Scene Classification Algorithm Based on Covariance Descriptor

Li Xingsheng
Information Dept. of
Naval Headquarters
Beijing, China
netlxs@163.com

Wu Zemin
College of Communication Engineering
PLA University of Science and Technology
Nanjing, China
wuzemin_ice@163.com

Tan Wei
Communication network technology
Management Center, General Staff
Beijing, China
tanziyu081210@sina.cn

Yu Jiang
Office of Communication
Unit 73691 of PLA
Nanjing, China
yujiang@sina.com

Abstract—Scene classification has been a hot topic in the field of computer vision. Under the premise of image segmentation, this paper proposes a novel scene classification algorithm, combining pixel location, color characteristics, Gabor features, and local binary features (LBP) to form a covariance descriptor, and then converting it to the Sigma-point characteristics into a European Space, to complete the scene description and SVM training. To compare performance with some of the classic classification algorithms, we simulate the algorithm on standard Image SUN Database, and besides we construct data set with noise to validate their tolerance in dealing with noise and robustness. The results show that the proposed algorithm not only has a strong advantage on computation time, feature dimension and classification performance, but also has good fault tolerance and robustness.

Keywords- Image Segmentation; Covariance Descriptor; Sigma Points Feature; Scene Classification

I. INTRODUCTION

Scene classification is an important research direction in the field of computer vision and image understanding. Scene category not only includes a general understanding of images, but also is great significant in target detection and recognition, visual monitoring, remote control navigation and other applications. Scene description is the foundation for classification, which forms different scene elements and identifies them by quantifying and analyzing pixels in the image. Currently, the research of scene classification has achieved certain results, and the methods of scene description have become hot field of computer vision. Many researchers have committed to building a model capable of describing the scene, which has made great progress. However, to achieve efficient and accurate classification and superior performance, robustness scene description is still a challenging problem.

The traditional description methods are to describe the features of the underlying image, by extracting the points of interesting [1-3] (such as color, shape, texture, etc.) to characterize the scene. The accuracy of such methods largely depends on the selection of points of interest. To address the problem of lacking of relation between the scene content and interesting points, a method of image segmentation [4] was proposed by Rother, in which the image was divided into several manual annotated regions, to achieve supervised description and classification.

In recent years, researchers have pay more attention to the semantic information in the scene, and proposed a series of methods based on semantic description for classification, which are mainly divided into semantic object segmentation, local and global semantic modeling methods. Semantic object segmentation methods [5] require the "semantic target" segmentation and semantic information labeling in the scene, which are labor-intensive and can not achieve automatic classification. But local and global semantic modeling methods are not need for segmentation or labeling. Local methods using descriptors around points of interest to achieve scene modeling and classification, such as the typical "feature of bag" model (BOF) [6] and the pyramid matching model (SPM) [7]. Global methods regard the scene as a whole global descriptor using global features to complete classification, such as holistic Gist features proposed by Oliva [8].

In this paper, combining the image segmentation and local semantic modeling methods of underlying characteristics, we propose a local Sigma semantic points based on segmentation to model the semantic targets in the scene. First on the basis of conventional image segmentation, we segment the foreground objects of the scene, and then construct covariance descriptor to form semantic representation of objectives, combining with the pixel position, color [9], Gabor features [10] and LBP features [11], and finally convert the descriptor into the Sigma Points feature under Euclidean space for standard SVM learning and classification.

© 2014. The authors - Published by Atlantis Press

II. THE SIGMA POINTS METHOD

Different scenarios include particular targets which represent the properties of the scenarios. If these targets can be effectively described through semantic descriptions, we can use these descriptions, which indicate the attribute of the scene, to achieve scene understanding and classification. Image segmentation algorithms for classification based on the underlying characteristics [4], can split the foreground object, but need to mark divided regions artificially and cost much labor-intensive and time to infer scene category by inter-regional marked relations. In this paper, based on [4], we only use GraphCut algorithm to segment the foreground objects in the scene, and on this basis conduct the covariance descriptor representing semantic structure of foreground target regions. Eventually we convert it into the Sigma points feature in Euclidean space, and complete the scene description and classification task.

A. Covariance Descriptor

2007, Tucel etc [12-13] firstly proposed such a covariance matrix based on a simple characterization methods to achieve a fast target discovery and classification. Indeed, the covariance descriptor does integrate features in various channel (such as color, filter response, etc.) and calculate their correlation coefficient, forming a low-dimensional description vector. Assuming I is a three-dimensional color image, we can get the $W \times H \times d$ dimensional covariance description of the whole image, that is $F(x,y) = \phi(I,x,y)$, W、H is the width and height of the image , the mapping function ϕ that is the integration of multiple characteristic channel, expands each pixel in the image into d dimensional features. Thus, for any rectangular area R in the image, let $\{z_k\}_{k=1...n}$ equal all the d dimensional pixels in R , then according to formula (1), where u is the mean of d-dimensional feature pixels, we can calculate the $d \times d$ dimensional covariance matrix, called the region covariance descriptor.

$$C_R = \frac{1}{n-1}\sum_{k=1}^{n}(z_k - \mu)(z_k - \mu)^T \tag{1}$$

In traditional literatures [12-14], people used to construct covariance feature set with pixels position, RGB color features and gradient information. But for the targets in the scene, there are problems of illumination, viewing angle and non-rigid, which directly affect the scene judgment. Traditional covariance feature set can not solve these problems. Therefore, for these influencing factors, this paper proposes a new covariance feature set as follows:

$$F(x,y) = \left[x, y, \frac{O_1(x,y)}{O_2(x,y)}, \frac{O_2(x,y)}{O_3(x,y)}, \left|\frac{\partial I(x,y)}{\partial x}\right|, \left|\frac{\partial I(x,y)}{\partial y}\right|, Gabor(x,y), LBP(x,y) \right] \tag{2}$$

Besides retaining information of gradient and location, we change the description of color channels, and add Gabor filters to extract contour information and texture information (LBP). In formula (2), $\frac{O_1(x,y)}{O_2(x,y)}$ and $\frac{O_2(x,y)}{O_3(x,y)}$ represent the invariant characteristics of color space [9], that ensure the stability of color features under different conditions.

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{3}$$

$Gabor(x,y)$ [10] means that using m-scales and n-directions Gabor filters to filter grayscale image $f(x,y)$. So we can get the image with each pixel of $m \times n$ dimensional features, that is:

$$\begin{aligned} &Gabor(x,y) = f(x,y) * g_{mn}(x,y) \\ &g_{mn}(x,y) = a^{-m} g(x',y') \\ &g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left[-\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right)\right] \times \cos(2\pi f_0 x + \varphi) \\ &x' = a^{-m}(x\cos\theta + y\cos\theta), y' = a^{-m}(-x\sin\theta + y\cos\theta) \\ &\theta = n\pi/(n+1) \end{aligned} \tag{4}$$

a^{-m} is the scale factor, σ_x and σ_y are the standard deviation of the Gaussian function, f_0 is the filter center frequency, θ is filter direction, m and n represent the numbers of scale and direction, where m is 3 and n is 4. Using Gabor filters for image feature extraction, the real fact is to detect the frequency information corresponding to direction to construct salient feature. For the foreground objects segmented in the scene, it can form a robust, compact semantic feature indicating the target contour.

$LBP(x,y)$ [11] means binary results of eight neighbors around pixel (x,y), and for the segmented images, it can effectively describe texture features of the foreground objects, which distinguishes with others strongly and is invariant for illumination changes, and enrich the spatial information of covariance descriptor to enhance targets semantic description in the scene.

B. Feature Description of Sigma Points

Since the covariance descriptor is a matrix statistic, not the eigenvectors in European Space. The standard machine learning methods such as SVM, which need to create a sentence by computing similarity, can not directly use the covariance descriptor. In the literature [15], Julier, etc. proposed a nonlinear transformation method - UT transform. Since approximate function substituted by probability density of approximately nonlinear function, and statistics of random vectors characterized with a set of sampling points, it does complete space mapping. With UT transform, we can transform d-dimensional covariance descriptor into the 2d +1 feature points under Euclidean space, where each point contains d-dimensional feature, called Sigma points [15]. Expressed as follows:

$$s_0 = \mu \qquad s_i = \mu + \alpha(\sqrt{\Sigma})_i \qquad s_{i+d} = \mu - \alpha(\sqrt{\Sigma})_i \tag{5}$$

$i = 1...d$, μ is the mean value of d-dimensional features of all pixels. $\sqrt{\Sigma}$ represents the standard deviation of the covariance matrix, and $(\sqrt{\Sigma})_i$ defines the i-th column of

the required matrix square root $\sqrt{\Sigma}$. The scalar α defines a constant weighting for the elements in the covariance matrix and is set to $\sqrt{2}$.The construction pipeline for the set of Sigma Points is summarized as follows:

TABLE I Construction of Sigma Points

Algorithm 1 Construction of feature representation based on Sigma Points

Require: Covariance matrix Σ^k and mean vector μ^k

1: Perform a simple regularization $\Sigma^k = \Sigma^k + \varepsilon I$, ε =1e-6, I is the identity matrix;
2: The efficient Cholesky factorization can be applied to compute the matrix square root by decomposing $\Sigma^k = \sqrt{\Sigma}\ (\sqrt{\Sigma})^T$, $\sqrt{\Sigma}$ is a lower triangular matrix;
3: Compute Sigma Points s_i^k according to (5), $i = 0 \ldots 2d$;
4: Construct the final feature set by catenation of the Sigma Points, $S^k = (s_0^k, s_1^k \cdots s_{2d}^k)$

The final feature representation captures both first and second order statistics, which are given by mean and covariance information. Each of these generated vectors describe Euclidean space, therefore, these vectors can be learned directly by SVM to build a reliable surface for judgment and classification.

III. EXPERIMENTAL RESULTS AND ANALYSIS

A. Scene Database

The algorithms for scene classification are the current issues in computer vision. Usually we choose the standard scene database for classification algorithms testing and performance analysis, such as Caltech101,Caltech256, etc. 2010, researchers in Brown University created SUN Database[16], which contains 899 artificial marked categories and 130,519 scene images, and more accurate experiments were done to evaluate the performance of some classical algorithms to prove that SUN Database is a basic and exhaustive scene database and verify the performance of the algorithms for the classification of universal significance. Therefore, in this article compared with [8], we select the same 7 categories in SUN Database as standard test set to analyze the performance with some classical algorithms. While we do smooth and add noise with the standard set to construct a new fuzzy scene set to evaluate the proposed algorithm and the classical algorithms robustness more accurately.

B. Description of Test Process

Testing process is divided into two parts: First, we regulate the size of each image to 256x256 pixels in the standard test set. On the basis of the literature [4], we can segment the semantic targets in the scene with CraphCut algorithm, and then calculate the Sigma Points set of targets, which is described as the final feature vector of the scene. For the fully learning of all scenes in standard set, we randomly selected half of the categories as the training set, and the remaining as the test set; then we select LIBSVM for all-vs-all multi-class learning and linear function as kernel function selection for the testing of the proposed in this paper, and compared with the typical algorithms such as Gist[8], LBP[11], HMAX[17], BOF[6] and SPM[7]. Second, based on the standard set, we giving add salt and pepper noise to each image and do smooth to create a fuzzy set. When testing the algorithms overhead, we still select the half of the standard categories as training set. But we switch the fuzzy set as testing set that we can compare fault-tolerance and robustness of Sigma Points with other classical algorithms more accurately.

C. Experimental results and analysis

TABLE II Comparison of six algorithms

Algorithms	Accuracy	Time for single judgment
SPM	80.86%	1.99 s
BOF	66.29%	1.61 s
Gist	63.43%	1.64 s
LBP	60.57%	1.23 s
HMAX	56.86%	1.42 s
Sigma	71.14%	1.28 s

As is showed in Table 2, SPM integrating the characteristic dictionary and spatial information is the best one on the accuracy; in this paper, Sigma Points based on the covariance can accurately describe the characteristics of the targets in the scene to get the high accuracy, which is only lower than SPM; although Gist and BOF consider semantic information, their effects are not as good as SPM and Sigma Points; the texture description of LBP and HMAX for simulating the optic nerve are the worst two algorithms. Considering time cost, SPM integrate SIFT features and k-means cluster to form a 4,200-dimensional feature vector, so it has the largest time cost for single judgment; BOF is the basis of the SPM and also need to calculate SIFT features and clustering dictionary; besides Gist is based on human perception and integrate multi-scale and multi-directional filters for block cascaded and feature extraction to form a 960-dimensional vector, and these two algorithms also cost much time for single judgment; HMAX with two layers of neurons in the visual analogs filtering ideas, and constructs 972-dimensional features, which costs little time; LBP, which is very simple, describe the texture features of global scene roughly and it is the fastest one for single judgment; Sigma Points algorithm proposed in this paper only partially describes the target area with covariance forming a 741-dimensional feature vector.

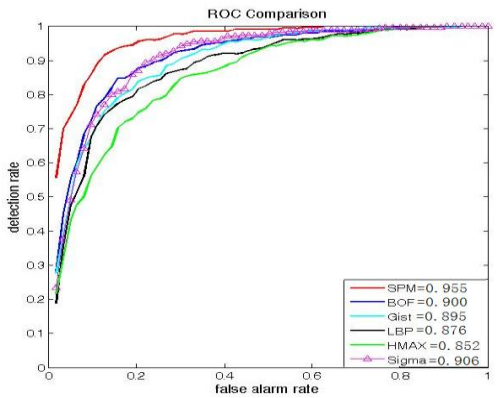

Figure 2. Standard ROC curves

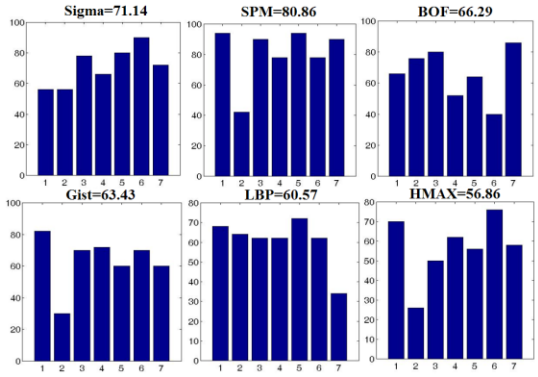

Figure 3. confusion tables

ROC curve can represent recognition and rate false alarm rate well, which is used as quality standard to determine the classification performance. The larger value of area is, the better performance of the algorithm is. As can be seen in Figure 1, the roc value of SPM is maximum of 0.955, which is the best one for classification of all algorithms; Sigma Points algorithm in this paper, the roc value is 0.906, which is superior to other classical algorithms and just below than SPM with excellent performance. Confusion table is a description of the degree of accuracy and misclassification. With Figure 2 and Figure 3, it can be seen that SPM has the highest accuracy, but its stability for classification is not enough with polarization on accuracy. The accuracy of Sigma Points algorithm is superior to other classical algorithms and just below than SPM, but it has most stable accuracy for classification, since the accuracy for each type of scene is almost average. In summary, Sigma Points algorithm has excellent classification performance and accuracy, which is better than other classical algorithms and just below SPM ; but it has the absolute advantage on stability on classification and time-costing for judgment, which will make up for the deficiencies of SPM.

IV. CONCLUSION

Scene classification based on understanding of the content of the scene is the current issue to the computer vision and pattern recognition researches, and how to select reasonable features to describe the scene has a direct impact on the final classification accuracy. In this paper, under the premise of image segmentation, we construct the covariance descriptor including pixel location, color characteristics, Gabor features and texture features (LBP) that describes the of the targets in the scene, in order to characterize the type of scene. However, the covariance descriptor is a statistic matrix that can not be learned by standard machine to build decision surface. To solve this problem, we use UT transformation to map the covariance matrix under the Riemannian space to Euclidean space, and form the final Sigma Points to describe semantic targets in the scene. With the experiments using SVM, we summarize that this algorithm has obvious advantage in the time-costing for calculation and decision and is able to accurately express the scene information, so it has a higher classification accuracy for the scene where the targets are clear; meanwhile, it can stably describe local semantic objectives with good robustness to local noise and fuzzy. The next we want to combine Sigma Points and features of the dictionary, and use the segmentation as prior knowledge to establish the feature dictionary of targets regarded as a feedback to more accurately extract the targets both in training and testing scenes, finally we use the histogram of target dictionary to establish the decision surface to achieve classification, and it is desirable to improve the classification performance and accuracy of Sigma Points algorithm.

REFERENCES

[1] A.Vailaya, M.Figueiredo, A.Jain and H.J.Zhang, "Image classify-cation for content-based indexing". IEEE Trans. on Image Processing, vol.10, pp.117-130, January 2001.

[2] E.Chang, K.Goh, G.Sychay and G.Wu, "CBSA: content-based soft annotation for multimodal image retrieval using Bayes point machines". IEEE Trans. on Circuits and Systems for Video Technology, vol.13, pp.26-38, January 2003.

[3] S.Belongie, J.Malika and J.Puzicha, "Shape matching and object recognition using shape contexts". IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.24, pp.509-522, June, 2003.

[4] C.Rother, V.Kolmogorov and A.Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts". ACM Transactions on Graphics, vol.23, pp.309-314, Mach, 2003.

[5] H.H.Cheng and R.S.Wang, "Semantic modeling of natural scenes based on contextual Bayesian networks", Pattern Recognition, vol.43, pp.4042-4054, December, 2010.

[6] Nowak E, Jurie F, Triggs B. Sampling Strategies for Bag-of-Features Image Classification[C]. European Conference on Computer Vision. 2006: 409-503.

[7] S.Lazebnik and C.Schmid, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", IEEE Conference on Computer Vision and Pattern Recognition, IEEE Press, May, 2006, pp.2169-2178.

[8] A.Oliva and A.Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope", International Journal of Computer Vision, vol.42, pp.145-147, March, 2001.

[9] K.E.Sande, T.Gevers and C.Snoek. "Evaluation of color descript-tors for object and scene recognition", IEEE Conference on Computer Vision and Pattern Recognition, IEEE Press, May, 2008: pp.1-8.

[10] Y.Pang, Y.Yuan and X.Li. "Gabor-Based Region Covariance Matrices for Face Recognition", IEEE Trans. on Circuit and System for Video Technology, vol.18, pp.989-993, June, 2008.

[11] Z.H.Guo and L.Zhang, "A Completed Modeling of Local Binary Pattern Operator for Texture Classification", IEEE Trans. on Image Processing, vol.19, pp.1657-1663, June, 2010.

[12] Tuzel, O. Porikli, F. Meer. Human detection via classification on Riemannian manifolds[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2007: 1-8.

[13] O.Tuzel, F.Porikli and F.Meer. "Learning on lie groups for invariant detection and tracking", IEEE Conference on Computer Vision and Pattern Recognition, IEEE Press, May, 2008, pp.1-8.

[14] C.Yinghao, T.Valtteri and P.Matti, "Matching groups of people by covariance descriptor", IEEE International Conference on Pattern Recognition, IEEE Press, August, 2010, pp.2744-2747.

[15] S.Julier and J.K.Uhlman, "A general method for approximating nonlinear transformations of probability distributions", Robotics Research Group, Department of Engineering Science, University of Oxford, Oxford, OC1 3PJ United Kingdom, Tech. Rep, 1996.

[16] J.X.Xiao, J.Hays and K.Ehinger, "SUN Database: Large-scale Scene Recognition from Abbey to Zoo", IEEE Conference on Computer Vision and Pattern Recognition, IEEE Press, May, 2010, pp.3485-3492.

[17] M.Riesenhuber and T.Poggio, "Hierarchical models of object recognition in cortex", Nature Neuroscience, vol.2, pp.1019-1025, November, 1999.