Identification for water quality based on color characteristics

Jiangang Wang¹,²,³, Zhengang Zhai¹,², Yunya Zhu¹,², Li Zhang¹,² and Xusheng Fang¹,²

¹ The 36th Research Institute of China Electronics Technology Group Corporation, 314000 Jiaxing, Zhejiang, China
² Jiaxing Key Laboratory of Water Ecology Intelligence, 314000 Jiaxing, Zhejiang, China
³ E-mail: wangjg@jec.com.cn

Abstract. It's great significance for protection of water ecological and water resources to identify water quality rapidly and conveniently. In the past time, water quality was test and monitored with traditional laboratory methods, which was hard to meet the requirements of urgent demand. A rapid and convenient method for the identification of water quality based on machine learning was used in this study. By sampling and photographing, the image of water was acquired. Then nine dimensional digital information features of the color information were obtained by the moment method. Based on the historical data and expert experience, a support vector machine (SVM) model was successfully built and well trained. Then the model was verified with the test data, and the accuracy reaches 95%, which proves this method has good effect and high precision. This work will generate fresh insight into water quality identification and contribute to water resources protection.

1. Introduction

With the rapid development of economy, water pollution has gradually become a more and more serious problem, which has a serious threat on water resources [1-3]. There is no doubt that water quality identification is an essential part of water pollution monitoring and water resources protection. Therefore, it is very important to explore a fast water quality identification method.

In recent years, the field of machine learning had developed rapidly, and its analysis and recognition technologies were widely used in image recognition, detection, and early warning, and have excellent effects [4, 5].

There have been many researches on water quality identification using various machine learning methods, and support vector machine (SVM) was one of the commonly used methods. For example, Ahmed (2019) used supervised machine learning method to study water quality index and water quality grade identification based on water quality parameters (temperature, turbidity, pH and total dissolved solids), and found that polynomial regression method and multi-layer perceptron method had better effects [6]; Haghiabi (2018) used artificial neural network (ANN), group method of data handling (GMDH) and SVM to predict the water composition of tireh River in southwest Iran [7]; Chou (2018) used machine learning (ANN, SVM, classification and regression trees, And Linear regression,) to predict the water quality indexes of the reservoir [8]; Lu (2020) used two improved decision tree models to predict six water quality indicators in Turadin River [9]; Wen (2007) built a SVM comprehensive evaluation model for water quality based on a variety of water quality indicators [10]; Ladjal (2016) used ANN and SVM in the Tilesdit dam (Algeria) area, combined with Dempster-
Shafer evidence theory (DSTE) to evaluate surface water quality [11]; Bouamar (2007) evaluated the effects of ANN and SVM on water quality classification through experiments [12]; Khotimah (2015) used the improved prediction model of smooth support vector machine (SSVM) to predict aquaculture water quality [13]. These studies proved the good results of SVM in water quality identification.

However, most of the data used in these studies were obtained using laboratory analysis methods, which is very accurate, but time-consuming, expensive and environmentally limited. The purpose of this research is to identify water quality through a kind of water data that is easy to obtain.

Water image is a kind of ideal data, whose acquisition speed is fast, the price is low, is not limited by space environment conditions. At the same time, water quality about water image is also supported by related theories, and there are also some researches in the field of remote sensing water quality evaluation.

Color feature is an important feature that reflects the properties of objects [14-16]. It can reveal some properties of objects. Water can't shine. The light perceived by the observer is reflected and scattered as a result of the interaction between light and substances in the water. It is closely related to the absorption and scattering of chlorophyll [17, 18].

The observation of water color can be traced back to the 1930s. The Forel-Ulescale has divided the natural water color into 21 levels ranging from dark blue to reddish brown. It is used to record the color of global oceans and inland waters, and then evaluate the water situation. After the satellite era, researchers developed the water color record on satellite images, and obtained the water color level Forel-Ule Index based on the water reflectance of the image, and then developed the water color remote sensing [19-29].

Bilge et al. (2003) used water images acquired by Landsat satellites to study the relationship between image spectra and water parameters, the water index in the Porsuk dam reservoir was estimated [30]; Mushtaq (2017) et al. used Landsat 8 OLI satellite data to estimate the water quality parameters of the Ural Lake (including pH, COD, DO, TDS, total suspended solids, etc.) [31]; Yu (2016) used Moderate Resolution Imaging Spectrometer (MODIS) data to estimate the concentration of dissolved inorganic nitrogen (DIN) [32].

This study uses the color of the water image for analysis to explore a new method --- a quick and easy water quality identification method. Based on the collection of multiple types of water samples, the image color features were obtained, and the SVM classifier was used to classify the samples, and finally a water color-based rapid identification method of water quality was built.

2. Methods

2.1. Sample collection and feature extraction

2.1.1. Sampling and photographing. Water samples in this study were collected from rivers, lakes and ponds, each sample collecting water of 500 ml. Figure 1 shows the photos of some sampling sources.

Figure 1. Some sampling sources.
Images of water samples were obtained with camera (Canon EOS 40D). A total of 203 original images were collected, and the image format was JPEG, and the original resolution of each water sample was 2352 × 1568 pixels. Some pictures show in Figure 2.

![Figure 2. Some sampling pictures.](image1)

All water samples were also analyzed in the laboratory, and classified into 5 classes according to China's Environmental quality standards for surface water (GB 3838-2002) [33].

2.1.2. Image processing and feature extraction. Irrelevant image areas were removed in this study so the original image was clipped. The specific method is to intercept the uniform water area in the center of each image, and the minimum size of the captured image is 101×101. Some clipped pictures show in Figure 3.

The water image is composed of data in three dimensions: red, green and blue. Extracting image features is a good method to explore data information. Digital image features mainly include color features, texture features, shape features, spatial relations and so on. Compared with other features, color features are more robust but insensitive to the size and direction of objects, which makes it perfect for feature extraction.

Stricker proposed a simple and effective method to represent color features: color moments [29]. The advantages of this method are: no space quantization is required, the dimension of feature vector is low, and it is very simple. Commonly used are the first, second and third moment, which are sufficient to express the color features of the image [29].

![Figure 3. Original picture (left) and cropped picture (right).](image2)

The first, second and third color moment can be interpreted as \(E_i\), \(H_i\), \(S_i\), and their formulas can be described as follows (1), (2), (3).
\[ E_i = \sum_{j=1}^{N} \frac{1}{N} p_{ij} \]  

(1)

\[ H_i = \sqrt{\left(1 - \frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^2 \right)} \]  

(2)

\[ S_i = \sqrt{\left(1 - \frac{1}{N} \sum_{j=1}^{N} (p_{ij} - E_i)^3 \right)} \]  

(3)

In these formulas, \( i \) is the color channel, \( N \) is the total number of pixels and \( p_{ij} \) is the value of the \( j \)-th pixel of the image at the \( i \)-th color channel [29].

Table 1 shows some of the input data.

**Table 1.** Part of the sample data.

| number | Class | \( E_1 \) | \( H_1 \) | \( S_1 \) | \( E_2 \) | \( H_2 \) | \( S_2 \) | \( E_3 \) | \( H_3 \) | \( S_3 \) |
|--------|-------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
| 1      | 2     | 0.492    | 0.508    | 0.223    | 0.009    | 0.006    | 0.013    | 0.007    | 0.008    | 0.010    |
| 2      | 2     | 0.585    | 0.531    | 0.213    | 0.009    | 0.007    | 0.011    | -0.004   | 0.004    | 0.006    |
| 3      | 1     | 0.543    | 0.547    | 0.306    | 0.010    | 0.007    | 0.010    | 0.006    | 0.004    | 0.006    |
| 4      | 2     | 0.563    | 0.537    | 0.172    | 0.008    | 0.006    | 0.011    | 0.004    | 0.002    | -0.005   |
| 5      | 1     | 0.544    | 0.566    | 0.296    | 0.007    | 0.005    | 0.010    | 0.002    | 0.001    | -0.008   |
| 6      | 0     | 0.577    | 0.539    | 0.281    | 0.020    | 0.015    | 0.012    | 0.010    | -0.005   | 0.007    |
| 7      | 3     | 0.463    | 0.473    | 0.164    | 0.008    | 0.006    | 0.011    | -0.004   | -0.004   | -0.007   |
| 8      | 2     | 0.550    | 0.538    | 0.229    | 0.008    | 0.008    | 0.012    | -0.003   | -0.003   | 0.003    |

Table 1 shows the final organized data. In the table, there are data of 8 samples (203 in total), in which the first column of data describes the number and the second describes water quality type of samples (there are 5 types in total, represented by numbers 0–4). \( E_i \), \( H_i \) and \( S_i \) are the first, second and third moment of the original band data. The subscripts 1, 2, and 3 represent the three original color bands: red, green, and blue, respectively.

2.2. SVM model for water quality recognition

According to the demand of rapid identification for water quality, to meet the need to identify multiple types of water quality, the multi-classification SVM-SVC (Support Vector Classification of Support Vector Machine) algorithm was used to build an intelligent water quality identification model to quickly identify water quality in a wide range of areas and at multiple scales. The whole recognition model is shown in Figure 4.

![Figure 4. Model for water quality identification.](image)

The whole water quality identification model can be divided into four stages: sample collection, data processing, modeling training and result evaluation. After collecting water samples, data
processing was carried out. It mainly includes sample data photographing, image cutting processing, color moment feature extraction, data format organization and normalization. Then, the modeling training was carried out, and the SVM-SVC multi-classification model was built by Python language and relevant machine learning methods, and the training data and testing data were used for model training and verification. Finally, in the result evaluation stage, the results of the model training test are evaluated, and the quality of the results, the feasibility of the method and the future application prospect were evaluated, and the whole study was analyzed reasonably.

2.2.1. SVM. The original SVM was a binary classifier. For a given training sample set \( D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \), \( y_i \in \{-1, +1\} \), the basic idea is to find a partition hyperplane in the sample space based on training set \( D \) and separate samples of different classes [34].

Any hyperplane can be written as the set of points \( x \) satisfying:

\[
\mathbf{w}^T \mathbf{x} + b = 0
\]  

(4)

\( \mathbf{w}, b \) are the normal vector and intercept of the hyperplane.

Obviously, the key problem can be expressed as

\[
f(x) = \text{sgn}\left(\sum k_i y_i (x_i x) + b\right)
\]

(5)

\( k_i \) is Lagrange multiplier, \( x_i \) and \( y_i \) are support vectors, \( b \) is intercept.

For nonlinear classification problems, the samples are generally mapped to a high-dimensional space first, and an appropriate mapping function (kernel function) \( K(x, x) \) can be selected to make the sample mapped linearly separable. Common kernel functions include Polynomial (homogeneous), Polynomial (inhomogeneous), Gaussian radial basis function, etc [34].

2.2.2. SVM-SVC. In practical application scenarios, more problems are multi-classification, same in this study. The multi-classification problem based on SVM currently has a variety of solutions, such as one-versus-rest (1-v-r SVMs) and one-versus-one (1-v-1 SVMs), Hierarchical Support Vector Machines (H-SVMs), Directed Acyclic Graph SVMs (DAG-SVMs), etc.

This paper uses the constructed SVM-SVC method for processing. In this method, if \( k \) is the number of class, there will be \( k(k-1)/2 \) SVMs. Each class is scored in it and the class with the highest score is the sample class. Figure 5 shows the classification mode of the SVM-SVC multi-classification algorithm.

![Figure 5. Diagram of 1-V-1 SVM-SVC model.](image)

2.2.3. Model implementation. In this study, the model uses SVM-SVC method in Sklearn module and is implemented by Python programming language. The entire program was written by calling the
scikit-learn machine learning library, supplemented by related file processing libraries (such as numpy, etc.) and drawing tools.

2.3. Model training
Data were randomly selected from 80% of the samples, and the remaining 20% were used as testing data. The input data is 9 dimensions, and the sample labels are in 5 classes. After classification, confusion matrix method was used to evaluate the accuracy.

3. Result analysis and discussion

3.1. Result analysis
The data of model training and verification were recorded and drawn into graphs. The confusion matrix of the evaluation results is shown in the Figure 6 below.

![Figure 6. Confusion matrix of training data (left) and testing data (right).](image)

Calculating the index of the recognition results of each class. The results of TP (True positives), TN (True negatives), FP (False positives), FN (False negatives) are shown in Table 2 and Table 3 below.

| metrics class | TP | TN | FP | FN | Precision | TPR/Recall | Accuracy |
|---------------|----|----|----|----|-----------|------------|----------|
| 0             | 38 | 122| 2  | 0  | 0.950     | 1.000      | 0.988    |
| 1             | 32 | 128| 0  | 2  | 1.000     | 0.941      | 0.988    |
| 2             | 64 | 94 | 0  | 4  | 1.000     | 0.941      | 0.975    |
| 3             | 17 | 141| 3  | 1  | 0.850     | 0.944      | 0.975    |
| 4             | 4  | 156| 2  | 0  | 0.667     | 1.000      | 0.988    |
| Total         |    |    |    |    |           |            | 0.951    |

| metrics class | TP | TN | FP | FN | Precision | TPR/Recall | Accuracy |
|---------------|----|----|----|----|-----------|------------|----------|
| 0             | 10 | 30 | 1  | 0  | 0.909     | 1.000      | 0.976    |
| 1             | 11 | 30 | 0  | 0  | 1.000     | 1.000      | 1.000    |
| 2             | 14 | 25 | 0  | 2  | 1.000     | 0.875      | 0.951    |
| 3             | 3  | 37 | 1  | 0  | 0.750     | 1.000      | 0.976    |
| 4             | 1  | 40 | 0  | 0  | 1.000     | 1.000      | 1.000    |
| Total         |    |    |    |    |           |            | 0.957    |
In the index statistics Table 2 of the training data, the accuracy of each class and total sample perform good (all greater than 0.95), and the recall rate are also higher (both greater than 0.94). Although the precision of class 3 and class 4 is relatively low, indicating not good enough, considering the small number of samples (only 17 samples of class 3 and 4 samples of class 4, too less in the whole 164 samples), it can be considered that the overall training is good.

In the index statistics Table 3 of the test data, the accuracy of each class and total sample perform good (all greater than 0.95). Although the Precision of class 3 and the Precision of class 2 are not good enough, but others are very good and their data too less, it can also consider the overall result is good.

3.2. Discussion
Through the analysis of the experimental results, the accuracy rate of water quality identification experiment is higher than 0.95, which can be considered as a success. However, the number of samples in this experiment is small, and different class is unbalanced, and even some class have little data, so the credibility of the experiment is worth further exploring.

The main purpose of this experiment is to explore a convenient way to evaluate water quality. Using this method, rapid water quality identification can be performed in various environments. After all, taking photos is a fast, cheap and adaptable way to get data. Further research is expected to provide sufficient support for this idea.

4. Conclusions
In this experiment, based on the information acquisition and feature extraction of the water samples, the SVM was used to identify the water quality. The results of proved that the accuracy was better than 95%. It proves that color moment can express water characteristics and can be applied to water quality identification. This water quality identification method based on color moment using SVM multi-classification method has certain reference value for water protection detection, can be performed in various environments.

Acknowledgements
In this paper, the research was sponsored by the Major R & D Plan of Science and Technology of Zhejiang (Project No. 2019C03099, No. 2020C05005) and the Civil product project MK2101S. Thanks to Zhejiang Jiake Information Technology Co., Ltd. and Zhejiang Jiake Electronics Co., Ltd. for their help in data acquisition.

References
[1] Cai L N, Liu P B and Zhi C G 2008 Discussion on Remote Sensing Based on Water Quality Monitoring Methods Geomatics & Spatial Information Technology 31(4) pp. 68-73
[2] Bierman P, Lewis M, Ostendorf B and Tanner J 2011 A review of methods for analysing spatial and temporal patterns in coastal water quality Ecological Indicators 11(1) pp. 103-114
[3] Schaeffer B A, Schaeffer K G, Keith D, et al. 2013 Barriers to adopting satellite remote sensing for water quality management[J]. International Journal of Remote Sensing 34(21) 7534-7544
[4] Ma H L, Xu J and Wang P C 2014 Water resource utilization and China’s urbanization Resour. Sci 36(2) pp. 0334-0341
[5] Zhang X 2014 Trend of and the governance system for water pollution in China China Soft Science 10 11-14
[6] Ahmed U, Mumtaz R, Anwar H, Shah A A, Irfan R and García-Nieto J 2019 Efficient water quality prediction using supervised machine learning Water 11(11) 2210
[7] Haghiabi A H, Nasrolahi A H and Parsaie A 2018 Water quality prediction using machine learning methods Water Quality Research Journal 53(1) 3-13
[8] Chou J S, Ho C C and Hoang H S 2018 Determining quality of water in reservoir using machine learning Ecological Informatics 44 57-75
[9] Lu H and Ma X 2020 Hybrid decision tree-based machine learning models for short-term water
quality prediction Chemosphere 249 126169
[10] Wentao Ma 2007 SVM Regression Method for the Surface Water Quality Assessment and Application Mathematics in Practice and Theory
[11] Ladjal M, Bouamar M, Djeriou M, et al. 2016 Performance evaluation of ANN and SVM multiclass models for intelligent water quality classification using Dempster-Shafer Theory[C]. 2016 International Conference on Electrical and Information Technologies (ICEIT) IEEE 191-196
[12] Bouamar M and Ladjal M 2007 Evaluation of the performances of ANN and SVM techniques used in water quality classification[C]. IEEE International Conference on Electronics IEEE
[13] Khotimah W N 2015 Aquaculture water quality prediction using smooth SVM IPTEK Journal of Proceedings Series I(1)
[14] Stricker M A and Orenge M 1995 Similarity of color images. In Storage and retrieval for image and video databases III International Society for Optics and Photonics Vol. 2420 pp. 381-392
[15] Lehmann M K, Nguyen U, Allan M, et al. 2018 Colour classification of 1486 lakes across a wide range of optical water types[J]. Remote Sensing 10(8) 1273
[16] Wyszecki G and Stiles W S 2000 Color Science: Concepts and Methods, Quantitative Data, and Formulae, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA
[17] Kirk J T O 1994 Light & Photosynthesis in Aquatic Ecosystems, 2nd ed.; Cambridge University Press: Cambridge, UK
[18] Jerome J H 1994 Colours of natural waters: 1. Factors controlling the dominant wavelength[J]. Northwest Sci 68 43-52
[19] Barysheva L F 1987 On the issue of intercorrespondence of colour scales used in limnology In Remote Monitoring of L arge L akes. (Leningrad:Nauka) pp. 60-65
[20] Schaeffer B A, Schaeffer K G, Keith D, et al. 2013 Barriers to adopting satellite remote sensing for water quality management[J]. International Journal of Remote Sensing 34(21) 7534-7544
[21] Mobley C D 1994 Light and Water: Radiative Transfer in Natural Waters; Academic Press: Cambridge, MA, USA [Google Scholar]
[22] IOCCG. Remote Sensing of Inherent Optical Properties: Fundamentals, Tests of Algorithms, and Applications. In Reports of the International Ocean-Colour Coordinating Group (IOCCG); IOCCG: Dartmouth, NS, Canada, 2006
[23] Pope R M and Fry E S 1997 Absorption spectrum (380–700 nm) of pure water. II. Integrating cavity measurements Appl. Opt. 36 8710-8723
[24] Bukata R P, Bruton J E and Jerome J H 1983 Use of Chromaticity in Remote Measurements of Water-Quality Remote Sens. Environ. 13 161-177
[25] Zhang B, Li J S, Wang Q and Shen Q 2012 Hyperspectral remote sensing of inland waters Beijing: The Science Publishing Company 130(131) pp. 130-131
[26] Piech K R and Walker J E 1971 Aerial color analyses of water quality[J]. Journal of the Surveying and Mapping Division 97(2) 185-197
[27] Gardner J R, Yang X, Topp S N, et al. 2021 The color of rivers[J]. Geophysical Research Letters 48(1) e2020GL088946
[28] Knerr S, Personnaz L and Dreyfus G 1990 Single-layer learning revisited: a stepwise procedure for building and training a neural network Neurocomputing 41-50 doi:10.1007/978-3-642-76153-9_5
[29] Stricker M A and Orenge M 1995 Similarity of color images[C]. Storage and retrieval for image and video databases III. International Society for Optics and Photonics 2420 381-392
[30] Bilge F, Yazici B, Dogeroglu T, et al. 2003 Statistical evaluation of remotely sensed data for water quality monitoring[J]. International Journal of Remote Sensing 24(24) 5317-5326
[31] Mushtaq F and Nee Lala M G 2017 Remote estimation of water quality parameters of Himalayan lake (Kashmir) using Landsat 8 OLI imagery[J]. Geocarto International 32(3) 274-285
[32] Yu X, Yi H, Liu X, et al. 2016 Remote-sensing estimation of dissolved inorganic nitrogen concentration in the Bohai Sea using band combinations derived from MODIS data[J]. *International Journal of Remote Sensing* **37**(2) 327-340

[33] Ministry of Ecology and Environment of the People's Republic of China, Environmental quality standards for surface water[S]. 2002, GB 3838-2002.

[34] Zhou Z 2016 Machine Learning[M]. Tsinghua University Press