RAMAN SPECTRUM CLASSIFICATION OF CINNABAR AND CINNABAR-CLAM WHITE BASED ON DATA AUGMENTATION AND CONVOLUTIONAL NEURAL NETWORK**

Taotao Mu [1*], Wenbo Qi [1], Shaohua Chen [1], Haoru Li [2]

[1] Beijing Information Science and Technology University, Beijing, China; e-mail: mfjmtt@163.com
[2] Beijing University of Posts and Telecommunications, Beijing, China

Mineral pigments are commonly used in cultural relics, which makes the analysis of mineral pigments helpful in such research. It is complicated and time-consuming work to establish the data set of mineral pigment Raman spectra, so it is necessary to study the method of data augmentation. In this paper, two methods of augmenting Raman spectra data are explored – translation transformation, adding noise; expanding the size of the data set from 20 to 320 – then a convolutional neural network model is proposed and trained with the expanded data set. Experimental results showed that the accuracy of the model can reach 100% when the SNR of the test set is not less than 40 dB.

Keywords: *Raman spectrum, convolutional neural network, data augmentation, mineral pigments.*

КЛАССИФИКАЦИЯ СПЕКТРОВ КОМБИНАЦИОННОГО РАССЕЯНИЯ МИНЕРАЛЬНЫХ ПИГМЕНТОВ НА ОСНОВЕ АУГМЕНТАЦИИ ДАННЫХ И СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ

T. Mu [1*], W. Qi [1], S. Chen [1], H. Li [2]

УДК 535.375.5

[1] Пекинский университет информационных наук и технологий, Пекин, Китай; e-mail: mfjmtt@163.com
[2] Пекинский университет почты и телекоммуникаций, Пекин, Китай

(Поступила 9 марта 2022)

Исследованы методы аугментации данных спектров комбинационного рассеяния минеральных пигментов — трансляционное преобразование и добавление шума, увеличивающие размер набора данных от 20 до 320. Предложена модель сверточной нейронной сети и проверена на расширенном наборе данных. Показано, что точность модели может достигать 100 % при отношении сигнал/шум не менее 40 дБ для тестового набора.

Ключевые слова: *спектр комбинационного рассеяния света, сверточная нейронная сеть, аугментация данных, минеральный пигмент.*

Introduction. Mineral pigments are commonly used dyes in ancient times. The analysis of mineral pigments can be used as the basis for cultural relic identification, so it is necessary to develop a nondestructive method for mineral pigment detection. Nondestructive testing techniques commonly used at present include infrared spectroscopy [1–4], Raman spectroscopy [5–8], X-ray [9], etc. These technologies are widely used in chemical detection, medicine, and biology. Herein, the determination of mineral pigments by Raman spectroscopy was carried out.

Raman spectroscopy is a detection technology based on the Raman effect, which is a kind of scattering spectrum that detects substance composition by analyzing molecular vibration and rotation information. Many scholars use Raman spectroscopy to analyze pigments [10, 11]. Daniel Cosano et al. detected the pigments in the Annunciation Sculptural Group via Raman spectroscopy [12]. The artificial analysis of Raman

** Full text is published in JAS V. 90, No. 2 (http://springer.com/journal/10812) and in electronic version of ZhPS V. 90, No. 2 (http://www.elibrary.ru/title_about.asp?id=7318; sales@elibrary.ru).

spectroscopy is used, which requires the operator to consult a lot of data and have a certain amount of experience. The deep learning algorithm can solve this problem well.

As a branch of deep learning algorithms, convolutional neural network (CNN) is widely used in medicine [13–15], natural language processing [16, 17], image recognition [18, 19], and other fields. Kele Xu et al. proposed a multi-channel CNN model to solve the problem of audio scene classification [20], whereas Toshiaki Hirasawa et al. used the CNN model to detect gastric cancer in endoscopic images [21]. However, as a model based on data, insufficient data has always been one of the key problems that many scholars need to solve when using this model. Insufficient data may cause overfitting; therefore, it is necessary to extend the data set.

In this paper, a CNN model was proposed to classify the Raman spectra of the cinnabar and cinnabar-clam white mixture. Only 10 spectra of each sample were collected as training and validation sets. Data augmentation was carried out on the spectral data set by translation transformation and noise addition. The model obtained has good antinoise ability. When the signal-to-noise ratio (SNR) of the test set is no less than 40 dB, the accuracy can reach 100%.

Experimental. Cinnabar and cinnabar-clam white mixture were used to make samples. The mass ratio of cinnabar and clam white was 1:1. Because the sample belongs to the color material, the ordinary Raman spectrometer will burn the sample under the power mode of 500 mW, and thus cannot measure the spectrum. Therefore, the portable Raman spectrometer based on Digital Mirror Device and Lissajous pattern modulation was used in this experiment to collect Raman spectrum [22]. The specific parameters of the spectrometer were as follows: the laser wavelength was 785±0.02 nm; the resolution is 8–11 cm^{-1}; the laser power is set to 500 mW; the spectral range is 200–1800 cm^{-1}. In all, 60 spectra were collected for each sample, 120 spectra were collected in total, and each spectrum was 801*1 dimensional data. Figure 1 shows the collected spectral data, in which the characteristic peak of clam white is difficult to be identified, posing a challenge to the classification of the model.

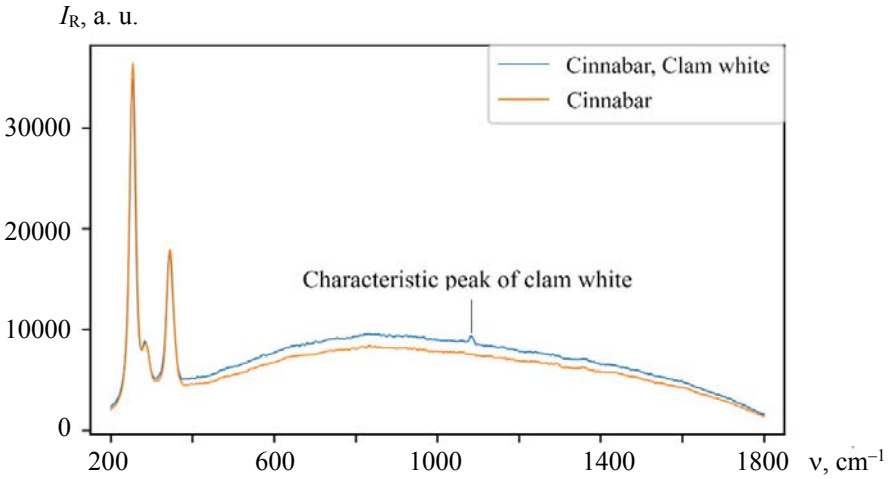

Fig. 1. Raman spectra of cinnabar and cinnabar-clam white mixture.

Deep learning is a model with strong data dependence. The size of the data set directly affects the performance of the model, whereby too little data leads to overfitting of the model (poor ability to classify unknown data). Common data augmentation methods include random scale, random crop, horizontal/vertical flip, color jittering, noise, etc. However, most of these methods are for image data, thus it is unreasonable to apply methods such as horizontal/vertical flips to spectral data.

In this experiment, to verify the performance of the data augmentation method, only 10 (20 in total) data were extracted from each type of spectrum as the original data of the training set and validation set, and the training set and validation set were expanded to 320 data by translation transformation and the noise method [23, 24]. To study the antinoise capability of the model, White Gaussian Noise was added to the test set so that the SNR of the test set was 10–70 dB. Table 1 shows the processing method and size of the final training set, validation set, and test set:

TABLE 1. Processing Methods and Size of the Training Set, Validation Set, and Test Set

Data set	Translation transformation	White Gaussian Noise (SNR, dB)	Size
Training set	✓	50–70+ non-Noise	240
Validation set	✓	50–70+ non-Noise	80
Test set	×	10–70+ non-Noise	800

Raman spectra under the influence of instrument error and measurement environment could lead to Raman offset. Therefore, translation transformation cannot solely expand the spectrum data sets, it must also be compatible with the offset to some extent. The experiments on each of the Raman spectra are at random deleted first and tail, a total of three data points, to achieve the translation of the spectrum transform. Then, the data set is expanded to be four times as large as the original dataset (80 in total).

Raman spectrum inevitably contains noise, and noise reduction has been the focus of many scholars, but there is no method to perfectly eliminate the noise at present. For the deep learning model, adding noise to the training data can improve the antinoise interference ability of the model and enhance its robustness. White Gaussian noise was added to the data after translation transformation, so that the SNR of data was 50, 60, and 70 dB, and the data set was expanded 4 times (320 in total). The ratio of the training set and validation set was 3:1.

Results and discussion. A portable Raman spectrometer has the advantages of convenient use and fast detection speed, but it is more susceptible to the influence of ambient light [25]. The CNN model has been applied in spectral analysis by many scholars [26–28]. This paper proposes a CNN model that can accurately classify the spectra collected by a portable Raman spectrometer. The CNN model proposed in this paper consists of four convolutional layers, four pooling layers, one flattened layer, two fully connected layers, and one output layer. Figure 2 shows the structure of this model. Among them, due to the shared parameters of the convolution kernel, the convolutional layer has translation invariance and is compatible with Raman offset. The calculation formula for the convolutional layer is as follows:

$$y_j^n = f\left(\sum_i x_i^{n-1} * k_{ij}^n + b_j^n\right) \quad , \tag{1}$$

where x_i^{n-1} is the ith input feature graph of the n-1th layer, * is the one-dimensional convolution operation, y_j^n is the jth output feature graph of the nth layer, k_{ij}^n is the convolution kernel used for the n-1th layer, and nth layer operations, b_j^n is the bias of the jth feature graph of the n-h layer and f is the activation function.

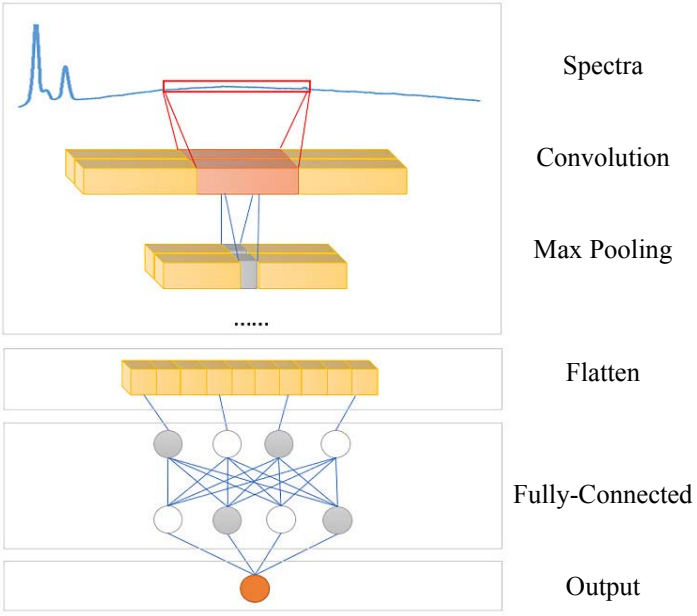

Fig. 2. Structure diagram of CNN model.

The pooling layer adopts the max pooling layer with a step size of 3 to reduce the length of data. The combination of multiple convolutional layers and pooling layers can extract the important features of Raman spectra. Flatten layer can flatten the data into the shape of a one-dimensional vector, which is convenient to input into the fully connected layer for classification. Each fully connected layer has 512 neurons; the number of neurons in a fully connected layer was determined by a large number of experiments and tests. To avoid overfitting, the dropout technology is applied to each fully connected layer. The dropout rate is set to 0.5.

$$y = f(Wx + b), \tag{2}$$

$$W = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{pmatrix}, \tag{3}$$

where W is the parameter matrix; x, y and b are the input vector, output vector, and bias vector, respectively; n is the number of neurons; and m is the length of the input vector. Rectified Linear Unit (ReLU) activation function was used to delinearize each convolutional layer and the fully connected layer, and the Sigmoid activation function was used to obtain the predicted probabilities for the output layer. Root Mean Square Prop (RMSProp) is used as the optimizer and binary cross-entropy of the loss function.

Raman spectrum is affected by an instrument error, ambient light, and other factors, which will produce noise and baseline drift; therefore, it is necessary to preprocess the spectral data before training the model. Common preprocessing steps are smoothing, baseline removal, and normalization. In this experiment, the Savitzky–Golay algorithm was used to reduce spectral noise. The Adaptive iteratively reweighted Penalized Least Squares (Air_PLS) algorithm was used to remove the baseline of the spectrum (where lambda, the parameter that controls baseline smoothness, takes the value 15), and then the spectrum is normalized to make it easy to compare features of different orders of magnitude. After several simulation experiments, the hyperparameters of the model are determined as batch size: 100 and learning rate: 0.0001. Table 2 summarizes the specific parameters of the proposed CNN model and Fig. 3 shows the loss curve of the training model. It is found that the model has no overfitting phenomenon.

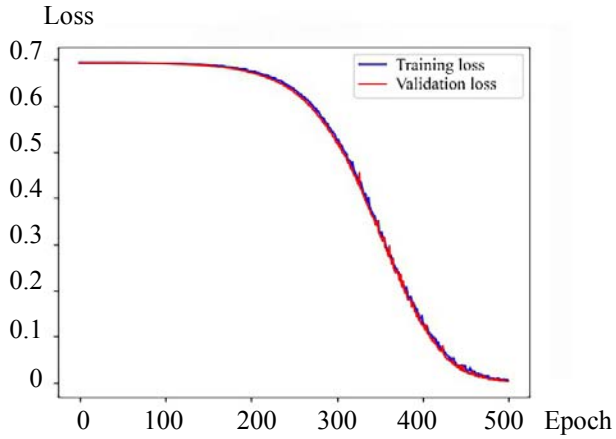

Fig. 3. Loss curve of the model.

TABLE 2. Output Shapes, Filter Size, and Stride at Each Layer of the Proposed CNN Model

Layers	Type	Output shape	Filter size	Stride	Layers	Type	Output shape	Filter size	Stride
1	Convolution	790*2	9	1	7	Convolution	25*16	3	1
2	Max-pooling	263*2	3	3	8	Max-pooling	8*16	3	3
3	Convolution	257*4	7	1	9	Flatten	128	–	–
4	Max-pooling	86*4	3	3	10	Fully connected	512	–	–
5	Convolution	82*8	5	1	11	Fully connected	512	–	–
6	Max-pooling	27*8	3	3	12	Output	1	–	–

Currently, commonly used model evaluation indicators include precision, sensitivity, specificity, accuracy, and F1Score. The calculation formula for these indicators is as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+ FP}}, \tag{4}$$

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP+ FN}}, \tag{5}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+ FP}}, \tag{6}$$

$$\text{Accuracy} = \frac{\text{TP+ TN}}{\text{P+ N}}, \tag{7}$$

$$\text{F}_1\text{Score} = \frac{2 * \text{Precision* Recall}}{\text{Precision+ Recall}}, \tag{8}$$

where TP is the number of positive samples correctly classified, FP is the number of positive samples incorrectly classified, FN is the number of negative samples incorrectly classified, TN is the number of negative samples correctly classified, P is the number of positive samples, and N is the number of negative samples. In this study, cinnabar and cinnabar-clam white mixture were designated as positive and negative samples, respectively. The test set was divided into eight pieces according to the SNR. (Note, to ensure the reliability of test results, the test set is not translated.) The test set without noise was used to evaluate the model. Precision, sensitivity, specificity, accuracy, and F1Score were all 100%, which shows that the model can correctly classify the two mineral pigments.

We tested the performance of the model using the test set with an SNR of 10–70 dB and found that the model performed well in a test set with an SNR of 40–70 dB, with precision, sensitivity, specificity, accuracy, and an F1Score reaching 100%. In a 30 dB SNR data set, its precision, sensitivity, specificity, accuracy, and F1Scores were 58, 100, 70.42, 79, and 73.42%, respectively. Twenty-one cinnabar spectra were misclassified. The model lost the ability of cinnabar classification in the SNR of 10–20 dB – precision, sensitivity, specificity, accuracy, and F1Score were 0, 0, 50, 50, and 0, respectively, and all cinnabars were misclassified.

Raman spectrum analysis showed that the characteristic peak of clam white was very small and could not be identified by the naked eye in the Raman spectrum with an SNR of 10–20 dB. According to the experimental results, it is speculated that the model is classified by analyzing whether there are peaks in the position of clam white characteristic peaks. The model mistook the noise in the cinnabar spectrum as the characteristic peak of clam white, so cinnabar was wrongly classified as cinnabar-clam white. Nevertheless, the model has demonstrated excellent performance.

Conclusions. The main research content of this paper is divided into two parts: 1) Two spectral data augmentation methods are explored, and in addition, a CNN model was proposed. A small amount of data (20 spectra) could be used to train the model. The precision, sensitivity, specificity, accuracy, and F1Score of the model could reach 100%. 2) The anti-noise capability of the model is investigated. The results show that the precision, sensitivity, specificity, accuracy, and F1Score of the model reach 100% when the SNR of the Raman spectrum is not less than 40 dB. Although the model loses the classification ability of cinnabar when the spectral SNR is less than 30 dB, the manual method of spectral analysis has become invalid at this time.

In this paper, a feasible solution is proposed using a small amount of Raman spectra for deep learning models, which has important application value in those cases where it is difficult to collect spectral data. Although the Raman spectrum is used as the data for the experiment, due to the similarity of the spectral data structures, the method proposed in this paper is expected to be applied to other spectra, providing a theoretical basis for the related research of automatic spectroscopy analysis in the future.

REFERENCES

1. J. A. Cayuela, J. F. Garcia-Martin, *LWT – Food Sci. Technol.*, **88**, 103–108 (2018).
2. D. Kim, M. H. Choi, H. J. Shin, *Agriculture*, **11**, 135 (2021).
3. D. C. Leite, A. A. P. Correa, L. C. Cunha, K. M. G. de Lima, C. D. M. de Morais, V. F. Vianna, G. Teixeira, A. O. Di Mauro, S. H. Uneda-Trevisoli, *J. Food Compos. Anal.*, **91**, 103536 (2020).
4. B. Abu Izneid, M. I. Fadhel, T. Al-Kharazi, M. Ali, S. Miloud, *J. Food Sci. Technol.*, **51**, 3244–3252 (2012).

5. C. Hartmann, M. Elsner, R. Niessner, N. P. Ivleva, *Appl. Spectrosc.*, **74**, 193–203 (2020).
6. L. Mandrile, S. Rotunno, L. Miozzi, A. M. Vaira, A. M. Giovannozzi, A. M. Rossi, E. Noris, *Anal. Chem.*, **91**, 9025–9031 (2019).
7. D. R. Zhang, H. B. Pu, L. J. Huang, D. W. Sun, *Trends Food Sci. Technol.*, **109**, 690–701 (2021).
8. X. L. Li, C. J. Sun, L. B. Luo, Y. He, *Sci. Rep.*, **5**, 15729 (2015).
9. Y. Abe, R. Shikaku, I. Nakai, *J. Archaeolog. Sci. Rep.*, **17**, 212–219 (2018).
10. C. Gurin, M. Mazzuca, J. G. Otero, M. S. Maier, *Archaeolog. Anthrop. Sci.*, **13**, 54 (2021).
11. O. Petrova, D. Pankin, A. Povolotckaia, E. Borisov, T. Krivul'ko, N. Kurganov, A. Kurochkin, *J. Cult. Herit.*, **37**, 233–237 (2019).
12. D. Cosano, D. Esquivel, C. M. Costa, C. Jimenez-Sanchidrian, J. R. Ruiz, *Spectrochim. Acta, A*, **214**, 139–145 (2019).
13. J. Jendeberg, P. Thunberg, M. Liden, *Urolithiasis*, **49**, 41–49 (2021).
14. S. A. Lee, H. C. Cho, H. C. Cho, *IEEE ACCESS*, **9**, 51847–51854 (2021).
15. G. H. Lian, Y. Peng, J. He, J. Yi, Y. N. Yin, X. W. Liu, F. Zeng, *Results Phys.*, **22**, 103912 (2021).
16. M. Gimnez, J. Palanca, V. Botti, *Neurocomputing*, **378**, 315–323 (2020).
17. W. Huang, M. Huang, *Int. J. Simulation and Process Modelling*, **15**, 120 (2020).
18. Y. F. Li, X. Y. Feng, Y. D. Liu, X. C. Han, *Sci. Rep.*, **11**, 16618 (2021).
19. S. Lingwal, K. K. Bhatia, M. S. Tomer, *Multimedia Tools Appl.*, **80**, 35441–35465 (2021).
20. K. L. Xu, D. W. Feng, H. B. Mi, B. Q. Zhu, D. Z. Wang, L. L. Zhang, H. X. Cai, S. W. Liu, *Adv. Multimedia Inform. Proc*., **11166**, 14–23 (2018).
21. T. Hirasawa, K. Aoyama, T. Tanimoto, S. Ishihara, S. Shichijo, T. Ozawa, T. Ohnishi, M. Fujishiro, K. Matsuo, J. Fujisaki, T. Tada, *Gastric Cancer*, **21**, 653–660 (2018).
22. Y. Wang, T. T. Mu, Y. G. Li, W. B. Qi, S. H. Chen, *Anal. Lett.*, **54**, 2423–2430 (2021).
23. M. Kazemzadeh, C. L. Hisey, K. Zargar-Shoshtari, W. L. Xu, N. G. R. Broderick, *Opt. Commun*., **510**, 127977 (2022).
24. F. L. Yue, C. Chen, Z. W. Yan, C. Chen, Z. Q. Guo, Z. X. Zhang, Z. Y. Chen, F. B. Zhang, X. Y. Lv, *Photodiagn. Photodyn. Ther*., **32**, 101923 (2020).
25. M. Jermyn, J. Desroches, J. Mercier, M. A. Tremblay, K. St-Arnaud, M. C. Guiot, K. Petrecca, F. Leblond, *J. Biomed. Opt.*, **21**, 094002 (2016).
26. J. Liu, M. Osadchy, L. Ashton, M. Foster, C. J. Solomon, S. J. Gibson, *Analyst*, **142**, 4067–4074 (2017).
27. D. Y. Ma, L. W. Shang, J. L. Tang, Y. L. Bao, J. J. Fu, J. H. Yin, *Spectrochim. Acta, A*, **256**, 119732 (2021).
28. Y. Zhao, S. Tian, L. Yu, Z. Zhang, W. Zhang, *J. Appl. Spectrosc.*, **88**, 441–451 (2021).