# A SEMANTIC REGION DESCRIPTOR FOR LOCAL FEATURE BASED IMAGE CATEGORIZATION

*Teng Li, In-So Kweon*

Korea Advanced Institute of Science and Technology
373-1 Guseong-dong Yuseong-gu Daejeon, Korea
Email: tengli@rcv.kaist.ac.kr, iskweon@kaist.ac.kr

## ABSTRACT

Region descriptor has proved to be very important for local feature based image categorization. Previous region descriptors are usually based on the statistics of low level features, such as intensity, edge response, and etc. In this paper a novel descriptor named Local Texton Statistics (LTS) that explores the high level semantic statistical characteristics of image regions is presented. Perceptual information is obtained by applying Gaussian filter banks and the image regions are described by the statistics of different 'texton's. Using the Bag of Words as classification algorithm, experiments show that the proposed descriptor is superior to the previous popular SIFT descriptors on the Wang dataset. The combination of these two descriptors shows high performance for categorization on both the Wang dataset and the fifteen Scene categories dataset.

***Index Terms***—Image classification, Image region analysis, Image texture analysis

## 1. INTRODUCTION

Image categorization is one of the most important and active topics in computer vision and image processing, with many applicable areas such as image retrieval, and etc. Recently, local feature based method has become popular in image categorization due to its flexibility, simplicity, and good performance [1, 2]. Usually local regions are detected or extracted from the images and converted to descriptors. Image categorization is then taken based on the local feature descriptors. Recent researches show that interest regions detector is not crucial for categorization and instead they just randomly or evenly extract local patches from the image [3, 4], while local region descriptor can influence the performance significantly, using different descriptors means extracting different information for representing images.

A number of algorithms for describing image regions have been reported in the literatures. The most straightforward way is using the color or intensity value in the region directly. Usually Principal Component Analysis (PCA) is adopted to reduce the dimension and improve the efficiency [5]. This descriptor was demonstrated for object categorization and showed good performance.

Many complex descriptors have also been introduced. The differential descriptor of [6] calculates a set of image derivatives up to a given order for image matching and retrieval. The complex filters of [7] are orthogonal, and the Euclidean distance between complex filters provides a lower bound on the Squared Sum Differences (SSD) between corresponding image patches. Van Gool [8] introduces the Generalized Color Moments describing the shape and the intensities of different color channels in a local region. Lowe [9] proposes Scale Invariant Feature Transform (SIFT) descriptors, which is computed by sampling the magnitudes and orientations of local image gradients and building smoothed orientation histograms. This description provides robustness against localization errors and small geometric distortions. PCA-SIFT [10], CSIFT [11] and Gradient Location and Orientation Histogram (GLOH) [12] are proposed to extend SIFT by applying PCA, incorporating color information and applying PCA after changing the location grid.

There are also other various descriptors, such as spin images [13], and etc. In [12], an experimental evaluation of several different descriptors for matching was reported, where the SIFT descriptor shows to be the best.

However, previous descriptors are only computed based on low level features. For image patches, human vision system can get some perceptual information or in other words, semantic meaning. In this paper we propose a novel descriptor that explores the high level characteristics of local patches named as Local Texton Statistics (LTS) to improve the discriminative power for image categorization. Wu, et al. [14] also utilizes the perceptual information for the texture descriptor, but they use it for texture classification and our algorithm is different from theirs in computing process.

The paper is organized as follows. In section 2, we briefly introduce our local feature based categorization framework. In section 3, the new LTS descriptor is presented and section 4 gives our experimental results for image categorization task and the comparison with previous works. We conclude this paper in section 5.

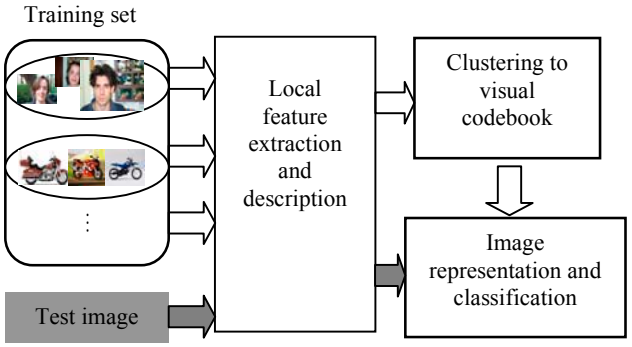


**Figure 1.** The framework of the BOW image categorization

## 2. LOCAL FEATUER BASED IMAGE CATEGORIZATION

The key issue of local feature based image categorization is how to represent the images given the extracted local features, on which lots of methods have been proposed. Among them, recent Bag of Words (BOW) algorithm [15] shows to be very effective for image categorization. We use the BOW as our image categorization algorithm.

Figure 1 shows the overall process of the BOW categorization. First, we extract local features from the labeled training set. A visual codebook is then learned by clustering training feature descriptors. By assigning the local features to the codebook with a vector quantization algorithm, and counting the number of features assigned to each code, we can get a distribution (histogram) for each image, which records how many times the features corresponding to a code occur in the image. We then apply the classification algorithm directly. Here we use the Support Vector Machine (SVM), which has shown state of the art performance in many classification problems, and Chi-Square is chosen as the kernel type.

For local feature detection, recent research shows that extracting local patches by evenly sampling can yield good performance [4]. We extract the regions by evenly scanning the images and calculate the descriptor of each region. SIFT descriptor is the most popular in previous works. We use SIFT, the proposed LTS descriptor which will be detailed in the next section, and their combination in our experiments. SIFT is computed for 8 orientation planes and each gradient region is sampled over a 4x4 grid of locations. The dimension of resulting descriptor we use is 128.

## 3. A SEMANTIC REGION DESCRIPTOR

The Marr's theory supports that in the early stages of the vision process, there are cells that respond to stimulus of primitive shapes, such as corners, edges, bars, etc [16]. These cells are modeled by Gaussian derivative functions and Yokono and Poggio have shown the excellent performance achieved by features created with filters based on Gaussian functions, applied to the problem of object recognition [17].

Inspired by this, we capture the perceptual information by applying a number of Gaussian filter banks to the images, and then we generate a set of 'texton's by clustering the response value of the pixels. The region is described using statistics (histogram) of different textons it contains.

Texton was originally used for texture classification, where the textons are considered as the local feature directly. Differently, we use the statistics of the textons to describe local regions, thus the local descriptors can contain high level information.

### 3.1. Filter banks for modeling perceptual functions

We use a number of different filter-banks made of isotropic Gaussians, first and second order derivatives of Gaussians. The functions used are defined by the following equations:

a) Isotropic Gaussian

$$G^0(x,y) = \frac{1}{2\pi\sigma^2}\exp(-\frac{x^2+y^2}{2\sigma^2}) \quad (1)$$

b) First order Gaussian derivative

$$G^1(x,y) = -\frac{y}{2\pi\sigma_x\sigma_y^3}\exp(-\frac{x^2}{2\sigma_x^2}-\frac{y^2}{2\sigma_y^2}) \quad (2)$$

c) Laplacian of Gaussian

$$LoG(x,y) = -\frac{(x^2+y^2-2\sigma^2)}{2\pi\sigma^6}\cdot\exp(-\frac{x^2+y^2}{2\sigma^2}) \quad (3)$$

Our filter-banks are made of 3 isotropic Gaussians (with $\sigma$=1, 2, 4), 4 Laplacian of Gaussians (LoG) (with $\sigma$=1, 2, 4, 8) and 4 first order derivatives of Gaussians (with $\sigma$ =2, 4 and into x, y directions). Therefore, for grey value images, each pixel is associated with an 11-dimensional vector. For color images, the 3 isotropic Gaussians are applied to each CIE L, a, b channel and thus a 17-dimensional vector is generated for each pixel.

### 3.2. Region descriptor computing process

The computing process of the proposed region descriptor is described in the following steps.

1. For each of the training images, apply the filter banks and for each pixel concatenate the response of filter banks to form a 17 or 11 dimensional vector.
2. The K-means clustering algorithm is applied to the training vectors to generate a set of textons. The number of textons K is set as 200. The K-means algorithm finds a local minimum of the following sum-of-square distance error:

$$Err = \sum_{i=1}^{N}\sum_{k=1}^{K} q_{ik}\left\|x_i - c_k\right\|^2 \quad (4)$$

Where


**Figure 2.** Example images of 10 categories from the Wang dataset.

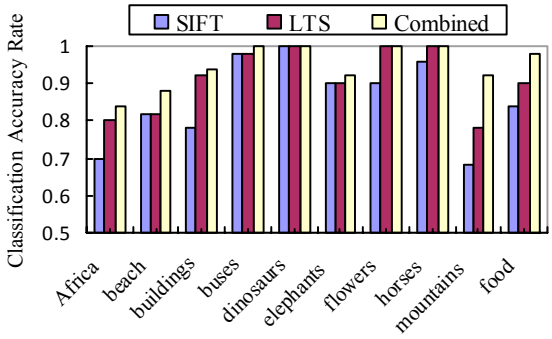


**Figure 3.** Classification performance comparison of LTS descriptor, SIFT descriptor and their combination on the Wang database.

$$q_{ik} = 1 \quad \text{if } \|x_i - c_k\|^2 < \|x_i - c_j\|^2$$
$$\forall j = 1,..., K \text{ and } j \neq k$$
$$q_{ik} = 0 \quad \text{otherwise}$$

$N$ denotes the number of training pixels; $x_i$ is the concatenated filter response vector of $i_{th}$ pixel and $c_k$ is the appearance vector for the $k_{th}$ texton. The K-means is initialized by random samples from all the data vectors.

3. Given a region to describe, associate each pixel, whose filter response vector is $x_t$, with a texton $c_k$ by:

$$\|x_t - c_k\|^2 < \|x_i - c_j\|^2 \qquad \forall j = 1,..., K \text{ and } j \neq k$$

4. Count the occurrence number of the textons in the region to form a K-dimensional vector, as the descriptor.

## 4. EXPERIMENTS

In this section, we evaluate the proposed LTS descriptor on two scene image databases: the Wang database [18] and the fifteen scene categories database [2]. We compare with SIFT descriptor in the BOW categorization framework. The measure of performance is the percentage of the images assigned to their correct classes. For SIFT descriptor, we extracted 18×18 pixel patches over a grid with spacing of 9 pixels in images as the local features; for LTS descriptor, 12 ×12 pixel local patches over a grid with spacing of 6 pixels are extracted. The codebook size of the BOW algorithm for each single descriptor is set to 500.

Combination of the two descriptors is done on the level of image representation. For an image, the histogram of each descriptor is calculated individually and then concatenated together as the combined representation. The dimension problem is handled through SVM learning.

### 4.1. Wang Database

This database contains 10 natural image categories. Each category contains 100 images, which makes a total of 1,000 images. Figure 2 shows the example images and the name of categories they belong to. We randomly divided each category set into a training set and a test set, each with 50 images.

Results are presented on Figure 3 as the classification accuracy of each category using SIFT, LTS and the combined descriptors. We can observe that the proposed LTS descriptor shows to be better than SIFT descriptor, with average rates of 91% and 85.6%, respectively. The combination of these two descriptors yields the best accuracy rate as 94.8%. To the best of our knowledge, the highest accuracy which had been previously reported on this database was 92.8% [1] using both the color and SIFT descriptors in a complex classification framework. However, their reported result using the BOW algorithm is 87%.

### 4.2. Fifteen Scene Categories Database

This database is composed of fifteen scene categories: store, office, tall building, street, open country, mountain, inside city, highway, forest, coast, living room, kitchen, industrial, suburb and bedroom. Each category has 200 to 400 images, and average image size is 300 × 250 pixels. The major sources of the pictures in the dataset include the COREL collection, personal photographs, and Google image search. This is one of the most complete scene category dataset used in the literature thus far. It is public available on internet [2], so we do not show the example images due to the paper length limitation. We randomly chose 100 images per class for training and the rest for test.

**Table 1.** Classification results on the scene categories.

| *References* | *Feature Descriptor* | *Categorization algorithm* | *Result* |
|---|---|---|---|
| Fei-Fei [3] | SIFT | pLSA | 65.2% |
| Lazbinik[2] | SIFT | Spatial Pyramid Matching | 81.4% |
| Lazbinik[2] | SIFT | BOW | **74.8%** |
| This paper | SIFT | BOW | **71.5%** |
| This paper | LTS | BOW | **67.2%** |
| This paper | Combined | BOW | **79.2%** |

The last three rows of table 1 show the classification results on this dataset using SIFT descriptor, LTS descriptor and their combination, respectively. On this dataset SIFT descriptor shows to be better than the proposed LTS descriptor. However, these two descriptors can be complementary to each other, and the combination of the two descriptors shows the best performance when use the BOW categorization algorithm, achieving the accuracy rate of 79.2%. Some previous reported results on this dataset are also given in table 1 for comparison. The best result reported to our knowledge is 81.4% of S. Lazebnik et al. [2] using the SIFT descriptor and their Spatial Pyramid Matching categorization algorithm, while they obtained 74.8% using the baseline BOW. Fei-Fei and Perona [3] obtained the accuracy rate of 65.2% for 13 scene classes from the 15 using the SIFT descriptor and their classification method. Thereby we can see the effectiveness of the proposed descriptor.

## 5. CONCLUSIONS

This paper has presented a novel local region descriptor, named LTS. Different from existing methods, the proposed approach characterizes local regions by using high level semantic information. The proposed descriptor is discriminative for different visual categories. Evaluation results for the local feature based image categorization task proved its effectiveness. It shows superior or comparable performance to SIFT descriptor on two images datasets. High performance is achieved when we combine LTS with SIFT descriptor.

## 6. ACKNOWLEDGEMENTS

Thanks Jihwan Woo's implementation of the filter banks. This research has been supported by the Korean MOST for NRL Program (Grant numberM1-0302-00-0064), by the MIC for the project, "Development of Cooperative Network-based Humanoids' Technology" of Korea, and by Agency for Defense Development.

## 7. REFERENCES

[1] F. Peronnin, C. Dance, G. Csurka and M. Bressan, "Adapted Vocabularies for Generic Visual Categorization", In *Proc. of ECCV*, 2005.

[2] S. Lazbinik, C. Schmid, J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories", In *Proc. of CVPR*, 2006.

[3] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. of CVPR*, 2005.

[4] E. Nowak, F. Jurie, and B. Triggs, "Sampling Strategies for Bag-of-Features Image Classification", In *ECCV,* 2006.

[5] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning", In *Proc. of CVPR*, 2003.

[6] J. Koenderink and A. van Doorn, "Representation of local geometry in the visual system", *Biological Cybernetics*, 1987.

[7] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets", In *Proc. of ECCV*, 2002.

[8] L. V. gool, T. Moons and D. Ungureanu, "Affine photometric invariants for planar intensity patterns", In *ECCV*, 1996.

[9] D. Lowe, "Distinctive image features from scale invariant keypoints", In *IJCV* 60(2):91-110, 2004.

[10] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for a local image descriptors", *CVPR,* 2004.

[11] Alaa E. Abdel-Hakim and Aly A. Farag, "CSIFT: A SIFT Descriptor with Color Invariant Characteristics", *CVPR*, 2006.

[12] K. Mikolajczyk, C. Schmid, "A performance evaluation of local descriptors". In *PAMI* 27(10):1615-1630.

[13] S. Lazebnik, C. Schmid, and J. Ponce, "Sparse texture representation using affine-invariant neighborhoods", In *Proc. of CVPR*, 2003.

[14] P. Wu, B. S. Manjunanth, S. D. Newsam, and H. D. Shin, "A Texture Descriptor for Image Retrieval and Browsing", In *Proc. of CVPR*, 1999.

[15] G. Csurka, C. Dance, L. Fan, J. Williamowski, and C. Bray "Visual Categorization with Bags of Keypoints", In *Proc. of SLCV Workshop, ECCV*, 2004.

[16] David Marr, *Vision*. W. H. Freeman and Co., 1982.

[17] J. J. Yokono and T. Poggio, "Oriented filters for object recognition: an empirical study". In *Proc. of IEEE FGR*, 2004.

[18] Y. Chen, and J. Z. Wang, "Image categorization by learning and reasoning with regions", *Journal of Machine Mearning Research*, 5(2004):913–939, 2004.