COVARIANCE MATRIX BASED FEATURE FUSION FOR SCENE CLASSIFICATION

Nanjun He[a,b], Leyuan Fang[a], Shutao Li[a], Antonio J. Plaza[b]

a. College of Electrical and Information Engineering, Hunan University, China
b. Department of Technology of Computers and Communications, University of Extremadura, Spain

ABSTRACT

In this paper, a covariance matrix based feature fusion (CMFF) framework is proposed to combine two low-level visual features i.e., the Gabor feature and color feature for scene classification. Generally, the proposed method consists of following three steps. Firstly, the Gabor feature and color feature are extracted from original image and stacked together. Then, a covariance matrix is extracted to fuse these two low-level visual features. Each nondiagonal entry in the covariance matrix stands for the correlation of two different feature dimensions. Finally, the obtained covariance matrix is handled by a kernel linear discriminative analysis algorithm followed with nearest neighboring classifier for label assignment. The proposed method is tested on a public 21-classes UC Merced land use data set and compared with mid-level visual feature oriented method and the high-level feature oriented methods. The experimental results demonstrate that the proposed CMFF framework can not only improve the classification performance of the low-level visual feature (the Gabor feature and the color feature), but also can outperform the conventional mid-level visual feature oriented methods.

Index Terms— Scene classification, feature representation, feature fusion,

1. INTRODUCTION

With the advance of the satellite sensors, a large number of high spatial resolution (HSR) and hyperspectral images have become available [1, 2]. How to understand and recognize these huge amount of images effectively become a critical task. Nevertheless, the (HSR) images often present the characteristics of complex spatial construction with high intraclass and low interclass variabilities, which makes the recognition of HSR images become a very challenging problem. Under this context, many scene classification methods have been proposed over the past years.

Generally, the existing scene classification methods can be categorized into three classes. The low level visual feature (LLF) oriented method, the mid-level visual feature (MLF) oriented method and the high level visual feature (HLF) oriented method. For LLF oriented method, the classical local or global feature descriptor is firstly extracted to represent the test images. Then the obtained feature are sent into a supervised classifier (e.g., the support vector machine (SVM)) for label assignment. In [3], Yang and Newsam hybrid the Gabor text feature with the maximum a posteriori model (MAP) for scene classification where each test image is represented by a vector consisting of the mean and standard deviation of corresponding Gabor feature. Moreover, the global color histogram is used to characterize the image and the SVM is then utlized to classify the obtained feature vectors [4].

By taking into account that there may be semantic gap between LLFs and high level semantic meaning of images, the MLF oriented methods are proposed bridge these two levels. In [4], Yang *et al.* use the bag of visual word (BoVW) model to quantized the scale invariant feature transformation (SIFT) descriptor for MLF extraction. The MLFs are finally fed into a SVM with intersection kernel for classification. To further take the spatial construction into account, the the spatial pyramid matching (SPM) is used to extend the BoVW model [5, 6].

Recently, inspired by the success of deep learning in the computer vision community. The convolution neural network (CNN) has been extended for HLF oriented scene classification and achieved the state of the art classification performance. In [7], two popular CNN architectures (i.e., the CaffeNet and GoogleNet) are applied on the HSR image for HLFs extraction. Both the CaffeNet and GoogleNet are applied with three different strategies, train from scratch, pre-train and fine-tune, and feature extraction. The work in [7] demonstrates that pre-train and fine-tune strategy gains the best classification performance on both CaffeNet and GoogleNet.

In this paper, we focus on the LLF oriented method. To this end, we propose a covariance matrix based feature fusion (CMFF) framework for scene classification. The proposed CMFF framework includes the following three steps, i.e., Feature extraction, covariance matrix construction and KLDA based label assignment. In the first step, two LLFs (Gabor feature and color feature) are extracted from original HSR image and stacked together. In the second step, the covariance matrix (CM) is constructed among the sacked fea-

This paper is supported by the National Natural Science Fund of China for International Cooperation and Exchanges under Grant 61520106001, the National Natural Science Foundation for Young Scientist of China under Grant No. 61501180, and the Fund of Hunan Province for Science and Technology Plan Project under Grant 2017RS3024.

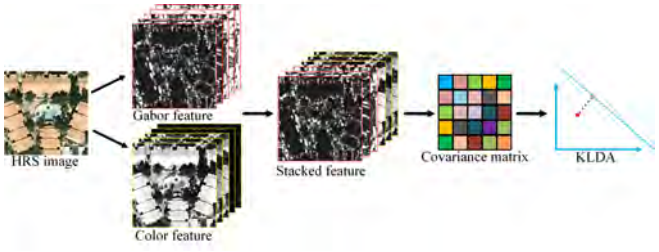

Fig. 1. The flowchart of the proposed CMFF framework.

ture. The diagonal entries of CM represent the variance of each feature and the nondiagonal entries reflect their respective correlations which offer a natural way of fusing different features without using blending weights. Since the CM lies on the manifold space, the kernel LDA followed by nearest neighboring rule is finally adopted to classify the CM. Though, the CMFF method relies on the LLF feature, the CMFF framework can gain much better classification performances than the classical and LLF and MLF oriented method. Moreover, the CMFF method is a generalization framework and can be easily extended to fuse the MLF and HLF, which will be our ongoing works.

The rest of this paper is organized as follows. Section II describes the proposed CMFF framework. Section III gives the result conducted on a real 21-classes aerial data sets and makes comprehensive comparison between the proposed CMFF framework and several classical LLF oriented methods, MLF oriented methods and HLF oriented methods, respectively. The final conclusion and some of our future works are summarized on Section IV.

2. PROPOSED COVARIANCE MATRIX BASED FUSION FRAMEWORK

The flowchart of proposed CMFF framework are shown on Fig. 1 which consists of three main steps: Feature extraction, covariance matrix construction and KLDA based label assignment.

2.1. Feature Extraction

2.1.1. Gabor Feature

Given a gray image $\boldsymbol{I}(x, y)$, the Gabor feature are extracted as follows:

$$Gabor_{uv}(x, y) = |\boldsymbol{I}(x, y) * \varphi_{u,v}(x, y)| . \quad (1)$$

Where $\varphi_{u,v}(x, y)$ are a series of Gabor filters. u and v denotes the orientation and scale of the Gabor filters, respectively. The $*$ denote the convolution operation. The $|\cdot|$ is a magnitude operator. In our method, the Gabor filters are construed with 6 orientation and 10 scales and then conducted on the intensity image. Thus, each pixel (x, y) is represented by $6 \times 10 = 60$ dimensions vector.

2.1.2. Color Feature

Given a color image with three color channels i.e., $\boldsymbol{R}$, $\boldsymbol{G}$ and $\boldsymbol{B}$, the color feature can be extracted as follows which is similar to [8],

$$Color(x, y) = [\phi(\boldsymbol{R}(x, y)), \phi(\boldsymbol{G}(x, y)), \phi(\boldsymbol{B}(x, y))]^T, \quad (2)$$

where $\phi(\boldsymbol{R}(x, y)) = [\boldsymbol{R}(x, y), |\boldsymbol{R}_x(x, y)|, |\boldsymbol{R}_y(x, y)|, |\boldsymbol{R}_{xx}(x, y)|, |\boldsymbol{R}_{yy}(x, y)|]$. $\phi(\boldsymbol{G}(x, y))$ and $\phi(\boldsymbol{B}(x, y))$ are likewise. The $|\boldsymbol{R}_x(x, y)|$, $|\boldsymbol{R}_y(x, y)|$, $|\boldsymbol{R}_{xx}(x, y)|$, $|\boldsymbol{R}_{yy}(x, y)|$ denotes the norms of first-order derivative and second-order derivatives of intensities in x and y direction, respectively. As a consequence, a $5 \times 3 = 15$ dimensions color feature is built for each pixel.

Finally, the Gabor feature and color feature are stacked together to represent the original image.

2.2. Covariance Matrices Construction

Assume the obtained stacked feature of a image is $\boldsymbol{S} \in \mathbb{R}^{M \times N \times D}$, where M and N are two spatial dimensions and D denotes the feature dimension (In this paper $D = 75$), the covariance matrix of the stacked feature is constructed as follows:

$$\boldsymbol{C} = \frac{1}{K-1} \sum_{k=1}^{K} (\boldsymbol{s}_k - \boldsymbol{\mu})(\boldsymbol{s}_k - \boldsymbol{\mu})^T, \quad (3)$$

where $\boldsymbol{s}_k$ is a D-dimension feature vector and $\boldsymbol{\mu}$ denotes the mean vector of feature vectors set $\{\boldsymbol{s}_k\}_{k=1,...,K}$, $K = M \times N$. Note that, though the CM is quite simple, it has been widely used for many vision tasks, such as texture classification [8] and image set classification [9] due to its powerful representation ability.

2.3. KLDA Based Label Assignment

Since the CM lies on a manifold space which can not be directly processed by the algorithm designed for vector space, the KLDA method is introduced to deal with the CM. Assuming the $\{\boldsymbol{C}_j\}_{j=1,...,m}$ is a set of training samples which belongs to c classes. Each class has m_k samples and $\sum_{k=1}^{c} m_k = m$. Let $\psi(\cdot)$ denotes a nonlinear mapping which can transform the CM into a high dimension space F, a kernel function can be define as the inner product in the F, i.e., $k(\boldsymbol{C}_i, \boldsymbol{C}_j) = \langle \psi(\boldsymbol{C}_i), \psi(\boldsymbol{C}_j) \rangle$. The KLDA is a supervised algorithms which aims to find $c-1$ project planes $\boldsymbol{P} = [\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_{c-1}]$ to enlarge the distance among interclass while reducing the distance intraclass by solving following maximum problem:

$$\boldsymbol{A}_{opt} = \underset{\boldsymbol{A}}{\operatorname{argmax}} \frac{\boldsymbol{A}^T \boldsymbol{K} \boldsymbol{W} \boldsymbol{K} \boldsymbol{A}}{\boldsymbol{A}^T \boldsymbol{K} \boldsymbol{K} \boldsymbol{A}}, \quad (4)$$

Fig. 2. The examples of UC Mecerd land use data set.

where $\boldsymbol{K}$ is the kernel matrix and $\boldsymbol{K}_{ij} = k(\boldsymbol{C}_i, \boldsymbol{C}_j)$ and $\boldsymbol{W}$ is a affine matrix which is defined as below.

$$\boldsymbol{W}_{ij} = \begin{cases} \frac{1}{m_k}, & \text{if both } \boldsymbol{C}_i \text{ and } \boldsymbol{C}_j \text{ belong to } k\text{-th class} \\ 0, & \text{otherwise} \end{cases}. \quad (5)$$

The optimal $\boldsymbol{A}$ can be obtained by grouping the c-1 eigenvectors associate with c-1 largest eigenvalues of matrix $(\boldsymbol{KK})^{-1}(\boldsymbol{KWK})$ and $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_{c-1}]$ and then the $\boldsymbol{P}$ can be obtained by the representer theorem as follow [9].

$$\boldsymbol{P} = [\psi(\boldsymbol{C}_1), \psi(\boldsymbol{C}_2), ..., \psi(\boldsymbol{C}_m)]\boldsymbol{A}. \quad (6)$$

For a given test sample $\boldsymbol{C}_t$, the corresponding $c-1$ dimension projection $\boldsymbol{f}_t$ on the discriminative subspace can be obtained by:

$$\boldsymbol{f}_t = \boldsymbol{A}^T \boldsymbol{K}_t, \quad (7)$$

where $\boldsymbol{K}_t = [k(\boldsymbol{C}_1, \boldsymbol{C}_t), k(\boldsymbol{C}_2, \boldsymbol{C}_t), ..., k(\boldsymbol{C}_m, \boldsymbol{C}_t)]^T$. The projection is conducted on both the training and test sets. The nearest neighboring with Euclidean distance is used to assign the label for the test samples. Note that, in our method, the Log-Euclidean distance bases Gaussian kernel is adopted which is defined as bellow [10]:

$$k(\boldsymbol{C}_i, \boldsymbol{C}_j) = \exp(-\beta \left\| \text{logm}(\boldsymbol{C}_i) - \text{logm}(\boldsymbol{C}_j) \right\|_F^2), \quad (8)$$

where the logm denotes the matrix logarithms operation and $\text{logm}(\boldsymbol{C}) = \boldsymbol{U}\text{log}(\boldsymbol{\Sigma})\boldsymbol{U}^T$. $\boldsymbol{C} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^T$ is the eigen-decomposition of $\boldsymbol{C}$. β is a scalar parameter which fixed to be 0.02 in this paper.

3. EXPERIMENTAL RESULTS

3.1. Data Sets

To evaluate the performance of the proposed method, we conduct the experiments on a widely used land-used data set i.e., the UC Merced land use data set [4]. This data set is composed of 21 land use scene classes including the agricultural, airplane, baseball diamond and son on. Each class consists of 100 aerial images measuring 256×256 pixels, with a spatial resolution of 0.3m per pixel in the red green blue color space. Some examples of this data set are shown in Fig. 2. For each class, 80 samples are randomly selected for training, while reminding samples for test.

3.2. Comparison Methods

To validate the performance of proposed CMFF framework. Following methods are used for comparison.

- Firstly, two variants of CMFF are considered, i.e., only Gabor feature or only color feature are used to construct the CM, while reminds other operations unchanged. The two variants of methods are denoted by Gabor+CM and Color+CM, respectively.

- Then, two typical LLF oriented methods, i.e., Gabor feature (Gabor-LLF)[4] and color histogram (CH-LLF)[4] are considered. For the Gabor-LLF, the HSR image are represented by mean and standard deviation extracted of Gabor feature. For color histogram, the histogram of RGB color space are used to characterize the HSR image.

- Thirdly, two classical MLF oriented methods are investigated, the SPCK [5] and PSR [6]. For the SPCK, the BoVW model with SPM is used to extract MLFs from the keypoint SIFT feature. For the PSR, the BoVW with pyramid of spatial relatons is utilized to extract MFLs from the dense SIFT feature.

- Finally, we compared the proposed method with two popular HFL oriented framework [7]. i.e., the CaffeNet and GoogleNet. Both two networks are trained from scratch or fine-tune strategy, respectively.

3.3. Results Comparison

Table 1 shows the overall classification results of all test methods. From Table 1, we can observe that firstly, the CMFF method shows obvious improvements over the Gabor-CM method and Color-CM method with the improvements of 7.12% and 7.11% respectively. This is mainly due to that Gabor feature and color feature can characterize the image from different aspects (the Gabor feature mainly focuses on texture information, while color feature represent the color intensity information) which can offer complementary information and with the CMFF framework, the complementary information among the Gabor feature and color feature are sufficiently utilized. Moreover, Fig. 3 shows per class accuracies obtained by the proposed method and Gabor-CM and Color-CM. As can be seen, the proposed CMFF method can outperform the Gabor-CM and Color-CM almost over all classes. For instance, over 5th class (buildings), both the classification results of Gabor-CM and Color-CM are less than 60%, while the accuracy of CMFF is over 70%, which

Table 1. The average accuracy (in %) of ten repeated experiments obtained from different methods.

Method	Gabor-LLF	CH-LLF	SPCK	PRS	CaffeNet		GoogleNet		Gabor-CM	Color-CM	CMFF
					From scratch	Fine-tune	From scratch	Fine-tune			
Accuracy	76.91	76.71	77.38	89.10	85.71	95.48	92.86	97.10	84.02	84.03	**91.14**

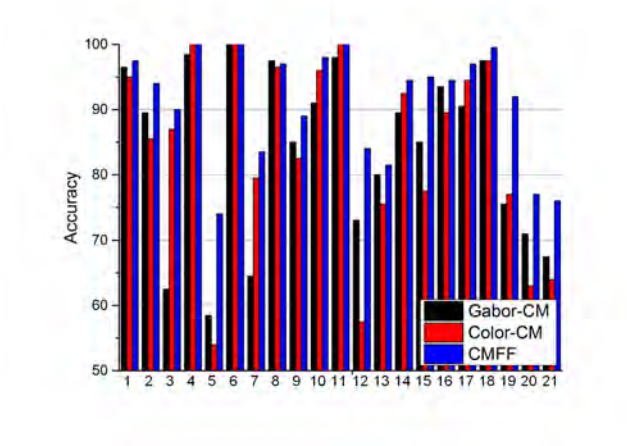

Fig. 3. Per-class accuracies comparison among the proposed CMFF, the Gabor-CM and Color-CM.

can also verify that the CMFF can utilize the complementary information among these LLFs.

In addition, from Table 1 the proposed CMFF method is superiority to the Gabor-LLF, CH-LLF, SPCK, and PSR. For example, the accuracy of SPCK and PSR are 77.71% and 89.10% respectively, while the CMFF can achieve 91.14%. Indeed, both the CaffeNet and GoogleNet with fine-tune strategy show better result than the proposed method. We highlight that, the CMFF is a generalization framework which can be extended to combine the HLF as well, which would be a very interesting future work.

4. CONCLUSION

In this paper, we propose a new fusion framework that uses the covariance matrix for remote sensing scene image classification. The proposed method consists of following three steps: Feature extraction, covariance matrix construction and KLDA based label assignment. In the first stage, both the Gabor and color feature are extracted. Then, the covariance matrix is extracted to fuse these two types of feature. Finally, the KDLA is adopted to search transformation planes to separate the CMs. Experiments on the UC Mecerd land use data demonstrate that the CMFF method can not only enhance the classification performance of the LLF, but also outperform several classical MLF oriented methods. In the future, we will work towards combining the CMFF with MLF and HLF.

5. REFERENCES

[1] L. Fang, N. He, S. Li, P. Ghamisi, and J. A. Benediktsson, "Extinction profiles fusion for hyperspectral images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1803–1815, Mar. 2018.

[2] L. Fang, N. He, S. Li, A. J. Plaza, and J. Plaza, "A new spatial-spectral feature extraction method for hyperspectral images using local covariance matrix representation," *IEEE Trans. Geosci. Remote Sens.*, to be published, 2018.

[3] Y. Yang and S. Newsam, "Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery," in *Proc. IEEE Int. Conf. Image Proces.*, 2008, pp. 1852–1855.

[4] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. Int. Conf. Advances Geographic Inf. Systems*, 2010, pp. 270–279.

[5] Y. Yang and S. Newsam, "Spatial pyramid co-occurrence for image classification," in *Proc. Int. Conf. Computer Vision*, 2011, pp. 1465–1472.

[6] S. Chen and Y. Tian, "Pyramid of spatial relatons for scene-level land use classification," *IEEE Trans. Geosc. Remote Sens.*, vol. 53, no. 4, pp. 1947–1957, Apr. 2015.

[7] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, "Land use classification in remote sensing images by convolutional neural networks," *arXiv*, vol. 1508, pp. 1–11, Aug. 2015.

[8] Oncel Tuzel, Fatih Porikli, and Peter Meer, "Region covariance: A fast descriptor for detection and classification," in *Proc. Eur. Conf. Comput. Vision*, 2006, pp. 589–600.

[9] R. Wang, H. Guo, L. S. Davis, and Q. Dai, "Covariance discriminative learning: A natural and efficient approach to image set classification," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2012, pp. 2496–2503.

[10] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel methods on the riemannian manifold of symmetric positive definite matrices," in *Proc. IEEE Conf. Comput. Vision Pattern Recog.*, 2013, pp. 73–80.