LETTER

How the Number of Interest Points Affect Scene Classification

Wenjie XIE[†a)], ***Student Member***, **De XU**[†], ***Member***, **Shuoyan LIU**[†],
and **Yingjun TANG**[†], ***Student Members***

SUMMARY This paper focuses on the relationship between the number of interest points and the accuracy rate in scene classification. Here, we accept the common belief that more interest points can generate higher accuracy. But, few effort have been done in this field. In order to validate this viewpoint, in our paper, extensive experiments based on bag of words method are implemented. In particular, three different SIFT descriptors and five feature selection methods are adopted to change the number of interest points. As innovation point, we propose a novel dense SIFT descriptor named Octave Dense SIFT, which can generate more interest points and higher accuracy, and a new feature selection method called number mutual information (NMI), which has better robustness than other feature selection methods. Experimental results show that the number of interest points can aggressively affect classification accuracy.

key words: *bag-of-words, feature selection, SIFT*

1. Introduction

Scene classification is an important aspect for computer vision, and has received considerable attention in recent years. As a scene composed of several entities is often organized in an unpredictable layout, scene classification is much more difficult than conventional object classification, and is still a challenging question.

In scene classification, how to represent scenes is a critical component. Early work on scene classification used low-level global features extracted from the whole image to classify images into a small number of categories [1]. Recently, a method called bag-of-words representation has been widely uesd [2]. This method uses interest points to model scenes as a collection of points labeled by a codebook which is constructed by quantizing these interest points using local invariant features. Recent works have shown that local features represented by bag-of-words model are suitable for scene classification and show impressive levels of performance [2].

What's the relationship between the number of interest points and accuracy in bag of words model? It is commonly believed that more interest points can generate higher accuracy in scene classification. In order to validate the correctness of the viewpoint, extensive experiments are designed in this paper to change the number of interest points.

On one hand, how to extract interest points containing distinctive invariant features from images is the most critical component. In this paper, the Scale Invariant Feature Transform (SIFT) is adopted as the prototype descriptor, which is first proposed by David G. Lowe [3]. In Paper [4], Mikolajczyk et al. proved that the robustness and the distinctive character of the SIFT descriptor can generate better performance than other related descriptors. In this paper, we propose a new dense SIFT descriptor called octave dense SIFT which can extract more interest points and generate higher accuracy. Experiments with three different SIFT descriptors, which can generate diverse number of interest points, are implemented and prove that descriptor with more points can bring higher accuracy.

On the other hand, feature selection methods are adopted to reduce the number of interest points and prove the common belief in another aspect. In our paper, a new feature selection method called number mutual information (NMI) is proposed, which is an improved method based on mutual information. In feature selection stage, beside NMI, another four feature selection methods are used to verify the high robustness of NMI and analyze the relationship between aggressive reduction of interest points and the accuracy rate in scene classification.

2. Our Approach

In this section, we will introduce the proposed dense SIFT descriptor (octave dense SIFT) and feature selection method NMI respectively.

2.1 Octave Dense SIFT

Currently, many different techniques for describing local image regions have been developed and in [4], Mikolajczyk et al. have proved that SIFT was the most suitable descriptor for scene classification. As to SIFT descriptor [3], first, the initial image is incrementally convolved with Gaussians function to produce blurred images, which compose an octave. Then, the resolution of all blurred images of octave is changed gradually by taking every second pixel in each row and column to form new octaves, and extreme points are detected using different-of-Gaussian function within all octaves of different resolutions. Last, each extreme point is described to a vector of 128 dimensions. As a result, an image can be denoted as a matrix with the size of $n \times 128$, where n is the number of interest points. SIFT descriptor

Manuscript received September 24, 2009.
Manuscript revised December 4, 2009.

[†]The authors are with Institute of Computer Science and Engineering, Beijing Jiaotong University, Beijing China.
a) E-mail: xiewenjiebj@126.com

DOI: 10.1587/transinf.E93.D.930

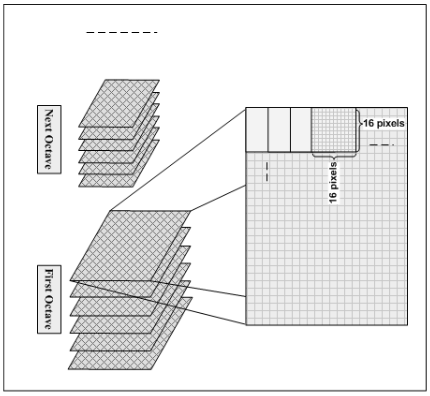

Fig. 1 Octave dense SIFT.

preferably resolve some transformations, including rotation, scaling, affine stretch, change in brightness and contrast, addition of image noise and generate better performance than other descriptors. Therefore, in our paper, we choose SIFT descriptor as the technique for extracting distinctive invariant features.

However, the original SIFT descriptor only describes the extreme points and the remainder points are disregarded, which is suitable for image matching and object recognition. But in scene classification, one scene may comprise lots of objects, and interest points from arbitrary object may contain important information for classification. So using the original SIFT will lose some of features, which results in low accuracy of scene classification. Based on analysis above, in our paper, we propose a novel SIFT called octave dense SIFT. With octave dense SIFT, the resolution of initial image is changed by taking every second pixel in each row and column to form a set of images with different resolution. In each image of the set, instead of finding extreme points, descriptor is computed on a regular grid with the size of 16 × 16 pixels. The grid begins at the left-up corner of image, and shifts 8 pixels every time to right or bottom respectively. The interest points are evenly distributed in the image with an interval of 8 pixels, as Fig. 1 shows. Using the octave dense SIFT descriptor, more interest points are extracted to provide more important information for scene classification.

2.2 Feature Selection

Feature selection is a process that chooses a subset from the original feature set according to some criterions. The selected feature retains original physical meaning and provides a better understanding for the data and learning process. In this subsection, five feature selection methods are used to analyze the relationship between aggressive reduction of visual words which are the cluster center of interest points and the accuracy in scene classification. Several researchers has performed related work on scene classification [6], [7], but, our purpose is to validate the relationship between the number of interest points and the accuracy rate, rather than the feature selection methods themselves. In our paper, we first introduce the proposed feature selection method called number mutual information (NMI), which an improved method based on mutual information (MI). Then, extensive experiments using five feature selection methods are implemented in scene classification research domain.

2.2.1 Number Mutual Information (NMI)

We know that Mutual Information (MI) is a method which can measure the dependence between two random variables. The mutual information method between a visual word t and a category c is defined as:

$$MI(t,c) = \sum_{i=1}^{m} P(c_i) \log \frac{P(t,c_i)}{P(t)P(c_i)} \tag{1}$$

where m is the number of category.

In our paper, we propose a new feature selection method called Number Mutual Information (NMI). Comparing to MI, NMI considers not only the probability that a visual word t and a category c co-occur, but also the number of the visual word t when they co-occur. So, NMI can measure the relationship between the visual word t and the category c quantitatively. NMI method between a visual word t and a category c can be defined as:

$$NMI(t,c) = \sum_{i=1}^{m} P(c_i) \log \frac{Sum(t)}{n_i P(t)P(c_i)} \tag{2}$$

where $Sum(t)$ means the total number of t when a visual word t and a category c co-occur, m is the number of image category. n_i is the number of images in category c_i.

Experimental results testify to the close link between the MI or NMI value of visual words and their effect for classification; the visual words with higher MI or NMI value do more contribution for scene classification.

2.2.2 Other Feature Selection Methods

Beside MI and NMI methods, another three feature selection methods including document frequency (DF), term frequency (TF), X^2 test (CHI), are adopted in our experiments.

Document Frequency (DF). It is the number of images in which a visual word appears. In our experiment, words with low DF are removed, as the basic assumption is that rare words are either non-informative for category prediction or not influential in global performance. In [6], Jun Yang also proved that feature selection with frequent visual words outperformed of that with rare ones.

Term Frequency (TF). It is the total number of a visual word in all images. Like DF method, words with higher frequent would do more contribution to classification. So, we choose the visual words with high TF value in feature selection.

X^2 test (CHI), which can also measure the level of dependence between two random variables. A large value of $X^2(t,c)$ indicates a strong correlation between word t and category c, and vice versa. CHI method between a visual word t and a category c can be defined as:

$$\mathrm{CHI}(t) = \sum_{i}^{m} P(c_i) \frac{N(P(t, c_i)P(\bar{t}, \bar{c}_i) - P(t, \bar{c}_i)P(\bar{t}, c_i))^2}{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)} \tag{3}$$

where N means the number of all images, m is the number of image category.

3. Experimental Setup and Result

In this paper, we use Lazebnik dataset, which contains 15 category of natural scenes: bedroom, suburb, industrial, kitchen, living room, coast, forest, highway, inside city, mountain, open country, street, tall building, office and store. Each scene category is divided randomly into two separate set of images: 100 images for training and remaining images for testing. Experiment is run with Matlab 7.0 by using computer with Pentium 4 3.0 GHz processor. In [5], Svetlana Lazebnik et al. have proved that classification rate of accuracy based on bag-of-words representation is much higher than PLSA and LDA models in database containing a variety of categories. So in our paper, experiments are implemented based on bag-of-words model. SIFT serves as the prototype descriptor. In bag of words model, the size of codebook is set to 300 and the same number of visual words are generated.

First, in order to validate common belief that more interest points can generate higher accuracy and the effectiveness of our proposed octave dense SIFT, experiments compared with original SIFT, normal dense SIFT and octave dense SIFT are implemented. As to normal dense SIFT, it is computed like octave dense SIFT, but it only works on the images of original resolution. The classification accuracy is shown in Table 1.

Using the octave dense SIFT, images in different resolution are considered and more interest points can be extracted. Using parse SIFT, about 300 interest points can be extracted from each image, and normal dense SIFT is around 900, octave dense SIFT can reach about 1700 interest points. From the Table 1 we can see that more points generate higher accuracy and the highest accuracy rate is obtained by octave dense SIFT. The classification result using octave dense SIFT is shown as Fig. 2.

As we know, not only the number of interest points, but also how they are extracted affects the performance. So, extended experiment is implemented to compare octave dense SIFT with different feature extraction methods, including random sampling, grid sampling and spatial pyramid. The result is shown in Table 2.

For all feature extraction methods, we chose 128-dim SIFT as descriptor.

Random Sampling (RS), 300 randomly sampled patches are extracted from each image. The size of the patch is set to 16 × 16 pixels to suit SIFT descriptor. The number of interest points in Random Sampling approximates that of Parse SIFT. However, in Parse SIFT, interest points are extracted based on gray value changing and usually locate on the salience position and carry more semantic information. So, Parse SIFT is out performance of Random Sampling.

Table 1 Performance comparison between different SIFT.

	Parse SIFT	Dense SIFT	Octaves Dense SIFT
accuracy	46.18%	57.69%	61.71%

Grid Sampling (GS), the image is evenly segmented to patches with the size of 16 × 16 pixels and about 400 interest points are extracted in each image. The points extracted by Grid Sampling are evenly distributed in image. Actually, Grid Sampling is a simplification procedure of Octave Dense SIFT where patches are extracted in different resolution images and the occlusion of patches is allowed. So, comparing to Grid Sampling, Octave Dense SIFT exacts more semantic information and generates higher accuracy.

Spatial Pyramid (SP), each image is respectively segmented to 1 × 1, 2 × 2, 3 × 3 patches, and histograms based on different segmentation are concatenated to form a high dimension vector. Although Spatial Pyramid method generates the highest accuracy with Lazebnik 15 dataset, Spatial Pyramid method makes use of absolute spatial information and is lack of the robustness with respect to partial occlusion, clutters, and changes in viewpoint and illumination. Beside, large size of codebook or excessive segmentation may lead to "curse of dimensionality".

Second, in feature selection stage, as Table 1 shows that octave dense SIFT descriptor get better performance than other SIFT descriptors, in this part experiment, we adopt the octave dense SIFT descriptor to verify the performance of the five feature selection methods. Every visual word of codebook will get a value based on each feature selection method. We remove 10% visual words (that is, 30) according to the value sorted by ascending, and new codebook (containing 270 visual words) is reused to compute the accuracy rate. The process above is implemented nine times (up

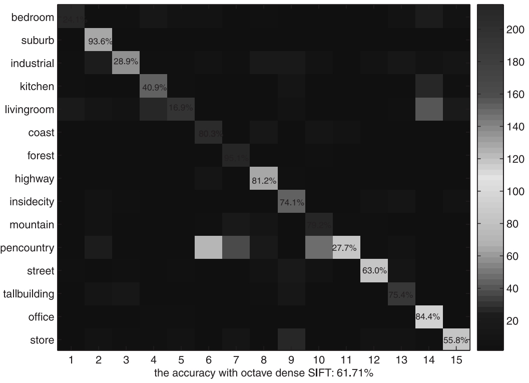

Fig. 2 Classification accuracy with octave dense SIFT.

Table 2 Performance comparison between different feature extraction methods.

RS	GS	SP	Octaves Dense SIFT
41.7%	52.12%	81.4%	61.71%

Table 3 Accuracy of experiments with different feature selection methods.

	10%	20%	30%	40%	50%	60%	70%	80%	90%
tf	61.41	60.74	**59.90**	**59.36**	**57.62**	**56.95**	**54.27**	**50.79**	41.64
df	61.11	59.97	58.83	57.45	56.55	54.30	51.62	47.67	43.99
mi	60.87	59.70	59.40	56.85	56.45	54.70	51.62	47.67	43.99
chi	**61.64**	**61.44**	**60.70**	59.16	57.45	55.34	51.62	47.67	43.99
nmi	**61.51**	**61.01**	59.80	**59.36**	**57.96**	**56.92**	**54.61**	**50.32**	**44.25**

to 90% visual words are removed) to visual words. The results of accuracy rate are shown in Table 3. Visual words are the cluster center of interest points, and removing a certain visual word is equivalent to eliminating this cluster of interest points' contribution to classification. So, the accuracy drops down along with the number of visual words descends with all feature selection methods, which proves the common belief in another aspect. As Table 3 shows, no matter how the percentage of reduction of interest points change, the proposed NMI can get the highest accuracy or the accuracy only second to the highest one. But other feature selection methods' performances vary along with the changing of percentage of reduction of interest points: CHI is suit to low percent reduction of interest points and DF is fit for high situation. So, although using NMI only improves the accuracy slightly, it has more robustness than other feature selection methods.

The result provided in Table 3 is different from that of object recognition, where the rate of accuracy raises by feature selection. We infer that images in object recognition usually contain only one object, using feature selection methods can effectively remove redundancy points that may be regarded as noise, and get better performance. But in scene classification, each image contains lots of object, and interest points extracted from arbitrary object may do contribution to classification. So, feature selection removing lots of interest points will bring down the accuracy in scene classification.

4. Conclusion

In this paper, two approaches are used to analyze the relationship between the number of interest points and the accuracy rate in scene classification. First, three SIFT descriptors are adopted in bag of word model. Experimental results show that our proposed SIFT descriptor which extracts more interest points can generate better performance and prove the common belief. Second, five feature selection methods are used to reduce the number of interest points. Results show that the accuracy drops down gradually along with the diminishment of visual words and prove the common belief in another aspect. Furthermore, a new feature selection method called NMI is put forward, which gets the highest accuracy or the accuracy only second to the highest one and shows more robustness than other feature selection methods. Last, from all the experimental result, we can come to the conclusion that more interest points can generate higher accuracy.

Acknowledgement

This work is supported by National Nature Science Foundation of China 60803072 and 90820013.

References

[1] A. Vailaya, A. Jain, and H. Zhang, "On image classification: City vs landscapes," Pattern Recognit., vol.31, no.12, pp.1921–1935, 1998.

[2] A. Bosch, X. Munoz, and R. Mart, "Which is the best way to organize/classify images by content?," Image Vis. Comput., vol.25, no.6, pp.778–791, 2007.

[3] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.

[4] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.10, pp.1615–1630, 2005.

[5] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.2, pp.2169–2178, New York, June 2006.

[6] J. Yang, Y.-G. Jiang, A.G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," Proc. International Workshop on Multimedia Information Retrieval, Augsburg, Bavaria, Germany, Sept. 2007.

[7] Y.H. Yang, P.T. Wu, C.W. Lee, K.H. Lin, W.H. Hsu, and H.H. Chen, "ContextSeer: Context search and recommendation at query time for shared consumer photos," Proc. ACM Multimedia, 2008.