Research on the Classification of Image Semantic Scene

Zhang Fang [a], Guo Huiling[b] and Jia Lingshan[c]

Environmental Management College of China, Qinhuangdao, China, 066004

[a]fanny_zh117@163.com, [b]16951203@qq.com, [c]63786960@qq.com

Keywords: Regular super pixel grid block, Semantic visual vocabulary SVV, scene categorization

Abstract. This paper mainly aims at the algorithm of scene categorization. Firstly, putting forward by using the edge information of training set image to modela probabilistic generative model, and getingt the shape prior distribution of each scene category by learning, and inferencing to get the boundary distribution image,so to finish the regular super pixel grid block.Then, clustering the local word package description data of sub region, to construct the semantic visual vocabulary SVV, combining with the regularized super pixel grid to descript image; finally, finishing the task.

1 Introduction

With the rapid development of computer and network technology and the rapid popularization of digital equipment, the scale of the image data has rapidy expaned.How to find what we really need information from the multitude of resources of image data, and how to effectively classify and organize the data, has become the problem of urgently need to solve [1,2]. Changes and differences on the content of the image itself, can lead to the inconsistency of the classobjects within scene. It can cause the visual similarity of the scene class. Therefore, the task of scene classification is very difficult, at present is still in the exploration and the preliminary study stage.

In the tasks of scene classification, image feature extraction is the first step. How to not only keep the invariant of the characteristics in the description of the image, but also embodie the characteristics of spatial information, realize the integrated complementary of local features and global features, is the key of visual feature extractionis [3,4]. This paper presents the mothed of doing dense SIFT sampling in multi scale pixel interval and regional conditions, combined with the upper and lower scale and spatial neighborhood information,structing context characteristics of the interested region, and synthesizing local information and spatial information, making the visual features more in line with the characteristics of human visual awareness.

The classification performance of the algorithm depends heavily on how to define the visual vocabulary [5]. The existing methods of definition visual vocabulary can not provide the real local semantic concepts. This paper presents that doing the feature clustering to generate the primary visual vocabulary at first, and then doing the super pixel grid block to the image, for each sub region using the bag of words model to describ, formating the histogram representation of each sub region; all the histogram representation of each sub region are expressed as new input datas,carring on two times of clustering, to generate a new kind of visual vocabulary--semantic visual vocabulary (Semantic visual vocabulary, SVV).

2 Approaches

2.1 Regular super pixel grid image block

2.1.1 Scene shape distribution priors

The edge of the image set by $\mathbf{X}=[x_1,\cdots, x_p]^T$, a total of p pixels, each element value is 0 or 1. Assuming that x_p obeys the parameters for the two items of the distribution of y_p, then the vector $\mathbf{y}=[y_1,\cdots, y_p]^T$, represents boundary distributionimage. The Scene shape prior distribution is studied by CLT (clustered latent trait) graph model, as shown in Fig.1. Assuming that the distribution of **y** is decided by K clusters, each set consists of a sub space, so for image $\mathbf{y}_i$:(1) The existence of a discrete hidden variable c, indicating those K set that coming into being data;(2) The existence

of a continuous latent variable $\boldsymbol{h}$, said that the sub space position of above-mentioned set. The matrix $\mathbf{F}_k=[f_{1k},\cdots, f_{Jk}]$ said the J basis functions of the k set. The definition of $\mathbf{a}=[a_1,\cdots, a_p]^T$ describes a pixel in the image orientation, and$\mathbf{a}=\mu_c+\mathbf{F}_c\mathbf{h}$, which μ_c said the ensemble mean of vector c. The use of Eq. 1 will be converted **a** to **y** with probability:

$$y_p = \sigma(a_p) = \frac{1}{1+\exp(-a_p)} \tag{1}$$

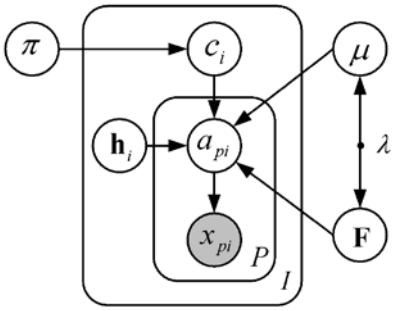

Fig.1 The CLT diagram model

2.1.2 Regular super pixel grid block

(1)Inhomogeneous Strip: On the basis of the distribution of image boundary to compute each inhomogeneous strip of the longitudinal and transverse direction, ensuring to find the boundary path accordance with the probability distribution of boundary in the strip. The Calculation of the longitudinal strip.

(2) The boundary path: After the calculation of the boundary strip, first through an affinetransformation matrix mapping the strip to a transformation space, into alinear band, to obtain the path follows the boundary shape; Then, calculating the minimum cost path in the alternated strip, and reflecting the bands to the original space, so the minimum cost path that followes the boundary shapeis obtained.

(3) Regular super pixel grid block: According to the pre-set resolution ratio of the regular super pixel grid, and based on the distribution prior of the scene shape, to get the boundary distributionimage by learning, and using the above method to find the best path, complete the regular super pixel grid image block.

2.2 Robust context feature extraction

2.2.1 SIFT feature extraction

In order to make the feature points have better uniquenessr, this paper used the method of grid sampling to extract the dense SIFT feature extraction of the image :(1) Using the grid densely sampling to the training set image, obtained the corresponding grid sampling points, and the grid sampling interval is 8 pixels;(2) Calculating the SIFT characteristic for the size of 16 * 16 pixel area around each grid sampling points. SIFT features were represented by the regional gradient direction histogram, is a 128 dimensional vector.

2.2.2 The construction of context features

Doing the dense SIFT sampling at multiple scales of pixel interval and multi scaleregional conditions, the multi scale is defined as:s=1,2,...,S.Taking the consider of computation,setting S=3.S=3 said that the sampling interval is 32 pixels, and the statistical regional block size is 64 x 64.

In order to obtain the context information of the region of interest (region of interest, ROI), combined with its neighborhood information and information center for the ROI on a scale at each scale, the combination of the threecan contain contextual information, and takes it as the new feature, called context features, as shown in Fig. 2.

In that, f_s denote the SIFT features of ROI at scale s, f_{s+1}, the SIFT features of the region having the same center as the ROI but at a coarser scale level andf_{s_n} having the SIFT featuresof ROI neighbors at scales. The contextual descriptor of the ROI is built by Eq. 4:

$$f = [\, f_s ; w_C f_{s+1}; w_N f_{s\_n}]$$

(2)

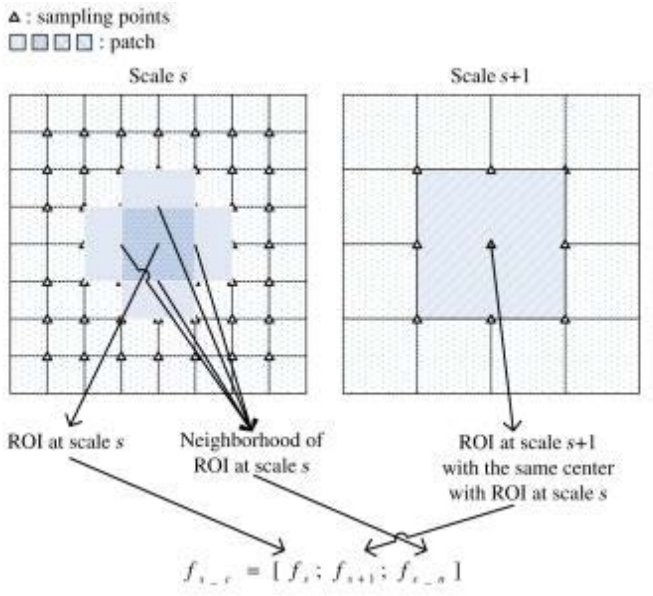

Fig. 2 The schema of the contextual features at scale *s*

2.3 The construction of semantic visual vocabulary SVV

2.3.1 The primary visual vocabulary

With the $\lambda=\{w_i, \mu_i, \sum_i, i=1,2,\ldots N\}$ denoting the parameters set of the Gauss mixture model,in it, w_i, μ_i, and $\sum_i$ respectively represent the Gauss's weight of i, the mean vector and the covariance matrix, the number of N for Gauss; Each of the Gauss model is a word in the visual vocabulary of $\lambda=\{v_i, \ i=1,2,\ldots N\}$, w_i is the relative frequency of v_i, μ_i is the mean, $\sum_i$ is the variance, and$\sigma_i^2=diag(\sum_i)$. q is the observed values implicit mixed variable x,so:

$$p(x/\lambda)=\sum_{i=1}^{N} w_i p(x/q=i,\lambda) \tag{3}$$

$$p(x/q=i,\lambda)=\frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}}\exp\{-\frac{1}{2}(x-\mu_i)^T\sum_i^{-1}(x-\mu_i)\} \tag{4}$$

Among them, dis the characteristic dimension. To estimate λ by using the expectation maximization algorithm EM:

E-step：

$$\gamma_t(i)=p(q_t=i\,|x_t,\lambda)=\frac{w_i p(x_t|q_t=i,\lambda)}{\sum_{j=1}^{N} w_j p(x_t|q_t=j,\lambda)} \tag{5}$$

2.3.2 Semantic visual words SVV

Based on the Image block of the regular super pixels grid technology, each sub region is a significant area. Each sub block according to the primary visual vocabulary be mapped, and by using the word package model to formation image description of the respective region. Suppose the number of the words of primary visual vocabulary for m, the resolution of regularsuper pixel grid is $h\times v$, then the number of image sub block is $h\times v$, and the sub block image region can be described as m dimension vector, it's number is $h\times v$. The entire region discription of sub block can be as a new data set, using the Gauss mixture model to cluster, each cluster center corresponds to new semantic visual words, thus forming SVV.

2.4 Image semantic description and scene classification

To train and test images by the supported vector machine classifier of libSVM based on the histogram intersection kernel HIK. Assume that X and Y are the vector of two D dimensional space, then the definition of HIK is:

$$k_{HI}(X,Y)=\sum_{j=1}^{D}\min(x_j,y_j) \tag{6}$$

The semantic description of the scene class image as the input data of the classifier can complete the training and testing of the scene classification.

3 Conclusions

This paper used the method of regular super pixel grid block to divide an image into different sub zones, clustered the description of the primary words bag to generate SVV of each sub region, and combined the regular super pixel grid, increasing global spatial information of image, finally, the Semantic description of the image send to libSVM classifier to complete the scene classification. After analysis, this method is feasible and effective.

Acknowledgements

The author appreciates the generous financial support from The natural science foundation of Hebei Province for Projects No. Q2012047.

References

[1] Datta R: Semantics and aesthetics inference for image search: statistical learning approaches [D]. The Pennsylvania State University. PhD dissertation, 2009:1-20

[2]Gao J, Xie Zh: *The theory and method of image understanding* [M]. Science Press, 2009: 399-430

[3]Qin J, Yung N: Feature fusion within local region using localized maximum- margin learning for scene categorization [J]. Pattern Recognition, 2012, 45(4): 1671-1683

[4] Jiang Y, Chen J, Wang R S: Fusing local and global information for scene classification [J]. Optical Engineering, 2010, 49(4): 047001-047010

[5] Yang X, Xu D, Feng S H: Scene categorization with classified codebook model [J]. IEICE Transactions on Information and Systems, 2011, 94D (6):1349-1352