Object Recognition Under Difficult Conditions Based on Superpixel

Martin Klinkigt Koichi Kise
Graduate School of Engineering, Osaka Prefecture University
klinkigt@m.cs.osakafu-u.ac.jp kise@cs.osakafu-u.ac.jp

1 Introduction

In computer vision the task of object recognition is to recognize a certain object in an provided image. For this task a description about the object of interest is learned from images. This process often involves the use of local features like SIFT [1] which can be extracted reliably from images, even if the resolution or the lighting conditions change. The drawback of such local features are that they may lose discriminative power to distinguish between similar objects or the object from the background.

We proposed a system that utilizes superpixel [2]. The system is not only working on local features rather it packs the features belonging to one such superpixel and reject the whole area, if it is ambiguous. A superpixel is ambiguous if the system can not name an object with a high confidence to which object this superpixel could belong. By doing so we have achieved an improvement of over 10% on a difficult dataset.

2 Related Work

In recent years, researches put more interest on superpixel or sometimes also called image patches. One of the first approaches was proposed by Ren et al. in [2] and more recent Plath et al. [3]. While Ren et al. analysed the calculation of such superpixel, Plath et al. mainly focus on the segmentation of the object from the background. Our purpose is to utilize such superpixel directly to improve recognition performance.

3 Voting Schema in Object Recognition

Object recognition on the base of local features is commonly used in computer vision. One of the major difficulties with local features is how to store them and the search of similar features. Kise et al. [4] achieved a reduction of the memory by using PCA-SIFT [5] and only store the occurrence of an feature and not its concrete values. A search is done with the help of hashing which applies Kise et al. to work even in real-time.

The process to recognize an object can be summarized as follows. First from the training images local features were extracted and stored in a database. During recognition the same type of features were extracted from the query image and for each feature the nearest neighbor is search in the database. The object from which this feature is extracted during training receives one vote. This is done for all features from the query image. The object which gathers the most votes is considered to be shown in the image.

4 Superpixel

Normally images are processed on the basis of pixel. Superpixel are larger connected regions of pixel and sometimes also called image patches. There are various approaches proposed to calculate such superpixel, e.g., Felzenszwalb et al. [6] or Mori [7]. For our implementation we have chosen Felzenszwalb et al. approach, since it is faster to compute. Figure 1(b) shows such a resulting segmentation. Individual patches are indicated by different colors.

5 Superpixel for SIFT features

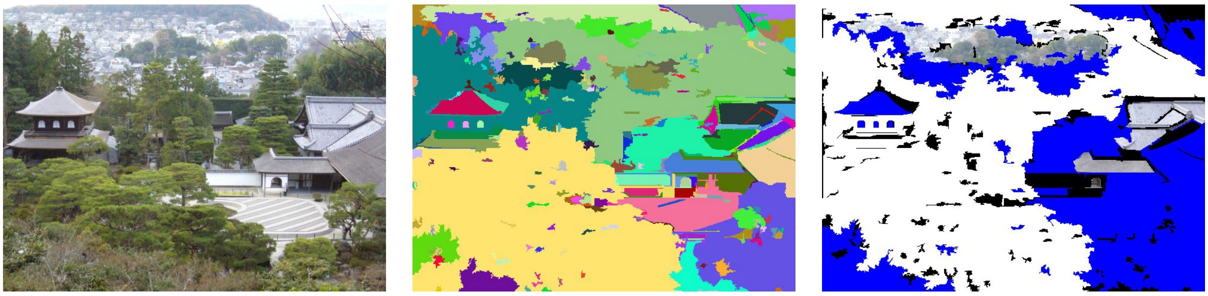
After the necessary elements are explained we give the details of our proposed method. We keep the training phase simple by just extracting the features from the images and storing them into the database. During recognition we apply first the segmentation algorithm described in Section 4. As parameters we choose the following setting: $\sigma = 0.3$, $k = 600$ and as minimum size of the areas 50 pixel. With this setting we achieve on one hand a reasonable segmentation into patches and on the other hand useful patch sizes. If patches are too large, we end up with the problem that patches may contain information about the object and the background at the same time. If patches are too small, then they may not contain enough information (features) to analyse the object they are showing.

In the next step we accumulate the features laying in each individual region to one set s. For these features the system has to decide to which object they belong. Here we apply a simple threshold based on the number n of objects which have some similarity with features from the patch and n_{top} be the number of objects with high confidence. The patch will be used only if $n_{\text{top}} = 1$ or $n_{\text{top}}/n < 3/4$.

6 Experiments

Our evaluation was performed on a dataset consisting of shrines and temples. This dataset is difficult in various ways. First, the objects the system has to distinguish looking quite similar, since the are traditional Japanese buildings and second, a high amount of background clutter is involved as we can see from Fig. 1(a). These 107 images were used as query images, while the database itself was trained with the images provided by Wikipedia [8]. More concrete we trained 84 objects with 819 images. For further analysis we downloaded distractor images from Flickr [9]. These images are used to increase the database with the purpose to provoke false matching of features.

Figure 1(c) indicates a result of our system. For regions in blue it could not even find matching features in



(a) A query image

(b) Image segmented into patches indicated with different colors.

(c) Areas in one color are not used for recognition. Blue areas contain no matching features, white regions are ambiguous.

Figure 1: Example segmentation of a query image.

Table 1: Results for the temple dataset. Shown is the mean average precision (mAP) comparing the simple voting approach (SIFT) and second the patch discarding with the help of superpixel (SP). All values are in percentage.

number of	Ginkaku-ji		Kinkaku-ji		Kiyomizu-dera	
dist. img.	SIFT	SP	SIFT	SP	SIFT	SP
0	22.75	26.49	45.56	42.95	19.05	22.22
2500	12.92	19.33	30.63	27.36	13.08	14.29
5000	10.91	18.53	25.52	24.56	11.87	14.05
7500	9.59	21.11	21.85	20.43	10.98	13.65

the database. So features from these regions are not considered at all. White color mark discarded regions based on the above explained threshold. We can see that these are mainly regions showing trees. Also some parts of the object are discarded which have become a part of the background. Here the segmentation algorithms fails to separate between the object and the background. However, for this example the simple voting schema fails to return the correct result, since we have too many incorrect matching features from white areas, while our proposed method return the correct result. We can also see that the object is mainly detected by its side building at the right of the image.

From the results in Table 1 we notice that by working in superpixel (column named “SP“) the recognition performance is mainly improved. Only for one object (Kinkaku-ji) the simple voting (column named “‘SIFT‘) performs slightly better. This is may be due to the unique characteristic of this object. Most interesting are the results of Ginkaku-ji. With an increasing database the difference in the performance becomes more significant. This can be explained again with the characteristic of the object. It has only less unique properties and, therefore, is hard to distinguish from others. When the database increases, the probability of false matching also increases. Our proposed method can address this problem.

7 Conclusion

We have proposed a system for object recognition working on superpixel rather than in pixel level. Via a threshold we can detect ambiguous regions and ignore them in further calculations. For a difficult dataset we improved the performance significantly and perform even better on artificially increased databases.

Further research will focus on a better pre-segmentation of the images and an analysis of characteristics of ambiguous regions.

Acknowledgment

This research was supported in part by the Grant-in-Aid for Scientific Research (B)(22300062) from Japan Society for the Promotion of Science (JSPS) and A-Step(AS2111193A) from Japan Science and Technology Agency (JST).

References

[1] D.G. Lowe, “Object recognition from local scale-invariant features,” Proc. of ICCV, p.1150, 1999.

[2] X. Ren and J. Malik, “Learning a classification model for segmentation,” Proc. of ICCV, vol.1, pp.10–17, 2003.

[3] N. Plath, M. Toussaint, and S. Nakajima, “Multi-class image segmentation using conditional random fields and global classification,” Proc. of ICML, pp.817–824, New York, NY, USA, ACM, 2009.

[4] K. Kise, K. Noguchi, and M. Iwamura, “Robust and efficient recognition of low-quality images by cascaded recognizers with massive local features,” Proc. of WS-LAVD2009, pp.2125–2132, Oct. 2009.

[5] Y.K. Rahul, Y. Ke, and R. Sukthankar, “Pca-sift: A more distinctive representation for local image descriptors,” Proc. of IEEE CVPR, pp.506–513, 2004.

[6] P.F. Felzenszwalb and D.P. Huttenlocher, “Efficient graph-based image segmentation,” Int. J. Comput. Vision, vol.59, no.2, pp.167–181, 2004.

[7] G. Mori, “Guiding model search using segmentation,” Proc. of ICCV, vol.2, pp.1417–1423, 2005.

[8] “Wikipedia,” `http://www.wikipedia.org`, 2010.

[9] “Flickr,” `http://www.flickr.com`, 2010.