Classification of clouds in the Japan Sea area using NOAA AVHRR satellite images and self-organizing map

Classification of clouds in the Japan Sea area using NOAA AVHRR satellite images and self-organizing map

Mamoru Kubo and Ken-ichiro Muramoto
Kanazawa University, Kakuma, Kanazawa 920-1192, JAPAN
E-mail: kubo@ec.t.kanazawa-u.ac.jp

***Abstract*—This paper introduces a method for cloud classification using NOAA AVHRR satellite images. AVHRR (Advanced Very High Resolution Radiometer) data consists of five-channel multi-spectral images. To reduce the dimensionality of the data, principal component analysis (PCA) is calculated for each channel separately. The most significant principal component values are then composed into an image feature vector. Finally, the feature vectors are clustered using self-organizing map (SOM). This method is applied for the study of winter season clouds in the Japan Sea area.**

***Index Terms*—Satellite images, AVHRR, principal component analysis, self-organizing map, image clustering**

I. Introduction

Since the utilization of man-made satellites the amount of available satellite image data has been increasing rapidly. Therefore it has become necessary to develop un-supervised data-mining techniques to handle the ever growing databases. In this paper, we a propose a novel self-organizing map (SOM) based clustering method and show that it has the ability to distinguish seasonal variance within the satellite data.

Satellite image clustering has been used as a part of classification [1] and time-series prediction tasks [2]. It also provides a researcher a means to quickly gain a better insight to the data without the need to actually browse thought the entire dataset. There exist many good un-supervised clustering algorithms such as k-NN, fuzzy clustering and genetic algorithms [3]. However, since 1980s Kohonen's self-organizing map [4] has become a popular method due to its robustness and good ability to generalize.

Generally speaking, in order to develop a successful satellite image clustering method, one needs to tackle with the problem of feature extraction. It is necessary to reduce dimensionality; the shorter the feature vector, the easier the clustering phase becomes. But discovering a vector that can express the desired features of the data in compact form is often very difficult. We propose that multispectral principal component analysis (PCA) can satisfy both of these goals.

II. NOAA AVHRR Satellite Image Data

AVHRR multi-spectral images consist of five channels ranging from visible light wavelengths to thermal infrared [5]. Spatial resolution is 1.09 km. The channel bandwidths are shown in Table I. AVHRR data is well suited for sea surface temperature measurement, cloud detection and water-land border detection.

TABLE I
Sensor bandwidths for NOAA AVHRR channels. The channel 3 has two optional wavelengths, although channel 3a is available only in satellites NOAA15, 16 and 17 [5]. In this study, only 3b data is utilized

Channel	Bandwidth (μm)	
1	0.58 - 0.68	Visible
2	0.725 - 1.10	Near IR
3a	1.68 - 1.64	Mid IR
3b	3.55 - 3.93	Thermal IR
4	10.3 - 11.3	Thermal IR
5	11.5 - 12.5	Thermal IR

IR: Infrared

The satellite images are acquired from the Japan Image Database (JAIDAS) of Tohoku University [6]. It consists of daily AVHRR images of northern Japan area. The size of each image is 512x512 pixels for all the channels. An example image is presented in Fig. 1. For this study, images between years 2001 and 2005 are used. After removing corrupted or partial images, the dataset consists of total 1276 images. To concentrate on studying clouds in the Japan Sea area, a fixed window of size 128x96 pixels is used. The window is shown in Fig. 1.

III. Method for Satellite Image Clustering

Fig. 2 shows the diagram of satellite image clustering. Firstly, in order to emphasize cloud features the difference signal between channels 3b and 4 (denoted as $D34$) as well as channels 4 and 5 ($D45$) are computed. These signals have been successfully used for cloud detection [7] [8]. The $D34$ image emphasizes most typical clouds while the $D45$ signal can be used to detect optically thin cirrus clouds.

A. Feature Vector

Because the data set consists of five channel images of considerable size, it is essential to reduce the data dimensionality before clustering. For this purpose PCA is calculated for all channels and the difference images separately. Fig. 3 shows an example of the first 30 PCA eigenimages for channel 2. The

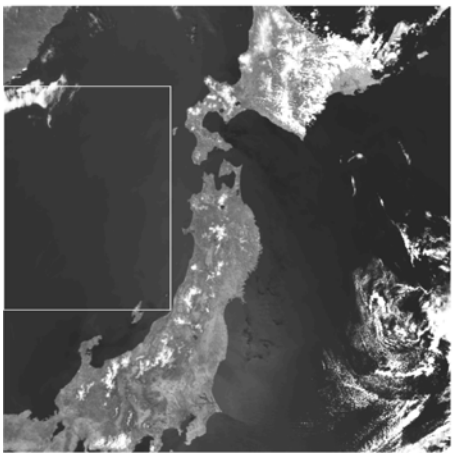

Fig. 1. An example of JAIDAS channel 2 image with the applied window.

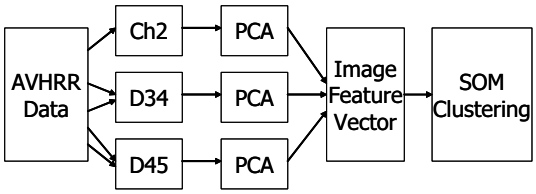

Fig. 2. The diagram of satellite image clustering.

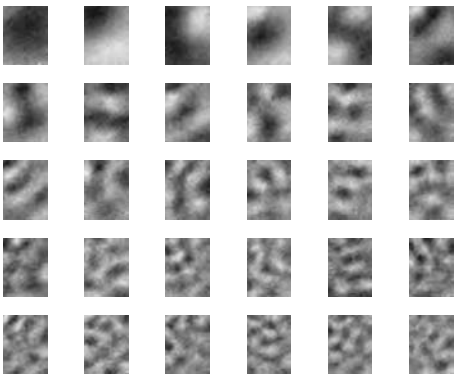

Fig. 3. First 30 eigenimages of AVHRR channel 2. Ordered from upper-left to right. High-frequency content is coded by the higher PCA components.

advantage of using separate PCAs is that the multispectral nature of the data can be preserved during the dimensionality reduction.

A combination of the obtained PCA parameters is then stored into a feature vector $\mathbf{t}_i$. Finally, the set of feature vectors is used as a training set for the SOM.

As a feature vector in this study, several different combinations of the principal components were experimented. The best performance was obtained by using vectors

$$\mathbf{t}_i = [PCA_1^{\mathrm{Ch2}}(\mathbf{x}_i), \quad PCA_1^{D34}(\mathbf{x}_i), \quad PCA_1^{D45}(\mathbf{x}_i)],$$

i.e. the first principal component for channel 2 and difference images $D34$ and $D45$, respectively. The performance of each combination was evaluated by visually examining how well the SOM is able to separate summer and winter images.

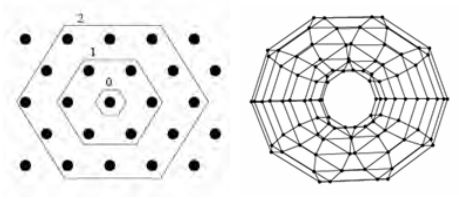

Fig. 4. SOM: hexagonal grid and toroidal map topology.

B. SOM Parameters

The self-organizing map [4] is a unsupervised learning algorithm based on neural networks. A SOM consists of a set of neurons arranged in some topological map. Fig. 4 shows hexagonal grid and toroidal map topology. Each neuron j has a codebook vector $\boldsymbol{\omega}_j$ which lies in the same space as the training examples $\mathbf{t}_i$.

The training data $\mathbf{t}_i$ is presented to the SOM several times in random order and the codebooks are updated recursively:

$$\begin{aligned} &\forall\ j: \quad \boldsymbol{\omega}_j = \boldsymbol{\omega}_j + \eta \cdot e^{-\sigma \cdot N(j,j_0)}(\mathbf{t}_i - \boldsymbol{\omega}_j), \\ &j_0 \ = \ \{\ j \mid \ |\boldsymbol{\omega}_j - \mathbf{t}_i| \ \leq \ |\boldsymbol{\omega}_k - \mathbf{t}_i|,\ k \neq j\}, \end{aligned}$$

where j_0 stands for the winning neuron, $N(j, j_0)$ is a distance metric in the map topology, η is a learning parameter and σ is a width parameter for the Gaussian function. During the training, both the learning rate parameter and the Gaussian width parameter are decreased to ensure convergence. A receptive field of a neuron is defined as

$$R_j \ = \ \{\ \mathbf{t}_i \mid \ |\boldsymbol{\omega}_j - \mathbf{t}_i| \ \leq \ |\boldsymbol{\omega}_k - \mathbf{t}_i|,\ k \neq j\}$$

and it contains all the training vectors in the proximity of the codebook $\boldsymbol{\omega}_j$.

Several different setups of SOM were experimented. The best performance was obtained using a 6-by-8 SOM with sequential training algorithm. A toroidal map topology seemed to give slightly better results in comparison to normal hexagonal grid. Again, the performance was evaluated by visually examining the SOM clusters.

IV. Results

Fig. 5 shows the result of trained SOM. In order to evaluate how well the SOM can distinguish different seasons the images were separated into 12 bins according to the month when the image was taken. Fig. 6 shows these monthly histograms for each receptive field of the SOM. As one can easily observe, the SOM is able to separate seasonal variance, because there are some receptive fields that contain only winter or summer season images. And what is more, the winter season neurons seem to be concentrated in the lower left side of the SOM.

Fig. 7 shows all the images belonging to the receptive field of the neuron 36 (See Fig. 6). All of these images are typical daytime winter images. A larger version of one of the images is presented in Fig. 8. The narrow convective clouds, also

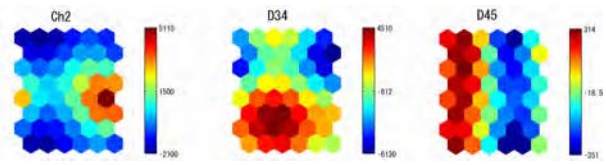

Fig. 5. The result of trained SOM.

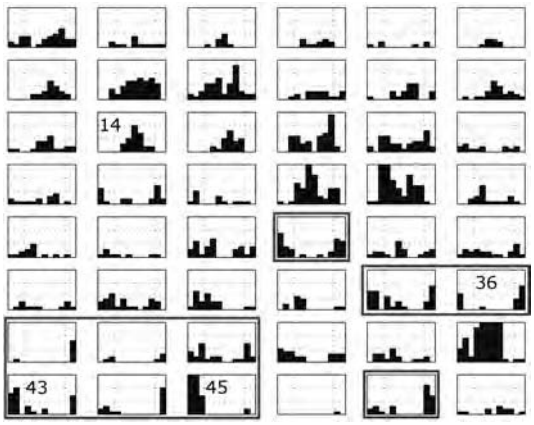

Fig. 6. Monthly image histograms of the SOM. The bins are arranged from January to December, starting from left. Vertical axis ranges from 0 to 15. Some bins contain more than 15 images.

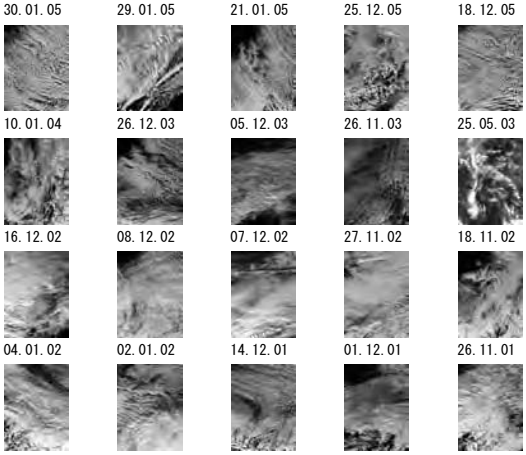

Fig. 7. An example of a winter node. All the images belonging to the receptive field of SOM node 36. The dates of each image are shown in format `day.month.year`. Channel 2.

known as snow bands [9], are clearly visible. The fact that the convective snow clouds are concentrated on certain areas of the SOM is very interesting because these clouds are known to cause heavy snowfall, which is characteristic of the winter in the Japan Sea area.

The amount of snow band images was high for other winter season neurons as well. For the framed neurons in Fig. 6, roughly 50 per cent or more of the images contained snow bands. To point out some characteristics of other neurons, neuron 43 seems to contain partial cloudy daytime images.

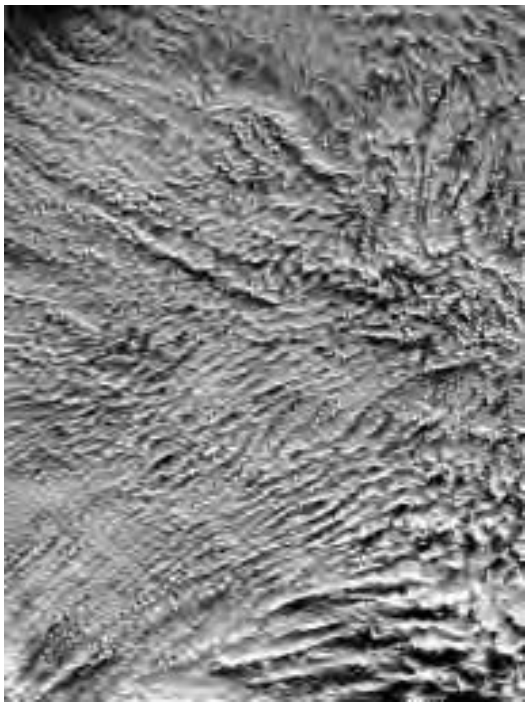

Fig. 8. An example of typical winter clouds (18th December 2005). The band-like convective snow clouds are clearly visible. Channel 2.

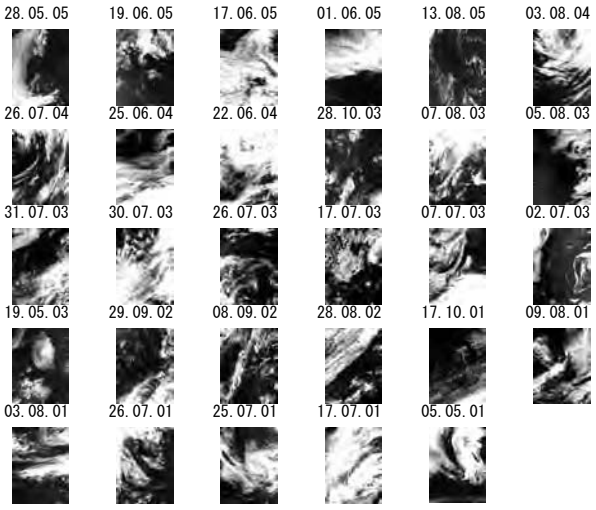

Fig. 9. An example of a summer node. All the images belonging to the receptive field of SOM node 14. Channel 2.

Snow band ratio was approximately 60 per cent. Neuron 45, on the other hand, contains a total 54 fully or partially cloudy night time images. In this case the snow band ratio was 75 per cent.

For reference, the images of a typical summer neuron 14 are also plotted in Fig. 9. The difference between the cloud types is clear when comparing to Fig. 7.

V. Conclusions

We have introduced a novel method for satellite image clustering. The results show that by using a suitable feature vector SOM is able to distinguish seasonal variance within the data acquired in the Japan Sea area. What is more, there exist clear clusters that contain typical snow band images.

In future, the SOM clustering involves several parameters that could be tuned for better performance. Also, the obtained results should be analyzed more thoroughly based on the physical properties of clouds.

References

[1] Honda R., Konishi O., et. al., "Data mining system for planetary images - crater detection and categorization - ", *Proceedings of the International Workshop on Machine Learning of Spatial Knowledge in conjunction with ICML*, Stanford, CA, p.103–108, 2000.

[2] A. Kitamoto, "Data Mining for Typhoon Image Collection," *Proceedings of the 2nd International Workshop on Multimedia Data Mining*, pp.68-77, Aug 2001.

[3] Jain A., Murty M., Flynn P., "Data Clustering: A Review", *ACM Computing Surveys*, Vol. 31, No. 3, September 1999.

[4] T. Kohonen, "Self-Organizing Maps," *Springer Series in Information Sciences*, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995, 1997, 2001.

[5] "The NOAA KLM User's Guide," September 2000 revision, [Online], Available: http://www2.ncdc.noaa.gov/docs/klm/index.htm.

[6] "Japan Image Database (JAIDAS) homepage," [Online], Available: http://asiadb.cneas.tohoku.ac.jp/jaidas/index-E.html.

[7] Minnis P., Chakrapani V.,Doelling D., et al., "Cloud coverage and height during FIRE ACE derived from AVHRR data," *Journal of Geophysical Research*, Vol. 106, No. D14, pp.15215-15232, July 2001.

[8] Katagiri S., Nakajima T., "Radiative Characteristics of Cirrus Clouds as Retrieved from AVHRR," *Journal of the Meteorological Society of Japan*, Vol. 86, No. 1, pp.81-99, 2004.

[9] Tsuboki K., "High Resolution Modeling of Multi-scale Cloud and Precipitation Systems Using a Cloud-Resolving Model," *Annual Report of the Earth Simulator Center April 2004 - March 2005*, pp.79-84, 2005.