RAA 2015 Vol. **15** No. **9**, 1604–1612 doi: 10.1088/1674–4527/15/9/016
http://www.raa-journal.org *http://www.iop.org/journals/raa*

Research in Astronomy and Astrophysics

A new segmentation algorithm for lunar surface terrain based on CCD images

Hong-Kun Jiang, Xiao-Lin Tian and Ao-Ao Xu

Lunar and Planetary Science Laboratory/Space Science Institute, Macau University of Science and Technology, Macau 999078, China; *jiang.hongkun@qq.com*

Received 2014 October 17; accepted 2014 December 30

Abstract Terrain classification is one of the critical steps used in lunar geomorphologic analysis and landing site selection. Most of the published works have focused on a Digital Elevation Model (DEM) to distinguish different regions of lunar terrain. This paper presents an algorithm that can be applied to lunar CCD images by blocking and clustering according to image features, which can accurately distinguish between lunar highland and lunar mare. The new algorithm, compared with the traditional algorithm, can improve classification accuracy. The new algorithm incorporates two new features and one Tamura texture feature. The new features are generating an enhanced image histogram and modeling the properties of light reflection, which can represent the geological characteristics based on CCD gray level images. These features are applied to identify texture in order to perform image clustering and segmentation by a weighted Euclidean distance to distinguish between lunar mare and lunar highlands. The new algorithm has been tested on Chang'e-1 CCD data and the testing result has been compared with geological data published by the U.S. Geological Survey. The result has shown that the algorithm can effectively distinguish the lunar mare from highlands in CCD images. The overall accuracy of the proposed algorithm is satisfactory, and the Kappa coefficient is 0.802, which is higher than the result of combining the DEM with CCD images.

Key words: Moon — methods: data analysis — techniques: image processing

1 INTRODUCTION

Due to the rapid growth of satellite equipment, a large amount of planetary data from satellites is available for aerospace and planetary research. Automatic terrain classification is one important area of research. Nowadays, automatic terrain classification is usually applied to two tasks: one has the goal of increasing the off-road navigation capabilities and autonomy of unmanned ground vehicles (Brooks & Iagnemma 2005; Fujita & Ichimura 2011; Sancho-Pradel & Gao 2010; Shankar et al. 2008); the other is intended to be used in mapping lunar geological features and selecting landing sites from remote sensing data (Zhou et al. 2011). However, regarding the remote sensing data, most published works have focused on the Digital Elevation Model (DEM) or combining data (DEM and CCD) to distinguish lunar terrain (Zhou et al. 2011). Traditional algorithms use the elevation data and slope to define relief when representing the features with data pixel values from images. This

paper presents a new segmentation algorithm for terrain classification that is only based on lunar CCD images. The traditional algorithm uses DEM data to distinguish lunar mare from highlands. However, the DEM generated from Chang'e-1 data has a precision of 500 meters per pixel which is lower than the precision of Chang'e-1 CCD image data. The traditional algorithm also has a shortcoming in that it only directly uses features extracted from the CCD image values in calculations with the algorithm. Images from different orbits were spliced into one CCD image in some regions. However, images from different orbits had different lighting conditions and the pixel values were affected by changes in light conditions. The traditional algorithm cannot necessarily show geological features when it only uses CCD image pixel values. The CCD image data and the DEM data have different precisions. When the algorithm combines DEM data with CCD image data, it may cause an error in image registration. The new algorithm uses data from a single source image that has higher precision and it adds some image features which consider the correlation among various pixels to increase classification accuracy.

Geomorphological features on the lunar surface can be divided into lunar highlands (lunar terrae), lunar mare and volcanic landforms. Regions with lunar mare and lunar highland cover about four-fifths of the lunar surface area. Highland regions have many mountain ranges. A highland is the oldest one of the lunar geological units. A highland region is mainly composed of light plagioclase, so it has high reflectivity. There are some large dark areas which are called lunar mare. Lunar mare is a major lunar geological unit, and collectively they make up of about one fourth of the total lunar surface. The vast majority of lunar maria are distributed on the near side of the Moon that faces Earth and its surface is covered with a lot of basalt. There are 22 lunar maria known today. Most of them are two thousand meters lower than the mean elevation of the Moon. Several lunar maria are six thousand meters lower than their surrounding areas. The newly proposed automatic terrain classification algorithm has the goal of distinguishing between areas of lunar mare and highland. Chang'e-1 CCD data have been used to test the new algorithm; results have been checked by comparing with geological data published by the U.S. Geological Survey (USGS), which has shown that the new algorithm performs well in images used for terrain classification.

2 THE NEW ALGORITHM FOR LUNAR TERRAIN CLASSIFICATION

The new algorithm includes two parts: in the first, the algorithm divides the image into small blocks with the same size, then it calculates every block's image feature values. Two new features and one texture feature have been used in the new algorithm. The two new features that are incorporated are specifically designed to respond to properties of light reflection in a sub-image and the dispersion of solar radiation in an incident CCD image. Because the CCD image is just a gray scale image, the new features are based on the relationship between every point's gray scale, adjacent points and the entire sub-image. The second part of the new algorithm uses weighted Euclidean distance and Ward's method to classify regions of the image into clusters based on their image feature values. Finally, the new algorithm compares the maximum number of pixels in the corresponding image histogram of the last two classes. The one with smaller values will be classified as lunar mare and the other one will be classified as highland. A flow chart for the new classification algorithm is shown in Figure 1.

3 SELECTION OF FEATURES USED IN CLUSTERING

Feature selection is important for any clustering application. The method described in this article identified features with an enhanced image histogram, with one image texture feature (contrast) and the feature based on properties of light reflection that is used in clustering to distinguish lunar highlands from maria.

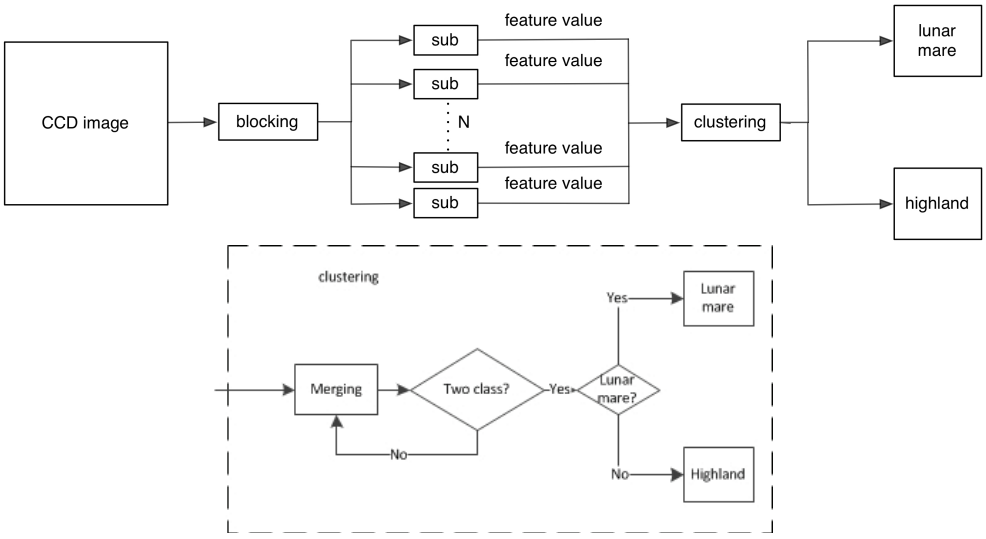


Fig. 1 The flow chart of the newly proposed algorithm.

3.1 Selecting Features with an Enhanced Image Histogram

On the lunar surface, the highlands are lighter in a CCD image because they are made of light-colored anorthosite, so most of the pixels may be distributed in areas with higher gray-level values; however lunar mare is made of basalt, which is darker than soil from the highland, so most of these pixels are distributed in areas with lower gray-level values. In this paper, a feature has been designed to represent the different gray level properties based on an enhanced gray-level histogram. The raw gray-level histogram has 256 bins. To reduce the number of bins in the histogram, an algorithm that clusters every 32 successive gray levels into one bin is applied, generating 8 bit pixels. Then the new feature F_{hist} has been defined as the median value in the range which has the maximum number of pixels in the modified histogram.

Figure 2 shows different types of areas that have different F_{hist}values, for example (a) and (b) are highland areas, but (c) is lunar mare. From Figure 2, it can be seen that the F_{hist} value in the mare area is smaller than the values in highland areas.

3.2 Selecting Features with Properties of Light Reflection

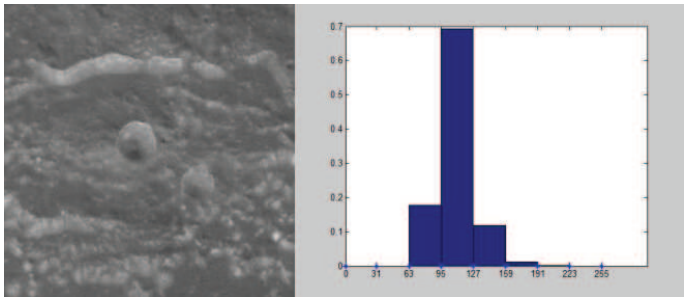
As the lunar mare and lunar highland have different geological features, the lunar highland has larger variations in terrain, and the lunar mare is relatively flat. Under the same condition of illumination, the CCD image in an area of lunar highland can produce mare bright and dark areas in different directions. Because the lunar highlands are also not flat, the area which fluctuates obviously produces shaded parts and lighter parts in different directions (bright and dark areas). However, this is not true for areas in the lunar mare.

According to the following description, we construct a new feature based on the reflection properties of light.

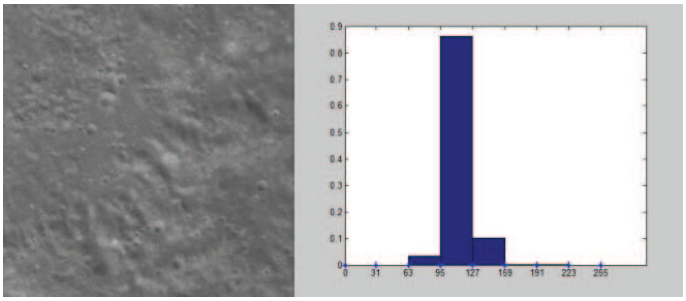
First, it calculates the gradient value for every point in the CCD image.

$$F_x(i,j) \;=\; [F(i,j+1) - F(i,j-1)]/2\,, \tag{1}$$

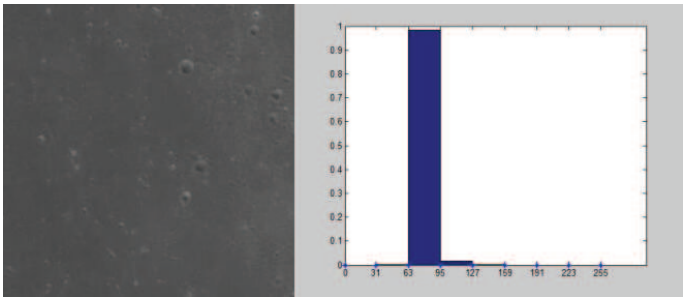
$$F_y(i,j) \;=\; [F(i+1,j) - F(i-1,j)]/2\,, \tag{2}$$



(a)



(b)



(c)

Fig. 2 The examples of feature F_{hist} with an enhanced image histogram. (a) $F_{\text{hist}} = 111$; (b) $F_{\text{hist}} = 111$; (c) $F_{\text{hist}} = 79$.

where F_x is the transverse gradient and F_y is the longitudinal gradient.

Then, after changing the gradient to angle θ, the algorithm builds a new image where the pixel value is the angle value.

$$\theta_{(i,j)} = \begin{cases} 0 & F_x(i,j) \geq 0 \ \& \ F_y(i,j) = 0 \\ \arcsin\left(\frac{F_y(i,j)}{\sqrt{F_x(i,j)^2 + F_y(i,j)^2}}\right) & F_x(i,j) \geq 0 \ \& \ F_y(i,j) > 0 \\ 180 - \arcsin\left(\frac{F_y(i,j)}{\sqrt{F_x(i,j)^2 + F_y(i,j)^2}}\right) & F_x(i,j) < 0 \\ 360 + \arcsin\left(\frac{F_y(i,j)}{\sqrt{F_x(i,j)^2 + F_y(i,j)^2}}\right) & F_x(i,j) \geq 0 \ \& \ F_y(i,j) < 0 \end{cases} . \quad (3)$$

Second, it calculates the global standard deviation of the new image as below

$$F_{\mathrm{aSD}} = \sqrt{\frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} (\theta_{(i,j)} - \overline{\theta})}, \tag{4}$$

where $\bar{\theta}$ is an average value of θ.

The global standard deviation of the new image is a measure of the light reflection properties. When the value related to the light reflection properties is greater, the area is not flat.

3.3 Deriving Texture from the Light Reflection Properties

Texture is an important concept in image analysis and recognition. Texture is related to the physical properties of the surface. Tamura et al. (1978) defined the following features: coarseness, contrast, directionality, regularity and roughness. After some testing and analysis, contrast has been selected to distinguish between lunar mare and highland.

Tamura's contrast is a global variable in an image. It shows the distribution of gray levels in pixels. The contrast can be calculated by the following steps:

(1) Calculate kurtosis (Joanes & Gill 1998) of the gray values in an image

$$\alpha_4 = \frac{\mu_4}{\sigma^4}\,, \tag{5}$$

where μ_4 is the fourth moment about the mean and σ^2 is the variance.

(2) Calculate global image contrast

$$F_{\mathrm{con}} = \frac{\sigma}{(\alpha_4)^n}\,, \tag{6}$$

where $n = 1/4$ is an empirical value used in image processing.

Figure 3 shows that different types of areas have different F_{con} and F_{aSD} values, for example (a) and (b) are highland areas, while (c) is lunar mare.

4 LUNAR DATA CLASSIFICATION

The objective of data clustering is to divide image data into meaningful or useful groups using some type of similarity measure. Data clustering does not necessarily rely on training or supervision.

After data normalization, the blocks are clustered by a merging algorithm, as Figure 4 shows.

4.1 Data Normalization

Before clustering, data x should be normalized (Kreyszig 1979). The goal is to equalize the size or magnitude and the variability of these features.

$$Z(x) = (x - \overline{x}) \Big/ \left(\sqrt{\frac{(x - \overline{x})^2}{n}} \right), \tag{7}$$

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i\,, \tag{8}$$

where $\bar{x}$ is the mean value of the data and n is the size of the original data. $Z(x)$ is the normalized result.

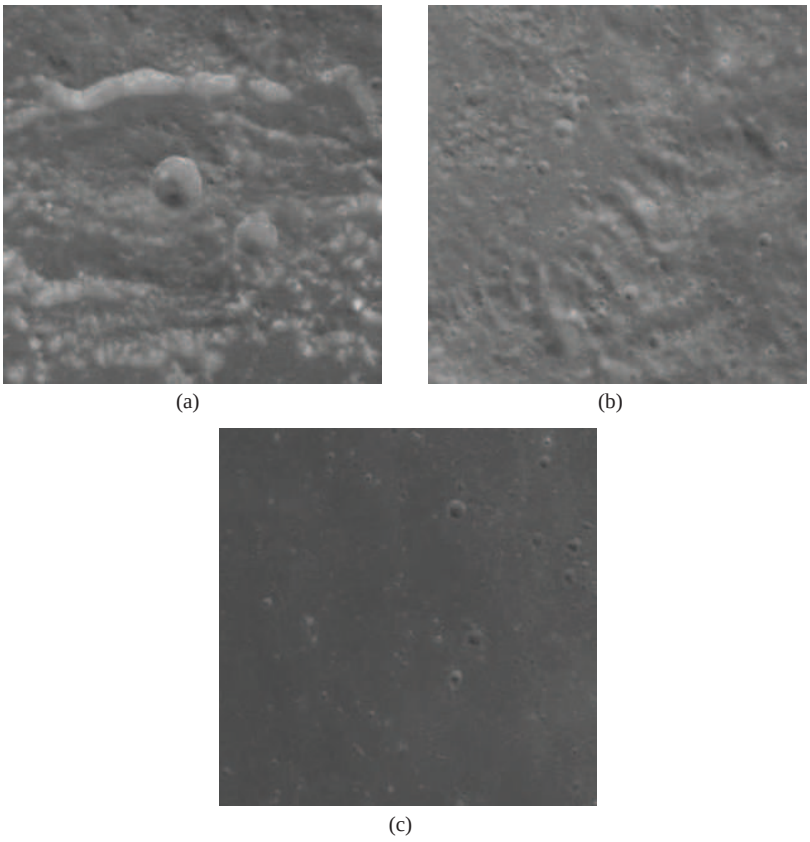


Fig. 3 Examples of features F_{con} and F_{aSD}. (a) $F_{\mathrm{con}} = 10.7860$, $F_{\mathrm{aSD}} = 99.1317$; (b) $F_{\mathrm{con}} = 7.6294$, $F_{\mathrm{aSD}} = 99.0094$; (c) $F_{\mathrm{con}} = 2.9216$, $F_{\mathrm{aSD}} = 94.7617$.

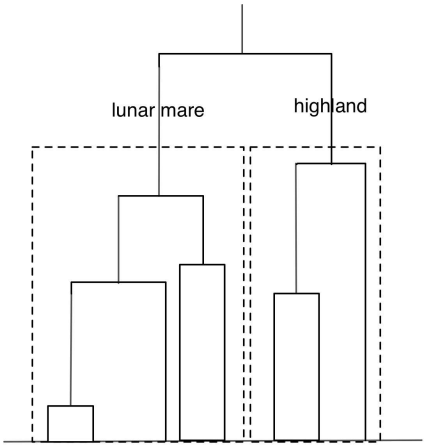


Fig. 4 A diagram illustrating the process of hierarchical clustering with the newly proposed algorithm.

4.2 Data Clustering

4.2.1 Clustering method

The merging algorithm (De & Marques 2001) is computationally easier and quicker than the splitting algorithm and produces similar solutions, so we decided to use the merging algorithm. The merging algorithm consists of the following steps:

(1) Each data point x_i is considered to be a singleton cluster as $\omega_i = \{x_i\}$, and setting $c = n$, c is the total number of initial patterns.
(2) While $c \geq 1$ do:
  (2.1) Determine the two nearest clusters ω_i and ω_j using an appropriate similarity measure and decide which one will be combined according to the criterion defined in 4.2.2 below.
  (2.2) Merge ω_i and $\omega_j : \omega_{ij} = \omega_i, \omega_j$, thereby obtaining a solution with $c - 1$ clusters.
  (2.3) Decrease c $(c = c - 1)$.

Notice that in step (2.1) the determination of the nearest clusters depends both on the similarity measure and the rule used to assess the similarity of pairs of clusters. Figure 4 shows a merging procedure used to classify lunar maria and highlands.

4.2.2 Clustering rule

Selecting an appropriate clustering distance

When clustering is applied, distance information between data from objects represents the degree of resemblance among samples. Resemblance not only depends on the samples' level of similarity, but it also depends on the properties of objects, in that there is some difference in the importance of each data point. Data points are assigned weights in light of their importance. In this paper, the algorithm uses Euclidian distance with weight defined as (Song et al. 2007)

$$d_{ab} = \sqrt{\sum_{k=1}^{n} \frac{(x_{ak} - x_{bk})^2}{s_k}}, \tag{9}$$

where $s_k (k = 1, 2, 3)$ has been used in this paper, which is inversely proportional to a feature's weight.

Table 1 shows different values of $S_k (k = 1, 2, 3)$, where S_1 is for feature $F_{\rm hist}$, S_2 is for feature $F_{\rm con}$, and S_3 is for feature $F_{\rm aSD}$.

Table 1 Values of S_k for different features

	$F_{\rm hist}$	$F_{\rm con}$	$F_{\rm aSD}$
Values of S_k	1	1.5	2

Selecting the appropriate similarity measure

Five different similarity measures (Ward & Joe 1963; Szekely & Rizzo 2005; Zhang et al. 2012, 2013; Zhao & Tang 2009) have been tested in several independent applications of Chang'e-1 data. Ward's method (Ward & Joe 1963) was selected here, since it shows the best classifying results among the similarity rules that were tested. In Ward's method, the sum of the squared within-cluster distances is computed

$$D(\omega_i, \omega_j) = \frac{1}{n_i + n_j} \sum_{x \in (\omega_i, \omega_j)} \|x - m\|^2, \tag{10}$$

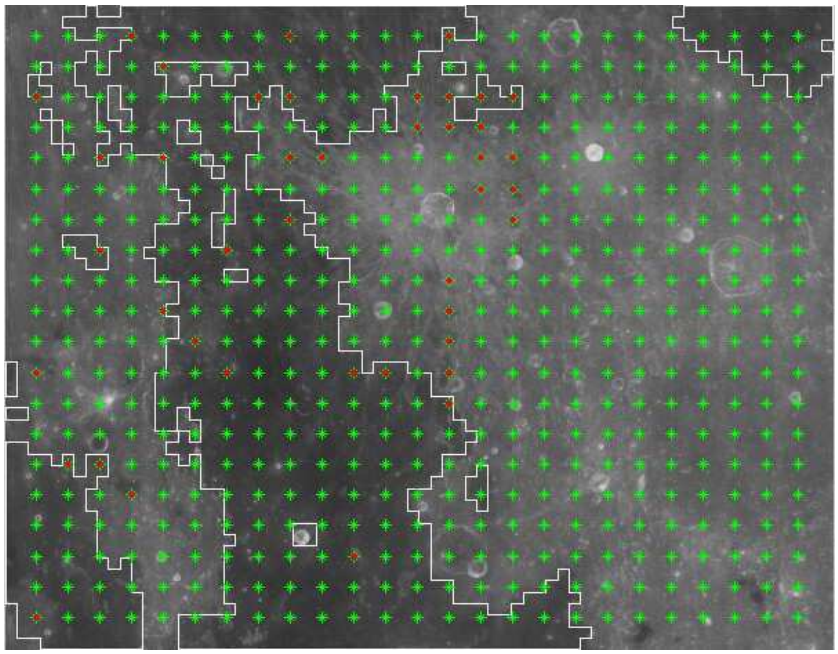


Fig. 5 The result of a test (the red dots represent misclassified points, and the green dots represent correctly classified points).

where m is the centroid of the merged clusters. Then it merges the pair of clusters with the minimum in the overall sum of squares.

5 TESTING RESULTS

The new algorithm has been tested for the purpose of comparison and applied to Chang'e-1 CCD data from the area $H010$ ($0°\mathrm{S} - 14°\mathrm{S}, 0°\mathrm{W} - 18°\mathrm{W}$), which has also been tested with a previous algorithm (Zhou et al. 2011). In the test, the size of the initial sub-block is 50×50. The total number of blocks is 4032. The material can be classified into two categories based on geological mapping as mare material and non-mare material. In total, 500 points were evenly sampled: 25 points per line for 138 lines in the horizontal direction and 20 points per column for 133 columns in the vertical direction. The correlations have been analyzed between the test result and geological data which are published by USGS.

Figure 5 shows a comparison of the results. The red dots represent misclassified points, and the green dots represent correctly classified points.

To evaluate the result, the Kappa coefficient of agreement for the newly proposed algorithm has been calculated. The Kappa coefficient of agreement is 0.802. It is better than the previous algorithm (Zhou et al. 2011) where the Kappa coefficient of agreement was 0.78.

6 DISCUSSION AND CONCLUSIONS

A new algorithm has been proposed and tested. The results show that the algorithm can accurately distinguish between images of the lunar mare and those of lunar highland. The automatic clustering results are almost the same as the result of manual classification and only very few points have been

misclassified. The new algorithm's classification accuracy is better than the traditional algorithm, and the new algorithm uses single data points, which could not only avoid the registration error between different sources of data, but also could improve classification accuracy and save processing cost.

At the border of lunar mare and highland, the algorithm can be improved further. The trouble is due to the fact that the algorithm is based on blocking, and the size of sub-blocks is not small enough. So, the sub-image at the border may contain both lunar mare and highland. Another problem is that some small areas of lunar mare are surrounded by large areas of highland, which also cannot be distinguished well.

In future work, the new algorithm will be optimized further mainly in two aspects. In the first aspect, it will try to increase classification accuracy. It will improve the image features that are applied by using the relationship between the image characteristics and geologic characteristics; not only that, dynamic sub-graphs will be used to improve classification accuracy. The second aspect will be to try to distinguish more detailed structures. It will add more image texture features. For example, a combination of image roughness features and texture orientation features may be used to recognize a mountain in the highlands and future research may be able to judge the range of the mountain range after further alignment.

Acknowledgements This work is supported by the Science and Technology Development Fund, Macao SAR, China (No. 048/2012/A2).

References

Brooks, C. A., & Iagnemma, K. 2005, Robotics, IEEE Transactions on, 21, 1185
De, S., & Marques, J. P. 2001, Pattern Recognition: Concepts, Methods, and Applications (Springer)
Fujita, K., & Ichimura, N. 2011, in AIAA Guidance, Navigation, and Control Conference Series, 411, 251
Joanes, D. N., & Gill, C. A. 1998, Journal of the Royal Statistical Soci ety: Series D (The Statistician), 47, 183
Kreyszig, E. 1979, Advanced Engineering Mathematics (10th edn.: Wiley Press)
Sancho-Pradel, D. L., & Gao, Y. 2010, Journal of the British Interplanetary Society, 63, 206
Shankar, U. J., Shyong, W.-J., Criss, T., & Adams, D. 2008, in Aerospace Conference, 2008 IEEE, 1
Song, Y. C., Zhang, Y. Y., & Meng, H. D. 2007, Computer Engineering and Applications, 43, 179
Szekely, G. J., & Rizzo, M. L. 2005, Journal of Classification, 22, 151
Tamura, H., Mori, S., & Yamawaki, T. 1978, Systems, Man and Cybernetics, IEEE Transactions on, 8, 460
Ward, J., & Joe, H. 1963, Journal of the American Statistical Association, 58, 236
Zhang, W., Wang, X., Zhao, D., & Tang, X. 2012, in Computer Vision–ECCV 2012, Lecture Notes in Computer Science (Springer), 428
Zhang, W., Zhao, D., & Wang, X. 2013, Pattern Recognition, 46, 3056
Zhao, D., & Tang, X. 2009, in Advances in Neural Information Processing Systems 21, eds. D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Curran Associates, Inc.), 1953
Zhou, Z. B., Cheng, W. M., Zhou, C. H., Wan, C., & Cao, Y. Y. 2011, Chinese Sci Bull (Chinese Ver), 56, 18