mCLOUD: A Multiview Visual Feature Extraction Mechanism for Ground-Based Cloud Image Categorization

YANG XIAO AND ZHIGUO CAO
National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan, China

WEN ZHUO
School of Electrical Engineering, Southwest Jiaotong University, Emeishan, China

LIANG YE AND LEI ZHU
National Key Laboratory of Science and Technology on Multi-Spectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan, China

(Manuscript received 31 December 2014, in final form 4 October 2015)

ABSTRACT

In this paper, a novel Multiview CLOUD (mCLOUD) visual feature extraction mechanism is proposed for the task of categorizing clouds based on ground-based images. To completely characterize the different types of clouds, mCLOUD first extracts the raw visual descriptors from the views of texture, structure, and color simultaneously—specifically, the scale invariant feature transform (SIFT), the census transform histogram (CENTRIST), and the statistical color features are extracted, respectively. To obtain a more descriptive cloud representation, the feature encoding of the raw descriptors is realized by using the Fisher vector. This is followed by the feature aggregation procedure. A linear support vector machine (SVM) is employed as the classifier to yield the final cloud image categorization result. The experiments on a challenging cloud dataset termed the six-class Huazhong University of Science and Technology (HUST) cloud demonstrate that mCLOUD consistently outperforms the state-of-the-art cloud classification approaches by large margins (at least 6.9%) under all the different experimental settings. It has also been verified that, compared to the single view, the multiview cloud representation generally enhances the performance.

1. Introduction

Ground-based cloud observation is an important way to acquire information, such as cloud height, cloud cover (or cloud fraction), and cloud type (Zhuo et al. 2014). These cloud attributes are essential for many cloud-related research studies. Moreover, the China Meteorological Administration specifies that these basic meteorological elements should be observed and recorded by weather stations (CMA 2003).

To date, cloud height (Allmen and Kegelmeyer 1996; Seiz et al. 2002; Kassianov et al. 2005) and cloud cover

(Pfister et al. 2003; Long et al. 2006; Kalisch and Macke 2008) can be automatically measured and estimated accurately. However, cloud type categorization is still a challenging task that is currently performed by humans. This is why automatic categorization can reduce the cost significantly. Recently, many efforts (Isosalo et al. 2007; Calbó and Sabburg 2008; Heinle et al. 2010; Liu et al. 2011; Zhuo et al. 2014) have been devoted to address this problem. Unfortunately, the performance is still not satisfactory enough. Indeed, the classification accuracy of all the methods proposed cannot exceed 80% in some challenging cases (Zhuo et al. 2014). Thus, accurate cloud type classification is still an active research area.

Ground-based cloud image categorization can be divided in three steps: image preprocessing, feature extraction, and classification. The work presented in this paper focuses on the feature extraction step. Multiview CLOUD (mCLOUD), a novel multiview visual feature
extraction mechanism for ground-based cloud image categorization, is presented. It successfully captures the cloud characteristics from texture, structure, and color simultaneously. It is worth noting that, in this context, the term view specifically indicates the visual feature extraction perspective (i.e., texture, structure, or color). Compared to the previous methods, mCLOUD leads to a more descriptive cloud representation that significantly increases the classification performance.

Our key research motivation is based on the observation that texture, structure, and color can be complementary powerful clues for cloud type representation. This is demonstrated in experiments. It can therefore be argued that, to completely characterize the different kinds of clouds, using the extracting visual feature from the single view only is not sufficient.

Nevertheless, some of the existing cloud feature extraction methods rely on this single-view paradigm. For instances, Singh and Glennen (2005) and Isosalo et al. (2007) applied texture features including local binary patterns (LBP), local edge patterns (LEP), a co-occurrence matrix-based feature, and a run-length encoding feature, among others, to cloud classification. Although many different kinds of texture features have already been employed for cloud description, their performance was still not satisfactory enough. Indeed, the classification accuracy of the texture features presented above is generally lower than 73% on our test datasets (Zhuo et al. 2014). It seems that the cloud type cannot be well characterized only by texture. On the other hand, Liu et al. (2011) investigated the effectiveness of structure features (i.e., cloud fraction, edge sharpness, and cloud maps and gaps) for infrared cloud categorization. Although many different kinds of texture features have already been employed for cloud description, their performance was still not satisfactory enough. Indeed, the classification accuracy of the texture features presented above is generally lower than 73% on our test datasets (Zhuo et al. 2014). It seems that the cloud type cannot be well characterized only by texture. On the other hand, Liu et al. (2011) investigated the effectiveness of structure features (i.e., cloud fraction, edge sharpness, and cloud maps and gaps) for infrared cloud categorization. Unfortunately, the accuracy is still relatively low. To enhance the performance, researchers have tried to use the multifeature perspective. Specifically, besides the texture features, Buch et al. (1995), Calbó and Sabburg (2008), and Heinle et al. (2010) also employed the structure (e.g., fractional sky cover, cloud brokenness, and position information) and the color descriptors (e.g., color intensity, color skewness, and color difference) to represent the clouds. More recently, Zhuo et al. (2014) proposed to extract the texture–structure cloud feature.

That is, the census transform histogram (CENTRIST) (Wu and Rehg 2011) is applied by Zhuo et al. (2014) to capture the cloud’s texture and structure information simultaneously, combing it with the multichannel feature extraction mechanism (Xiao et al. 2014). Zhuo et al. (2014) have achieved the state-of-the-art ground-based cloud image classification performance. It somewhat demonstrates the effectiveness of multiview cloud description. However, according to the conclusion drawn by Wu and Rehg (2011), CENTRIST tends more to capture the rough structures within an image rather than texture.

As a result, CENTRIST does not reveal the cloud’s texture information well. Thus, Zhuo et al.’s (2014) proposition cannot be fully regarded as a multiview feature extraction paradigm.

To address the drawbacks of the existing feature extraction approaches, mCLOUD is proposed to jointly extract the texture, structure, and color visual descriptors. More specifically, the scale invariant feature transform (SIFT) (Lowe 2004) is employed to describe texture, CENTRIST is extracted to characterize structure, and we also introduce some statistical color features to use the color information. After achieving the multiview raw visual descriptors, we propose to further leverage their descriptive power via Fisher vector (FV) encoding (Sánchez et al. 2013). To our knowledge, we are the first to apply FV to cloud categorization. The Fisher vectors extracted from the different feature perspectives are then fused by concatenation to generate the final cloud visual representation for the classifier.

It should be noted that the type of classifier used will also affect the final cloud classification performance. According to Zhuo et al. (2014), the support vector machine (SVM) (Cortes and Vapnik 1995) is the optimal choice for cloud categorization. This is why we use a linear SVM to execute the final decision-making. The main technical pipeline of mCLOUD is shown in Fig. 1.

Finally, mCLOUD is tested on a challenging ground-based cloud image dataset—the six-class Huazhong University of Science and Technology (HUST) cloud
collected by Zhuo et al. (2014)—to verify its effectiveness and robustness. A wide-range comparison with state-of-the-art approaches has been made. In addition, the advantage of characterizing the cloud types from multiple feature perspectives and FV encoding is also investigated.

The main contributions of this paper can be summarized as follows:

- **mCLOUD** is a novel multiview visual feature extraction mechanism that captures a cloud’s texture, structure, and color information simultaneously. It significantly outperforms state-of-the-art approaches.

- The effectiveness of the Fisher vector for acquiring cloud representation is demonstrated.


The rest of this paper is organized as follows. Section 2 illustrates the methodology. Section 3 presents the experiment and a discussion. Section 4 concludes the paper and discusses future work.

## 2. Methodology

The main pipeline of the proposed cloud image categorization method is shown in Fig. 1. For the unlabeled cloud images, mCLOUD first extracts the multiview cloud representation. Then, the cloud features are used as input into an SVM to predict the cloud type. In the following subsections, mCLOUD is illustrated in detail, and the use of the SVM classifier and the implementation method are introduced.

### a. mCLOUD: A novel multiview visual feature extraction mechanism

The multiview raw visual descriptor extraction and the FV-based cloud representation calculation are the two main components of mCLOUD.

#### 1) MULTIVIEW RAW VISUAL DESCRIPTOR EXTRACTION

During the phase of raw visual descriptor extraction, SIFT, CENTRIST, and statistical color features are extracted. This phase captures a cloud’s texture, structure, and color information simultaneously. This feature extraction strategy is mainly inspired by Wu and Rehg (2011); that is, between SIFT and CENTRIST—the two visual descriptors widely used for image classification—SIFT tends to emphasize texture, while CENTRIST focuses more on the rough structures. Among them, the structure feature reveals the global cloud properties (e.g., the cloud shape). While the texture feature is good at depicting the finescale local cloud characteristics (e.g., the coarseness of cloud surface), the color feature is intrinsically complementary to both structure and texture. In other words, extracting these three descriptors simultaneously should fully characterize the clouds.

Moreover, to capture finer cloud details, these raw descriptors are extracted in a densely sampled way (Li and Fei-Fei 2007), as shown in Fig. 2. Under this paradigm, a subwindow of size \( PW \times PH \) will slide over the whole cloud image, with the horizontal stride \( SX \) and the vertical stride \( SY \). Consequently, a series of local regions (e.g., the patches of the dashed black edges in Fig. 2) can be acquired. As we can see from Fig. 2, these patches are able to reflect the fine local cloud properties. SIFT, CENTRIST, and statistical color features are extracted for each patch. Next, we will briefly introduce the definition of these raw visual descriptors.

First, SIFT is extracted by equally dividing each patch into \( 4 \times 4 = 16 \) cells, following a Gaussian smooth filtering (Nixon and Aguado 2008). Within each cell, the gradient magnitude and gradient orientation of a certain pixel are calculated as

\[
m(i,j) = \sqrt{m_x(i,j)^2 + m_y(i,j)^2},
\]

and

\[
\theta(i,j) = \tan^{-1}[m_x(i,j)/m_y(i,j)],
\]

where

\[
m_x(i,j) = I(i+1,j) - I(i-1,j),
\]

\[
m_y(i,j) = I(i+1,j) - I(i,j-1),
\]

and \( I(i,j) \) denotes the grayscale intensity of the pixel located at \( (i,j) \). Term \( \theta(i,j) \) is then softly quantized into eight orientation bins that equally divide the orientation range \([0 \ 2\pi]\), being weighted by its corresponding gradient magnitude. As a consequence, a gradient
magnitude–weighted histogram of gradient orientation (HOG) of eight bins can be acquired for each cell. By concatenating the HOGs from all 16 cells, the SIFT descriptor of $8 \times 16 = 128$ dimensionality for the sliding patch is finally extracted. The main SIFT extraction procedure is shown in Fig. 3. Since SIFT divides the cloud path into 16 cells and the HOGs are extracted from them individually, SIFT tends to describe the finescale local texture information. The details of SIFT extraction can be found in Lowe (2004).

Second, CENTRIST is extracted from the whole patch directly. It is the histogram of the census transform (CT) (Zabih and Woodfill 1994) values within an image (or image patch) that characterizes both local and global properties (Wu and Rehg 2011). For certain pixels, the census transform compares its intensity value with its eight neighboring pixels, which is illustrated in Eq. (5):

$$
\begin{array}{c|c|c|c|c|c|c|c|c|c}
65 & 10 & 29 & 0 & 1 & 0 \\
25 & 25 & 5 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\
55 & 14 & 20 & 0 & 1 & 1 \\
\end{array}
$$

(5)

That is, if the intensity value of the central pixel is higher than (or equal to) one of the neighbors, then bit 1 is set at the corresponding position; otherwise, 0 is set. Next, the eight bits are concatenated from top left to bottom right and converted to a base 10 number range in $[0, 255]$ as the CT value. For each pixel, its corresponding CT value will be calculated accordingly. CENTRIST is then yielded as a 256-dimensional histogram descriptor by calculating the occurrence probability of the CT values within the cloud patch. The main CENTRIST extraction procedure is shown in Fig. 4. The two main reasons for why CENTRIST can capture the global structure are twofold. First, CENTRIST is globally extracted from the whole cloud patch, rather than the local regions. Second, as explained by Wu and Rehg (2011), the neighboring CT values keep strong structural constraints among them. The intrinsic structure characteristics can be propagated via the CT values, which are embedded in CENTRIST. For more details, readers are encouraged to turn to Wu and Rehg (2011).

Last, the color feature is also holistically extracted from the whole cloud patch. Red–green–blue (RGB), hue, saturation, and value (HSV), Lab, and the opponent color space (Xiao et al. 2014) are employed simultaneously; that is, $3 \times 4 = 12$ different color channels are available for feature extraction. Within each color channel, the mean and standard deviation values of the patch are calculated as

$$
\text{mean} = \frac{\sum_{i=1}^{PW} \sum_{j=1}^{PH} f(i,j)}{PW \times PH},
$$

(6)
introduced in the last subsection; let from a cloud image, for example, a set of SIFT descriptors denote a set of \(v_i\), where \(u_i\), \(\mu_i\), and \(\sigma_i\) are the mixture weight, mean vector, and standard deviation vector of Gaussian \(u_i\), respectively. Following Sánchez et al. (2013), a Gaussian mixture model (GMM) is employed as the generative model that reveals the generative process of the data by us. In particular, GMM can be regarded as a "probabilistic visual vocabulary." For our cloud classification task, let \(X = \{x_i, i = 1, \ldots, L\}\) denote a set of \(D\)-dimensional local descriptors extracted from a cloud image, for example, a set of SIFT descriptors introduced in the last subsection; let \(u_i = \sum_{i=1}^{N} \omega_i u_{i_l}\) be an \(N\)-component GMM with diagonal covariance matrices and parameters, and \(\lambda = \{\omega_i, \mu_i, \sigma_i, i = 1, \ldots, N\}\), where \(\omega_i\), \(\mu_i\), and \(\sigma_i\) are the mixture weight, mean vector, and standard deviation vector of Gaussian \(u_i\), respectively. Following Sánchez et al. (2013), \(\lambda\) is estimated on a large number of training local descriptors, using the expectation–maximum (EM) algorithm (Dempster et al. 1977). Since the partial derivatives w.r.t. \(\omega_i\) is not discriminative enough (Sánchez et al. 2013), they are ignored. Let \(\gamma_i(i)\) represent the soft assignment of \(x_i\) corresponding to \(u_i\); the gradients w.r.t. \(\mu_i\) and \(\sigma_i\) are calculated as follows:

\[
G_{\mu_i} = \frac{1}{L \sqrt{\omega_i}} \sum_{i=1}^{L} \gamma_i(i) \frac{(x_i - \mu_i)}{\sigma_i},
\]

and

\[
G_{\sigma_i} = \frac{1}{L \sqrt{2 \omega_i}} \sum_{i=1}^{L} \gamma_i(i) \left[ \frac{(x_i - \mu_i)}{\sigma_i} \right]^2 - 1, \tag{8}
\]

It is worth noting that within Eqs. (8) and (9), the feature aggregation procedure is executed over the local descriptors by sum pooling (Sánchez et al. 2013) to obtain the holistic image representation. FV can be consequently generated by concatenating the vectors defined in Eqs. (8) and (9) from all the Gaussians as

\[
G_{\lambda} = [G_{\mu_1}, \ldots, G_{\mu_N}, G_{\sigma_1}, \ldots, G_{\sigma_N}]^T. \tag{10}
\]

Thus, the resulting FV is of \(2DN\) dimensionality. FV can map the \(D\)-dimensional raw descriptor to a much higher feature space. This property is beneficial to leverage the performance of the linear classifier (e.g., a linear SVM) (Vinyals et al. 2012). Following Sánchez et al. (2013), square root and L2 normalization are executed on FV to further enhance its performance. Overall, the main pipeline of FV encoding for FV-based cloud representation extraction is shown in Fig. 5. Returning to cloud classification, FV encoding is first applied to SIFT, CENTRIST, and the color feature. Then, the resulting multiview FVs are concatenated to form the final cloud representation for classification.

b. SVM classifier

As aforementioned, multiview cloud representation can be acquired through mCLOUD. The next critical issue is how to choose the classifier to decide the cloud type. For the ground-based cloud classification task, Zhuo et al. (2014) experimentally demonstrated that, among \(k\)-nearest neighbor (KNN) (Cover and Hart 1967), an SVM (Cortes and Vapnik 1995), and an artificial neural network (Singh and Glennen 2005)
multiclass cases by applying a regularization issue. However, it can be easily extended to classify the data samples into two categories; that is, for a certain unlabeled sample $x_i \in \mathbb{R}^n$, an SVM predicts its class label $y_i$ as

$$y_i = \text{sign}(\omega \times x_i + b),$$

(11)

where $\omega$ and $b$ are the pretrained SVM parameters that characterize the hyperplane for classification. An SVM is originally proposed to address the two-class categorization issue. However, it can be easily extended to multiclass cases by applying a one-versus-one or one-versus-all strategy. Readers are encouraged to turn to Cortes and Vapnik (1995) and Chang and Lin (2011) for more details on the SVM and the multiclass classification strategy.

c. Implementation details

In this subsection, some essential implementation details on mCLOUD and our cloud image categorization system will be introduced:

- mCLOUD is implemented mainly using the MATLAB programming language.
- Within mCLOUD, the multiview raw visual descriptors are densely extracted from the subwindows of three scales. The sizes are $12 \times 12$, $16 \times 16$, and $20 \times 20$. The sliding strides $PW$ and $PH$ are empirically fixed as 8 in all the cases. For the raw descriptors corresponding to each subwindow scale, FV encoding is individually executed to them.
- SIFT is implemented using the SIFT MATLAB source code released by Wang et al. (2010; http://www.itp.illinois.edu/~jyang29/LLC.htm). And, CENTRIST is constructed based on the mCENTRIST MATLAB source code released by Xiao et al. (2014; http://cs.nju.edu.cn/~upload/tpl/00/ed/237/template237/projects/mCENTRIST/mCENTRIST_v1.0.zip). They are extracted only from the grayscale channel.
- The number of Gaussians in FV is set to 50 to achieve the nonlinear ones but with much lower computational cost. Following this, a linear SVM is applied to cloud classification by us. The LIBSVM (Chang and Lin 2011; http://www.csie.ntu.edu.tw/~cjlin/libsvm) package is used to build the SVM classifier. The one-versus-one multiclass classification strategy is applied; that is, suppose the number of cloud classes is $K$, $\binom{K}{2} = \frac{K \times (K - 1)}{2}$; two-class SVM classifiers will be constructed for cloud categorization. Additionally, the penalty factor $C$ is empirically set to 1 in all the cases.

3. Experiments

To verify the effectiveness of mCLOUD, it was tested on a challenging ground-based cloud dataset, the six-class HUST cloud collected by Zhuo et al. (2014). The dataset consists of 1231 cloud images from six classes: cumulus, cirrus and cirrostratus, cirrocumulus and alto-cumulus, clear sky, stratocumulus, and stratus and altostratus. Some sample images from the different classes are shown in Fig. 6. This dataset was captured from August 2010 to May 2011 in Beijing, China, by a sky camera and stored in RGB JPEG format. The employed sky camera is composed of an ordinary charge-coupled device (CCD) and an auto iris lens (not a fish-eye lens), with the approximate viewing angle of 60°. To cover the whole sky, it is fixed on a pan–tilt platform and scans the sky with a horizontal interval of 40° and a vertical interval of 30°, yielding 28 local sky region images. In particular, the elevation angles of the test images contain 30°, 60°, and 90° from horizon to zenith; 0° is excluded because much ground scene information (e.g., buildings, trees, or grass) is captured simultaneously. Compared to the fish-eye lens cameras [i.e., whole-sky imager (WSI), total-sky imager (TSI) and all-sky imager (ASI)], this sky camera can capture cloud images similar to the human eye without the so-called fish-eye distortion (Zhuo et al. 2014). In addition, it somewhat restricts the sun disk effect. The ground-truth cloud types are mainly labeled by a meteorological observer with more than 10 years of sky observation experience, according to the cloud type characterization addressed by Heinle et al. (2010). A full description of the dataset is out of the scope of this paper. Interested readers should refer to Zhuo et al. (2014).

Following the experimental setup in Zhuo et al. (2014), the available samples were randomly split into a training set and a test set for 10 times. The average accuracy and standard deviation were computed to investigate the effectiveness and the robustness of the different methods. For a fairer comparison, the experiments were divided into five folders, corresponding to the different number of...
training samples per class. Specifically, 5, 10, 20, 40, and 80 training samples per class will be used within the five folders. For efficiency, all the images were resized to be no larger than $300 \times 300$ pixels during the phase of raw visual descriptor extraction.

The following experiments demonstrate mCLOUD’s effectiveness. Its performances were compared with the state-of-the-art approaches using two cloud categorization principles. The advantage of the multiview visual feature extraction mechanism and FV were also investigated.

a. Comparison to the state-of-the-art approaches

A wide-range comparison with the state-of-the-art ground-based cloud categorization approaches was executed. Specifically, the methods proposed by Zhuo et al. (2014), Liu et al. (2011), Heinle et al. (2010), Calbó and Sabburg (2008), and Isosalo et al. (2007) were employed for comparison. Since the experimental setup is the same, we cite their performance on the six-class HUST cloud dataset listed in Zhuo et al. (2014) directly. The classification results of the different methods are listed in Table 1. The following are observations drawn from Table 1.

- On the challenging six-class HUST cloud dataset, mCLOUD can achieve good results; that is, it significantly outperforms the state-of-the-art methods on average accuracy by large margins (at least 6.9%, and at most 10.1%), corresponding to all the training sample numbers. This result demonstrates the effectiveness and generality of mCLOUD for the ground-based cloud image categorization task. Since the performance advantage of mCLOUD is remarkable, the Student’s $t$ test (O’Mahony 1986) was not executed as by Zhuo et al. (2014) to justify mCLOUD’s effectiveness.

- Compared to the other methods, mCLOUD achieves a comparable standard deviation, without obvious superiority. Considering mCLOUD’s high

![Fig. 6. Typical ground-based cloud images from the different cloud categories. Specifically, in the nine-class cloud categorization principle, the three compounded classes—cirrus and cirrostratus, cirrocumulus and altocumulus, and stratus and altostratus—in the six-class principle are further split into six individual classes: cirrus, cirrostratus, cirrocumulus, altocumulus, stratus, and altostratus.](image-url)

<table>
<thead>
<tr>
<th>Training samples</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isosalo et al. (2007)</td>
<td>39.8 (±3.1)</td>
<td>47.7 (±4.8)</td>
<td>58.3 (±2.1)</td>
<td>66.5 (±1.3)</td>
<td>72.2 (±1.0)</td>
</tr>
<tr>
<td>Calbó and Sabburg (2008)</td>
<td>37.4 (±2.4)</td>
<td>43.3 (±3.8)</td>
<td>50.9 (±2.0)</td>
<td>57.6 (±3.0)</td>
<td>63.8 (±1.2)</td>
</tr>
<tr>
<td>Heinle et al. (2010)</td>
<td>34.4 (±2.3)</td>
<td>37.7 (±3.1)</td>
<td>44.9 (±1.9)</td>
<td>51.5 (±1.5)</td>
<td>56.8 (±1.6)</td>
</tr>
<tr>
<td>Liu et al. (2011)</td>
<td>29.0 (±2.5)</td>
<td>30.6 (±2.0)</td>
<td>33.8 (±1.5)</td>
<td>38.1 (±1.1)</td>
<td>41.1 (±1.9)</td>
</tr>
<tr>
<td>Zhuo et al. (2014)</td>
<td>45.2 (±3.6)</td>
<td>53.5 (±2.8)</td>
<td>66.2 (±2.1)</td>
<td>74.7 (±1.0)</td>
<td>79.8 (±1.2)</td>
</tr>
<tr>
<td>mCLOUD</td>
<td>55.3 (±3.9)</td>
<td>63.4 (±3.9)</td>
<td>74.6 (±2.3)</td>
<td>81.6 (±1.6)</td>
<td>87.5 (±1.6)</td>
</tr>
</tbody>
</table>

Table 1. Cloud categorization results (%) of the six-class HUST cloud dataset. The best performance of each training sample number is shown in boldface. Standard deviations are listed in parentheses. It is worth noting that, because of the same test dataset and experimental setup, the performance of the methods for comparison is directly listed as the results reported in Zhuo et al. (2014).
classification accuracy, its robustness is still acceptable for applications.

- With the increment of training samples, both of mCLOUD’s effectiveness and robustness are generally boosted, which is consistent with the conclusion drawn in Zhuo et al. (2014) that the training sample amount is a vital factor that affects the final performance. It also happens to all the other approaches. This finding is helpful to guide the construction of a practical ground-based cloud categorization system; that is, besides proposing the effective classification method, collecting sufficient training samples is another fundamental issue that should be addressed.

b. Effectiveness of multiview visual feature extraction mechanism

To demonstrate the superiority of the proposed multiview visual feature extraction mechanism over single-view paradigms, we investigated the cloud categorization performance on the six-class HUST cloud dataset using visual features from single and multiple perspectives. Table 2 lists the comparison results. The experimental setup and feature extraction pipeline were the same as in Table 1. The following are observations drawn from Table 2.

- Evidently, compared to the single-view features, describing the clouds from multiple perspectives leverages the performance remarkably in most cases. When SIFT, CENTRIST and color features are combined together, the best performance can be achieved for all the training samples. This strongly supports our key proposition that, the cloud types had better be characterized from the perspectives of texture, structure, and color simultaneously.

- Among the three kinds of visual descriptors, SIFT generally outperforms the other two, especially when the training sample amount is relatively small. It reveals the importance of texture clue for cloud characterization. Another issue worth noting is that SIFT consistently goes beyond the method proposed by Zhuo et al. (2014) that gave the best result before our work. It can demonstrate the effectiveness of the proposed feature extraction pipeline with the densely sampled sliding windows and the FV-based cloud representation. On the other hand, Zhuo et al. (2014) employed multiple color channels (i.e., opponent color space: Xiao et al. 2014) to improve the performance, SIFT and CENTRIST are extracted only from the grayscale channel by us. Thus, if additional color channels are applied to SIFT and CENTRIST, further performance enhancement is expected. However, the computational memory consumption should be first restricted into a feasible range. At the current stage, mCLOUD generally uses more than 8 GB of memory during the SVM training.

- It is interesting that, although the proposed color features are simply defined, they still yield competitive results. For the high real-time demanding applications, the color features can be regarded as the best choice. CENTRIST tends to be inferior to SIFT and color features when the training sample number increases. The reason may be that it is originally proposed as a holistic visual descriptor. Under our local extraction framework, its descriptive power seems to be weakened.

c. Effectiveness of FV encoding

In mCLOUD, FV encoding is employed to project the raw visual descriptors to the feature space of much higher dimensionality to leverage the performance. To justify the effectiveness of FV encoding, the experiments were conducted in two parts on the six-class HUST cloud dataset. In the first case, the raw visual descriptors were aggregated via sum pooling directly from the grayscale channel by us. Thus, if additional color channels are applied to SIFT and CENTRIST, further performance enhancement is expected. However, the computational memory consumption should be first restricted into a feasible range. At the current stage, mCLOUD generally uses more than 8 GB of memory during the SVM training.

[Table 2. Demonstration of the effectiveness of multiview feature extraction mechanism. SIFT, CENTRIST, and color correspond to texture, structure, and color clue, respectively. The best performance of each training sample number is shown in boldface.]

<table>
<thead>
<tr>
<th>Training samples</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIFT</td>
<td>51.3 (±2.1)</td>
<td>58.3 (±2.0)</td>
<td>68.1 (±1.6)</td>
<td>74.9 (±1.1)</td>
<td>80.3 (±1.8)</td>
</tr>
<tr>
<td>CENTRIST</td>
<td>48.6 (±3.0)</td>
<td>55.0 (±3.2)</td>
<td>63.3 (±3.3)</td>
<td>68.2 (±1.4)</td>
<td>73.2 (±1.1)</td>
</tr>
<tr>
<td>Color</td>
<td>39.5 (±6.1)</td>
<td>53.2 (±3.5)</td>
<td>64.9 (±2.3)</td>
<td>74.7 (±1.5)</td>
<td>82.7 (±2.2)</td>
</tr>
<tr>
<td>SIFT+CENTRIST</td>
<td>52.0 (±2.2)</td>
<td>59.1 (±3.2)</td>
<td>69.1 (±2.2)</td>
<td>75.8 (±1.1)</td>
<td>80.0 (±1.6)</td>
</tr>
<tr>
<td>SIFT+CENTRIST+Color</td>
<td>55.3 (±3.9)</td>
<td>63.4 (±3.9)</td>
<td>74.6 (±2.3)</td>
<td>81.6 (±1.6)</td>
<td>87.5 (±1.6)</td>
</tr>
<tr>
<td>Zhuo et al. (2014)</td>
<td>45.2 (±3.6)</td>
<td>53.5 (±2.8)</td>
<td>66.2 (±2.1)</td>
<td>74.7 (±1.0)</td>
<td>79.8 (±1.2)</td>
</tr>
</tbody>
</table>
in both single-view and multiview cases, corresponding to the different amounts of the training sample. The experimental results are listed in Table 3. The following are observations drawn from Table 3.

- In almost all the cases using the same linear SVM, FV outperforms the raw visual descriptors by large margins. This result evidently demonstrates the effectiveness of FV encoding for the performance gain.

- Compared to the high-dimensional feature mapping approach with the RBF kernel, FV encoding consistently demonstrates a significant performance advantage from the perspectives of effectiveness and stability. It is also worth noting that, the RBF kernel generally performs worse than the linear kernel. This could be due to two main reasons. First, the RBF kernel parameters are hard to tune. Although a large parameter space has been searched, the performance is still not satisfactory. Second, the amount of training samples is not sufficient enough to avoid overfitting (Chang and Lin 2011). Thus, for cloud categorization FV encoding is the most appropriate choice for high-dimensional feature mapping.

**d. Comparison with finer categorization principle**

Following Zhuo et al. (2014), we refine the three compound classes: cirrus and cirrostratus, cirrocumulus and altocumulus, and stratus and altostratus into six individual categories, including cirrus, cirrostratus, cirrocumulus, altocumulus, stratus, and altostratus. The sample images from the six newly defined classes are shown in Fig. 6. As a result, the cloud images will be categorized into nine classes in all. The finer categorization principle actually imposes more challenges to cloud classification. The classes are easy to confuse with each other. As a matter of fact, the average accuracy of Zhuo et al.’s (2014) method drastically drops from 74.7% to 64.1% under the condition that 40 training samples are used per class. mCLOUD is also tested according to the new categorization way. In particular, one round experiment was carried out. For each of the nine classes, 40 samples were randomly selected for training and the rest are used as a test, which was the same as Zhuo et al. (2014). The state-of-the-art methods in Table 1 were reemployed for comparison. Moreover, the classification results of each class and the overall accuracy were reported simultaneously. The specific results listed in Table 4 are summarized below.

- It is shown that mCLOUD outperforms all the other methods by large margins, both from the perspectives of single-class and overall performance. Concerning the average accuracy, mCLOUD (81.2%) is significantly superior to Zhuo et al.’s (2014) method (the second place one with 64.1%) by 17.1%. Compared to its result (81.6%) using the six-class categorization rule, mCLOUD’s performance drop (0.4%) is trivial. For all the individual classes, mCLOUD surpasses the other methods by at least 8.3% and at most by 23.9%. The promising results given above further justify the effectiveness of mCLOUD, as well as its robustness to the variation of the categorization principle. Thus,

**Table 3. Demonstration of the effectiveness of FV encoding.** The aggregated raw visual descriptors with linear kernel, RBF nonlinear kernel, and FV are compared. In particular, “raw” indicates the raw visual descriptors, “linear” denotes the linear SVM, and “RBF” represents the nonlinear SVM with the RBF kernel. For each descriptor or descriptor combination, the best performance of each training sample number is shown in boldface.

<table>
<thead>
<tr>
<th>Training samples</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>40</th>
<th>80</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SIFT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Raw” + “Linear”</td>
<td>29.8 (±2.9)</td>
<td>34.3 (±3.1)</td>
<td>41.2 (±2.4)</td>
<td>47.5 (±1.3)</td>
<td>53.2 (±1.8)</td>
</tr>
<tr>
<td>“Raw” + “RBF”</td>
<td>28.1 (±2.8)</td>
<td>32.1 (±1.8)</td>
<td>33.1 (±3.4)</td>
<td>34.3 (±2.5)</td>
<td>39.0 (±1.8)</td>
</tr>
<tr>
<td>FV + “linear”</td>
<td><strong>51.3 (±2.1)</strong></td>
<td><strong>58.3 (±2.0)</strong></td>
<td><strong>68.1 (±1.6)</strong></td>
<td><strong>74.9 (±1.1)</strong></td>
<td><strong>80.3 (±1.8)</strong></td>
</tr>
<tr>
<td><strong>CENTRIST</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Raw” + “linear”</td>
<td>43.9 (±3.2)</td>
<td>50.8 (±3.3)</td>
<td>57.3 (±1.5)</td>
<td>61.6 (±2.1)</td>
<td>65.1 (±1.0)</td>
</tr>
<tr>
<td>“Raw” + “RBF”</td>
<td>41.5 (±6.6)</td>
<td>46.6 (±8.3)</td>
<td>60.6 (±6.9)</td>
<td>65.5 (±4.7)</td>
<td>73.1 (±1.3)</td>
</tr>
<tr>
<td>FV + “linear”</td>
<td><strong>48.6 (±3.0)</strong></td>
<td><strong>55.0 (±3.2)</strong></td>
<td><strong>63.3 (±3.3)</strong></td>
<td><strong>68.2 (±1.4)</strong></td>
<td><strong>73.2 (±1.1)</strong></td>
</tr>
<tr>
<td><strong>Color</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Raw” + “linear”</td>
<td>46.2 (±3.7)</td>
<td>49.8 (±3.4)</td>
<td>54.7 (±2.3)</td>
<td>57.4 (±1.9)</td>
<td>59.7 (±1.7)</td>
</tr>
<tr>
<td>“Raw” + “RBF”</td>
<td>43.7 (±6.9)</td>
<td>50.5 (±7.1)</td>
<td>58.7 (±3.2)</td>
<td>63.1 (±6.2)</td>
<td>71.3 (±4.4)</td>
</tr>
<tr>
<td>FV + “linear”</td>
<td>39.5 (±6.1)</td>
<td><strong>53.2 (±3.5)</strong></td>
<td><strong>64.9 (±2.3)</strong></td>
<td><strong>74.7 (±1.5)</strong></td>
<td><strong>82.7 (±2.2)</strong></td>
</tr>
<tr>
<td><strong>SIFT + CENTRIST</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Raw” + “linear”</td>
<td>39.8 (±1.6)</td>
<td>49.7 (±3.2)</td>
<td>58.8 (±1.8)</td>
<td>63.8 (±1.8)</td>
<td>67.9 (±1.7)</td>
</tr>
<tr>
<td>“Raw” + “RBF”</td>
<td>37.9 (±3.8)</td>
<td>48.4 (±3.1)</td>
<td>53.9 (±7.0)</td>
<td>65.4 (±3.0)</td>
<td>72.5 (±2.0)</td>
</tr>
<tr>
<td>FV + “linear”</td>
<td><strong>52.0 (±2.2)</strong></td>
<td><strong>59.1 (±3.2)</strong></td>
<td><strong>69.1 (±2.2)</strong></td>
<td><strong>75.8 (±1.1)</strong></td>
<td><strong>80.0 (±1.6)</strong></td>
</tr>
<tr>
<td><strong>SIFT + CENTRIST + Color</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>“Raw” + “Linear”</td>
<td>43.4 (±2.7)</td>
<td>54.3 (±2.9)</td>
<td>64.8 (±2.5)</td>
<td>70.6 (±1.2)</td>
<td>74.6 (±1.1)</td>
</tr>
<tr>
<td>“Raw” + “RBF”</td>
<td>37.8 (±5.1)</td>
<td>47.5 (±7.6)</td>
<td>55.7 (±9.2)</td>
<td>65.5 (±8.2)</td>
<td>75.4 (±5.4)</td>
</tr>
<tr>
<td>FV + “Linear”</td>
<td><strong>53.3 (±3.9)</strong></td>
<td><strong>63.4 (±3.9)</strong></td>
<td><strong>74.6 (±2.3)</strong></td>
<td><strong>81.6 (±1.6)</strong></td>
<td><strong>87.5 (±1.6)</strong></td>
</tr>
</tbody>
</table>
mCLOUD possesses stronger potentiality for the practical applications.

- For the classes cumulus, cirrus, cirrocumulus, and stratocumulus, mCLOUD’s performance is relatively low (less than 80%). The corresponding confusion matrix for all the classes is shown in Fig. 7. In our opinion, the performance drop mainly ascribes to the finer categorization principle. Under it, the cloud classes tend to be more confused with each other. Another issue is the inherent difficulty in distinguishing between all the categories depending only on the visual information. For instance, the visual appearance of altostratus, cirrostratus, and stratus is evidently similar. Adding other discriminative clues (e.g., cloud height) would be a way to further leverage the performance. Actually, it is actually one of our future research avenues.

e. Sensitivity to color balance variation

Within mCLOUD, the color feature is used for cloud representation. Its sensitivity to the camera color balance variation needs to be addressed for the practical applications. In this subsection, we will investigate this issue from the perspectives of color gamma adjustment and white balance. In particular, the experiments were carried out in three parts on the six-class HUST cloud dataset. In the first experimental setting, the white balance method proposed by Weng et al. (2005) with randomly selected parameters were executed individually. Second, the color gamma adjustment was imposed on the images in the dataset using the randomly

<table>
<thead>
<tr>
<th>Cloud types</th>
<th>Altocumulus</th>
<th>Altostratus</th>
<th>Cirrocumulus</th>
<th>Cirrostratus</th>
<th>Cirrus</th>
<th>Cumulus</th>
<th>Stratocumulus</th>
<th>Stratatus</th>
<th>Clear sky</th>
</tr>
</thead>
<tbody>
<tr>
<td>Josse et al. (2007)</td>
<td>29.0</td>
<td>21.1</td>
<td>26.0</td>
<td>35.9</td>
<td>30.1</td>
<td>30.2</td>
<td>26.0</td>
<td>38.7</td>
<td>30.1</td>
</tr>
<tr>
<td>Calbo and Söderbäck (2008)</td>
<td>25.2</td>
<td>23.0</td>
<td>23.7</td>
<td>23.7</td>
<td>43.7</td>
<td>42.6</td>
<td>42.7</td>
<td>60.5</td>
<td>60.5</td>
</tr>
<tr>
<td>Heinle et al. (2010)</td>
<td>43.7</td>
<td>42.7</td>
<td>52.1</td>
<td>60.5</td>
<td>58.3</td>
<td>58.2</td>
<td>52.1</td>
<td>60.5</td>
<td>43.7</td>
</tr>
<tr>
<td>Liu et al. (2011)</td>
<td>60.5</td>
<td>60.5</td>
<td>52.1</td>
<td>60.5</td>
<td>58.3</td>
<td>58.2</td>
<td>52.1</td>
<td>60.5</td>
<td>60.5</td>
</tr>
<tr>
<td>Zhuo et al. (2014)</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
</tr>
<tr>
<td>mCLOUD</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
<td>75.8</td>
</tr>
</tbody>
</table>

Fig. 7. mCLOUD’s confusion matrix, which corresponds to Table 4. Only elements more than 0.1 are shown.
chosen gamma values range [0.7, 1.3]. Third, white balance and color gamma adjustment happened simultaneously with the randomly appointed parameters. The sample images that correspond to the three color balance settings are shown in Fig. 8. We can see that, due to the color balance variation, the adjusted cloud images of the same scene possess very different color appearances. It actually imposes challenges to cloud categorization. The corresponding classification results are listed in Table 5. The following are observations drawn from Table 5.

- Generally, the color balance variation will weaken the performance of the color feature and mCLOUD significantly. Compared to color gamma adjustment, the color variation yielded by the white balance leads to more serious performance drops.
- When the color balance variation is yielded by color gamma adjustment and white balance simultaneously, the categorization accuracy is lowest in all the cases.

Thus, to ensure the performance, suitable color balance calibration is required for the practical applications. On the other hand, how to address the effect of color balance variation during the phase of feature extraction is also an issue we want to focus on in the future.

4. Conclusions

In this paper, mCLOUD is proposed as a novel multi-view visual feature extraction mechanism for ground-based cloud image categorization. To leverage the performance, mCLOUD characterizes the clouds from the perspectives of texture, structure, and color simultaneously. It leads to more complete cloud descriptions. In particular, SIFT, CENTRIST, and statistical color features are extracted in a densely sampled way and are used as the raw visual descriptors. To the best of our knowledge, it is the first time that Fisher vector encoding is applied to acquire more descriptive cloud representations. Finally,

![Table 5. Cloud categorization results that correspond to the different color balance conditions and original color value space. In particular, “white balance” indicates that the color balance variation is yielded by white balance adjustment. And, “color gamma” denotes that the color balance variation is triggered by color gamma adjustment. The best performance of each training sample number is shown in boldface.](http://journals.ametsoc.org/doi/pdf/10.1175/JTECH-D-15-0015.1)
receiving the cloud features generated through mCLOUD as input, an SVM classifier is employed to construct the cloud image classification system.

On the challenging six-class HUST cloud dataset, mCLOUD significantly outperforms the state-of-the-art methods in all the experimental settings. It is worth noting that, using the finer cloud categorization principle (the nine-class rule), mCLOUD’s advantage is further highlighted. The promising experimental results demonstrate the effectiveness and robustness of mCLOUD. The superiority of multiview visual feature extraction mechanism and FV encoding was also investigated. However, for the practical applications, mCLOUD still needs improvement (e.g., sensitivity to the color balance variation).

mCLOUD is indeed a general multiview feature extraction mechanism that can be further extended. In future work, we intend to integrate more discriminative information (e.g., cloud height via binocular stereo vision, and extracting SIFT and CENTRIST from multiple imaging channels) into it. Moreover, mCLOUD is relatively computationally expensive in terms of memory consumption. We plan to address this issue in the future. Another potential research direction is to apply more powerful machine learning technologies to address the fundamental visual feature extraction issue.

Acknowledgments. This work is jointly supported by the National Natural Science Foundation of China under Grant 61502187, the Fundamental Research Funds for the Central Universities (HUST: 2014QNRC035 and 2015QN036), and the National High-Tech Research and Development Program of China (863 Program) (Grant 2015AA015904).

We thank the China Meteorological Administration for providing the ground-based cloud image dataset. We also thank Prof. Shuqing Ma and the anonymous reviewers for their constructive discussion and suggestions during the entire phase of this research.

Last, we thank Dr. Aryel Beck for his feedback on an earlier version of this paper.

REFERENCES


