Freak Descriptor With Spatial Pyramid Kernel For Scene Categorization

Qiong Yao
School of Computer Engineering
University of Electronic Science and Technology
ZhongShan, China
2542807@qq.com

Xiang Xu
School of Computer Engineering
University of Electronic Science and Technology
ZhongShan, China
467796538@qq.com

Abstract—Spatial Pyramid Kernel performs good on challenging scene categorization tasks for combining multi-resolution and spatial information of image features , while the popular feature extraction still be the global SIFT or GIST, neglecting more recent , effective and efficient descriptors, such as FREAK ,which is faster to compute , more compact , and with lower memory. We partition image into fixed sub-regions, compute dense freak descriptor over fixed subregions' center points instead of on interest points, then use the spatial pyramid kernel recognition method for scene categorization. We make two comparisons, the first is between the FREAK (Fast Retina Keypoint) and the SIFT descriptor under the same scene categorization framework and the second is compute dense freak descriptors over fixed subregions' center points versus on interest points. The results show that the dense freak descriptor ensures a very competitive categorization performance with lower computational cost and less memory . It indicates that the FREAK descriptor is avaiable alternative to the SIFT descriptor for the problem of scene categorization and particularly useful for real-time-recognition systems for low-resource devices such as mobile phones.

Keywords-Spatial Pyramid Kernel; FREAK; Sift; Scene Categorization; SVM

Ⅰ. INTRODUCTION

The performance of scene categorization system depends mainly on two components. First , a suitable bag-of-features representation and second a powerful SVM kernel[10] on this representation. Despite the fact that a lot of advancements have been made in the area of keypoint descriptors over the last years, the literature on scene categorization for the most part still focuses on established descriptors, such as Lowe's SIFT [1] and torralba's "GIST" [2] descriptors, and largely neglects more recent descriptors , such as the FREAK descriptor [3] which is faster to compute , more compact while remaining robust to scale, rotation and noise . On the other hand , SVMs [9] are standard techniques in machine learning and excel by their ability to control the regularization, but they need a suitable kernel in order to work well. The spatial pyramid kernel [4] works perfect by computing rough geometric correspondence on a global scale using an efficient approximation technique adapted from the pyramid matching scheme of Grauman and Darrell[5], the recognizing scheme involves repeatedly subdividing the image and computing histograms of local features at increasingly fine resolutions.

We partition image into fixed sub-regions, compute dense freak descriptor over fixed subregions' center points instead of on interest points, then use the spatial pyramid kernel recognition method for scene categorization. We make two comparisions, the first is between the FREAK (Fast Retina Keypoint) and the SIFT descriptor under the same scene categorization framework and the second is compute dense freak descriptors over fixed subregions' center points versus on interest points. The results show that the dense freak descriptor ensures a very competitive categorization performance with lower computational cost and less storge space . It indicates that the FREAK descriptor is avaiable alternative to the SIFT descriptor for the problem of scene categorization and particularly useful for real-time-recognition systems for low-resource devices such as mobile phones.

In this paper we try to make a experimental comparision between the dense SIFT and dense FREAK desctiptors over subregions' center points instead of on interest points for the task of scene categorization. We use the spatial pyramid matching framwork presented in[4], and apply the framework to images databases from the fifteeen scenes[6] and the caltech101[7]. The results of our experiments suggest that the FREAK descriptor represents an appealing alternative to the SIFT descriptor with lower computational cost and storage space.

The rest of the paper is structured as follows. In Section Ⅱwe introduce the Spatial Pyramid kernel. In Sections Ⅲ we briefly describe the theory of the FREAK descriptors. In Section Ⅳ we present the experimental results and our main findings . We conclude the paper with some final comments and direction for future work in section Ⅴ.

Ⅱ. SPATIAL PYRAMID KERNEL

Grauman and Darrell [5] propose pyramid matching kernel to find an approximate correspondence between two sets. Let X and Y be two sets of vectors in a d-dimensional

feature space , construct a sequence of grids at resolutions $0,\cdots, L,$ such that the grid at level l has 2^l cells along each dimension, for a total of $D=2^{dl}$ cells. Let H_X^l and H_Y^l denote the histograms of X and Y at this resolution, so that $H_X^l(i)$ and $H_Y^l(i)$ are the numbers of points from X and Y that fall into the i th cell of the grid. Then the number of matches at level l is given by the histogram intersection function:

$$I\left(H_X^l, H_Y^l\right)=\sum_{i=1}^{D}\min\left(H_X^l(i), H_Y^l(i)\right) \tag{1}$$

The weight associated with level l is set to $\frac{1}{2^{L-l}}$,which is inversely proportional to cell width at that level , and penalizing matches found in larger cells because they involve increasing dissimilar features , the pyramid match kernel is definied by :

$$k^L(X,Y)=I^L+\sum_{l=0}^{L-1}\frac{1}{2^{L-l}}\left(I^l-I^{l+1}\right)$$

$$=\frac{1}{2^L}I^0+\sum_{l=1}^{L}\frac{1}{2^{L-l+1}}I^l \tag{2}$$

Then quantizing all feature vectors into M discrete types , and make the simplifying assumption that only features of the same type can be matched to one another. Each channel m gives us two sets of two-dimensional vectors, X_m and Y_m representing the coordinates of features of type m found in the respective images . The final kernel is then the sum of the separate channel kernels:

$$K^L(X,Y)=\sum_{m=1}^{M}k^L(X_m,Y_m) \tag{3}$$

Ⅲ. FREAK DESCRIPTOR

A. Retinal sampling pattern

The sampling pattern adopted by the FREAK approach is biologically inspired by the retinal pattern in the eye . Before the descriptor is computed , the N sample points located around the given keypoints are smoothed with a Gaussian kernel. Here, the size of the kernel is varied with respect to the location of the sampling point to simulate the behavior of the human retina. In analogy to the human visual system, the sampling points of the FREAK descriptor , hence , represent the centers of the receptive fields. Mathematically, this can be defined as follows:

$$P_i=P(x_i,y_i)=L_{r_i}(x_i,y_i) \tag{4}$$

Where

$$L_{r_i}(x,y)=I(x,y)*G_{r_i}(x,y,\sigma_{r_i}) \tag{5}$$

In the above equations $I(x,y)$ stands for the input image, $G_{r_i}(x,y,\sigma_{r_i})$ denotes the Gaussian kernel for the i -th receptive field ($i=1,2,\cdots,N$) and $L_{r_i}(x,y)$ represents the smoothed version of the input image. The i -th sampling point P_i correspoinding to the center of the i -th receptive field r_i and is defined with the predefined coordinates (x_i,y_i) from the sampling pattern, where $i=1,2,\cdots,N$.

B. Building the descriptor

Considering a pair of sampling points $P_a=(P_i,P_j)$, where $i,j\in\{1,2,\cdots,N\}$ and $i\neq j$, the FREAK approach defines a binary encoded intensity comparision $s(P_a)$ on this pair as

$$s(P_a)=\begin{cases}1 & if\ P_i>P_j\\ 0 & \text{otherwise}\end{cases} \tag{6}$$

The presented comparision forms the basis for building the FREAK descriptor F as a N-dimensional bit string:

$$F=\sum_{0\le a<N}2^a s(P_a) \tag{7}$$

C.Orientation normalization

The orientation of the FREAK descriptor is estimated based on 45 selected sampling-point pairs that are arranged symmetrically with respect to the center of the sampling pattern. Let G be the set of all the selected pairs and assume that local gradients have been computed for all the selected sampling points, then the orientation o of the given keypoint can be computed as:

$$o=\frac{1}{M}\sum_{\substack{P_i,P_j\in G\\ i\neq j}}(P_i-P_j)\frac{T(P_i)-T(P_j)}{\left\|T(P_i)-T(P_j)\right\|} \tag{8}$$

Where M is the number of pairs in G and $T(P_i)$ denotes a function returning the 2D vector of the spatial coordinates of the center of receptive field, i.e. the vector of coordinates of the k-th sampling point $T(P_k)=[x_k,y_k]$.

Ⅳ. EXPERIMENTS

In this section, three comparision experiments are carried out . First, extract dense SIFT descriptors of

16×16 pixel patches computed over a grid with spacing of 8 pixels , second, extract dense freak descriptors on the same grid, third, extract freak descriptors on the interest points[8]. We follow the experimental setup of [4] , Typical vocabulary sizes are $M=200$, set pyramid levels are L=2 .The experiment image databases are fifteen scene categories[6] and Caltech-101[7]. Our first dataset is composed of fifteen scene categories[6], each category has 200 to 400 images, and average image size is 300×250 pixels. The second set of experiments is on the Caltech-101[7],this database contains from 31 to 800 images per category. Most images are medium resolution,i.e., about 300×300 pixels.

TABLE Ⅰ THE MEAN AND STANDARD DEVIATION OF CALSSIFICATION RESULTS FOR THE DATABASES

Image database	**Dense SIFT features(%)**	**Dense Freak features(%)**	**Freak on keypoint(%)**
15scene	92.52 ± 1.09	94.16 ± 0.57	48.67 ± 1.55
Caltech101	91.85 ± 1.11	93.99 ± 0.58	46.53 ± 1.58

TABLE Ⅱ AVERAGE CLASSIFICATION RATES FOR INDIVIDUAL CLASSES ON THE FIFTEEN SCENE DATABASE

Classes	**Dense SIFT features(%)**	**Dense Freak features(%)**	**Freak on keypoint(%)**
01bedroom	93.52	96.76	54.63
02CALsuburb	99.59	99.59	74.27
03industrial	80.06	88.75	26.37
04kitchen	94.29	97.14	59.52
05livingroom	98.27	96.54	34.60
06MITcoast	80.56	92.50	46.11
07MITforest	98.17	100	96.04
08MIThighway	87.31	95.38	50.77
09MITinsidecity	92.53	94.81	22.40
10MITmountain	88.77	91.44	37.70
11MITopencountry	90.24	92.68	33.66
12MITstreet	97.60	93.15	70.55
13MITtallbuilding	87.36	87.08	46.35
14PARoffice	93.95	96.74	36.28
15store	99.05	100	80.95

We train on 100 images per class on first database and 30 images per class on second database and test on the rest. All experiments are repeated ten times with different randomly selected training and testing images, the average of per-class recognition rates is recorded for each run. The final result is reported as the mean and standard deviation of the results from the individual runs.

Table Ⅰ shows detailed results of respective classification results for two databases. Dense freak descriptor performs best, achieve 94.16% and 93.99% average classification rate on two image databases respectively, dense sift descriptor achieve 92.52% and 91.85% average classification rate on two image databases respectively, freak descriptor on keypoints performs worst ,only achieve 48.67% and 46.53% respectively. Table Ⅱ shows average classification rates for individual classes on the

TABLEⅢ TOP FIFTEEN LOW AVERAGE CLASSIFICATION RATES FOR INDIVIDUAL CLASSES ON THE CALTECH-101 DATABASE

Classes	**Dense SIFT features(%)**	**Classes**	**Dense Freak features(%)**	**Classes**	**Freak on keypoint(%)**
saxophone	60	pigeon	68.89	ewer	27.06
starfish	61.63	BACKGROUND_Google	78.37	Ferry	28.36
chair	66.13	scissors	79.49	nautilus	30.91
scorpion	67.86	buddha	80	BACKGROUND_Google	31.69
butterfly	68.13	beaver	80.43	butterfly	31.87
pigeon	71.11	stop_sign	81.25	camera	32
mandolin	72.09	mayfly	82.50	sea_horse	33.33
flamingo	76.12	menorah	82.76	chair	33.87
gramophone	76.47	umbrella	84	strawberry	34.29
strawberry	77.14	revolver	84.15	menorah	34.48
Electric_guitar	78.67	scorpion	84.52	crab	36.99
laptop	79.01	stegosaurus	84.75	brontosaurus	37.21
BACKGROUND_Google	79.87	ceiling_fan	85.11	gramophone	39.22
beaver	80.43	lamp	86.89	lotus	39.39
umbrella	81.33	bass	87.04	Dragonfly	39.71

fifteen scene database.Table Ⅲ shows top fifteen low average classification rates for individual classes on the Caltech-101 database.

Ⅴ. DISCUSSION

This paper we repeatedly subdividing an image and computing histograms of dense freak features over the resulting subregions, has shown promising results on two large-scale,diverse datasets compared with dense sift and sparse freak on keypoints. For the FREAK descriptor is faster to compute , more compact , and with lower memory. It indicates that the FREAK descriptor is avaiable alternative to the SIFT descriptor for the problem of scene categorization and particularly useful for real-time-recognition systems for low-resource devices such as mobile phones.

ACKNOWLEDGMENT

This work was financially supported by the ZhongShan Science and Technology Planning Project (413S40), and Guangdong province science and technology plan projects（2013B090500035）.

REFERENCE

[1] D.G.Lowe. Distinctive image features from scale-invariant keypoints. IJCV,60,2, pp.91-110,2004.

[2] A.Torralba, K.P.Murphy, W.T.Freeman, and M.A.Rubin. Context-based vision system for place and object recognition. ICCV, pp.273-280, 2003.

[3] A .Alahi, R. Ortiz, P.Vandergheynst, FREAK:Fast Retina Keypoint. CVPR, pp. 510 – 517, 2012.

[4] S.Lazebnik, C.Schmid, J.Ponce . Beyond Bags of Features:Spatial Pyramid Matching for Recognizing Natural Scene Categories.CVPR, pp.2169-2178,2006.

[5] K.Grauman and T.Darrell. Pyramid match kernels: Discriminative classification with sets of image features. ICCV, pp.1458-1465, 2005.

[6] L.Fei-Fei and P.Perona. A Bayesian hierarchical model for learning natural scene categories. CVPR, pp.524-531, 2005.

[7] Information on http: //www.vision.caltech.edu /Image_Datasets/Caltech101 .

[8] C.Harris and M.Stephens. A combined corner and edge detector. In Alvey vision conference, pp. 50-57, 1988.

[9] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, pp.1—27, 2011

[10] J.Eichhorn and O.Chapelle. Object Categorization with SVM: Kernels for Local Features.Technical report,MPI for Biological Cybernetics,2004.