CLOUD CLASSIFICATION FOR GROUND-BASED SKY IMAGE USING RANDOM FOREST

Xinrui Wan [1], Juan Du [1]

[1] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China - 2019282130141@whu.edu.cn (X.W.); dujuan_rs@whu.edu.cn (J.D.)

Commission III, WG III/8

KEY WORDS: Cloud Classification, Total Sky Imager, Muti-feature, Random Forest

ABSTRACT:

The use of solar power as a renewable energy has grown rapidly over the last few decades. However, the amount of solar radiation reaching the ground vary significantly in the short term. Clouds are the main factor. In this paper, a novel cloud detection method for ground-based sky images is proposed. First, the multiple features from the sky images, including spectral, texture and colour features are combined into a feature set. Then, Random Forest with this feature set is used to classify different types of cloud and clear sky. The experimental results show that cumulus and cirrus clouds can be identified from sky images. Combined with random forest, three types of features and various feature combinations are used for cloud classification, respectively. The classification accuracy with multiple features is higher than that of single-type features and dual-type features.

1. INTRODUTION

Some researchers have predicted that human would run out of fossil fuels within 100 years (Lackner, 2002). Therefore, it is necessary to make full use of renewable and clean energy. Solar energy is one of the most essential renewable energies, which is rich in resources, free to use and no pollution to the environment. However solar radiation reaching the earth's surface can only be used in small amount. How to effectively utilize the solar energy has become a concern.

The amount of solar energy that eventually reaches the surface is affected by a number of factors, such as latitude, season, cloud cover, atmospheric pollutants, and the sun's altitude. However clouds is the greatest impact in the short term (Chenni et al., 2007). There are two ways to obtain cloud observation data, respectively satellite and ground equipment. The satellite data are often used to observe the large-scale area of cloud. The shortage of these data is the low temporal resolution, which is only able to predict solar irradiance for a few hours or days. The ground-based equipment collects the local region's cloud observation data in a certain location. The temporal resolution of these data is 30 seconds or 1 minute, which can be used for short-term solar irradiance prediction (Tapakis, Charalambides, 2013).

There have been many cloud detection studies of ground-based sky images. Generally, they are mainly divided into two categories: threshold method and classifier method.

Owing to the significant difference between the diameters of atmospheric particles and cloud particles, they generate Rayleigh scattering and Mie scattering to sunlight respectively which result in different spectral characteristics of clear sky and cloud. Earlier researchers (Johnson, 1991; Long et al., 2006) made use of the differences between cloud and clear sky in the red and blue bands and set a fixed threshold for RBR(Red-to-Blue Ratio) to discriminate clear sky, thin cloud, and opaque cloud. On the basis of NRBR (Normalized Red-to-Blue Ratio), the flexible thresholds were set to identify clouds (Li et al., 2011; Yang et al., 2012). Considering a single RGB threshold cannot be used effectively for thin clouds, the CSL (Clear Sky Library) threshold (Shields, 2009) was established as the benchmark of clear sky. A pixel was classified as cloud if its RBR was larger than the CSL threshold. Threshold methods are the simplest and fastest classification methods, which mainly use the spectral characteristics of the image. However, the threshold methods are not accurate enough because of the diversity and complexity of information in the cloud image (Ackerman et al., 1998).

In order to overcome the shortage of threshold methods, classifier methods integrate multi-feature into cloud detection of ground-based sky image. They can get better classification results than that of threshold methods. Several features, which were statistical features, features from the Fourier transform of the image, features of the thresholded image are extracted and are used by a simple classifier based on supervised parallelepiped technique. The accuracy is 68% when eight sky conditions were considered (Calbó, Sabburg, 2008). An automatic cloud classification algorithm are proposed for seven cloud-type based on a set of statistical features describing the color as well as the texture of an image and the KNN (k-Nearest-Neighbor) classifier to achieve the high accuracy about 75% (Heinle et al., 2010). The solar zenith angle, cloud coverage and the visible fraction of solar disk were taken into account and an improved KNN algorithm was presented. The average performance of the this classifier was 87.9% (Kazantzidis et al., 2012). The texture features, color features and shape features are gathered from four different sky conditions into ANN (Artificial Neural Network), KNN, hybrid method based on KNN and ANN severally (Xia et al., 2015). A novel duplex norm-bounded sparse coding method based on norm-bounded sparse coding was more accurate in overall accuracy than neural networks and support vector machine (Gan et al., 2017). Three new features based on the Fourier transform were introduced, and the classification technique was based on ANN with tree algorithm. This method was greatly effective in distinguishing clouds from non-clouds (Kliangsuwan, Heednacram, 2018).

In this paper, the multiple features from the sky images, including spectrum, texture and color features, were chosen. Random forest was used to classify different types of cloud and clear sky because of its advantages of high accuracy, fast

running speed and high robustness. The rest of paper is organized as follows. Section 2 states the details of TSI sky images and the preprocessing of them. Section 3 describes the major features used in random forest classifier. In section 4, the analysis of the classification results is shown. Finally, a summary and suggestions for future research are given in section 5.

2. DATA AND PREPROCESSING

2.1 Data

The TSI (Total Sky Imager) is located at the SGP (Southern Great Plains) atmospheric observatory (36.6060˚N, 97.4850˚W) in Oklahoma, United States. The device has a CCD (Charge Coupled Device) imaging camera that looks down at a heated hemisphere curved mirror (Morris, 2005). The mirror reflects information from the sky into the camera lens, and the shape of the curved surface can enlarge the scope of the observed sky as much as possible. There is a sun tracking shadow band that continuously shields the mirror from direct sunlight in order to protect the camera sensor from the sun's reflection and reduce image overexposure. TSI works at a 30-sec sampling interval and saves them to JPEG files with 288×352 pixels that creates the opportunity to achieve near real-time operation.

Figure 1 shows the three types of clouds on which the research focused. Cirrus clouds and cumulus clouds are the two most common types of clouds. Cirrus clouds have a filamentous structure, which is relatively thin and better light transmittance. Cumulus clouds are thicker and have clearer boundaries. Stratus clouds are also thick, but the main difference is that they usually cover the whole sky, and last for a long time.

(a) (b) (c)

Figure 1. Three different cloud types on the TSI images: (a) Cirrus cloud; (b) Cumulus cloud;(c) Stratus cloud

2.2 Image completion

According to the center of camera lens, the images were cropped to 257*257 pixels to remove the irrelevant areas of cloud observations. As mentioned before, a shadow band is moving followed the sun. Together with the arm that supports the camera, these two regions make the whole image lose about 10% of the information. The black band formed by the arm is fixed in the image, with a narrow width and a small coverage area, while the moving shading band has a wider width and a much larger shielding area. For better computation and prediction of cloud classification and cloud movement, it is necessary to restore the images to get the complete images.

Due to the shadow band moves with the sun, the first step is to compute the sun's exactly position on the image. The solar position can be calculated for a specific location on Earth given the date and time. Using the algorithm provided by Reda and Andreas(2004), the zenith angle θ and azimuth angle φ of sun at time instance i can be obtained. Then the coordinates of sun position (X_{sun}, Y_{sun}) on the image are determined by:

$$X_{\text{sun}} = X_0 + r_S \sin\theta \tag{1}$$

$$Y_{\text{sun}} = Y_0 - r_S \cos\theta \tag{2}$$

where r_S is the sun's radial distance from the center position on the image, which can be calculated according to the relationship with the solar zenith angle.

Because of the image distortion, the relationship between r_S and the tangent of the sun zenith angle θ is nonlinear. The relationship between r_i and θ is approximated by fitting pairs of the sun's apparent radial distance and the corresponding solar zenith angle to obtain a cubic polynomial. The polynomial coefficients are obtained using the Least Square Method for data computed from images of clear sky days in 2008. The coefficient is as {0.09563, -0.2583, 0.95124, -0.01094}.

According to the sun position on the image, the mask of shadow band can then be automatically created for the image restoration, shown in Figure 2(b). Next, a method for automatically guiding patch-based image completion (Huang et al., 2014) was used. The method got the information by estimating planar projection parameters, softly segmenting the known region into planes, and discovering translational regularity within these planes. Then the above information was converted into soft constraints for the low-level completion algorithm by defining prior probabilities for patch offsets and transformations. The filled image can lastly be obtained with several iterations, showed in Figure 2(c).

(a) (b) (c)

Figure 2. Image completion: (a) original image; (b) mask of shadow band; (c) restored image

3. CLOUD CLASSIFICATION BASED ON RANDOM FOREST

3.1 Image features

In this paper, spectral features, texture features and color features were simultaneously combined to obtain a set of feature images.

3.1.1 Spectral feature

Spectral features describe the information of the image's color and tonal variation (Heinle et al., 2010). The scattering of atmospheric molecules for visible light is inversely proportional to the wavelength of the visible band. Under the condition of sunny, Rayleigh scattering of atmospheric molecules for the blue channel is far greater than that of the red channel, so the sky is blue; while the cloud particles homogeneously scatter each band of visible light, so the cloud is mainly white. This makes a great difference in the spectrum between cloud and clear sky pixels.

1) Removal of atmospheric scattering

Due to scattering of sunlight and atmospheric molecules, the brightness of the clear sky pixels on the image are dissimilar. Yang et al.(2017) proposed RAS (Removal of atmospheric scattering) which is a new composite channel by operating RGB channels. It can impair inhomogeneous sky background on the whole image. The formula of RAS is as followed:

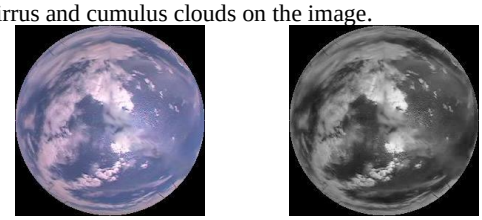
$$RAS = Y - (L - D) \tag{3}$$

$$Y = 0.299R + 0.587G + 0.114B \quad (4)$$

$$L = \max\{R, G, B\} \quad (5)$$

$$D = \min\{R, G, B\} \quad (6)$$

where Y is the panchromatic channel which is sensitive to all visible colors (Ford, 1998); The bright channel L denotes the most energy channel of each pixel in the RGB component while the dark channel D represents the least. RAS can greatly enlarge the difference between the clouds and the sky and highlight both cirrus and cumulus clouds on the image.



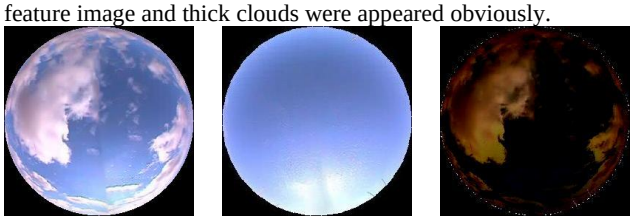
(a) (b)

Figure 3. RAS feature: (a) Original image and (b) RAS image

2) Real clear sky background

Circumsolar region is easily misclassified because of the high brightness of this region. An background removal method was used, which remove the clear sky image with the corresponding solar zenith angle from the cloud image to eliminate its impact (Yang et al., 2016).

A real clear sky background library was built by collecting monthly clear sky images with the solar zenith angle interval of 0.1°. Each cloud image found the clear sky image with the closest time and the smallest difference of solar zenith angle from this library. Usually cloud image and the corresponding clear sky image have similar zenith angle but different azimuth angle. It is necessary to rotate the clear sky image so that the azimuth angle of the two images are consistent. The non-background feature images of three band was obtained by calculating the difference between the cloud image and the corresponding clear sky image. As shown in Figure 4(c), the interference of the sunrays was successfully removed on the feature image and thick clouds were appeared obviously.



(a) (b) (c)

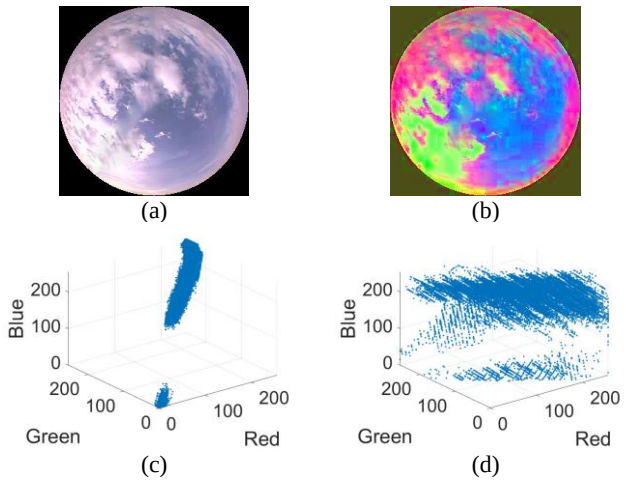
Figure 4. Non-background feature: (a) cloud image, (b) corresponding clear sky, (c) feature image obtained by subtracting (a) and (b)

3) Decorrelation stretch

DS (Decorrelation stretch) can effectively amplify the information with low correlation to enhance the color differences. Meanwhile, it remains the chrominance information. and the spectral characteristics without large distortion. It is showed in Figure 5.

$$\mathrm{b} = T * (a - m) + m\_desired \quad (7)$$

where a is a nBands-by-1 vectors which contains the value of a given pixel in each band of original image. m presents the average of each band in the image and $m\_desired$ is the mean of desired output value for each band. T is the linear transformation matrix.



(a) (b) (c) (d)

Figure 5. Comparison of DS result: (a) original image; (b) image after DS; (c) color scatterplot of (a); (d) color scatterplot of (b)

3.1.2 Texture feature

Texture is one of the most important characteristics used in identifying objects or regions of interest in an image. Different from spectral and colour features of the image, texture displays different objects by the gray distribution of pixels and their spatial neighborhood. The result of texture characteristics is the statistical value of an area containing multiple pixels.

1) Gray-level co-occurrence matrix

GLCM (Gray-level co-occurrence matrix) is the most commonly used method of texture feature extraction based on statistics. It characterizes the texture of an image by calculating how often a pixel with the gray value i occurs in a specific spatial relationship to a pixel with the value j, creating a probability matrix P, and then extracting statistical measures from this matrix. Haralick et al.(1973) proposed a total of 14 statistics calculated based on the GLCM, and we used four of them.

Contrast reflects the sharpness of the image and the depth of the texture groove. If there is a large amount of variation in an image, the contrast will be high.

$$Con = \sum_i \sum_j (i - j)^2 P(i, j) \quad (8)$$

Correlation measures the linear dependency of gray levels on those of neighboring pixels or specified points. Higher values of correlation can be obtained for similar gray-level regions Conversely, the correlation value is small when the pixel values differ greatly.

$$Con = \sum_i \sum_j \frac{(i - Mean)(j - Mean)}{Variance} P(i, j)^2 \quad (9)$$

$$Mean = \sum_i \sum_j i * P(i, j) \quad (10)$$

$$Variance = \sum_i \sum_j (i - Mean) P(i, j) \quad (11)$$

Entropy measures the randomness of the image texture and shows the complexity of the image. The value of entropy is maximum when all values of the co-occurrence matrix are equal. Therefore, a homogeneous image will result in a lower entropy

value, while a heterogeneous region will result in a higher entropy value.

$$ENT = -\sum_i \sum_j P(i,j)\log(P(i,j)) \quad (12)$$

IDM (Inverse different moment) reflects the clarity and regularity of texture and measures the local homogeneity of an image. Hence, inhomogeneous images have low IDM value while homogeneous images have relatively higher value.

$$IDM = \sum_i \sum_j \frac{P(i,j)}{1+(i-j)^2} \quad (13)$$

2) Tamura feature

Tamura features, which include coarseness, contrast, directionality, linelikeness, regularity and roughness, are designed in accordance with psychological studies on the human perception of texture (Tamura et al., 1978). Generally, coarseness, contrast, and directionality are especially important for image processing.

Coarseness relates to distances of notable spatial variations of gray levels and describes the roughness of the image texture. The higher the coarseness value is, the rougher is the texture.

Firstly, for each pixel *(i,j)*, six average intensity value A_k of the active window with the size of 2^k*2^k around the pixel were calculated:

$$A_k = \sum_{i=x-2^{k-1}}^{x-1+2^{k-1}} \sum_{j=y-2^{k-1}}^{y-1+2^{k-1}} \frac{P(i,j)}{2^{2k}} \quad (14)$$

where k=0, 1, 2, …, 5; P is the probability matrix which is the same as GLCM.

Secondly, the absolute intensity differences $E_k(x,y)$ between the pixel pairs of non-overlapping average in the horizontal and vertical directions is calculated respectively:

$$E_{k,h}(x,y) = |A_k(x+2^{k-1},y) - A_k(x-2^{k-1},y)| \quad (15)$$

$$E_{k,v}(x,y) = |A_k(x,y+2^{k-1}) - A_k(x,y-2^{k-1})| \quad (16)$$

The value of k that maximizes the difference $E_k(x,y)$ in either directions is set to the optimal size of the window. Finally, coarseness can be obtained by averaging $S_{best}(x,y)$ for the entire image:

$$S_{best}(x,y) = 2^k \quad (17)$$

$$F_{crs} = \frac{1}{m*n}\sum_{x=1}^{m}\sum_{y=1}^{n} S_{best}(x,y) \quad (18)$$

Contrast measures how gray levels vary in the image and to what extent their distribution is biased to black or white:

$$\alpha_4 = \frac{\mu_4}{\delta^4} \quad (19)$$

$$F_{con} = \frac{\delta}{\alpha_4^{1/4}} \quad (20)$$

where μ_4 is the normalized fourth-order moment of the gray level histogram, δ^2 is the variance, α_4 is the kurtosis.

Directionality shows how the texture is concentrated along certain directions. The gradient vector of each pixel is calculated, and the module $|\nabla Q|$ and direction θ of the vector are defined as:

$$|\nabla Q| = \frac{|\nabla H| + |\nabla V|}{2} \quad (21)$$

$$\theta = \tan^{-1}\left(\frac{\nabla V}{\nabla H}\right) + \frac{\pi}{2} \quad (22)$$

where ∇H and ∇V are the horizontal and vertical gray level differences between the adjacent pixels, respectively. The histogram H_D is relatively uniform for images without strong orientation and exhibits peaks for highly directional images. The directionality related to the sharpness of the peaks:

$$F_{dir} = \sum_{p=1}^{n^p} \sum_{\varphi \in W_p} (\varphi - \varphi_p) H_D(\varphi) \quad (23)$$

where p denotes the peak in histogram H_D, n^p is the number of peaks, W_p represents the range between valleys around the peak, φ is the quantized direction angle.

3) Gabor

The Gabor function is a filter that is used as an orientation and scale tunable edge and line detector. The two-dimensional Gabor filter has the characteristic of obtaining the optimal position in both the spatial and frequency domain, so it can well describe the local structural information corresponding to the spatial frequency, spatial position and directional selectivity. The equation of complex number, real number and imaginary number in the two-dimensional Gabor filter is described as followed:

$$g_1(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\exp\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \quad (24)$$

$$g_2(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\cos\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \quad (25)$$

$$g_3(x,y;\lambda,\theta,\psi,\sigma,\gamma) = \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\sigma^2}\right)\sin\left(i\left(2\pi\frac{x'}{\lambda}+\psi\right)\right) \quad (26)$$

$$x' = x\cos\theta + y\sin\theta \quad (27)$$

$$y' = -x\sin\theta + y\cos\theta \quad (28)$$

where λ is wavelength, θ specifies the direction of the Gabor function's parallel stripes and ψ is phase deviation ranging from -180° to 180°; γ is length-width ratio and σ is the standard deviation of the gaussian factor of the Gabor function.

4) Local Binary Pattern

LBP (Local Binary Pattern), which has significant advantages such as grayscale invariance and rotation invariance, is an effective texture descriptor for images. According to the previous studies, LBP is chosen as one of the texture features (Cheng, Yu, 2015). To adapt to the texture features of different scales, Ojala et al.(1996) improved the original LBP operator that the 3×3 neighborhood was extended to any neighborhood, and the square neighborhood was replaced by the circular neighborhood. The improved LBP operator with P sampling points in a circular region of radius R is calculated as below:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{p-1} 2^p s(i_p - i_c) \quad (29)$$

$$s(x) = \begin{cases} 1 & x > 0 \\ 0 & x \le 0 \end{cases} \quad (30)$$

where i_c denotes the gray value of center pixel (x_c, y_c) and i_p is the gray-level value of neighborhood point. It was found that the fine image texture can obtained by setting a small radius of circle region. The brightness of final feature image has a great relationship with the number of sampling points P. If P is small, the brightness of feature image will low.

3.1.3 Color feature

Color features are widely used visual features. The main reason is that color is usually closely related to the objects in the image. In addition, compared with other visual features, color features have less dependent on the size, direction and viewing angle of the image itself, so they are more robust.

1) HSV color space

According to the intuitive characteristics of color, HSV color space is also known as the Hexcone Model (Smith, 1978). The color parameters of this model are: hue (H), saturation (S), value (V). Hue, saturation and value separately represent the colour type, the intensity of the colour and the brightness of the color. Compared with RGB space, HSV space can directly express the tone and brightness of colors, which is benefit for the comparison between colors. The formula for converting from RGB space to HSV space are defined as:

$$h = \begin{cases} 0^\circ & \max = \min \\ 60^\circ * \dfrac{G-B}{\max - \min} + 0^\circ & \max = R, G \ge B \\ 60^\circ * \dfrac{G-B}{\max - \min} + 360^\circ & \max = R, G < B \\ 60^\circ * \dfrac{B-R}{\max - \min} + 120^\circ & \max = G \\ 60^\circ * \dfrac{B-R}{\max - \min} + 240^\circ & \max = B \end{cases} \quad (31)$$

$$s = \begin{cases} 0 & \max = 0 \\ 1 - \dfrac{\min}{\max} & \max \ne 0 \end{cases} \quad (32)$$

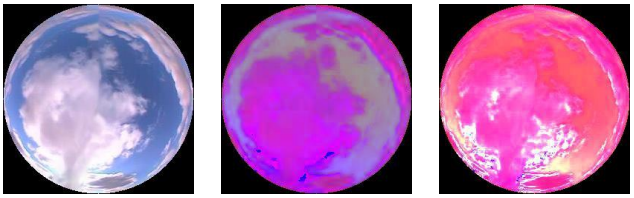
$$v = \max \quad (33)$$

2) HSL color space

H, S and L indicate hue, saturation and lightness in the HSL color model, respectively. H has the same meaning as HSV space, but the definitions of S and L are different. Saturation refers to variation of the color depending on the lightness. Lightness, also named luminary, carries both black and white information.

$$s = \begin{cases} 0 & l = 0, or \max = \min \\ \dfrac{\max - \min}{2l} & 0 < l \le \dfrac{1}{2} \\ \dfrac{\max - \min}{2 - 2l} & l > \dfrac{1}{2} \end{cases} \quad (34)$$

$$l = \frac{1}{2}(\max + \min) \quad (35)$$



(a) (b) (c)

Figure 6. Color feature: (a) RGB color space; (b) HSV color space; (c) HSL color space

3.2 Random forest

The above features are integrated into a feature set to train random forest model for the classification of various types of clouds and clear sky. Random Forest is the main representative of ensemble learning method (Breiman, 2001).

Compared with other machine learning methods such as decision tree, support vector machine, artificial neural network, random forest has some remarkable advantages in image classification. The first is that it could surpass other classification algorithms in terms of classification accuracy(La Rosa, Wiesmann, 2013; Pal, 2005). The second is that it can process thousands of feature variables without feature selection and calculate the importance of each feature. In the CART (Categorical and Regression Tree) algorithm of random forest, each node selects the optimal splitting tree according to the Gini index which represents the highest purity of each child node (Fu et al., 2019). When all the samples falling on one child node belong to the same category, the Gini index is the minimum, which means that it has the highest purity and the lowest uncertainty. The importance of each feature can be determined by Gini index. Last but not least, it only needs to set two parameters, namely *ntree* and *mty*. *ntree* is the number of decision trees in random forest, which is related to OOB (Out-Of-Bag) error. With the increase of the *ntree* value, the OOB error gradually decreases. When the value of *ntree* increases to a certain value, the OOB error tends to converge. Therefore, the most suitable *ntree* value can be obtained by the graph of *ntree* and OOB error. *mty* denotes the number of selected feature variables. When the value of *mty* is high, it will increase the correlation between any two decision trees, resulting in poor prediction result. But this value cannot be set too low, because it will reduce the classification ability of each tree.

The basic unit of random forest is the CART decision tree, and each decision tree is a small classifier. As for an input sample, N trees may have N classification results. The random forest can integrate the voting results of all decision trees for classification and specify the category with the most votes as the final output. The randomness of random forest mainly reflects in random sampling and random feature selection. This random sampling method is called bootstrap sample, which could ensure that the samples of each decision tree are different, but there are also duplicates between the decision trees. From the perspective of probability theory, the new sample set of each decision tree contains only 2/3 of the original samples. The remaining 1/3 samples are called OOB data and used for internal cross-validation to evaluate the classification accuracy of random forest (Feng et al., 2015). In terms of random feature selection, if there are M input feature variables, m ($m<<M$) feature variables are randomly selected from the M features. These m features are used to construct the best split node. The value of m is constant during forest growth (Breiman, 2001).

3.3 Postprocessing

The area near the sun is the most likely to be misclassified. One situation is the brightness of thin cirrus clouds and the haze around the sun is high because of forward scattering of sunlight by them. The other is the image overexposure of the clear sky around the sun, because the CCD is affected by sun reflection (Pfister et al., 2003).

After random forest classification, the images that exist a large proportion of cirrus clouds around the sun need to be improved. It causes circumsolar area to be whiter and brighter than other regions. As a result, the clouds in this region are always classified as cumulus. "SP (Sunshine Parameter)" defined by Pfister et al.(2003) is introduced as a parameter for processing cirrus in the solar region in this paper. The region is set up as a sector, the size of which is ±50 degrees of the solar azimuth angle. SP is the average of RBR of each pixel in this region. When the difference between SP and RBR of a pixel is greater than a fixed threshold, the pixel is regarded as cirrus cloud.

Regarding the problem that the clear sky pixels in the circumsolar region are easily misclassified, the pixels in a circular area are reclassified by threshold method. This area is centred on the sun and has a radius of 60 pixels. The circularity based on CSP (Circumsolar Saturated Pixels) proposed by Montero(2009) are used. The formula is as followed:

$$C = \frac{4\pi * S}{L^2} \tag{36}$$

where C denotes circularity, S is the area of CSP and L is the perimeter of CSP. Whether the sun is blocked is determined by the C value. When C is greater than 0.3, it can be considered that the sun is not blocked by clouds, and the clear sky pixels can be re-determined by threshold method.

4. RESULTS AND ANALYSIS

4.1 Training Samples Selection

From the images in 2008, a total of 400 sky images of different cloud types were selected, of which 300 were used for training and 100 were used for verification. Cumulus and cirrus clouds account for a large proportion of these images. Stratus clouds mainly appear on overcast days and last for a long time, they don't cause drastic fluctuations in the short-term solar radiation that will be carried out in the following research. Therefore, the focus of classification is to distinguish clear sky, cumulus clouds and cirrus clouds. We collected as many samples as possible for various situation to ensure that the training samples were typical.

4.2 Parameterization of Random Forest

We selected 21 features in total, including 7 spectral features, 9 texture features and 5 color features. According to the existing research (Rodriguez-Galiano et al., 2012), when random forest is applied to classification, the value of *mty* is suitable to be the square root of the feature number we used. Therefore, *mty* is set to 5 as all features are used; and *mty* is set to 3 when only spectral features are input.

Figure 7 showed the relationship between *ntree* and OOB error. It can be seen when *ntree* is increased to a certain value, the classification accuracy will not increase anymore, and the OOB error rate will not decrease.

In Figure 7, when *ntree* increases from 1 to 20, OOB error rate drops from 0.029 to 0.011. Since then, with the increase of *ntree*, the error rate has only decreased slightly. When *ntree* is greater than 80, the error rate remains at 0.0074. Hence, *ntree* can be set to 80, which can obtain better classification accuracy under less calculating pressure.

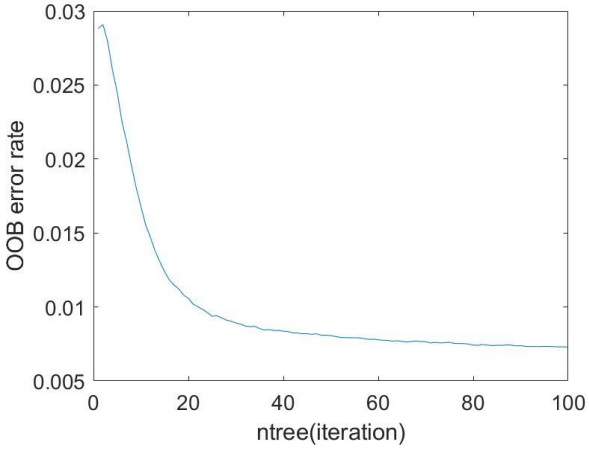


Figure 7. OOB error of random forest

4.3 Classification Results

In order to test the cloud classification results of different types of features for sky images, spectral feature, texture feature, color feature and their different combinations are put into the random forest model with 80 decision trees. The classification results of random forest are compared with the true value of the cloud types by visual interpretation to calculate the overall accuracy of classification, which is equal to the number of correctly classified pixels divided by the number of all pixels in a class.

	Overall Accuracy
spectral feature	76.80%
texture feature	59.46%
color feature	73.74%
Spectral + texture feature	77.31%
Spectral + color feature	79.76%
Texture + color feature	77.60%
Spectral + texture + color feature	80.51%

Table 1. The accuracy of single-type features and feature combinations

It can be seen from Table 1 that spectral feature plays a crucial role in the accuracy of cloud classification results. Only texture feature or color feature can just get limited classification accuracy. But when they are used in combination with spectral feature, the classification results are improved. The same conclusion as above could also be obtained from Figure 8.

Figure 8 is the feature importance computed by random forest. Random forest provides Gini index that represents the importance of features. The higher Gini value of the feature is, the greater the contribution of this feature to the classification is. In Figure 8, it can be seen that the contribution of each feature varies greatly. For cloud classification, spectral and color features have higher contribution, while texture features have lower importance. Therefore, it can only be used in combination with spectral or color features.

Table 2 is a comparison table of cloud classification accuracy before and after post-processing. The classification accuracy of clear sky, cumulus and cirrus without post-processing are 86.08%, 85.24% and 61.35%, respectively. Some cirrus clouds are misclassified into clear sky or cumulus clouds. Cirrus clouds are thin, wispy cloud which form in the upper levels of the troposphere. Due to they are thin enough to be transparent or very close to it, the difference between cirrus clouds and clear sky is small. Thus, the cirrus clouds are easily misclassified as

clear sky in the non-circumsolar areas. Cirrus clouds are misclassified as cumulus clouds occurring in the region around the sun. This kind of error can be effectively eliminated by the threshold method mentioned in 3.3.

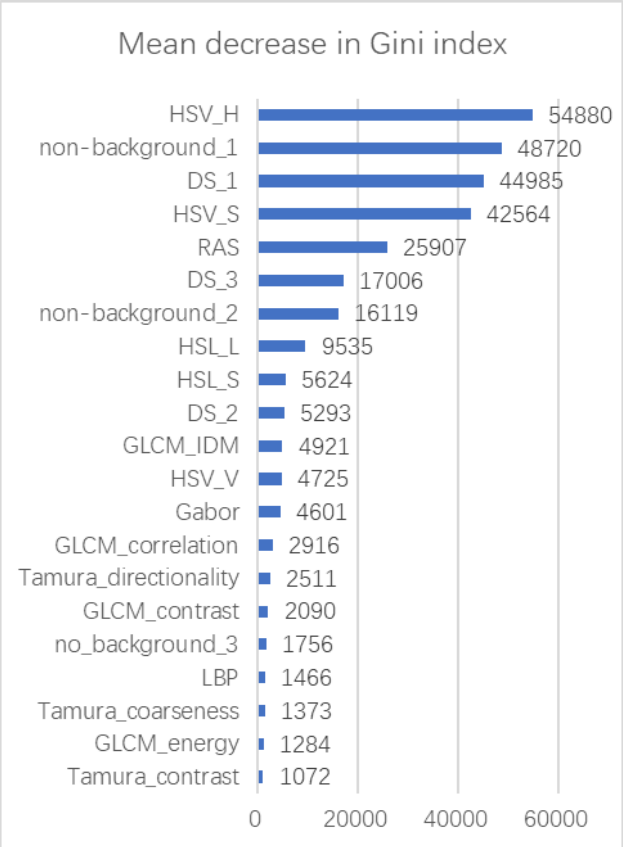


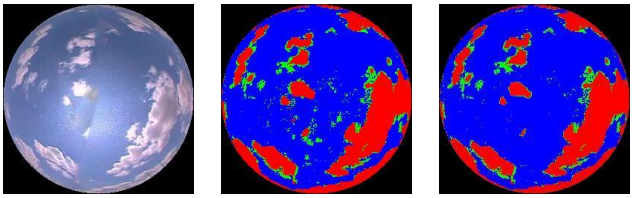
Figure 8. Feature importance

Clear sky and cumulus are also misclassified as cirrus clouds. In clear sky image, even without the scattering of clouds, the area around the sun still shows different characteristics with other part, forming a bright circle in the image. As the distance from the center of the sun increases, the image brightness gradually decreases to a normal level. Therefore, there are cases in this region where the clear sky pixels are misclassified as cirrus in this region. The edge of cumulus clouds is misclassified as cirrus clouds. The main reason is that the cloud edge may increase the scattering of the sun to make its brighter. This kind of error could be reduced by increasing the number of samples at the edge of cumulus.

On the basis of the classification result of random forest, additional threshold method is used for the post-processing in the cirrussolar region. The overall classification accuracy was improved by 0.56% to 81.07%. The new classification accuracy of cumulus and cirrus cloud are 85.47% and 61.68%, respectively. Since the area for post-processing is small, the overall classification accuracy was not improved much. Figure 9 is the cloud classification result, in which blue, red and green represent clear sky, cumulus and cirrus respectively. It can be seen from Figure 9 that the classification result after postprocessing was obviously better.

	Random forest without postprocessing	Random forest with postprocessing
Clear sky	86.08%	87.72%
Cumulus	85.24%	85.47%
Cirrus	61.35%	61.68%

Table 2. The classification accuracy before and after postprocessing



(a) (b) (c)

Figure 9. Classification result: (a) cloud image; (b) classification using random forest without postprocessing; (c) classification using random forest with postprocessing

5. CONCLUSION

A novel method combining multiple features and random forest classifier is applied to cloud classification of ground-based sky images. The whole method is divided into three steps: 1) In order to obtain a continuous and complete image, the patch-based method is used to repair the shadowed part of the original image; 2) Multiple features of the image such as spectrum, texture and color feature are extracted; 3) Random forest, combined with multiple features, are used to classify the various cloud and the clear sky. Next the cirrus solar area are postprocessed. The experimental results showed that cumulus and cirrus clouds can be identified from the image, and the cumulus classification is better with the accuracy of 85.47%.

At the same time, the proposed method has some limitations t: 1) The image completion method can be improved to make it more close to the real situation. The current experimental results showed that if the shadow band region is removed in the accuracy assessment, the final accuracy will be improved by 0.2%;2) When cumulus and cirrus cloud are mixed, the classification accuracy is affected to a certain extent. It is necessary to find more effective features to distinguish them. These limitations will provide ideas for future research.

ACKNOWLEDGEMENTS

This work was supported by Natural Science Foundation of Hubei Province (grant 2019CFB732).

REFERENCES

Ackerman, S.A., Strabala, K.I., Menzel, W.P., Frey, R.A., Moeller, C.C.,Gumley, L.E., 1998: Discriminating clear sky from clouds with MODIS. *Journal of Geophysical Research-Atmospheres*, 103, 141-157.

Breiman, L., 2001: Random forests. *Machine Learning*, 45(1), 5-32.

Calbó, J.,Sabburg, J., 2008: Feature Extraction from Whole-Sky Ground-Based Images for Cloud-Type Recognition. *Journal of Atmospheric and Oceanic Technology*, 25(1), 3-14.

Cheng, H.Y.,Yu, C.C., 2015: Block-based cloud classification with statistical features and distribution of local texture features. *Atmospheric Measurement Techniques*, 8(3), 1173-1182.

Chenni, R., Makhlouf, M., Kerbache, T.,Bouzid, A., 2007: A detailed modeling method for photovoltaic cells. *Energy*, 32(9), 1724-1730.

Feng, Q.L., Liu, J.T.,Gong, J.H., 2015: UAV Remote Sensing for Urban Vegetation Mapping Using Random Forest and Texture Analysis. *Remote Sensing*, 7(1), 1074-1094.

Ford, A.a.R., A., 1998. Colour Space Conversions. Westminster University, London.

Fu, H.L., Shen, Y., Liu, J., He, G.J., Chen, J.S., Liu, P., Qian, J.,Li, J., 2019: Cloud Detection for FY Meteorology Satellite Based on Ensemble Thresholds and Random Forests Approach. *Remote Sensing*, 11(1), 28.

Gan, J., Lu, W., Li, Q., Zhang, Z., Yang, J., Ma, Y.,Yao, W., 2017: Cloud Type Classification of Total-Sky Images Using Duplex Norm-Bounded Sparse Coding. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(7), 3360-3372.

Haralick, R.M., Shanmugam, K.,Dinstein, I., 1973: Textural Features for Image Classification. *Ieee Transactions on Systems Man and Cybernetics*, SMC3(6), 610-621.

Heinle, A., Macke, A.,Srivastav, A., 2010: Automatic cloud classification of whole sky images. *Atmospheric Measurement Techniques*, 3(3), 557-567.

Huang, J.B., Kang, S.B., Ahuja, N.,Kopf, J., 2014: Image Completion using Planar Structure Guidance. *Acm Transactions on Graphics*, 33(4), 129, 1-10.

Johnson, R., Hering W., Shields, J., 1991. Analysis and interpretation of simultaneous multi-station whole sky imagery. Marine Physical Laboratory, Scripps Institution of Oceanography, San Diego, USA.

Kazantzidis, A., Tzoumanikas, P., Bais, A.F., Fotopoulos, S.,Economou, G., 2012: Cloud detection and classification with the use of whole-sky ground-based images. *Atmospheric Research*, 113, 80-88.

Kliangsuwan, T.,Heednacram, A., 2018: FFT features and hierarchical classification algorithms for cloud images. *Engineering Applications of Artificial Intelligence*, 76, 40-54.

La Rosa, D.,Wiesmann, D., 2013: Land cover and impervious surface extraction using parametric and non-parametric algorithms from the open-source software R: an application to sustainable urban planning in Sicily. *Giscience & Remote Sensing*, 50(2), 231-250.

Lackner, K.S., 2002: Can fossil carbon fuel the 21st century? *International Geology Review*, 44(12), 1122-1133.

Li, Q.Y., Lu, W.T.,Yang, J., 2011: A Hybrid Thresholding Algorithm for Cloud Detection on Ground-Based Color Images. *Journal of Atmospheric and Oceanic Technology*, 28(10), 1286-1296.

Long, C.N., Sabburg, J.M., Calbo, J.,Pages, D., 2006: Retrieving cloud characteristics from ground-based daytime color all-sky images. *Journal of Atmospheric and Oceanic Technology*, 23(5), 633-652.

Montero, R.S.a.B., E., 2009: State of the art of compactness and circularity measures. *International Mathematical Forum*, 4(27), 1305–1335.

Morris, V.R., 2005. Total Sky Imager Handbook. Atmospheric Radiation Measurement Climate Research Facility, USA.

Ojala, T., Pietikainen, M.,Harwood, D., 1996: A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1), 51-59.

Pal, M., 2005: Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), 217-222.

Pfister, G., McKenzie, R.L., Liley, J.B., Thomas, A., Forgan, B.W.,Long, C.N., 2003: Cloud Coverage Based on All-Sky Imaging and Its Impact on Surface Solar Irradiance. *Journal of Applied Meteorology*, 42(10), 1421-1434.

Reda, I.,Andreas, A., 2004: Solar position algorithm for solar radiation applications. *Solar Energy*, 76(5), 577-589.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M.,Rigol-Sanchez, J.P., 2012: An assessment of the effectiveness of a random forest classifier for land-cover classification. *Isprs Journal of Photogrammetry and Remote Sensing*, 67, 93-104.

Shields, J., Karr, M., Burden, A., 2009. Research toward Multi-Site Characterization of Sky Obscuration by Clouds. Marine Physical Laboratory, Scripps Institution of Oceanography, San Diego, USA.

Smith, A.R., 1978: Color Gamut Transform Pairs. *Siggraph 78: Conference on Computer Graphics & Interactive Techniques*, 12(3), 12-19.

Tamura, H., Mori, S.,Yamawaki, T., 1978: Textural Features Corresponding to Visual Perception. *Ieee Transactions on Systems Man and Cybernetics*, 8(6), 460-473.

Tapakis, R.,Charalambides, A.G., 2013: Equipment and methodologies for cloud detection and classification: A review. *Solar Energy*, 95, 392-430.

Xia, M., Lu, W., Yang, J., Ma, Y., Yao, W.,Zheng, Z., 2015: A hybrid method based on extreme learning machine and k-nearest neighbor for cloud classification of ground-based visible cloud image. *Neurocomputing*, 160, 238-249.

Yang, J., Lu, W.T., Ma, Y.,Yao, W., 2012: An Automated Cirrus Cloud Detection Method for a Ground-Based Cloud Image. *Journal of Atmospheric and Oceanic Technology*, 29(4), 527-537.

Yang, J., Min, Q., Lu, W., Ma, Y., Yao, W.,Lu, T., 2017: An RGB channel operation for removal of the difference of atmospheric scattering and its application on total sky cloud detection. *Atmospheric Measurement Techniques*, 10(3), 1191-1201.

Yang, J., Min, Q., Lu, W., Ma, Y., Yao, W., Lu, T., Du, J.,Liu, G., 2016: A total sky cloud detection method using real clear sky background. *Atmospheric Measurement Techniques*, 9(2), 587-597.

Revised May 2020