

Received November 18, 2020, accepted December 12, 2020, date of publication December 25, 2020, date of current version January 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047387

Morphological Expression Model for Geographic Elements Based on the Shape-Tree Principle

LIHONG SHI¹, ZHONGXIANG WANG², HAO WU², BINBIN WANG^{1,3}, AND XINLIN QIAN¹

¹Government GIS Research Center, Chinese Academy of Surveying and Mapping, Beijing 100036, China

²National Geomatics Center of China, Beijing 100830, China

³Beijing Aerospace Hongtu Information Technology Company Ltd., Beijing 100195, China

Corresponding authors: Zhongxiang Wang (wangzhongxiang@ngcc.cn) and Hao Wu (wuhao@ngcc.cn)


This work was supported by the National Key Research and Development Project under Grant 2018YFB0505000 and Grant 2018YFC0807000.

ABSTRACT In recent years, with the continuous development of online geographic information services, the multiscale online expression of geographic elements and topological information has become a hot research topic. Currently, a large amount of research focuses mainly on the compression of vector line elements. The main shortcoming of current algorithms is that vector data compression does not consider the optimality of line simplification. In the face of terabytes or even petabytes of data, problems arise such as low processing efficiency and long transmission time. Based on the shape-tree principle, the proposed method aims to reconstruct the organizational structure of a geographic element and establish a hierarchical representation model of the element based on its structure, realizing multiscale shape recognition and matching and searching geographic elements. Finally, the proposed method is compared with a model based on the wavelet transform and discrete Fourier transform. The results demonstrate the applicability and effectiveness of the proposed method in morphological expression with geospatial big data.

INDEX TERMS Geographical element, hierarchical structure, morphological expression model, shape-tree principle.

I. INTRODUCTION

In recent years, with the continuous development of online geographic information services, the demand for real-time multiresolution visualization of geographic data has increased. The multiscale online expression of geometric and topological information of geographic elements has become a research hotspot. At present, a large amount of research focuses mainly on vector line element compression, regarding point and polygon elements as special linear patterns. By incorporating a compression algorithm and setting the limit difference, vector line elements can be simplified, linearized, and compressed; the same method can be used to simplify the point and area elements. The process of linear simplification realizes the multiscale representation expression of geographical elements. However, vector data compression does not consider the optimization of line simplification, and in the case of a data volume of terabytes or even petabytes, problems arise such as low processing efficiency and long transmission time.

The associate editor coordinating the review of this manuscript and approving it for publication was Stefania Bonafoni .

In morphology, shape is one of the most basic features of an object in the sense of visual perception. The tree structure is one of the most powerful structures in data organization [1]. The shape-tree method proposed by Felzenszwalb and Schwartz [2] can capture a hierarchical representation of the shape at multiple resolution levels by constructing a tree structure of object contours layer by layer, realizing the comprehensive description of multiscale graphics. Therefore, a new morphological expression mode based on the shape tree is proposed, which reconstructs the organization structure of a geographic element, establishes a hierarchical representation model of the element with a tree structure, and captures the layered representation of the shape at multiple resolution levels.

II. BACKGROUND

In the visualization of geographic information, data compression technology has received great attention as a supporting technology for enabling massive information storage and transmission. However, previous work on data compression has not considered the optimization of line simplification. The present work aims to enable this optimization. In a vector

data set, a line is composed of two endpoints and a series of points marking the shape of the line between them, and a surface is composed of connected, closed, and disjoint line segments. Therefore, in the element model, taking the vertex as the unit, polygon objects are defined from collections of line objects, and line objects are defined from their constituent vertex objects. The vertex as the central unit is an extremely important feature in the element model. This feature serves as an important concept to realize data compression and to understand the organizational structure of the geographic element in real life, through taking the vertex as the basic granularity, adopting a simplified algorithm to reduce the data volume of collection objects, revealing and dividing the vertex hierarchy and analyzing the vertices.

Currently, the commonly used algorithms for data simplification of vector line elements include the Douglas-Peucker algorithm [3], circle simplification [4], polygon fitting algorithm [5], McMaster-Jenks algorithm [6], [7], Delaunay triangulation algorithm [8], [9], GSC compression algorithm and de-GSC decomposition algorithm [10]–[12]. Wang [13] focused on the SPIHT compression algorithm and real-time visualization of large-scale scenes. Yu [14] proposed a GSC compression algorithm, the de-GSC decomposition algorithm, to realize the lossless compression processing of GML documents. Hu [15] studied vector geographic data compression according to the theory of integer transformation in terms of multilevel differential compression of vector data. On the basis of the STR-M tree, an innovative multiresolution organization model of vector geographic data was proposed, which was based mainly on spatially characterizing vector data, dividing the data accordingly, and reorganizing the resolution based on the partition unit. Chen *et al.* [16] proposed a compression method combining the Douglas-Peucker algorithm and intersection features, which can simplify road vector data and save intersection points and important feature points. Although these algorithms have their own advantages, they do not consider the optimization of line simplification. There are some shortcomings, such as low processing efficiency and long transmission time.

In terms of the morphological description of geographic elements, current research focuses mainly on the contours, sizes and directions of the element, as well as the element's attribute information, topological relations, and semantic relations [17]. Among these topics, research on spatial relationships based on graphical contour points is an active research area. The relevant methods mainly include the shape context method [18], [19] Gaussian mixture mode [20], and shape-manifold-based method [21]. Xin *et al.* [22] concluded that the shape is very important to the cognition of objects and is the key attribute to determine the object category. Ai and Chen [23] reported that multiscale representation of spatial data presents challenges with large data volumes, slow responses, conflicts between data representations and steep changes in scale ranges. Ang *et al.* [24] added the principle of manifolds to the shape description method, using the approximate measures calculated by the shape manifolds

to measure the degree of similarity and difference in shape; this method improves the element matching aspect substantially. Chui and Rangarajan [25] improved the use of thin-plate splines and proposed a new shape description matching method combining fuzzy-assign and deterministic annealing with these splines. Peter and Rangarajan [26] proposed a Gaussian mixture model method and a Riemannian geometry method to measure the distance. Zhang and Cai [27] proposed an inner distance measurement method based on the shape context method and obtained a new shape context descriptor through experiments that showed good matching. Wu [19] aimed to solve the problems that the shape context description method cannot achieve fast image retrieval and that the algorithm is time consuming. A new shape context simplification descriptor was proposed, which was used as the input of a neural network algorithm to realize a fast retrieval function. These methods focus mainly on the improvement of operators for distance measurement of spatial relations. These operators include the Euclidean distance [28], Riemannian geometric operators [29], shape manifold operators [30], and so on.

Van Oosterom proposed BLG-tree for line simplification. BLG-tree, which is based on the R-tree, was the first fully dynamic and reactive data structure. It showed the advantage of the node selection based on importance level and geometric position, which can be used for displaying the geographic object at interactive speed. The importance level was designed to facilitate insertion and deletion of point elements in line simplification [31]. Felzenszwalb and Schwartz [2] proposed a shape-tree description method on the basis of fully considering the shape features and the geometric position of the feature points, regarding the edge contour segment of the graph as the basic unit for analysis and processing, subdividing the boundary contour of the target graph continuously, and constructing the tree structure of the contour layer by layer. This method can be used for accurate shape matching. A shape-layered description can be used for layered representation of two-dimensional objects at multiple resolution levels. Hierarchical representation explicitly captures global shape information, with more accurate recognition results than other approaches. Hierarchical matching of shapes has been used to compare objects and to detect objects in cluttered images [2]. Jussi used a shape tree to visualize multivariate density estimation, which can determine the area of Euclidean space in the probability density set [31]. The application of shape trees in visualizing the distribution of stock index returns was also proposed. Experiments showed that the method has low computational complexity and good practicality, even in the case of shape stretching and contour deformation. The shape tree provides a basis for the multilevel expression of the geographic element in research, as well as the realization of fast compression and efficient data transmission and online multiresolution visualization of geospatial big data.

In addition, the demand for online real-time multiresolution visualization of geographic data continues to increase. The results from querying geographical features in spatial

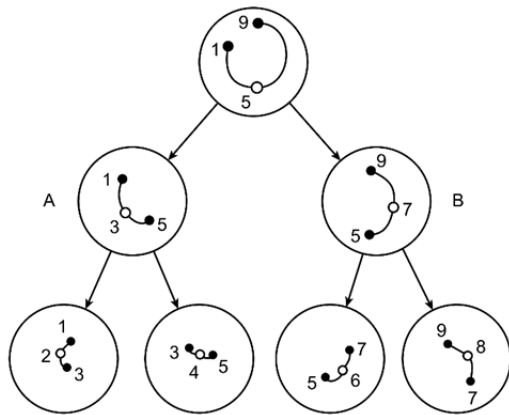


FIGURE 1. Shape-tree diagram.

databases are accurate and unique; however, the geographic element is allowed to only approximately meet accuracy requirements for different applications, and the present study of geographical elements has not solved the related problems well. The shape-tree algorithm constructs the tree structure of the contour layer by layer, which can capture the layered representation of the shape at multiple resolution levels. Therefore, it is very meaningful to study the method of geometric shape analysis of geographical elements based on shape trees in this paper.

III. SHAPE-TREE-BASED EXPRESSION MODEL

Based on the shape-tree principle, the morphological expression model of geographical elements is proposed. The shape-tree algorithm improves point sequence construction of a geographic element from a linear structure to a tree structure. Therefore, a vector line element or a polygon object can be expressed in a tree structure. This can facilitate global information and local information description for vectoring the element and has the characteristics of simplicity and efficiency. Geographic features expressed by shape trees can conveniently perform vertex sampling and capture global shape information of geographical elements hierarchically (Fig. 1).

As shown in Fig. 1, curve $L_{-1,5,9}$ are composed of curves $L_{-1,5}$, and $L_{-5,9}$. When taking the first layer of nodes as an example, node A stores the distance between the starting point L_{-1} and midpoint L_{-3} and the distance between the end point L_{-5} and midpoint L_{-3} . Node B stores the distance between the starting point L_{-5} and midpoint L_{-7} and the distance between the end point L_{-9} and midpoint L_{-7} . Thus, the subcurve information or local information can be described by nodes A and B, and the global information of the curve can be composed of the information of the two nodes.

In this paper, the shape-tree model is improved mainly based on the Douglas-Peucker algorithm for the selection of the vertex of the vector line element. First, the shape-tree structure is constructed, which needs to recursively subdivide the sequence of vertices into subsequences until only one vertex is included in each subsequence. Such a process of

recursive subdivision directly corresponds to a specific tree structure, where each tree node corresponds to a subsequence of vertices. In addition, the maximum distance generated by the Douglas-Peucker algorithm calculation in each recursive process is stored in the tree structure corresponding to the tree node. The reconstruction of the shape-tree description algorithm mainly uses weighted breadth traversal to reconstruct similar geographical elements by inverse calculation of geographical elements represented by a tree structure.

A. SHAPE-TREE HIERARCHICAL EXPRESSION

The principle of a shape tree is mainly to describe the hierarchical expression of a curve. This curve, $L = (P_0, P_1, \dots, P_n)$, consists of a number of vertices, P_i ($0 < i < n$). For any vertex, P_i ($0 < i < n$), the vertex is gradually sampled according to a certain criterion, and a sampled vertex sequence is formed. Then, a hierarchical expression is established according to the source sequence of the vertex and the sampled sequence of the vertex. The expression can represent the vertex hierarchy existing in the curve and can be directly used for curve reconstruction. This section takes a line element as an example and describes the process of generating a tree structure expression of the vertex sequence.

1) SHAPE-TREE ALGORITHM

$L = (P_0, P_1, \dots, P_n)$ can be specified as a curve with a series of sample points. Parameter k is the number of subdivisions for each level in the tree structure. This section takes k in the recursive subdivision as 2, and L is divided to form two subsequences each time. Select the midpoint of the vertex sequence as the partition point, set $i = \lfloor n/2 \rfloor$, and make P_i the midpoint of the vertex sequence. Then, $L = (P_0, P_1, \dots, P_n)$ is simplified as $L' = (P_0, P_n)$, which contains only two vertices. P_i is added to the simplified vertex set, and $L' = (P_0, P_i, P_n)$ is made as close as possible to L . $L'(P_i|P_0, P_n)$ represents the position of P_i in P_0, P_n . The positions of the first and last sample points can be used to define the coordinate system at which we measure the midpoint position. The first and last sample points define the gauge and direction, so the relative position $L'(P_i|P_0, P_n)$ is invariant. Then, according to the selection of P_i , the original curve is divided into two halves, and the two subsequences $L_{0,1} = (P_0, P_1, \dots, P_i)$ and $L_{1,1} = (P_i, P_{i+1}, \dots, P_n)$ are formed. $L_{0,1}$ and $L_{1,1}$ are recorded as the simplest form separately as $L'_{0,1} = (P_0, P_i)$ and $L'_{1,1} = (P_i, P_n)$. Again, the midpoints of the vertex sequences $L_{0,1}$ and $L_{1,1}$ are selected as the dividing points and are then added to the simplified vertex sets $L'_{0,1}$ and $L'_{1,1}$. This iterative process is continued until all the sequences are subdivided into only two vertices that cannot be further subdivided.

In the above process, the inclusion relationship formed by all vertex sequence subdivisions can be directly expressed by the shape-tree algorithm, so when the entire subdivision process is finished, a shape tree is also established, and the

tree nodes correspond to $L(i, h)$, where i indicates the serial number of the tree node in the same layer and h indicates the distance between the tree node and the root node, which is also called the height of the node. In this binary tree, a leaf node represents a sequence of two adjacent vertices.

B. VERTEX PARTITIONING STRATEGY BASED ON THE DOUGLAS-PEUCKER ALGORITHM

In this section, we analyze a closed geographic line element and approximate the vertex sequence. Set L as a simple sequence of vertices (P_0, P_1, \dots, P_n) ; the lines formed by the vertex sequence have no self-intersection. Set S as a line segment. If the vertex sequence expresses a closed line, then $P_0 = P_n$ for any $i \neq j, P_i \neq P_j$ and $S_{P_i, P_{i+1}} \cap S_{P_j, P_{j+1}} \in \{P_f\}, f = 0, \dots, n$, which means that all the vertices of the vector geographic line element are passed only once and that none of the adjacent vertex segments have an internal intersection. According to the Douglas-Peucker algorithm, the process of vertex division is as follows:

First, the vertex is reduced to the simplest form; that is, $L = (P_0, P_1, \dots, P_n)$ is simplified as $L' = (P_0, P_n)$, which contains only two vertices. Then, according to the Douglas-Peucker algorithm, the distance between any vertex $P_f (0 < f < n \text{ in } L_0)$ and the line S_{P_0, P_n} is $d(P_f, S_{P_0, P_n})$, and P_m is the vertex with the longest distance, satisfying $d(P_m, S_{P_0, P_n}) \geq d(P_f, S_{P_0, P_n}), 0 < m, f < n$.

Then, two subcurves of L are divided by the dividing point P_m . Setting the dividing point as P_i , the two subsequences are recorded as $L_{0,1} = (P_0, P_1, \dots, P_i)$ and $L_{1,1} = (P_i, P_{i+1}, \dots, P_n)$ and are reduced to the simplest form, that is, $L'_{0,1} = (P_0, P_i)$ and $L'_{1,1} = (P_i, P_n)$. Repeating the above calculation process, the maximum distances $d_{(0,1)}$ and $d_{(1,1)}$ are calculated. In detail, the vertex with the maximum distances is then associated with $L'_{0,1}$, and the maximum distances $d_{(0,1)}$ can be calculated. The shape description coefficient is associated with the vertex that has the longest corresponding distance in $L'_{0,1}$. In the same way, the vertex with the longest distance is associated with $L'_{1,1}$, and the maximum distance $d_{(1,1)}$ is calculated. The shape description coefficient is associated with the vertex with the longest corresponding distance in $L'_{1,1}$. Continue this iterative process until all sequences are subdivided to include only two vertices that cannot be further subdivided.

1) SHAPE COEFFICIENT CALCULATION BY THE HAUSDORFF DISTANCE

This section mainly introduces the Hausdorff distance calculation, which is used to calculate the shape coefficient.

The Hausdorff distance is defined as:

$$d_H(A, B) = \max_{a \in A} \{ \min_{b \in B} Eu(a, b) \} \quad (1)$$

where A and B represent two sets, a represents the original data vertex $a(x, y)$, b represents the reconstructed vertex $b(x, y)$, and $Eu(a, b)$ represents the Euclidean distance between the two vertices. Douglas-Peucker algorithm

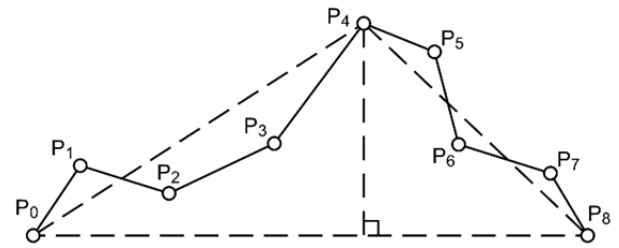


FIGURE 2. Distance calculation diagram.

sampling is the operation of selecting the vertices from the line element. The length of d_H is related to the point feature, the number of points, and the inverse reconstruction of the line element. Therefore, the Hausdorff distance definition is used in the vertex division process as follows:

$$d_H(L, L') = \max \{ d(P_f, S_{P_0, P_n}) | P_f \in P', 0 < f < n \} \quad (2)$$

The key characteristic of Douglas-Peucker sampling is to retain important vertices with large d_H values, delete vertices with small d_H values, and restore the geometric shape of the original line element as much as possible. To better reconstruct the line elements and estimate the error of the sampling vertices, it is necessary to calculate the values of each vertex d_H to form a vertex value table. The basic idea is to start with one vertex, and the process of calculating the maximum distance value at the vertex is shown as follows:

Enter: P_i, P_{i+1}, \dots, P_j

Step 1. If $i + 1 \geq j$, return zero;

Step 2. For the vertices P_{i+1}, \dots, P_{j-1} between P_i and P_j in the line, calculate the distance S_{P_i, P_j} , the maximum value of which is associated with L' ;

Step 3. Take the vertex P_m between P_i and P_j , which corresponds to the maximum distance value, divide the vertex sequence into two vertex subsequences P_i, P_{i+1}, \dots, P_m and $P_m, P_{(i+j)/2+1}, \dots, P_j$, and then recursively perform steps 1-3, comparing two distance values in $L'_{i,h}$, and take the longer one to associate with L' ;

Output: The value d_H in L' .

This process is shown in Fig. 2.

As shown in Fig. 2, the line element L contains a sequence of nine vertices. Starting from P_0 , the distances between the middle points and the line are calculated individually, and the maximum value is stored in the attribute table of the reconstructed line feature L' . The vertex P_4 , which corresponds to the maximum value, divides the vertex sequence of the original line element L into lines S_{P_0, P_4} and S_{P_4, P_8} . According to the above steps, the distances from each point in the middle of the corresponding line to lines S_{P_0, P_4} and S_{P_4, P_8} are calculated. The two maximum values calculated from lines S_{P_0, P_4} and S_{P_4, P_8} are stored in the attribute table of the reconstructed line feature L' . Continue to segment the vertex sequence until the Hausdorff distance of all intermediate nodes of the line L is obtained.

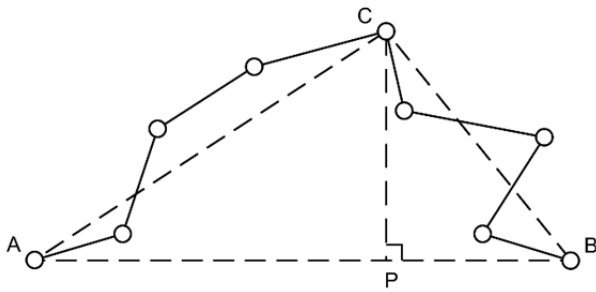


FIGURE 3. Relative position diagram.

2) CORRELATION VALUE CALCULATION OF THE RELATIVE POSITION AT THE DIVISION POINT

The distance $d(P_f, S_{P_0, P_n})$ between vertex $P_f, 0 < f < n$ in L and line segment S_{P_0, P_n} is obtained by the Douglas-Peucker algorithm. P_m represents the vertex with the longest distance, satisfying $d(P_m, S_{P_0, P_n}) \geq d(P_f, S_{P_0, P_n}), 0 < m, f < n$. Then, the relative position values between P_m and S_{P_0, P_n} are calculated. These values are used in the process of reconstructing the geographic element. The main steps are as follows:

The relative position correlation values P_m mainly include:

The perpendicular ratio of the virtual triangle, which is recorded as the direction vector $pRatio$.

The ratio of the projected vector to the segment-defined vector, which is recorded as the $hRatio$ value.

The Sgn function, which returns an integer variable indicating whether $pRatio$ is positive or negative.

In $Sgn(number)$, $number$ represents a numeric expression. If the return value is greater than 0, then Sgn returns 1, which indicates that the point P_m lies to the upper left of the vector S_{P_0, P_n} . If the number is equal to 0, then Sgn returns 0, which indicates that point P_m lies on vector S_{P_0, P_n} . If the number is less than 0, then Sgn returns -1, which indicates that the point P_m lies to the upper right of the vector S_{P_0, P_n} .

Fig. 3 schematizes the calculation of the associated value of the relative position at the division point.

According to the vertex sequence and the first selected partition points, a virtual triangle ΔABC is constructed. The coordinates of $A, B,$ and C are respectively represented as $A(x_a, y_a), B(x_b, y_b),$ and $C(x_c, y_c)$.

The virtual triangle vertical ratio $pRatio$ is defined as pr : the ratio of the maximum distance d_H and the line \overrightarrow{AB} . The result is as follows:

$$pr = \frac{d_H}{|\overrightarrow{AB}|} * Sgn \tag{3}$$

The ratio ($hRatio$) of the projected vector to the segment-defined vector is defined as hr : the projection vector \overrightarrow{AP} is constructed such that the vector \overrightarrow{AC} projects onto the vector, and the ratio of the vector \overrightarrow{AP} to the vector \overrightarrow{AB} is taken as hr .

$$\overrightarrow{AP} = \frac{\overrightarrow{AC} \cdot \overrightarrow{AB}}{|\overrightarrow{AB}|} \tag{4}$$

Therefore, hr can be defined as follows:

$$hr = \frac{\overrightarrow{AP}}{|\overrightarrow{AB}|} \tag{5}$$

$Sgn(number)$ function: the parameter $number$ is a numeric expression, which is used mainly to determine whether the point is on the left or right side or on the line by the cross product of the vector. Set two vectors \overrightarrow{CA} and \overrightarrow{CB} and then $\overrightarrow{CA}(x_c - x_a)$ and $\overrightarrow{CB}(x_c - x_b)$; the formula is as follows:

$$Sgn = \overrightarrow{CA} \times \overrightarrow{CB} \tag{6}$$

If the result $Sgn = \overrightarrow{CA} \times \overrightarrow{CB}$ is less than 0, the point C lies to the upper left of line \overrightarrow{AB} . If the result of $Sgn = \overrightarrow{CA} \times \overrightarrow{CB}$ is equal to 0, then the point C lies on line \overrightarrow{AB} . If the result of $Sgn = \overrightarrow{CA} \times \overrightarrow{CB}$ is greater than 0, then the point C lies to the right of line \overrightarrow{AB} .

Calculation process of relative position correlation values P_m :

The line element is divided into two subcurves by the dividing point P_m . Then, the dividing code is set as P_i , and the two subsequences formed are recorded as $L_{0,1} = (P_0, P_1, \dots, P_i)$ and $L_{1,1} = (P_i, P_{i+1}, \dots, P_n)$, which are reduced to the simplest form, that is, $L'_{0,1} = (P_0, P_i)$ and $L'_{1,1} = (P_i, P_n)$. Then, $pRatio, hRatio,$ and Sgn are calculated according to the above steps. These values are associated with the vertex with the longest distance in $L'_{0,1}$. Continue this iterative process until all sequences are subdivided to include only two vertices and cannot be further subdivided. The calculation process of the association value of the vertex relative position is as follows:

Enter: P_i, P_{i+1}, \dots, P_j

Step 1. If $i + 1 \geq j$, return zero;

Step 2. For vertex P_{i+1}, \dots, P_{j-1} , the distance to S_{P_i, P_j} is calculated, where vertex P_m corresponds to the maximum distance value.

Step 3. Based on the vertices $P_i, P_m,$ and P_j , the terms pr, hr and Sgn are calculated, and these values are associated with L' .

Step 4. Set P_m as the vertex corresponding to the maximum distance value. The vertex sequence is then divided into two subsequences P_i, P_{i+1}, \dots, P_m and $P_m, P_{(i+j)/2+1}, \dots, P_j$. Perform steps 1-4 recursively.

C. CONSTRUCTION OF THE SHAPE-TREE ANALYSIS MODEL FOR GEOGRAPHICAL ELEMENTS

1) INTRODUCTION OF THE BUILDING MODEL ALGORITHM

The vector element is generally stored by a vector data structure: a point P is represented as (x, y) , a line L is represented by a plurality of vertices $P_1, P_2 \dots P_n$ and the corresponding coordinate pairs, and polygon elements and combined objects are represented by a plurality of closed vertex sequences.

When the Douglas-Peucker algorithm is used to divide the line element, the threshold value ϵ is set to a negative number, so that the algorithm is terminated after all the vertices in the line element L_0 are selected in turn. The algorithm processing

without the qualification is called the complete Douglas-Peucker algorithm. The corresponding selection process can be regarded as the process of gradually dividing the vertices into the sampling vertex sequence from the original line element. Moreover, according to the structure based on the shape-tree principle, a line-to-tree morphological analysis model is constructed. The vertices that are obtained by the recursive process establish a shape tree with the selected order in the Douglas-Peucker algorithm. Then, $d_{\max(i,h)}$, pr , hr , and Sgn of the vertices are associated on each tree node, that is, the correlation value of the vertex corresponding to the process of the Douglas-Peucker algorithm.

In this section, based on the Douglas-Peucker algorithm and the shape-tree morphological analysis principle, the shape-tree morphological analysis model is constructed as follows:

Input: the vertex sequence $\{P_i, P_{i+1}, \dots, P_j\}$ of the line element L_0 .

Step 1. If $i + 1 < j$, go to step 2; otherwise, go to step 7.

Step 2. Calculate the distance from P_f ($i < f < j$) to the line S_{P_i, P_j} , and record the longest distance as $d_{\max(i,h)}$; the vertex corresponding to the longest distance is recorded as P_m .

Step 3. Based on the vertices P_i , P_m , and P_j , the terms pr , hr and Sgn are calculated.

Step 4. The vertex P_m is used to divide the vertex sequence $\{P_i, P_{i+1}, \dots, P_j\}$ into the two subsequences $\{P_i, P_{i+1}, \dots, P_m\}$ and $\{P_m, P_{m+1}, \dots, P_j\}$.

Step 5. Create a tree node $T_{i,j}$, and store d_H , pr , and hr in the tree node.

Step 6. Perform Step 1 to 5 on the vertex sequences $\{P_i, P_{i+1}, \dots, P_m\}$ and $\{P_m, P_{m+1}, \dots, P_j\}$ recursively, and record the generated tree nodes $T_{i,m}$ and $T_{m,j}$ as the left and right child nodes of node $T_{i,j}$.

Step 7. Return $T_{i,j}$.

Output: The root node $T_{i,j}$ of the generated tree structure $T = \varpi(L_0)$ of the line element L_0 .

2) MODEL BUILDING PROCESS

The shape-tree morphological expression model of a geographic line element describes the element as a shape-tree structure. Its operation and processing runtime complexity is of linear logarithmic order $O(n \log 2n)$. This complexity is very favorable, as it greatly reduces the time complexity of operation; thus, it is simple and efficient. The model construction process is shown in Fig. 4.

Taking the line element as an example, the shape tree is constructed for the vertex sequence $\{P_0, P_1, \dots, P_8\}$.

Step 1. Find the vertex P_5 as the one with the longest distance to the line element S_{P_0, P_8} ; the distance value is d_H ;

Step 2. Select vertex P_5 as the dividing point;

Step 3. Generate the root node of the tree with P_5 , and associate the distance values d , pr , and hr .

Perform the above operations on the vertex sequences $\{P_0, P_1, \dots, P_5\}$ and $\{P_5, P_6, \dots, P_8\}$, and take the generated node as the left and right child nodes of the root node.

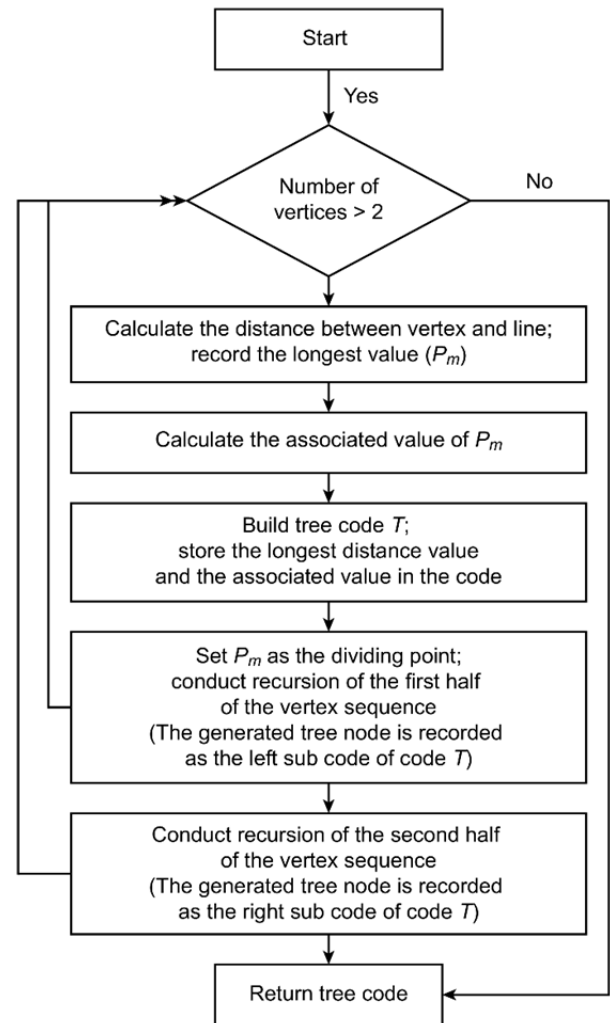


FIGURE 4. Construction of the shape-tree analysis model for the geographical element flow chart.

The result of shape-tree generation is shown in Fig. 5.

IV. EXPERIMENT AND DISCUSSION

A. EXPERIMENTAL DATASET

The experiment mainly uses the Globally Consistent Hierarchical High-resolution Shoreline Database (GSHHS), OpenStreetMap (OSM), Natural Earth-Vector, National Geographic Center of China (NGCC), GADM database and CENSUS data source (Table 1).

GSHHS contains high-resolution shoreline data. These data are obtained mainly through the University of Hawaii and the US Atmospheric and Oceanic Administration, with an independent source of continent and ocean data sets.

OSM is an online map collaboration plan. Use of the map image and vector data on the website is authorized by the Open Database License (ODbL). The main source of OSM data is the OSM Foundation. These data constitute global shoreline data from independent sources and are of high quality.

TABLE 1. The number of vertices of the experimental data.

	Number of vertices (data source)		
Mainland China (CN)	736313 (GADM)	14193 (NE)	9170 (NBGIC)
Taiwan Island, China (TW)	18327 (OSM)	406 (NE)	7337 (GSHHG)
Continental United States (CONUS)	2061606 (GADM)	36054 (NE)	30153 (CENSUS)
Island of Hawaii (HI)	25487 (OSM)	121 (NE)	3551 (GSHHG)
United Kingdom (UK)	644737 (OSM)	3707 (NE)	49153 (GSHHG)
Japan Hokkaido (JH)	105839 (OSM)	626 (NE)	18066 (GSHHG)
Japan Honshu Island (JHI)	508723 (OSM)	2479 (NE)	51035 (GSHHG)
Japan Kyushu (JK)	188113 (OSM)	976 (NE)	21814 (GSHHG)
Japan's Four Islands (JP)	132460 (OSM)	473 (NE)	11424 (GSHHG)
Eurasia (EA)	4317907 (OSM)	83917 (NE)	1160926 (GSHHG)
Australia (AU)	332408 (OSM)	9464 (NE)	213065 (GSHHG)
Africa (AF)	679067 (OSM)	13409 (NE)	257365 (GSHHG)
New Zealand (NZ)	157002 (OSM)	1922 (NE)	37219 (GSHHG)

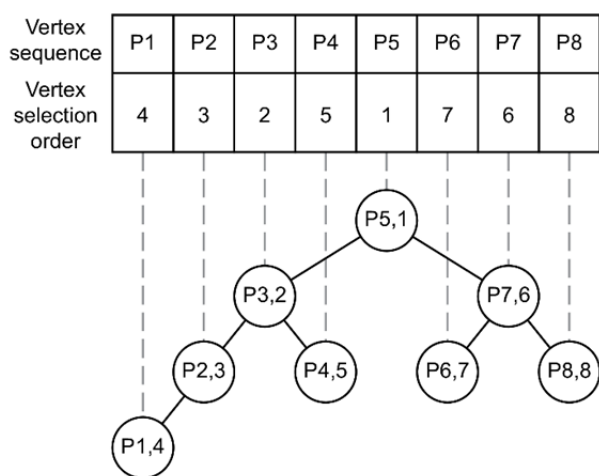


FIGURE 5. Shape-tree construction process.

The source of the Natural-Earth-Vector dataset is the North American Cartographic Information Association; this dataset comprises independent data from administrative divisions worldwide.

NGCC is an independent source within a Chinese administrative division.

The GADM database is a global administrative division database. It includes administrative division data for almost all countries and regions around the world.

The CENSUS database is an independent source associated with the US administrative border.

Table 1 summarizes the experimental data, constituting the data of administrative region lineation elements in China, the United States, Japan and other countries and regions.

These data are mainly from OSM, NE, GSHHS, GDAM, NBGIC and CENSUS.

B. EXPERIMENTAL METHOD AND ERROR CALCULATION METHOD

In this paper, the shape-tree geographic morphological expression model, Fourier algorithm model and wavelet algorithm model are used to describe and analyze the experimental data and reconstruct and visualize the results. Then, a comparison is conducted based on the error calculation of the three models. This approach mainly calculates the error between the reconstructed data and the original experimental data. The smaller the error is, the higher the similarity between the reconstructed data and the original experimental data, and the higher the feasibility of the proposed method.

The Hausdorff distance algorithm is used for the error calculation. The distance is calculated between the vertices of the original data and those of the reconstructed data. That is, each vertex of the original data is used to calculate the distance with all vertices of the reconstructed data. The minimum value is first selected, and then the maximum value in all minimum values is selected as the error value. The error values of the three methods are compared for further analysis of the models.

In this paper, the generalized Hausdorff distance is defined as formula (1).

C. EXPERIMENTAL RESULTS

According to the error calculation method introduced in the previous section, the selected dataset is tested by three models, and error analysis is conducted on the experimental

TABLE 2. Errors in the experimental results - OSM Data (Degree).

	Shape-tree model (S-T)	Wavelet transform (WT)	Discrete Fourier transform (FT)
Mainland China (CN) (GADM)	0.14517	0.52314	1.20730
Taiwan Island, China (TW)	0.35548	0.68051	1.63394
Continental United States (CONUS) (GADM)	0.08235	0.15027	0.42361
Island of Hawaii (HI)	0.95931	1.54225	2.03006
United Kingdom (UK)	0.80124	1.46703	1.79245
Japan's Four Islands (JP)	0.54323	0.91734	1.41238
Eurasia (EA)	0.61133	1.30025	1.62497
Australia (AU)	1.13051	2.05823	2.38234
Africa (AF)	0.44130	0.75349	1.56623
New Zealand (NZ)	0.79283	1.20305	1.98770

TABLE 3. Errors in the experimental results- NE Data (Degree).

	Shape-tree model (S-T)	Wavelet transform (WT)	Discrete Fourier transform (FT)
Mainland China (CN)	2.31169	3.41917	4.07865
Taiwan Island, China (TW)	1.86977	2.54028	3.26352
Continental United States (CONUS)	3.14521	4.76543	5.39714
Island of Hawaii (HI)	2.49070	3.26751	3.75716
United Kingdom (UK)	3.07625	4.30120	5.37152
Japan's Four Islands (JP)	1.06532	2.14397	2.908115
Eurasia (EA)	1.53986	2.01639	2.97614
Australia (AU)	2.37410	3.42185	3.84221
Africa (AF)	1.68434	2.34761	2.97462
New Zealand (NZ)	2.08743	3.21203	4.09873

TABLE 4. Errors in the experimental results-GSHHS Data (Degree).

	Shape-tree model (S-T)	Wavelet transform (WT)	Discrete Fourier transform (FT)
Mainland China (CN) (NGCC)	1.91573	2.41767	3.02341
Taiwan Island, China (TW)	1.03290	1.79006	2.60334
Continental United States (CONUS) (CENSUS)	1.53219	2.11374	2.85436
Island of Hawaii (HI)	1.95217	2.38325	2.72468
United Kingdom (UK)	2.07632	3.14060	3.71425
Japan's Four Islands (JP)	0.76532	1.10094	1.84510
Eurasia (EA)	0.98326	1.37408	1.52349
Australia (AU)	1.73210	2.39012	2.92453
Africa (AF)	0.79234	1.26208	1.79532
New Zealand (NZ)	1.01354	1.70329	2.38682

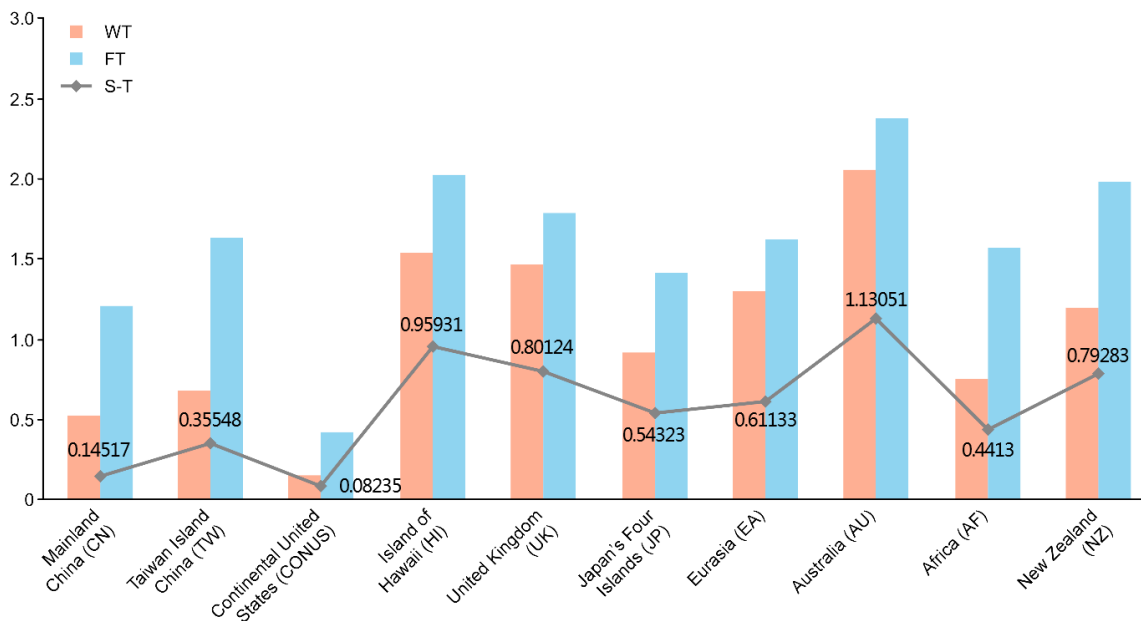


FIGURE 6. Error statistics diagram - OSM data.

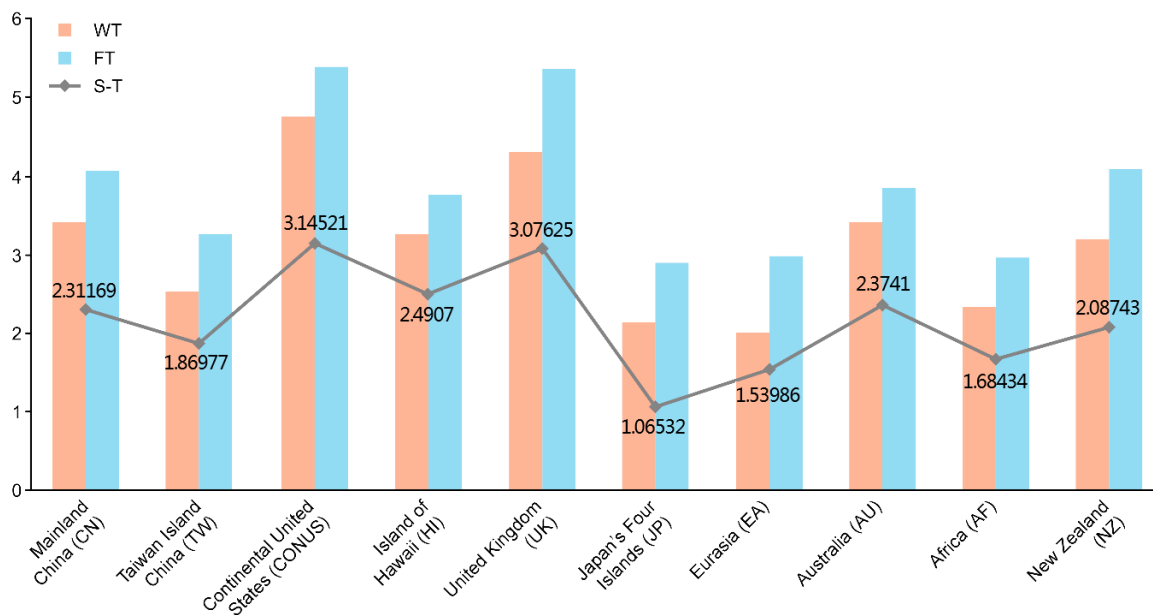


FIGURE 7. Error statistics diagram - NE data.

results. First, the three models are tested separately with GSHHS, OSM and Natural-Earth-Vector data. Each study area is modeled and reconstructed using three models, and then the error is calculated and unified.

Select the OSM data to carry out three model comparison experiments and obtain the error of the experimental results (Table 2).

The errors generated by the three models are visually displayed using the line and histograms in Fig. 6 for comparison.

The data source used for the Chinese mainland and the continental United States is GADM. According to the experimental error shown in Table 2 and the statistical analysis of the experimental error in Fig. 7, compared with the discrete Fourier transform model and the wavelet transform model, the shape-tree geographic morphological model has a smaller error, and the error calculated by the wavelet transform model is smaller than that calculated by the discrete Fourier transform model. The approximated image reconstructed by the

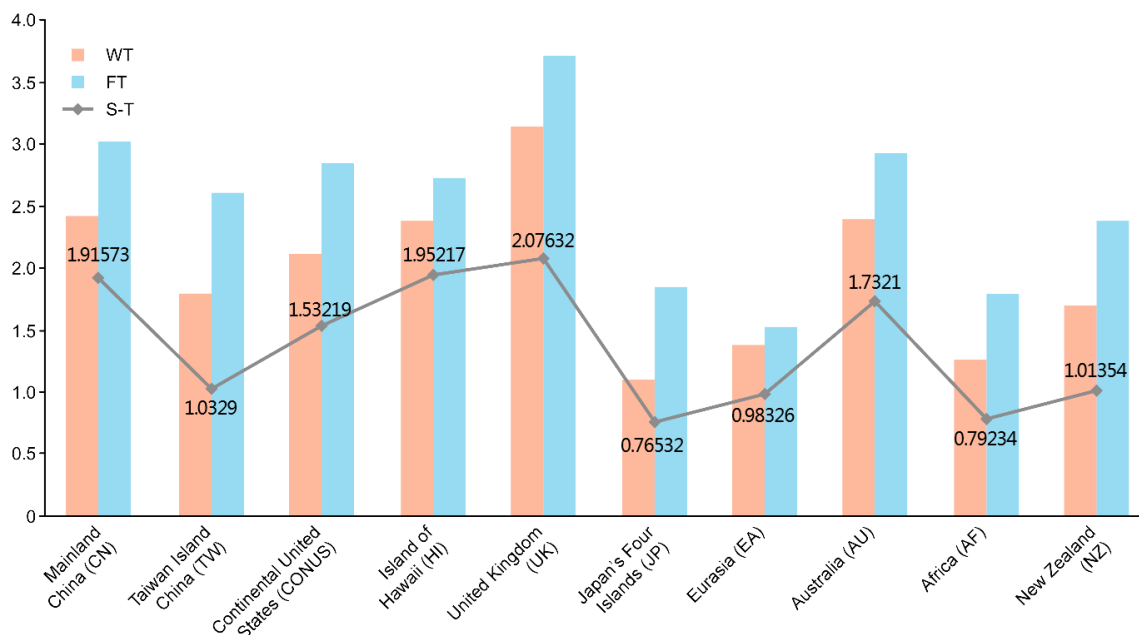


FIGURE 8. Error statistics diagram - GSHHS data.

proposed model has a higher similarity and smaller error than the original graphic and can achieve better matching of geographical element morphology. The model can be further studied and applied to multiresolution dynamic visualization.

The NE data are selected to carry out three model comparison experiments, and the errors of the experimental results are obtained (Table 3).

The errors generated by the three models are visually displayed using the line and histograms in Fig. 7 for comparison.

As shown in Table 3, compared with the discrete Fourier transform model and the wavelet transform model, the shape-tree geographic morphological model has a smaller error, and the error calculated by the wavelet transform model is smaller than that calculated by the discrete Fourier transform model. The shape-tree model is the best method to achieve the matching effect between the reconstructed data and the original data in the three methods. In addition, the error calculated by the NE data is larger than that of the OSM data.

By selecting the GSHHS data to carry out three model comparison experiments, the errors in the experimental results are obtained (Table 4).

The errors generated by the three models are visually displayed using the line and histograms in Fig. 8 for comparison.

As shown in Table 4, compared with the discrete Fourier transform model and the wavelet transform model, the shape-tree geographic morphological model has a smaller error. In addition, the error value calculated using GSHHS is larger than that using OSM but smaller than that using NE data. The experimental results also demonstrate that the greater the number of vertices of the vector data, the smaller is the experimental error.

V. CONCLUSION

In this paper, the shape-tree principle in morphology and the Douglas-Peucker algorithm are combined to construct a morphological expression model for a geographic element, which provides solutions for analyzing the element in the field of geographic information systems. Through the steps of hierarchical representation, vertex division, recursive subdivision, application processing, and correlation value calculation of the dataset, a multilevel expression corresponding to the original data and characterized by a tree hierarchy is established, which provides a basic data structure for the vector of the approximate query of the geographic element with a controllable data volume. The experimental results show that the shape-tree model can well describe and analyze geographic elements and can also achieve better matching results than other models. In summary, the shape-tree model describes the characteristics of the geographic element simply and efficiently. The hierarchical representation of the element explicitly captures global shape information and richer geometric information and obtains more accurate recognition results than other representations.

In the future, we plan to further analyze the application of the shape-tree model to the geographic element of polygon shape. We also plan to study the updating algorithm based on the local vertex sequence and its hierarchical subtree reconstruction, including the operations of vertex sequence insertion, deletion and modification.

REFERENCES

- [1] H. Zhang, S. Wang, X. Xu, T. W. S. Chow, and Q. M. J. Wu, "Tree2Vector: Learning a vectorial representation for tree-structured data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 11, pp. 5304–5318, Nov. 2018, doi: 10.1109/tnnls.2018.2797060.

- [2] P. F. Felzenszwalb and J. D. Schwartz, "Hierarchical matching of deformable shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.
- [3] L. Zhao and G. Shi, "A trajectory clustering method based on douglas-peucker compression and density for marine traffic pattern recognition," *Ocean Eng.*, vol. 172, pp. 456–467, Jan. 2019, doi: [10.1016/j.oceaneng.2018.12.019](https://doi.org/10.1016/j.oceaneng.2018.12.019).
- [4] C.-Y. Tsai, "Effect of graphic simplification and graphic metaphor on the memory and identification of travel map," *Int. J. Ind. Ergonom.*, vol. 61, pp. 29–36, Sep. 2017, doi: [10.1016/j.ergon.2017.05.016](https://doi.org/10.1016/j.ergon.2017.05.016).
- [5] J. Cao, Y. Xiao, Z. Chen, W. Wang, and C. Bajaj, "Functional data approximation on bounded domains using polygonal finite elements," *Comput. Aided Geometric Des.*, vol. 63, pp. 149–163, Jul. 2018, doi: [10.1016/j.cagd.2018.05.005](https://doi.org/10.1016/j.cagd.2018.05.005).
- [6] P. Sanzana, J. Gironás, I. Braud, F. Branger, F. Rodriguez, X. Vargas, N. Hitschfeld, J. F. Muñoz, S. Vicuña, A. Mejía, and S. Jankowsky, "A GIS-based urban and peri-urban landscape representation toolbox for hydrological distributed modeling," *Environ. Model. Softw.*, vol. 91, pp. 168–185, May 2017, doi: [10.1016/j.envsoft.2017.01.022](https://doi.org/10.1016/j.envsoft.2017.01.022).
- [7] M. Feng, S.-L. Shaw, Z. Fang, and H. Cheng, "Relative space-based GIS data model to analyze the group dynamics of moving objects," *ISPRS J. Photogramm. Remote Sens.*, vol. 153, pp. 74–95, Jul. 2019, doi: [10.1016/j.isprsjprs.2019.05.002](https://doi.org/10.1016/j.isprsjprs.2019.05.002).
- [8] S. Jiang and W. Jiang, "Reliable image matching via photometric and geometric constraints structured by delaunay triangulation," *ISPRS J. Photogramm. Remote Sens.*, vol. 153, pp. 1–20, Jul. 2019, doi: [10.1016/j.isprsjprs.2019.04.006](https://doi.org/10.1016/j.isprsjprs.2019.04.006).
- [9] A. D. Parakkat, U. B. Pundarikaksha, and R. Muthuganapathy, "A delaunay triangulation based approach for cleaning rough sketches," *Comput. Graph.*, vol. 74, pp. 171–181, Aug. 2018, doi: [10.1016/j.cag.2018.05.011](https://doi.org/10.1016/j.cag.2018.05.011).
- [10] B.-J. Jang, S.-H. Lee, and K.-R. Kwon, "Perceptual encryption with compression for secure vector map data processing," *Digit. Signal Process.*, vol. 25, pp. 224–243, Feb. 2014, doi: [10.1016/j.dsp.2013.09.013](https://doi.org/10.1016/j.dsp.2013.09.013).
- [11] C. Karri and U. Jena, "Fast vector quantization using a bat algorithm for image compression," *Eng. Sci. Technol., Int. J.*, vol. 19, no. 2, pp. 769–781, Jun. 2016, doi: [10.1016/j.jestech.2015.11.003](https://doi.org/10.1016/j.jestech.2015.11.003).
- [12] J. Uthayakumar, T. Vengattaraman, and P. Dhavachelvan, "A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications," *J. King Saud Univ.-Comput. Inf. Sci.*, to be published, doi: [10.1016/j.jksuci.2018.05.006](https://doi.org/10.1016/j.jksuci.2018.05.006).
- [13] W. T. Wang, *Research on Geographical Raster Date Compression and Geographical Scene Management*. Changsha, China: National Univ. Defense Technology, 2011.
- [14] L. Yu, *Research on GML Spatial Data Compression Mechanism*. Nanjing, China: Nanjing Normal Univ., 2013.
- [15] Q. X. Hu, *Research on Storage Optimized Multi-Resolution Vector Geographic Data Organization*. Nanjing, China: Nanjing Normal Univ., 2015.
- [16] T. Chen, S. Li, J. Liu, and H. Chen, "A vector data reduction algorithm based on intersection feature points," *Sci. Surveying Mapping*, vol. 2, pp. 58–62, Aug. 2018, doi: [10.1109/icma.2018.8484336](https://doi.org/10.1109/icma.2018.8484336).
- [17] A. G. Qiu, "In-memory distributed computing based approximate query processing on spatial data," *Acta Geodaetica Cartographica Sinica*, vol. 46, no. 12, p. 2044, Dec. 2017, doi: [10.1109/icoac.2017.8441297](https://doi.org/10.1109/icoac.2017.8441297).
- [18] Y. Gui, A. Su, and J. Du, "Point-pattern matching method using SURF and shape context," *Optik*, vol. 124, no. 14, pp. 1869–1873, Jul. 2013, doi: [10.1016/j.ijleo.2012.05.037](https://doi.org/10.1016/j.ijleo.2012.05.037).
- [19] J. L. Wu, *Study of Shape Matching Based on Modified Shape Context Algorithm*. Anhui, China: Anhui Univ. Technology, 2017.
- [20] A. Daemi, H. Kodamana, and B. Huang, "Gaussian process modelling with Gaussian mixture likelihood," *J. Process Control*, vol. 81, pp. 209–220, Sep. 2019, doi: [10.1016/j.jprocont.2019.06.007](https://doi.org/10.1016/j.jprocont.2019.06.007).
- [21] X. Li and J. Zhao, "An overview of particle-based numerical manifold method and its application to dynamic rock fracturing," *J. Rock Mech. Geotech. Eng.*, vol. 11, no. 3, pp. 684–700, Jun. 2019, doi: [10.1016/j.jrmge.2019.02.003](https://doi.org/10.1016/j.jrmge.2019.02.003).
- [22] R. Xin, T. Ai, X. F. Yan, and M. Yang, "Similarity measurement-based outline design of metaphor map," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 44, no. 4, pp. 625–632, 2019.
- [23] T. H. Ai and G. Chen, "Key issues of multi-scale representation of spatial data," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 30, no. 5, pp. 377–382, Jun. 2005.
- [24] Y. H. Ang, Z. Li, and S. H. Ong, "Image retrieval based on multidimensional feature properties," *Proc. SPIE*, vol. 2420, pp. 47–57, Mar. 1995.
- [25] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Comput. Vis. Image Understand.*, vol. 89, nos. 2–3, pp. 114–141, Feb. 2003, doi: [10.1016/s1077-3142\(03\)00009-2](https://doi.org/10.1016/s1077-3142(03)00009-2).
- [26] A. Peter and A. Rangarajan, "Shape analysis using the Fisher-Rao Riemannian metric: Unifying shape representation and deformation," in *Proc. Int. Symp. Biomed. Image*, Piscataway, NJ, USA, 2006, pp. 1164–1167.
- [27] G. M. Zhang and B. F. Cai, "Study on shape matching based on inner distance shape context feature," (in Chinese), *J. Nanchang Hangkong Univ., Natural Sci. Ed.*, vol. 30, no. 2, pp. 1–7, 2016, doi: [10.3969/j.issn.1001-4926.2016.02.001](https://doi.org/10.3969/j.issn.1001-4926.2016.02.001).
- [28] N. Madian, K. B. Jayanthi, D. Somasundaram, and S. Suresh, "Identifying centromere position of human chromosome images using contour and shape based analysis," *Measurement*, vol. 144, pp. 243–259, Oct. 2019, doi: [10.1016/j.measurement.2019.05.029](https://doi.org/10.1016/j.measurement.2019.05.029).
- [29] A. Srivastava, P. Turaga, and S. Kurtk, "On advances in differential-geometric approaches for 2D and 3D shape analyses and activity recognition," *Image Vis. Comput.*, vol. 30, nos. 6–7, pp. 398–416, Jun. 2012, doi: [10.1016/j.imavis.2012.03.006](https://doi.org/10.1016/j.imavis.2012.03.006).
- [30] T. P. Fries and D. Schöllhammer, "Higher-order meshing of implicit geometries, part II: Approximations on manifolds," *Comput. Methods Appl. Mech. Eng.*, vol. 326, pp. 270–297, Nov. 2017, doi: [10.1016/j.cma.2017.07.037](https://doi.org/10.1016/j.cma.2017.07.037).
- [31] J. Klemelä, "Visualization of multivariate density estimates with shape trees," *J. Comput. Graph. Statist.*, vol. 15, no. 2, pp. 372–397, Jun. 2006, doi: [10.1198/106186006x113007](https://doi.org/10.1198/106186006x113007).



LIHONG SHI was born in 1969. She received the M.S. degree in photogrammetry and remote sensing from Wuhan University, China, in 2006. She is currently a Researcher with the Chinese Academy of Surveying and Mapping, Beijing. Her research interests include mapping and GIS, and spatio-temporal big data analysis.



ZHONGXIANG WANG was born in 1969. He received the M.S. degree in cartography from the Wuhan Technical University of Surveying and Mapping, China. He is currently a Senior Engineer and the Director of the Business Department with the National Geomatics Center of China, Beijing. He is a member of the China Marine Information Professional Committee. His research interests include emergency surveying, mapping, management and support, island reef mapping, survey, and geography information application.



HAO WU was born in 1983. He received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, China, in 2012. He is currently a Senior Engineer with the National Geomatics Center of China, Beijing. His research interests include global mapping updating, verification, and interoperability, dynamic service computation, emergency surveying and mapping. He has been a Chair of ISPRS Working Group ICWG IV/III Global Mapping: Updating, Verification and Interoperability, since 2019.



BINBIN WANG was born in 1992. She received the M.S. degree in surveying and mapping from Liaoning Technical University, China, in 2018. She is currently an Algorithm Engineer with Beijing Piesat Information Technology Company Ltd. Her research interests include mapping and GIS, and analysis of spatial data mining.



XINLIN QIAN was born in 1980. He received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, in 2011. He is currently a Postdoctoral Researcher with the Chinese Academy of Surveying and Mapping, Beijing. His research interests include geospatial database and mapping services.

• • •