Uniqueness Filtering for Local Feature Descriptors in Urban Building Recognition

Giang Phuong Nguyen and Hans Jørgen Andersen

Department of Media Technology and Engineering Science,
Aalborg University, Denmark
{gnp,hja}@cvmt.dk

Abstract. Existing local feature detectors such as Scale Invariant Feature Transform (SIFT) usually produce a large number of features per image. This is a major disadvantage in terms of the speed of search and recognition in a run-time application. Besides, not all detected features are equally important in the search. It is therefore essential to select informative descriptors. In this paper, we propose a new approach to selecting a subset of local feature descriptors. Uniqueness is used as a filtering criterion in selecting informative features. We formalize the notion of uniqueness and show how it can be used for selection purposes. To evaluate our approach, we carried out experiments in urban building recognition domains with different datasets. The results show a significant improvement not only in recognition speed, as a result of using fewer features, but also in the performance of the system with selected features.

1 Introduction

Urban buildings usually have a lot of structural features such as windows, doors, and columns. For this reason local descriptors are widely used in building recognition to describe the content of captured images. A comprehensive overview of existing local descriptors can be found in [1]. Due to their stability under image transformation, partial occlusions and viewpoint changes, interest points are used in creating descriptors. The usual procedure in this approach is first to localize interest points, then to describe a region around each point to form a descriptor of that region. Well-known techniques in this regard are SIFT (Scale Invariant Feature Transforms) [2] and other SIFT-based approaches [3].

While most of the current systems that apply local descriptors employ SIFT in the original or a modified version, we have investigated a more recent technique known as Multi-Scale Oriented Patches (MOPS) that was developed by Brown et al. [4]. In previous work [5] we have analyzed the features of MOPS in detail and compared its performance in building recognition to that of SIFT. In this respect we concluded that MOPS performs better in urban building recognition, especially when large temporal variations are involved. In this paper, we also employ MOPS.

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 85–93, 2008.

A major issue in approaches that use local detectors is that they usually produce a large number of interest points, i.e. a large set of descriptors is extracted for each image. The number of descriptors greatly depends on the content of the image. For example, SIFT detectors on average create $\sim$ 2000 descriptors for an image of 500x500 pixels [2]. As mentioned at the outset, buildings in general are highly structured, and this means that the number of interest points can go much higher than this. Since the computational cost of matching is positively correlated with the number of descriptors extracted, this problem should be taken into account. Recently, more attention has been paid to recognition speed in application systems, since this is an essential requirement in applications such as robot tracking [6,7], and to run-time recognition with mounted devices [8]. To meet this speed requirement, different approaches have been proposed to improve local detectors. In [3], the authors present a method of reducing the dimensionality of SIFT descriptors, using the PCA dimensional reduction method which projects the original SIFT feature space from 128 dimensions to 20 dimensions. The PCA-SIFT method achieves significant space benefits and requires a third of the time in the matching phase compared to the original SIFT. A different approach is put forward in [9] where a vocabulary tree is used to index descriptors. The K-means algorithm is used to cluster all descriptors and place them in the correct branch. For each query image, extracted descriptors are traced down the tree, a score list is given for all leaves, and the one with the highest score is returned as the best match. This approach has proved to be very fast and scalable to a very large number of descriptors.

The above approaches do not alter the original number of features. However, not all features are equally important. Some detected features are irrelevant in the recognition phase. In such cases, having too many descriptors will reduce the recognition rate. For this reason attention should be focused only on those features that are informative. Examples of techniques used for this purpose can be found in [10,11,4], where the authors concentrate on selecting a subset from given descriptors. For instance, in [10], the authors apply a method designed to select discriminative features best suited for characterizing individual locations in indoor office environments. They produced a high location recognition rate using only 10% of the original detected features. They were therefore able substantially to speed up recognition. In this paper, we focus on developing a new technique for selecting a subset of descriptors that should meet two essential requirements. The first requirement is that of speed. The second is that, while reducing information from the original descriptors, the system should perform as well as or better than existing systems. Throughout the paper, we will use the two terms “features” and “descriptors” interchangeably.

The paper is organized as follows. First, in section 2, we give a brief description of the MOPS detector. This technique is used as the basis for our approach to extracting initial descriptors. In section 3, we present an overview of existing techniques for defining informative descriptors. We then propose our own method of selecting a subset of features. In the next section, we apply our technique to a real application in urban building recognition with two different datasets.

The results are evaluated and discussed in section 4. Finally, we present our conclusions in section 5.

2 Review of the MOPS Detector

Multi-scale oriented patches is a technique recently proposed by Brown et al. for detecting local image features [4]. Let us take an image set $\mathcal{I}$. Each input image $I_i \in \mathcal{I}$ is incrementally smoothed with a Gaussian kernel $\{\sigma_t\}_{t=1..n}$. An image pyramid is then constructed by down-sampling the image at rate r. In the second step, interest points are extracted using the Harris corners detector at each level of the pyramid. This step yields a set of points at locations where the corner strength is a local maximum of a 3×3 neighborhood and above a threshold of 10 [4]. In the next step, the sub-pixel precision is found by means of a Taylor expansion (up to the quadric term) at those extreme points. Each extreme point is described in terms of its orientation within a window of size 28×28 (corresponding to a Gaussian kernel with $\sigma = 4.5$), and through sampling of grey level values in a 40×40 neighborhood. The grey level values are sampled in a grid with a spacing of 5 pixels rotated according to the orientation. This gives a feature vector for each landmark consisting of 8×8 grey level values. Before matching, the feature vector is standardized by subtracting the mean and dividing it by its standard deviation. Then, as in [4], we perform a Haar wavelet transform on the 8×8 descriptor patch to form a feature vector of 64 dimensions F_j. Finally, for each image we obtain a set of descriptors $\mathcal{F}_{I_i} = \{F_j\}_{j=1..k}$, where k is the number of descriptors extracted from the image I_i.

3 Selecting a Subset of Descriptors

In developing techniques for selecting descriptors, it is generally assumed that certain descriptors are more important than others. The terms "discriminative" and "informative" are usually used to describe significant descriptors. In [10], the authors observe that certain features are more stable and thus better able to handle variations in scale and viewpoint. They therefore aim to select such features. For each feature extracted by means of the SIFT detector from each image at each location, they calculate a posterior probability. The probability values are used as ranking criteria. In [4], the authors present an adaptive non-maximal suppression (ANMS) algorithm that selects a subset of interest points based on their corner strength. The general idea of this algorithm is that for each point extracted through the process described above, they calculate the corner strength, then select points that are maximum within a neighborhood of radius k pixels. In all their experiments, the authors select a maximum of 500 points for each image. This means a set of 500 descriptors is used to describe the content of an image. Another technique for selecting informative (i-SIFT)descriptors,

using the SIFT detector, can be found in [11]. For each given image, informative descriptors are defined as those that appear in discriminative regions. These regions are detected on the basis of an entropy-coded image derived by calculating posterior distribution.

We propose a different way of defining discriminative descriptors. They should identify the most salient features of a given building. One salient property is rarity [12,13]. We therefore define discriminative descriptors as those that are almost unique. Such descriptors maximize discrimination between objects. We thus propose to select descriptors on the basis of their uniqueness i.e. their rarity within a descriptor set.

Definition: *A unique feature has an identifiable property that distinguishes it from other features in the image.*

In other words, in a feature space where all descriptors are located, the unique feature is that which has the fewest features within its ϵ-neighborhood. Embarking from this definition, we now present our method of selecting unique features. Given an image I, assume that I has k descriptors or feature vectors $\mathcal{F}_I = \{F_1, F_2, \ldots, F_k\}$. To calculate the uniqueness of each feature, we first compute the dissimilarity values between feature vectors. For the MOPS descriptors, L2 distance is used as the dissimilarity function. We have $S_{ij} = \sqrt{\sum_{l=1}^{t}(f_l^i - f_l^j)^2}$, where f_l^i and f_l^j are components of feature vectors F_i and F_j respectively, and $t = 64$. Each feature vector F_i is compared to the others $\{F_1, \ldots, F_{i-1}, F_{i+1}, \ldots, F_k\}$. We then obtain a set of dissimilarity values $\{S_{i1}, S_{i2}, \ldots, S_{i,i-1}, S_{i,i+1}, \ldots, S_{ik}\}$. To decide whether two feature vectors are similar or not, an ϵ neighborhood is established. If a feature point F_j in the given feature space falls within the ϵ-neighborhood of F_i, it is considered similar to F_i, i.e.

If $S_{ij} < \epsilon$ then F_i and F_j are similar,
else then F_i and F_j are dissimilar.

As indicated in our definition of a unique feature, we will select features that have the smallest number of similar features within their ϵ-neighborhood. This means that the greater the number of neighbors, the less unique the feature is. Hence, the uniqueness of a feature vector F_i in image I is formulated as:

$$\mathcal{U}_{F_i} = \|\{S_{ij} < \epsilon\}\|_{j=1..k, j\neq i}$$

4 Application in Urban Building Recognition

Building recognition, or object recognition in general, often involves dealing with great variations in image content due to differences of scale, orientation, illumination, and viewpoint. Moreover, different buildings often have very similar structures. Selecting unique features that are able to distinguish one building from another is an essential issue. In this section, we will present our approach to urban building recognition.

4.1 Experimental Setup

We select two different datasets. The first, AAU dataset, contains 135 images captured in the Aalborg University area. The set includes a total of 21 buildings, with an average of 6 images for each building. In this dataset, the buildings are of very similar architectural design. In the second dataset (Aalborg city center), we use a set of 442 images of 19 buildings in the center of Aalborg. The main difference between this dataset and the AAU set is that the images from the city center were taken at different times of day (morning, afternoon, and evening), on different days, and under different weather and seasonal conditions (sunny, cloudy, winter, summer). Moreover, the buildings in the center vary in structure, have more decoration and are more often occluded by passing vehicles or people. These factors together pose a considerable challenge. The following shows more detail on how we created this dataset:

16/11/2006: during daytime (a cloudy day).
17/11/2006: in the evening with electrical lights on.
28/11/2006: during daytime (with Christmas decoration).
29/11/2006: during daytime (with Christmas decoration).
05/12/2006: in the evening (with Christmas decoration).
03/05/2007: sunny day (with Danish flags decoration).

Once the two datasets were created, we applied the MOPS detector to each one. Default parameters of the MOPS were used, and all extracted descriptors stored. Next, we applied our method of selecting unique features of each image. These unique features were stored separately. Since the extraction of unique feature from the images in the datasets was done offline, we did not count the time needed for this step. Instead, we compared our performance with that achieved by other methods. To evaluate performance during the matching process, we reported precision values. Each image in the dataset was sequentially used as a query. The query was then compared to all other images in the corresponding dataset. The top five best matches were returned, and the precision values calculated for each of these. The baseline is the performance achieved by using the default MOPS detector on all extracted descriptors. We also applied the so-called adaptive non-maximal suppression (ANMS)selection method described in [4], which selects the strongest features on the basis of corner strength. For a fair comparison, we performed experiments with different numbers of selected features, namely 100, 200, 300, 500, 800, and 1000. In all experiments, we identified unique features using $\epsilon = 0.3$.

4.2 Experimental Results

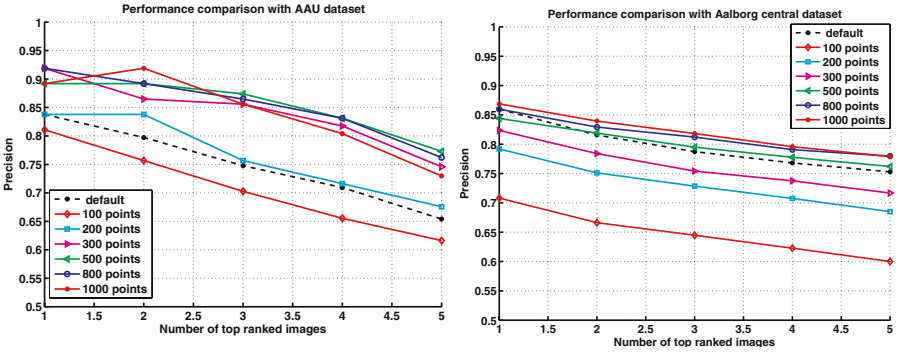
First, we present the results from different cases in which varying numbers of descriptors were used, as shown in table 2 and 1. k^* represents our uniqueness filtering method. In these tables, the last row is the default MOPS with all features taken into account. The second columns report the average number of descriptors per image. The third columns report the total descriptors of each

Table 1. AAU dataset with different numbers of unique features per image vs. default MOPS detector

Method	Features/image	Total features
k^*	100	13500
k^*	200	27000
k^*	300	40500
k^*	500	67255
k^*	800	108000
k^*	1000	130831
default	1470	198479

Table 2. Aalborg dataset with different numbers of unique features per image vs. the default MOPS detector

Method	Features/image	Total features
k^*	100	44200
k^*	200	88400
k^*	300	132600
k^*	500	221000
k^*	800	353600
k^*	1000	442000
default	2160	954409



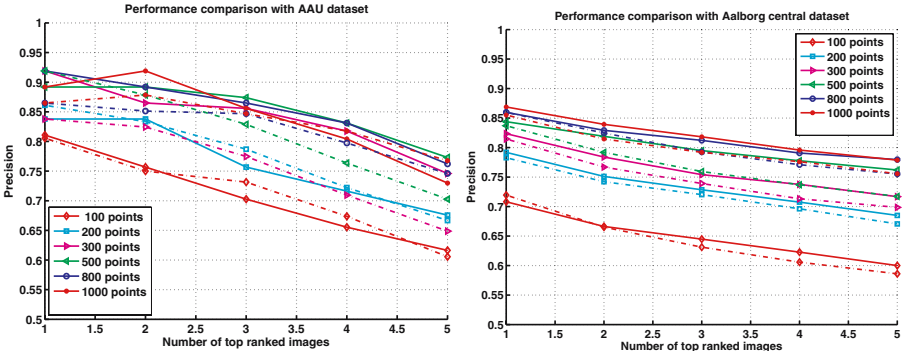
(a) (b)

Fig. 1. a. AAU dataset; b. Aalborg city center dataset. Precision vs. the number of top ranked images. Results show performance when the default MOPS with all extracted features is used, versus our approach which focuses only on unique features.

dataset. They show that, on average, the number of features extracted in the default cases is much higher, especially where the Aalborg city center dataset is concerned.

In figures 1a and 1b, we show recognition results for the two datasets. In these figures, we compare performance when different numbers of unique features are used, versus default performance when all extracted features are employed. The figures show that although the default MOPS has the highest number of descriptors, it performs less accurately than our method, in which smaller numbers of descriptors are used. The results achieved by using a certain number of unique features are significantly better, especially in the case of the AAU dataset. With this dataset, even selecting as few as 200 descriptors per image produces better results than the default approach. This means that instead of using all features, we can limit ourselves to only $\sim \frac{1}{7}$ of the total number. As noted in the above section, the Aalborg city center dataset presents a more difficult challenge. Here, at least 300 features were needed to enable reliable recognition of a given

building, although even where fewer features were used we still obtained a rather high recognition rate of 70% in the best match. However, with 500 unique features, i.e. fewer than $\sim \frac{1}{4}$ of the total descriptors, we achieved the same performance as the default approach. Further improvement is shown with 800 and 1000 unique features. We can also see from the two figures that there is a saturation point at which there is little improvement in performance between 800 or 1000 descriptors. This means more descriptors are unnecessary and may even reduce performance by creating disturbance.



(a) (b)

Fig. 2. a. AAU dataset; b. Aalborg city center dataset. Precision vs. the number of top ranked images. Results show performance using our approach vs. ANMS in selection of descriptors.

In the next experiment, we compared the performance achieved by our approach with that achieved when ANMS was used for selecting descriptors. Figures 2a and 2b show the results of the two approaches with different numbers of selected descriptors. The dotted lines represent the results when the ANMS approach was used, and the solid lines the results when our uniqueness method was applied. With 100 or 200 selected descriptors, the performance of the two approaches was comparable in both datasets. Where a higher number of descriptors was used, the results achieved by our method were superior. For example, when 300 descriptors are selected in the AAU dataset, we can achieve a recognition rate of $\{92\%, 87\%, 86\%, 82\%, 75\%\}$, while with ANMS we get $\{84\%, 82\%, 77\%, 71\%, 65\%\}$ for the top 1, 2, 3, 4, and 5 matches respectively.

In figure 3, we show some examples of buildings with 500 selected descriptors. The left column shows the results produced with our approach, and the right column the results produced with ANMS. In general, the descriptors selected by the two methods are quite close. ANMS selects features on the basis of their corner strength. Thus the features selected are mainly corners. In our case, we do not consider corners the most important features since they will recur in other similar areas. The features selected should be those that are least like other

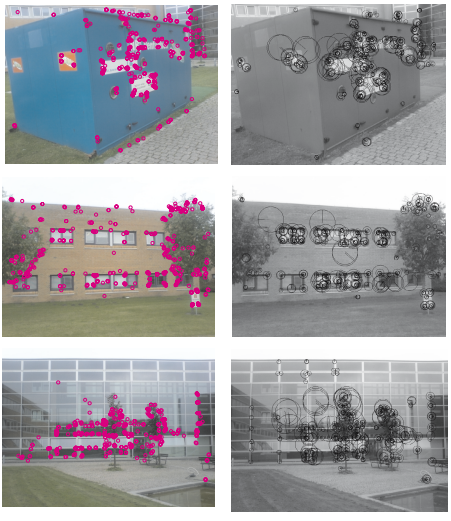


Fig. 3. Examples of buildings with 500 selected descriptors using uniqueness criteria vs. ANMS

features in the same image: i.e. where the smallest number of features in each image resemble them. Since corners are not our first priority in selection, other salient details will be chosen instead.

The proposed approach works well in the applications tested. In general, corner points are good features to detect. When it comes to distinguishing between different buildings, however, corners are not the most discriminative features. This has been shown by our experiments, where the selected points were not always corners. Further, the experiments with both datasets show that the default parameters of the MOPS detector are not optimal in building recognition. It is difficult to determine the number of descriptors that should be extracted per image. With fewer descriptors you certainly get the advantage of greater recognition speed, but on the other hand may not produce sufficient information for matching purposes. The results show that more descriptors are required to capture the content of a complex scene.

5 Conclusion

We have presented a new approach to selecting informative descriptors of an image, based on identifying unique features. Our experimental results with different datasets show that our approach improves recognition performance even where far fewer descriptors are used. Moreover, reducing the number of descriptors also speeds up the matching process, and this is an important factor for any run-time application. However, the time needed for the complex process of extracting all

the descriptors before the selection still remains. Finding informative features without extracting all descriptors is of interest to us in our future work.

Acknowledgement

This research is supported by the IPCity project (FP-2004-IST-4-27571), a EU-funded Sixth Framework program Integrated project on Interaction and Presence in Urban Environments.s

References

1. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis & Machine Intelligence 27(10), 1615–1630 (2005)
2. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
3. Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 506–513 (2004)
4. Brown, M., Szeliski, R., Winder, S.: Multi-image matching using multi-scale oriented patches. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 510–517 (2005)
5. Nguyen, G.P., Andersen, H., Christensen, M.: Urban building recognition during significant temporal variations. In: IEEE Workshop on Application of Computer Vision (2008)
6. Royer, E., et al.: Localization in urban environments: monocular vision compared to a differential gps sensor. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 114–121 (2005)
7. Kosecka, J., Li, F., Yang, X.: Global localization and relative positioning based on scale-invariant keypoints. Robotics and Autonomous Systems 52(1), 27–38 (2005)
8. Robertson, D., Cipolla, R.: An image based system for urban navigation. In: British Machine Vision Conference (2004)
9. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 2, pp. 2161–2168 (2006)
10. Li, F., Kosecka, J.: Probabilistic location recognition using reduced feature set. In: IEEE Intl. Conf. on Robotics and Automation, pp. 3405–3410 (2006)
11. Fritz, G., Seifert, C., Paletta, L.: Urban object recognition from informative local features. In: IEEE Intl. Conf. on Robotics and Automation, pp. 132–138 (2005)
12. Kadir, T., Brady, M.: Saliency, scale and image description. International Journal of Computer Vision 2(45), 83–105 (2001)
13. Schiele, B., Crowley, J.L.: Probabilistic object recognition using multidimensional receptive field histograms. In: International Conference on Pattern Recognition (1996)