

Combining Local Descriptors for 3D Object Recognition and Categorization

Andrea Selinger Salgian
Department of Computer Science
The College of New Jersey
salgian@tcnj.edu

Abstract

Various local descriptors have been used successfully in a variety of tasks including object recognition. Although different descriptors have been shown to have different strengths, they haven't been used in combination. In this paper we show that by combining local image descriptors at the feature level, we can significantly improve object recognition performance. Our system uses keyed context patches and SIFT, two descriptors that have been shown to have a somewhat uncorrelated performance [9]. By requiring hypotheses generated by both types of descriptors to satisfy the same consistency constraints, we were able to significantly reduce the error rate on recognition and categorization tasks.

1. Introduction

Many different local descriptors have been proposed in the computer vision literature, and they have been successful in a variety of applications, including object recognition [1], [2], [5], [8]. These descriptors can be computed efficiently, are resistant to clutter and partial occlusion, and are somewhat insensitive to pose, i.e. they change relatively slowly as the view of the object changes.

One of the most popular and successful local descriptors is SIFT (Scale-Invariant Feature Transform), introduced by Lowe [5]. SIFT features use smoothed weighted histograms of the image gradient. They are invariant to image scaling and rotation, and partially invariant to change in illumination and 3D camera viewpoint. After comparing a number of local image descriptors, Mikolajczyk and Schmid [6] found that SIFT-based descriptors performed best in image matching tasks.

Recently, Selinger-Salgian [9] used an object recognition setting to compare SIFT-based descriptors and keyed context patches, a feature that is centered around

prominent contour fragments extracted from the image [8]. She concluded that keyed context patches perform better overall, but the performance of individual descriptors is uncorrelated and can be improved through rank combination.

In this paper we answer a number of questions that were left open by [9]. We add a pose and scale consistency verification step to SIFT which was missing in that experiment, we combine the two descriptors at a lower level than rank fusion, and we extend the test cases to include not just recognition of six objects on cluttered backgrounds, but also object categorization using the ETH-80 database. We show that by using a combination of keyed context patches and SIFT, object recognition and categorization performance can be improved significantly.

2. The Descriptors

2.1. Keyed Context Patches

The keyed context patch method [8] starts by extracting contours from images using a stick growing method developed by Nelson [7]. The method uses both gradient magnitude and direction information to extract a set of boundary fragments terminated at corners (regions of high curvature).

In the second stage, keyed context patches are constructed by taking the prominent (i.e. longest) contour fragments (key curves) and embedding them in a local context consisting of a square image region, oriented and normalized for size by the key curve, which is placed at the center. Each keyed context patch contains a representation of all other segmented curves, key or not, that intersect it.

For object recognition, keyed context patches are extracted from the set of gallery images, and stored in a database together with the identity of the object that produced them, and the viewpoint they were taken from. The basic recognition procedure consists of four

steps. First, keyed context patches are extracted from the probe image. In the second step, these keyed context patches are used to access the database and retrieve information about what objects could have produced them. Verifying a match between a patch from the image and a stored patch is done using a form of directional correlation. Each context patch from the image may match zero, one or more context patches from view models in the database, and each match generates a hypothesis about the identity and pose of an object that could have produced it. Patches that are consistent with the same view model in the same configuration form a group that accumulates evidence for that configuration and view model. Finally, in the fourth step, after all features have been processed, the hypothesis corresponding to the group with the highest evidence score is selected.

2.2. SIFT

The SIFT method, as described in [5], consists of four major stages: scale-space extrema detection, keypoint localization, orientation assignment, and keypoint descriptor extraction. Matching gallery and probe features are found based on the Euclidean distance between features. A training feature will match a probe feature if the ratio between the distance to the closest probe feature and the second-closest probe feature is above a pre-defined threshold.

In [9] object recognition was performed by finding the gallery image that had the largest percentage of its features matched by features from the incoming probe image, without any additional consistency checks. Our approach is somewhat different, and is modeled on the keyed context patch approach described in the previous section.

The SIFT features extracted from a set of gallery images are stored in a database that contains the identity and viewpoint of the object that produced each feature. For recognition, SIFT features are extracted from the probe image and looked up in the database. Each database feature will match at most one image feature and each probe feature will match zero, one or more database features. Each matching database feature generates a hypothesis about the identity, pose and size of the object that may have generated it. Two hypotheses are consistent if the identity of the object is the same, the distance between the two hypothesized locations is less than 1/10 of the object size, the two hypothesized objects differ by less than 20% in scale, and their relative rotation in the image plane is less than 20 degrees. Consistent hypotheses accumulate evidence for the object identity, pose and size that they represent. The evi-

dence contributed by a single probe feature is computed based on the following:

- the frequency of the feature in the database, i.e. the number of database features that match the probe feature. Larger frequencies imply less distinctive features, so they contribute a smaller evidence.
- the number of features in the hypothesized training image. Objects and views with a larger number of features may end up accumulating more evidence than views with a smaller number of features. Normalizing by the number of features avoids this artifact.
- the size of the contributing feature in the probe image. Larger features are less likely to arise from noise and background clutter than smaller features, therefore they contribute more evidence.

After all features have been processed, the hypothesis corresponding to the group with the highest evidence score is selected.

2.3. Combining the descriptors

To combine the descriptors, we extracted the context patches and SIFT features from the probe and looked them up in the corresponding training database. Each context patch and SIFT feature contributed a number of hypotheses which were checked for consistency using the constraints described in Section 2.2, regardless of the type of descriptor that generated them. Each descriptor contributed evidence as described in Section 2.2, and evidences for consistent hypotheses were added regardless of descriptor type. Just like in the case of context patches and SIFT alone, the last step is to select the group with the highest evidence score.

This approach is more desirable than the rank fusion in [9] because the combined evidence scores are consistent in location, orientation and scale, and bring more information than simple rankings.

3. Experimental Results

3.1. Object Recognition

We tested object recognition performance on images of 6 objects that were easily distinguishable from each other: a cup, a toy bear, sports car, a toy rabbit, a plane and a fighter plane (see Figure 1).

The gallery set consisted of 583 clean, black background images, taken at about 20 degrees apart over the viewing sphere. We had 106 images per object (53 images per object hemisphere), except the sports-car that

had only 53 images (we covered only the top hemisphere).

We used two different sets of probe images: one with images taken on a clean, black background, and another one with images taken on a heavily cluttered background (Figure 1). Each set contains 24 images per object hemisphere (24 images total for the sports-car and 48 images for each of the other objects), positioned unevenly in between training views. The black background pictures were taken using the same setup as for the training images. The cluttered background images were taken by placing the objects on a colorful poster and moving them around to make sure that the clutter features did not repeat in the images.



Figure 1. Images of objects on black and cluttered backgrounds

All descriptors, alone or combined, perform perfectly or nearly perfectly on the clean, black background images, with recognition rates of 98.86% for context patches, 99.62% for SIFT, and 100% for the combined descriptors respectively.

The cluttered images proved to be more difficult. Overall, context patches achieves a performance of 71.97% and SIFT of 76.89%. This is a significant improvement over the SIFT performance reported in [9], and it brings the two descriptors to a comparable level. More interestingly, the combined descriptors yield a performance of 85.61%, a considerable improvement over either one of the descriptors alone, yielding a 37.73% reduction of the error rate.

A detailed examination of the error matrices (tables 1, 2, and 3) shows that some objects are recognized more easily using context patches (such as the cup and the plane), while others are easier recognized by SIFT (the toy-bear and the toy-rabbit). It is to no surprise, then, that by combining the two descriptors the performance improves so significantly.

class name	samples	0	1	2	3	4	5
cup	48	26	0	1	0	16	5
toy-bear	48	7	10	0	1	28	2
sports-car	24	0	0	23	0	1	0
toy-rabbit	48	1	0	1	37	8	1
plane	48	0	0	0	0	48	0
fighter	48	0	0	0	0	2	46

Table 1. Error matrix for keyed context patches, cluttered background.

class name	samples	0	1	2	3	4	5
cup	48	13	14	1	8	9	3
toy-bear	48	4	35	0	4	5	0
sports-car	24	0	0	23	1	0	0
toy-rabbit	48	2	4	0	40	2	0
plane	48	0	2	0	0	44	0
fighter	48	0	0	0	0	0	48

Table 2. Error matrix for SIFT, cluttered background.

3.2. Object Categorization

To test performance on object categorization, we used the ETH-80 database [4]. This database is targeted specifically to the task of object categorization, and contains 80 objects from 8 carefully chosen categories from the following superordinate areas: fruits and vegetables (apples, pears, tomatoes), animals (cows, dogs, horses), human-made, small and graspable (cups) and human-made, big (cars).

Each category contains 10 objects with 41 views per object, spaced equally over the upper viewing hemisphere, for a total of 3280 images. Figure 2 shows the objects from the database.

class name	samples	0	1	2	3	4	5
cup	48	29	2	2	3	9	3
toy-bear	48	6	34	1	1	4	2
sports-car	24	0	0	24	0	0	0
toy-rabbit	48	0	0	0	44	2	2
plane	48	0	0	1	0	47	0
fighter	48	0	0	0	0	0	48

Table 3. Error matrix for descriptor combination, cluttered background.

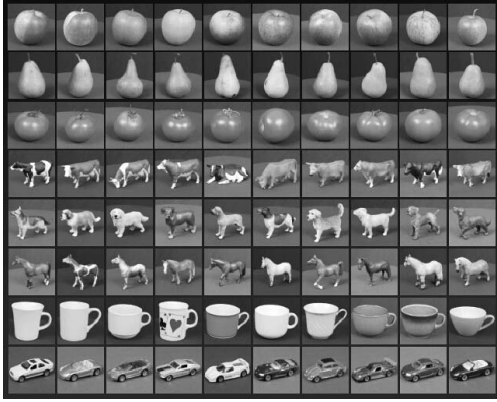


Figure 2. Objects in the ETH-80 database

We employed the test mode intended by the authors of the database, leave-one-out crossvalidation. This meant that we trained with 79 objects and tested with the one unknown object. Recognition was considered successful if the correct category label was assigned.

Table 4 shows the recognition performance for unknown objects using the two descriptors separately, as well as for the descriptor combination. Similarly to [4], results are averaged over the whole database and broken up per category. As can be seen, the overall performance of context patches and SIFT is again comparable, at 57.29% and 56.74% respectively. However, SIFT performs better than context patches on objects from the fruits and vegetables category, while the context patches perform better on human-made objects. There is no clear winner for the animals category. The combined descriptors achieve again a better performance than the individual descriptors alone. At 69.06%, this is a 27.56% reduction in the error rate.

class name	context patches	SIFT	combined
apple	34.15%	53.66%	57.56%
pear	83.66%	82.68%	91.22%
tomato	65.12%	77.32%	83.90%
cow	30.00%	44.63%	61.71%
dog	32.68%	30.00%	44.15%
horse	51.22%	20.98%	41.71%
cup	99.02%	86.59%	97.81%
car	62.44%	58.05%	74.39%
total	57.29%	56.74%	69.06%

Table 4. Recognition performance for categorization of unknown objects.

4. Conclusions

We presented an object recognition/categorization method that uses a feature-level combination of two local image descriptors: SIFT and keyed context patches. The system combines consistent object hypotheses regardless of the descriptor that generated them, yielding stronger matches.

Our experiments confirmed that the performance of the descriptor combination is higher than that of either of the descriptors alone. This is probably due to the fact that while the performance of SIFT and keyed context patches is comparable overall, they perform differently on the same object. Context patches achieved a better recognition rate on human-made objects, while SIFT performed better on fruits and vegetables.

In the future, we plan to add the performance of PCA-SIFT [3] to the comparison, since literature has shown that this descriptor outperforms SIFT in many cases. We are also interested in combining local descriptors with appearance and contour based methods.

References

- [1] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. In *9th International Conference on Computer Vision*, pages 634–640, Nice, France, 2003.
- [2] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation by image exploration. In *8th European Conference on Computer Vision*, pages 40–54, Prague, Czech Republic, 2004.
- [3] Y. Ke and R. Sukhtankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition*, pages 511–517, Washington, D.C., 2004.
- [4] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. In *Computer Vision and Pattern Recognition*, Madison, WI, 2003.
- [5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.
- [6] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. on Pattern Anal. and Machine Intell.*, 27:1615–1630, 2005.
- [7] R. C. Nelson. Finding line segments by stick growing. *IEEE Trans. on Pattern Anal. and Machine Intell.*, 16:519–523, 1994.
- [8] A. Selinger and R. C. Nelson. A perceptual grouping hierarchy for appearance-based 3d object recognition. *Computer Vision and Image Understanding*, 76, 1999.
- [9] A. Selinger-Salgian. Using multiple patches for 3d object recognition. In *2nd Beyond Patches Workshop, in conjunction with CVPR*, Minneapolis, MN, 2007.