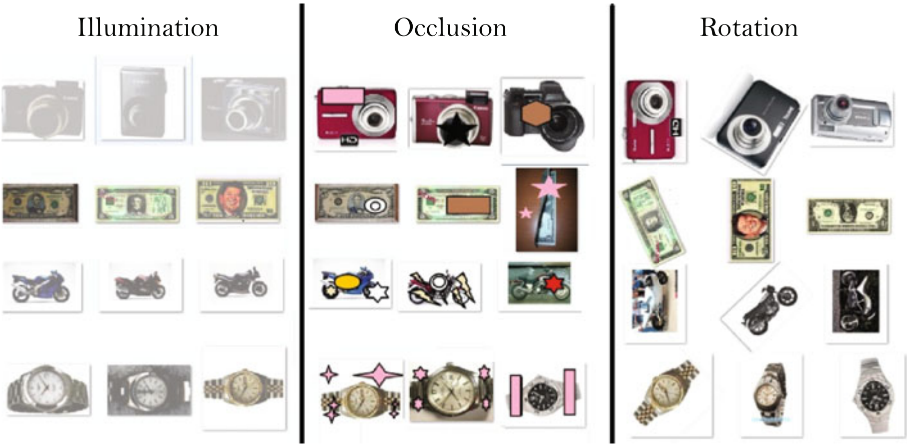

Fig. 5. Example images from the test set used for the Experiment 3. These images present partial occlusion and changes in illumination, rotation and scale.

Table 3. Performance indicators for Experiment 3

	Recall (%)	Precision (%)	TrueNegativeRate (%)	Accuracy (%)
Camera	94.8	92.0	97.5	96.5
Dollar bill	95.3	98.0	99.3	98.0
Motorcycle	92.5	94.0	97.9	96.5
Wristwatch	100	96.0	98.7	99.0
Average	95.6	95.0	98.3	97.5

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

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

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 per class. Example images from the test set used in this experiment are shown in Figure 5.

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 is 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 illumination changed images in the class camera.

7 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 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.

Acknowledgements

This project was supported in part by CONACYT grant No. 103876. L. Chang was supported in part by CONACYT scholarship No. 240251.

References

1. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. IEEE Trans. Pattern Anal. Mach. Intell. 26(11), 1475–1490 (2004)
2. Dorkó, G., Schmid, C.: Object class recognition using discriminative local features. Technical report, IEEE Transactions on Pattern Analysis and Machine Intelligence (2005)
3. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice-Hall, Inc., Upper Saddle River (1988)
4. Johnson, S.C.: Hierarchical clustering schemes. Psychometrika 2, 241–254 (1967)
5. Kadir, T., Brady, M.: Scale, saliency and image description. International Journal of Computer Vision 45(2), 83–105 (2001)
6. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: CVPR 2005: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), Washington, DC, USA, vol. 1, pp. 878–885. IEEE Computer Society, Los Alamitos (2005)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
8. Mikolajczyk, K., Leibe, B., Schiele, B.: Local features for object class recognition. In: ICCV 2005: Proceedings of the Tenth IEEE International Conference on Computer Vision, Washington, DC, USA, pp. 1792–1799. IEEE Computer Society, Los Alamitos (2005)
9. Mikolajczyk, K., Schmid, C.: Indexing based on scale invariant interest points. In: Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada, pp. 525–531 (2001)
10. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. 27(10), 1615–1630 (2005)
11. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: A survey. Foundations and Trends in Computer Graphics and Vision 3(3), 177–280 (2007)