ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 168

International Conference on Modeling, Analysis, Simulation Technologies and Applications (MASTA 2019)

Intelligent Identification of Traditional Chinese Medicine Materials Based on Multi-feature Extraction and Pattern Recognition

Rong-rong CHEN* and Ying-jun CHEN

School of Electronic and Electrical Engineering, Zhaoqing University, Zhaoqing, Guangdong, China

*Corresponding author

Keywords: Traditional Chinese Medicine (TCM) material, Feature extraction, Image recognition, K-Nearest Neighbor (KNN), Support Vector Machine (SVM).

Abstract. A discussion about image pattern recognition for Tradition Chinese Medicine (TCM) materials was explained in this paper. 150 images of each category of TCM materials were gathered, in total of five categories. 80% of the images were distributed as training samples randomly and the other 20% were used to test the pattern recognition algorithms. A multi-feature vector for each image was proposed including textual features, shape features and category labels to train pattern recognition methods K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) and test the recognition rates. Statistics of average recognition rates were made and indicated that the methods could classified the chosen five categories of TCM materials significantly with the accuracy of around 70% in average, providing a new solution for TCM materials intelligent identification.

Introduction

Traditional Chinese Medicine (TCM) materials had become the concern of medicine and pharmacology in the world. Modernize and modernize development of TCM may become the inevitable tendency; however, lack of standardized quality monitoring system restricted the development of TCM. Therefore, quality control appears particularly important, which depends on various identification methods. Observation, touch, smell and taste are the conventional discrimination methods of traditional Chinese medicinal materials by human senses to perform quality identification. In the history of China, there were several famous writings about TCM discrimination. Shen Nong's Classic of Materia Medica was the earliest existent work of TCM in the world, which summarized property and medicine before Han Dynasty. Li Shizhen's Compendium of Materia Medica, including 1892 medicines and attaching 1109 images of medicines, had been held in high esteem since it was first published. These ancient TCM writings proved that image identification done by human vision was an important method in TCM identification.

Limitations of these artificial identification methods may be the results relied on subjective factor of the judgment, such as physiology, experiences, mood, etc. On the other hand, environment luminance and visual fatigue may also be some influence factors. Hence, objective methods seemed to be necessary, including microscope [1], physical-chemical analysis [2, 3], mass-spectrum analysis [4], and fingerprint [5, 6], machine olfaction (e-nose) [7-10]. However, methods mentioned in reference [1-6] were time-consuming and the operation of the instrument was complex; samples had to be destroyed and continuous observation to the same sample could not be carried on. For microscope and machine olfaction methods, expensive equipment were required, which was hard to be wild spread applied.

Machine vision is a new technology imitating human vision to make a decision based on image processing and pattern recognition; it is simpler and more feasible for users than other methods mentioned above since photographic equipment and computer analysis were required. In application of TCM identification, O. Tao etc. had put forward a native Bayes and BP neural network identification model of traditional Chinese medicine based on texture feature parameters of slice image, and established a discrimination model of 18 kinds of traditional Chinese medicines [11]. C. Liu etc. proposed to extract the color, texture and shape features of Chinese medicinal materials, and combined with BP neural network to do classification [12]. These researches had selected TCM

slice to study commonly, while there are other forms of Chinese medicine, for example, seeds, fruits, leafs, flowers, and so on, whose identifications are also needed to be studied.

In this passage, a multi-feature extraction method was proposed including textual features, shape features and category labels to train and test pattern recognition models. Recognition rates statistics were made and discussed; a new solution for TCM materials intelligent identification was provided for TCM planting and processing companies or institutes.

Pattern Recognition Methods

K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) are two of the most famous pattern recognition algorithms, playing important roles in machine learning.

K-Nearest Neighbor (KNN)

The general idea of KNN was to computer the distances between a point *A* and other points, finding out the nearest *k* points and counting out the category with largest proportion among the *k* points, and then the point *A* was classified to this category [13,14]. The steps of KNN were as follows:

1. Distance computation: given a test object, calculating the distances to each object in the training set.
2. Finding neighbors: the nearest *k* training objects were selected as the nearest neighbors of the test object.
3. Classification: classifying the test objects according to the main categories of the *k* nearest neighbors.

The distance between two points in space was to measure the similarity of them: the greater the distance, the less similar the two points were. Common distance measure methods were Euclidean distance and Mahalanobis distance.

Euclidean distance of n dimension vectors $X_i\left(X_{i1}, X_{i2}, \cdots, X_{in}\right)$ and $X_j\left(X_{j1}, X_{j2}, \cdots, X_{jn}\right)$ was shown in (1):

$$D_{\text{euc}}\left(X_i, X_j\right)=\sqrt{\sum_{k=1}^{n}\left(X_{ik}-X_{jk}\right)^2}=\sqrt{\left(X_i-X_j\right)\left(X_i-X_j\right)^T} \tag{1}$$

Mahalanobis distance of vectors Xi and Xj was shown in (2):

$$D_{\text{mah}}\left(X_i, X_j\right)=\sqrt{\left(X_i-X_j\right)S^{-1}\left(X_i-X_j\right)^T} \tag{2}$$

S stands for the covariance matrix of data. If S is an identity matrix, the Mahalanobis distance of X_i and X_j equals to their Euclidean distance.

Support Vector Machine (SVM)

SVM is a learning machine suitable for high dimension and small sample data training introducing the kernel function. At present, there are several kernel functions in the study of kernel machine learning, and different kernel functions have their corresponding uncertain parameters [15]. The common kernel functions are shown as (3)-(5):

1. Polynomial Kernel function

$$K\left(x, x_i\right)=\left[\left(x \cdot x_i\right)+1\right]^q \tag{3}$$

Where q is the order of polynomials, and a q order polynomial classifier was obtain.

2. Radial Basis Function (RBF)

$$K\left(x, x_i\right)=\exp\left\{-\frac{\left|x-x_i\right|^2}{\sigma^2}\right\} \tag{4}$$

The center of each basis function corresponded to a support vector, which and the output weights are determined automatically by the algorithm. The inner product function in the form of radial basis was similar to the visual characteristics of human beings and was often applied in practical.

3. Sigmoid Kernel function

$$K\left(x, x_i\right) = \tanh\left[\gamma\left(x \cdot x_i\right) + c\right] \tag{5}$$

The SVM algorithm contained a hidden layer of multilayer perceptron network, both the weight of the network and the number of hidden layer nodes were decided by the algorithm automatically.

Experiment

Sample Selection

Because of the regional growth characteristics, the same TCM plants in different regions may show differences in shape, color, texture, especially effects, that's why genuineness was emphasized. Another problem must be noticed was that in China, with its multiplicity of dialects and cultures, different TCMs may have the same name in various regions. For someone with little background knowledge in TCM identification, it may cause confusion and serious consequence may emerge due to the very different effects. Thus, a kind of TCM may be divided to genuine medicinal materials, real materials but not genuine ones and fake ones. Fructus Amomi, which was called "Sha Ren" in Chinese, was one of the most famous TCM materials with prominent curative effect in stomach disease, with the form of dried fruits. However, only the one originated in Yangchun, Guangdong can be called "Yangchun Sha" as the genuine medicine material recorded in Chinese pharmacopoeia. The same materials produced in Hainan province was called "Hainan Sha" and the one produced in Burma or Yunnan, China was called "Lvke Sha", both of which were real materials. But there existed two other materials named Amomum kravanh and Alpinia zerumbet Pers with names of "Yan Shan Jiang" and "Dou Kou" in Chinese respectively, which were also called "Sha Ren" in some provinces of China. But the efficacy of the two medicines is completely different from Fructus Amomi. A common approach to distinguish these materials was necessary by image recognition.

In this research, five different kinds of medicine samples called "Sha Ren" mentioned before in daily Chinese were collected. The five kinds were identified as "Yangchun Sha", "Hainan Sha", "lvke Sha", "Dou Kou" and "Yan Shan Jiang" by certificated traditional Chinese pharmacist of Guangdong Mai Lin Ke Biological Technology Co. Ltd., labeled as C1, C2, C3, C4 and C5 respectively. Images of each category of material were photoed by a Cannon digital camera from a fixed position. A shot box with white background and fixed illuminate system was used to insure the photo condition. MATLAB R2015a was used to processing data and built image recognition algorithms. Typical images of these five materials were showed as Fig. 1(a)-(e).

(a) "Yangchun Sha" ----C1 (b) "Hainan Sha" ----C2 (c) "Lvke Sha" ----C3 (d) "Dou Kou" ----C4 (e) "Yan Shan Jiang" ----C5

Figure 1. Typical images of five TCM materials categories selected in this research

Images Preprocessing

150 images of each category of samples were collected and 80% of them were assigned randomly as training samples, others were used as testing samples; which meant that 120 images were distributed for training, while the other 30 images were distributed for testing. Each image was cut to the size of 800*800 pixels, gray level image, binary image, edge detection were proceed in turn

and the core area for feature extraction was defined finally. Substep results of one image for category C1 was shown in Fig. 2.

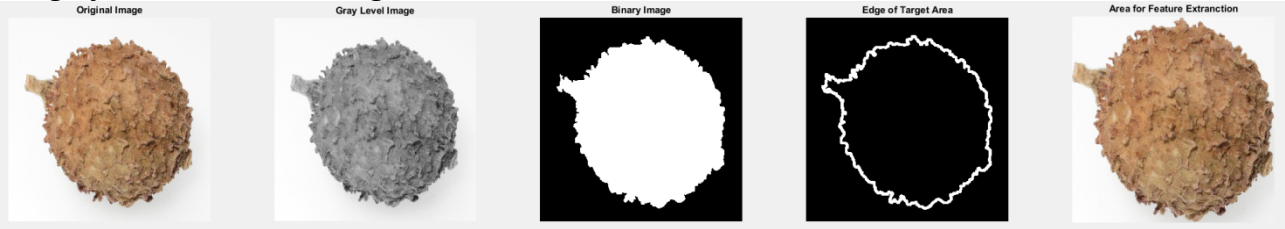

Figure 2. Substeps of image preprocessing

Feature Extraction

Next, textural features based on Gray level co-occurrence matrix (GLCM) were extracted, then a vector with eight feathers was generated [16]. According to the shape characteristic of the samples, ellipses were used to approximate the outline of the object in each image. The ratio of major and minor axis of the ellipse and the eccentricity would be used as the two shape features of the target area of the image. Together with a label from "1" to "5" to indicate the category of each kind of samples, a vector with eleven features was generated for each image, which could be called a "multi-feature vector". After feature extraction, a training feature matrix with 600*11 elements and a testing feature matrix with 150*11 elements were obtained. Both feature matrices were saved as ".mat" files.

Identification

In this passage, both KNN and SVM algorithms were modeled by MATLAB. 150 images were distributed randomly 10 times and repeated the process of training and testing, 10 groups of data for each algorithm were recorded and statistics were made. The average recognition rate for each category and each method were recorded in Table 1 and Table 2 respectively. The results are given to show the accuracy the algorithms.

Table 1. Numbers of testing results by KNN in 30 testing samples

Actual Category / Testing Result	C1	C2	C3	C4	C5
C1	**23.8**	1.3	2.3	2.6	0
C2	3.9	**18.875**	9.8	1.4	0.1
C3	5.1	9.9	**11.3**	3.2	0.3
C4	4.9	2	2.1	**19.4**	1.6
C5	0	0	0	1.2	**28.8**

Table 2. Numbers of testing results by SVM in 30 testing samples

Actual Category / Testing Result	C1	C2	C3	C4	C5
C1	**23.7**	1.9	3.1	1.3	0
C2	2.6	**20.7**	6.1	0.6	0
C3	4.5	10.8	**12.7**	2	0
C4	1.9	0.4	0.9	**26.8**	0
C5	0.2	0	0.1	0.2	**29.5**

The data on the diagonal line in Table 1 and Table 2 indicates the number of samples that were correctly identified for each category of samples in the average of 10 times for KNN and SVM; while the others were all misidentification, whose data reflecting the extent of confusion between the two types of samples. For example, Line 3 of Table 2 indicated that in 30 testing samples of C3, 4.5 of which were misidentified to be C1, 10.8 were misidentified to be C2 and 2 were misidentified to be C4. It means that C3 was easily to be misidentified to be C2. Thus, the recognition rate of C3 was 12.7/30=42.3%.

Table 3 shows the average recognition rate of five categories of samples by the two recognition methods, which could be expressed intuitively by a bar graph, as shown in Fig. 3.

Table 3. Average recognition rate statistics

Actual Category / Mathods	C1	C2	C3	C4	C5	Average Recognition Rate
KNN	79.3%	66.3%	37. 7%	64. 7%	96%	68.8%
SVM	79%	69%	42. 3%	89. 3%	98. 3%	75.6%

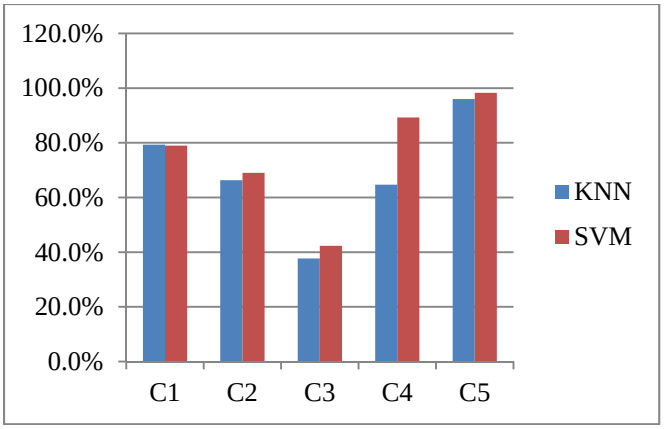

Figure 3. Bar graph of average recognition rate statistics

Conclusion

The results above indicated that both KNN and SVM could classify the five categories of TCM materials, providing a fast, objective, repeatable method for TCM discrimination. SVM performed a better recognition than KNN, especially for C4. C5 obtained the best average recognition rate because the appearance of C5 was quite different from other categories. Since C1, C2 and C3 can be considered as different sorts of “Sha Ren”, while C4 and C5 were totally different kinds, the value of data in the last two rows of Table 1 and Table 2 besides the ones located on the diagonal line was quite low, showing that these two methods could give an ideal performance for telling the true category of one material.

The recognition rate for C3 was relatively on a low side with the accuracy of about 40%, the reason may probably be: the shape and texture of C3 samples were not uniformed, since the shape, texture and withered degree of this kind of sample are not consistent, corresponding to the quality and price of C3 was the worse among C1, C2 and C3. A low accuracy may relative to low average quality of samples, which also provided an indication for discrimination. It can be referred that if C3 was removed from training and testing samples, the total recognition rate would increase, which could be considered as a further research.

Acknowledgement

This research was financially supported by Zhaoqing Scientific and Technological Innovation Guidance Projects, 2019.

References

[1] W. Mao, X. Wan, Liu Huijuan, H. Li, P. Li, Review and Applications of Microscopic Identification in Quality Standard of Chinese Herbal Medicines [J], World Science and Technology/Modernization of Traditional Chinese Medicine and Materia Medica, 2008, 16 (3): 538-542

[2] Z. Li, Y. Luo, T. Liu, Optimization of Enzymatic Extraction of Crude Polysaccharide from Rhizoma Polygonati Odorati by Response Surface Methodology [J], Journal of Guangzhou University of Traditional Chinese Medicine, 2013, 30(2): 218-221

[3] Z. Li, M. Liu, Z. He, T. Liu, Isolation, Purification and Monosaccharide Analysis of an Acidic Polysaccharide from Polygonatum odoratum [J], Chinese Journal of Experimental Traditional Medical Formulae, 2013, 19(9): 69-72

[4] M. Y. Wong, P. Soa, Z. Yao. Direct Analysis of Traditional Chinese Medicines by Mass Spectrometry. Journal of Chromatography B, 2016, 1026: 2–14

[5] D. Tang, X. Zheng, X. Chen, D. Yang, Q. Du. Quantitative and Qualitative Analysis of Common Peaks in Chemical Fingerprint of Yuanhu Zhitong tablet by HPLC-DAD–MS/MS [J], Journal of Pharmaceutical Analysis, 2014, 4(2):96–106

[6] C. Tang, L. Wang, X. Liu, M. Cheng, H. Xiao, Chemical Fingerprint and Metabolic Profile Analysis of Ethyl Acetate Fraction of Gastrodia Elata by Ultra Performance Liquid Chromatography/Quadrupole-time of Flight Mass Spectrometry [J], Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences, 2016, 1011: 233-239

[7] D. Luo, Y. Sun, J. Zhuang, H. H. Gholam, Classification of Hundred Grass Oil Samples Using E-nose[J], Journal of Computational Information Systems, 2013, 9(7): 2659-2666

[8] D. Luo, H. Chen, A Novel Approach for Classification of Chinese Herbal Medicines Using Diffusion Maps [J], International Journal of Pattern Recognition and Artificial Intelligence, 2015, 29 (1): 104-112

[9] D. Luo, H. H. Gholam, Application of ANN with Extracted Parameters From an Electronic Nose in Cigarette Brand Identification[J]. Sensors and Actuators B: Chemical, 2004, 99, (2): 253-257

[10] D. Luo, J. Wang, Y. Chen, Classification of Chinese Herbal Medicine based on SVM [C]. 2012 IET International conference on Information Science and Control Engineering(ICISCE 2012): 1191-1195

[11] O. Tao, Z. Lin, X. Zhang, Y. Wang, Y. Qiao, Research on Identification Model of Chinese Herbal Medicine by Texture Feature Parameter of Transverse Section Image [J], World Science and Technology/Modernization of Traditional Chinese Medicine and Materia Medica, 2014, 16(12): 2558-2562

[12] C. Liu, X. Wu, W. Xiong, Chinese Herbal Medicine Classification Based on BP Neural Network [J], Journal Of Software, 2014, 9(4): 938-943

[13] H.T. Cover, Nearest Neighbor Pattern Classification [J]. IEEE, 1967(1):21 - 27.

[14] H.T. Cover. Rates of Convergence for Nearest Neighbor Procedures [J]. Systems Sciences, 1968.

[15] Vapnik, V. SVM Method of Estimating Density, Conditional Probability, and Conditional Density [C]. Circuits and Systems, 2000 IEEE International Symposium on Circuits and Systems. Emerging Technologies for the 21st Century. Proceedings (IEEE Cat No.00CH36353), Geneva, Switzerland

[16] P. Jiao, Y. Guo, L. Liu, X. Wei, Implementation of Gray Level Co-occurrence Matrix Texture Feature Extraction Using Matlab [J], Computer Technology and Development, 2012, 22 (11): 169-171,175.