LETTER

Learning Discriminative Features for Ground-Based Cloud Classification via Mutual Information Maximization

Shuang LIU[†], *Nonmember*, Zhong ZHANG^{†a)}, *Member*, Baihua XIAO^{††}, and Xiaozhong CAO^{†††}, *Nonmembers*

SUMMARY Texture feature descriptors such as local binary patterns (LBP) have proven effective for ground-based cloud classification. Traditionally, these texture feature descriptors are predefined in a handcrafted way. In this paper, we propose a novel method which automatically learns discriminative features from labeled samples for ground-based cloud classification. Our key idea is to learn these features through mutual information maximization which learns a transformation matrix for local difference vectors of LBP. The experimental results show that our learned features greatly improves the performance of ground-based cloud classification when compared to the other state-of-the-art methods.

key words: *discriminative patterns, mutual information maximization, ground-based cloud classification*

1. Introduction

Ground-based cloud classification, as one of the major problems in cloud observations, has attracted much attention due to the potential value for practical applications, such as weather prediction, flight planning and disaster early warnings. To this day, ground-based clouds are still classified by human observers who have received professional training in meteorological station. However, this method relies heavily on expert domain knowledge and extensive human efforts. Moreover, it might suffer from classification ambiguities due to the diversity of multiple observers. Therefore, automatic ground-based cloud classification, which greatly reduces the human labor and yet achieves high classification accuracy, is highly desirable in the cloud observation field.

Cloud classification is a challenging issue because cloud images captured from outdoor environment are affected by weather, illumination, and noise. Based on the characteristics of cloud images, a lot of strategies have been proposed to represent the cloud images. Singh et al. [1] extracted several kinds of features including co-occurrence matrix and edge frequency for cloud classification. Calbó et al. [2] employed Fourier transformation to represent the cloud images. Heinle et al. [3] utilized spectral and textural features, such as energy and entropy for cloud classification. Liu et al. [4] proposed a tensor ensemble of ground-based

cloud sequences (eTGCS) model to represent the cloud sequences in a tensor manner. The performance of the existing classification methods rely heavily on the extracted visual features, but the discriminative ability of the features has not been carefully considered in previous studies.

Ground-based cloud images show rich high-frequency visual details and textures, which inspires us to exploit texture classification techniques for this specific problem. Recently, the bag-of-words (BoW) model has proven effective in providing texture features [5]. The BoW model focuses on the design of local texture descriptors capable of achieving local invariance [6]. These descriptors can be classified as dense or sparse. The sparse approaches such as RIFT [7] should detect the salient regions before representing the local patches. The dense approaches extracting local descriptors pixel by pixel are more popular, including MR8 [5], local binary patterns (LBP) [8], RP random features [9], and many others [10]. Among local texture descriptors, LBP [8] is one of the most promising candidates with some attractive properties. For example, it is easy to implementation, robust to illumination changes, and low computational complexity. Consequently, many LBP variants are proposed in recent literature [11], [12]. However, the above descriptors are defined in a hand-crafted way, which is unadaptable with the large variation of cloud images caused by illumination, climate and cloud deformation.

In this paper, we propose a novel method to learn discriminative features in a data-driven way for ground-based cloud classification. The idea is to learn the most discriminative features that maximize the mutual information between the local features and their labels. Concretely, we focus on LBP-like feature extraction and learn a transformation matrix for local difference vectors of LBP under the information theoretic framework. Hence, more useful cloud information for cloud classification is explored.

The rest of this paper is organized as follows. Section 2 gives a brief review of LBP. Section 3 presents our approach in detail. The experimental results and comparative evaluation are given in Sect. 4. Finally, in Sect. 5, we conclude the paper.

2. LBP Review

The LBP descriptor is a gray-scale texture operator that describes the spatial structure of the local image texture. The LBP descriptor labels each pixel in the image by computing the sign of the difference between the values of that pixel

Manuscript received December 25, 2014.

Manuscript publicized March 24, 2015.

[†]The authors are with College of Electronic and Communication Engineering, Tianjin Normal University, Tianjin, China.

^{††}The author is with State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

^{†††}The author is with Meteorological Observation Centre, China Meteorological Administration (CMA), Beijing, China.

a) E-mail: zhong.zhang8848@gmail.com

DOI: 10.1587/transinf.2014EDL8252

and its neighboring pixels. The result is a decimal number computed by the obtained binary string. Then the image can be represented by the histogram of these decimal numbers. Formally, given a pixel g_c , the LBP descriptor first computes the differences between its value and those of its neighboring pixels

$$\mathbf{d} = [d_0, \dots, d_{P-1}]^T = [g_0 - g_c, \dots, g_{P-1} - g_c]^T \quad (1)$$

where $\mathbf{d} \in \mathbb{R}^{P \times 1}$ is a local difference vector, P is the total number of involved neighboring pixels, and g_c and g_p are the gray values of the central pixel and neighboring pixels, respectively. The LBP value for the central pixel g_c is computed as

$$LBP_{P,R}^{ri} = \min_{0 \leq l < P} \left\{ \sum_{p=0}^{P-1} s(d_p) \times 2^{[(p+l) \bmod P]} \right\} \quad (2)$$

where R is the radius of the circle which determines the distance between the neighboring and the central pixels. The step function $s(x)$ is described with $s(x) = 1$ if $x \geq 0$ and $s(x) = 0$ otherwise. The minimum value in Eq. (2) denotes the label of the rotation invariant LBP at the central pixel.

Let N denote the total number of rotation invariant LBP patterns. According to the definition in Eq. (2), the value of N is determined by neighboring samples P . For example, when $P = 8$, $N = 36$. In order to reduce the interference of noise, Ojala et al. [8] defined the U value at each pixel as the number of bitwise transitions between 0 and 1 in the LBP

$$U(LBP_{P,R}) = |s(d_{P-1}) - s(d_0)| + \sum_{p=1}^{P-1} |s(d_p) - s(d_{p-1})| \quad (3)$$

The uniform LBP only includes those pixels with no more than two transitions (i.e. $U \leq 2$) in the adjacent binary presentation. For example, the LBP pattern '00000000' has a U value of 0 and '00010000' has a U value of 2. While the pattern '11010111' would not be included among the uniform LBP because it has four transitions. Accordingly, a uniform rotation invariant pattern is defined as

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s(d_p), & U(LBP_{P,R}) \leq 2 \\ P + 1, & \text{otherwise} \end{cases} \quad (4)$$

where the superscript *riu2* means uniform rotation invariant patterns with $U \leq 2$.

3. Approach

3.1 Maximizing Discrimination

The LBP descriptor, however, lacks of discriminative information due to its handcrafted design, which is unadaptable with the large variation of cloud images caused by illumination, climate and cloud deformation. Therefore, we incorporate the discriminative learning into LBP features. Specifically, we learn a transformation matrix for local difference vectors of LBP under the information theoretic framework

$$\mathbf{q} = M\mathbf{d} \quad (5)$$

where $M \in \mathbb{R}^{P \times P}$ is a transformation matrix. Intuitively, the purpose of discriminative learning is to find a transformation matrix M so that the local difference vectors of cloud images from different classes are enlarged. For convenience, we first discuss a two-class problem, and the multi-problem can be handled as a set of two-class problems using one versus all approach. With the help of mutual information, the problem of discriminant maximization is formulated as

$$\max_M I(\mathbf{q}_{all}; c) \quad (6)$$

where $\mathbf{q}_{all}$ is a set of labeled local difference vectors, and $c \in \{1, -1\}$ is the class label. I is the mutual information which measures the degree of dependence between two random variables. Equation (6) can be rewritten on the basis of the differential entropy H

$$\begin{aligned} I(\mathbf{q}_{all}; c) &= H(\mathbf{q}_{all}) - H(\mathbf{q}_{all} | c) \\ &= H(\mathbf{q}_{all}) - P(c = 1)H(\mathbf{q}_S) - P(c = -1)H(\mathbf{q}_T) \end{aligned} \quad (7)$$

where $\mathbf{q}_S$ and $\mathbf{q}_T$ are the set of local difference vectors computed from positive and negative samples respectively. To solve Eq. (7), we approximate differential entropy H using a set of $\mathbf{q}$. In the assumption of $\mathbf{q}$ drawn from a Gaussian distribution, we obtain

$$H(\mathbf{q}) = \frac{1}{2} \ln((2\pi)^P \det \Sigma) \quad (8)$$

where the covariance matrix Σ can be estimated from $\mathbf{q}$. In addition, we assume that the prior probabilities for the two classes are equal, so Eq. (7) can be approximated by

$$I(\mathbf{q}_{all}; c) \approx \ln \det \Sigma_{all} - \frac{1}{2} \ln \det \Sigma_S - \frac{1}{2} \ln \det \Sigma_T \quad (9)$$

where Σ_{all} , Σ_S , and Σ_T are covariance matrices computed from all labeled local difference vectors, the positive local difference vectors, and the negative local difference vectors respectively.

3.2 Optimization Algorithm

We present the optimization algorithm for the objective function Eq. (7), and denote it as $J(M)$. We employ a greedy axis-rotating algorithm which iteratively seeks the transformation matrix M so as to maximize $J(M)$. Concretely, a matrix $R(t) \in \mathbf{SO}(P)$ is introduced so that the objective value at step t is $M(t) = R(t)M(t-1)$, where $\mathbf{SO}(P)$ is the P -dimensional special orthogonal group. Here $\mathbf{SO}(P)$ corresponds to a set of rotation operations in $\mathbb{R}^P$, and therefore the resulting $M(t)$ will be orthonormal matrix as well. Essentially, we find $R(t)$ to provide a steep ascent in J . According to the Lie algebra, the optimal rotation direction for M can be found by:

$$R_n = \exp \left(n\mu \sum_{i,j} (E_{i,j} - E_{j,i}) \right) \quad (10)$$

Algorithm 1: Greedy Axis Rotation

Input: $M(0), \mu > 0, N > 0$
Output: $M(t)$

```

1 Initialize the initialization of  $M(0)$  is described in Sect. 3.3;
2 while 1 do
3   If  $M(t) = M(t-1)$ 
4     break;
5   a.  $n$  is calculated by
6      $n^* = \arg \max_{0 \leq n \leq N} J(R_n M(t-1))$ 
7   b. The optimal rotation direction  $R_n$  can be computed by
8     Eq. (10)
9   c.  $R(t) = R_{n^*}$ 
10  d.  $M(t) = R(t)M(t-1)$ 
11 end

```

where $2 \leq i \leq P$, and $i+1 \leq j \leq P$. Here μ is step length, n is the step number for searching optimal rotation direction, and $E_{i,j}$ is a matrix whose (i, j) -th element is one and all others are zero. n is obtained by

$$n^* = \arg \max_{0 \leq n \leq N} J(R_n M(t-1)) \quad (11)$$

The iterative algorithm terminates when $M(t) = M(t-1)$. The above process is illustrated in Algorithm 1. The mathematical principle behind this algorithm and the details can be found in [13].

3.3 Implementation Details

It is important to choose a good initialization for M . We employ the basis of principal subspaces of local difference vectors as the initialization of M . For a K -class cloud classification problem, we learn K binary one-against-all models as described above. After learning the transformation matrix M , we replace the local difference vectors $\mathbf{d}$ with $\mathbf{q}$ by Eq. (5). Afterwards, we calculate uniform rotation invariant patterns of all pixels for each cloud image by Eq. (4) and each cloud image is represented by the histogram of these uniform rotation invariant patterns. We use SVM as classifier. The cloud image label is determined by selecting the model whose SVM outputs the maximum response.

4. Experimental Results

In this section, we experimentally compare the proposed method with other state-of-the-art methods on two challenging databases: Kiel database and IapCAS-E database. The compared methods include 1) the fusion of texture and structure (FUS1) [3] which extracts spectral and simple textural features as the final representation; 2) the fusion of texture and structure (FUS2) [14] which utilizes the color census transform to capture texture and local structure information; 3) LBP [8] which has been introduced in Sect. 2; 4) LBP without discriminative learning (LBP_NDL), i.e. initialization of M . For the fair comparison, we use the same experimental setup for all experiments. Each cloud image is normalized to have an average intensity 128 and a standard

Fig. 1 (a) Cloud samples from the Kiel database. (b) Cloud samples from the IapCAS-E database. 1. cumulus; 2. cirrus and cirrostratus; 3. cirrocumulus and altocumulus; 4. clear sky; 5. stratocumulus; 6. stratus and altostratus; 7. cumulonimbus and nimbostratus.

Table 1 The average accuracy values (%) on the Kiel database.

	1/2	1/3	1/4
FUS1 [3]	82.64	81.55	73.69
FUS2 [14]	86.13	83.25	79.76
LBP [8]	85.35	82.27	77.31
LBP_NDL	88.62	87.35	83.40
Ours	93.31	91.83	90.26

	1	2	3	4	5	6	7
1	91.06	4.67	3.35	0.92	0.00	0.00	0.00
2	2.97	94.18	2.85	0.00	0.00	0.00	0.00
3	2.48	5.23	90.42	0.00	1.87	0.00	0.00
4	1.48	0.00	0.00	98.52	0.00	0.00	0.00
5	0.00	0.00	3.36	0.00	90.12	1.54	4.98
6	0.00	0.00	0.00	0.00	0.00	96.43	3.57
7	0.00	0.00	1.78	0.00	2.65	3.07	92.50

Fig. 2 Confusion table of our method on the Kiel database.

deviation 20. We set the radius of circle $R = 1$ and the number of neighboring pixels $P = 8$. In the experiments, 1/2, 1/3, 1/4 samples are randomly chosen from each class as training samples while the remaining ones are used for testing, and the partition process is implemented 50 times independently.

We first evaluate the effectiveness of the proposed method on the Kiel database provided by Kiel University in Germany. The Kiel database contains seven classes which is divided in terms of the international cloud classification system published in WMO. This database contains 1500 cloud images and each class samples are shown in Fig. 1 (a). The average accuracy values on the Kiel database are shown in Table 1, and the confusion table of recognition results on the Kiel database is shown in Fig. 2. The following four points can be drawn through analyzing the experimental results. First, the proposed method achieves the highest classification accuracy in each case. Second, the performance of our

Table 2 The average accuracy values (%) on the IapCAS-E database.

	1/2	1/3	1/4
FUS1 [3]	73.85	71.54	69.31
FUS2 [14]	79.09	74.26	70.65
LBP [8]	77.34	73.21	70.78
LBP_NDL	80.75	77.32	72.50
Ours	84.63	82.56	81.47

1	81.51	7.86	7.32	3.31	0.00	0.00	0.00
2	7.63	84.12	8.25	0.00	0.00	0.00	0.00
3	5.58	7.35	83.31	0.00	3.76	0.00	0.00
4	5.16	2.17	0.00	92.67	0.00	0.00	0.00
5	0.00	0.00	5.24	0.00	85.45	3.53	5.78
6	0.00	0.00	0.00	0.00	2.02	83.16	14.82
7	0.00	0.00	4.13	0.00	7.26	6.38	82.23

Fig. 3 Confusion table of our method on the IapCAS-E database.

method is over 7% higher than that of LBP in all the cases because the proposed method learns a transformation matrix for local difference vectors. Third, comparing the proposed method with LBP_NDL approach, we can see that the former is over 4% higher than the latter one on the accuracy. It demonstrates the effectiveness of discriminative learning for the transformation matrix under an information theoretic framework. Finally, when 1/4 samples are chosen as training samples, the improvement of the proposed method over other methods is more obvious than other cases. It is obvious that the proposed method can also obtain discriminative information even if the training samples are insufficient.

The IapCAS-E database, which is provided by the Institute of Atmospheric Physics, Chinese Academy of Sciences, contains seven classes and 2000 ground-based cloud images. The database is more challenging because it has large intra-class variation. Figure 1 (b) shows samples from different classes. The experiment results in Table 2 and Fig. 3 show that our method achieves the best results even on a challenging database. We draw the similar conclusions with that on the Kiel database, which proves the effectiveness of the proposed method once again.

5. Conclusions

This paper proposes a learning-based discriminative feature descriptor for ground-based cloud classification. It enhances the performance of cloud classification by introducing the discriminative learning into the step of LBP-like feature

extraction. Specifically, we learn a transformation matrix for local difference vectors under the information theoretic framework so that the mutual information between the local features and their labels is maximized. The experimental results clearly demonstrate the superiority of the proposed method over previous methods in ground-based cloud classification.

Acknowledgment

This work is supported by National Natural Science Foundation of China under Grant No. 61401309 and No. 61271429, and Doctoral Fund of Tianjin Normal University under Grant No. 52XB1405 and No. 5RL134.

References

- [1] M. Singh and M. Glennen, "Automated ground-based cloud recognition," *Pattern Analysis and Applications*, vol.8, no.3, pp.258–271, 2005.
- [2] J. Calbó and J. Sabburg, "Feature extraction from whole-sky ground-based images for cloud-type recognition," *Journal of Atmospheric and Oceanic Technology*, vol.25, no.1, pp.3–14, 2008.
- [3] A. Heinle, A. Macke, and A. Srivastav, "Automatic cloud classification of whole sky images," *Atmospheric Measurement Techniques*, vol.3, no.1, pp.557–567, 2010.
- [4] S. Liu, C. Wang, B. Xiao, Z. Zhang, and X. Cao, "Tensor ensemble of ground-based cloud sequences: Its modeling, classification, and synthesis," *IEEE Geosci. Remote Sens. Lett.*, vol.10, no.5, pp.1190–1194, 2013.
- [5] M. Varma and A. Zisserman, "A statistical approach to texture classification from single images," *International Journal of Computer Vision*, vol.62, no.1-2, pp.61–81, 2005.
- [6] M. Crosier and L. Griffin, "Using basic image features for texture classification," *International Journal of Computer Vision*, vol.88, no.3, pp.447–460, 2010.
- [7] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.27, no.8, pp.1265–1278, 2005.
- [8] T. Ojala, M. Pietikainen, and T. Maenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.24, no.7, pp.971–987, 2002.
- [9] L. Liu and P. Fieguth, "Texture classification from random features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.34, no.3, pp.574–586, 2012.
- [10] S. Liu, C. Wang, B. Xiao, Z. Zhang, and Y. Shao, "Soft-signed sparse coding for ground-based cloud classification," *International Conference on Pattern Recognition*, pp.2214–2217, 2012.
- [11] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol.19, no.6, pp.1635–1650, 2010.
- [12] J. Ren, X. Jiang, and J. Yuan, "Noise-resistant local binary pattern with an embedded error-correction mechanism," *IEEE Trans. Image Process.*, vol.22, no.10, pp.4049–4060, 2013.
- [13] B.C. Hall, *Lie groups, Lie algebras, and representations: An elementary introduction*, Springer, 2003.
- [14] W. Zhuo, Z. Cao, and Y. Xiao, "Cloud classification of ground-based images using texture-structure features," *Journal of Atmospheric and Oceanic Technology*, vol.31, no.1, pp.79–92, 2014.