International Journal of Geo-Information

MDPI

Article

A Multi-Scale Residential Areas Matching Method Using Relevance Vector Machine and Active Learning

Xinchang Zhang [1,2,*], Guowei Luo [1,3], Guangjing He [1] and Liyan Chen [1,4]

1 School of Geography and Planning, Sun Yat-sen University, Guangzhou 510275, China; bestlgw@163.com (G.L.); heguangjin1234@126.com (G.H.); jimigao@163.com (L.C.)
2 Guangdong Key Laboratory for Urbanization and Geo-simulation, Guangzhou 510275, China
3 School of Geography and Planning, Guangxi Teachers Education University, Nanning 530001, China
4 Guangzhou Urban Planning & Design Survey Research Institute, Guangzhou 510275, China
* Correspondence: eeszxc@mail.sysu.edu.cn; Tel.: +86-138-2221-3215

Academic Editors: Marinos Kavouras and Wolfgang Kainz
Received: 28 November 2016; Accepted: 1 March 2017; Published: 3 March 2017

Abstract: Multi-scale object matching is the key technology for upgrading feature cascade and integrating multi-source spatial data. Considering the distinctiveness of data at different scales, the present study selects residential areas in a multi-scale database as research objects and focuses on characteristic similarities. This study adopts the method of merging with no simplification, clarifies all the matching pairs that lack one-to-one relationships and places them into one-to-one matching pairs, and conducts similarity measurements on five characteristics (i.e., position, area, shape, orientation, and surroundings). The relevance vector machine (RVM) algorithm is introduced, and the method of RVM-based spatial entity matching is designed, thus avoiding the needs of weighing feature similarity and selecting matching thresholds. Moreover, the study utilizes the active learning approach to select the most effective sample for classification, which reduces the manual work of labeling samples. By means of 1:5000 and 1:25,000 residential areas matching experiments, it is shown that the RVM method could achieve high matching precision, which can be used to accurately recognize 1:1, 1:*m*, and *m*:*n* matching relations, thus improving automation and the intelligence level of geographical spatial data management.

Keywords: relevance vector machine (RVM); residential areas; entity matching; similarity; object merging

1. Introduction

Environmental protection, land resource management, emergency relief, and the construction of smart cities require reliable, applicable, and timely geographical spatial data for support. Therefore, an important issue that needs to be resolved in geographical information systems (GIS) is the immediate updating and integration of geographical spatial data [1,2]. The purpose of multi-scale data updating is to ensure that the different scales of spatial data reflect the latest situation. However, comprehensive updating that uses a large-scale data generalization method to produce small-scale data requires a great amount of work, and ensuring the consistency of multi-scale data is difficult. Multi-scale feature–cascade updating is a popular method for quickly updating spatial data in academic research [3–5]. This method uses incremental information from large-scale data to update small-scale data, and updates only the changed features. Therefore, it involves less work than other methods and is able to maintain consistency. The feature–cascade relationship forms the basis for multi-scale cascade updating. The establishment of this relationship relies on entity matching technology for recognition of different scale entities with the same name. Therefore, entity matching is the key technology for updating spatial data. Geospatial data fusion aims to automatically generate data that has higher

degree of accuracy and richer attribute information than multiple data sources [6]. To extract and merge information from different data sources, the corresponding relationships of different entities in the database must be identified, which relies on the technology of matching spatial entities [7]. Considering that multi-source spatial data invariably have different scales, the multi-scale features should be taken into account in the matching method.

Considering the current rapid urban and rural development, residential areas are among the fastest changing geographic objects and, thus, are an important data type that requires updating. The research objective of the present study is for multi-scale residential areas matching. A large amount of prior research has been conducted in matching of residential areas. The current methods can be classified as being a similarity-based matching method, a probability-based matching method, or an error-based matching method.

The similarity-based matching method for planar spatial entities realizes matching by analyzing the degree of overlap of buffer areas [8], distance [9,10], shape [11–13], topology, direction, semantics [14], and other shared characteristics. Similarity matching integrates multiple similar characteristics [15] with an optimal combination while considering many-to-many matching relationships [16]. Ai [11] proposed using Fourier shape descriptors to measure the shape similarity of residential areas to realize shape analysis and to match multi-scale residential areas. An and Sun [12] proposed describing the geometric shape step-by-step from an overall blueprint to specify details, and apply the multi-stage chord function and the center distance function to establish a common geometric similarity measurement model for multi-scale spatial data. These two shape similarity measurement methods are applicable to the matching of a single planar entity. Huh [17] proposed detecting the corresponding nodes of planar entities in multi-source data and conducting matching in terms of object outlines. Birgit [18] proposed extracting the skeleton of planar entities and conducting the matching of a multi-scale river network by calculating the similarity degree of skeleton characteristics. Kim [19] proposed an object-matching method based on the geographic environment. This method measures the geographic environment's similarity of the space between objects and the selected geographic landmarks to realize spatial data matching within different coordinate systems. Thus, this method is dependent on the selection of landmarks. Zhang and Ai [20] suggested the use of relaxation tag technology in combination with the overall information architecture to establish a compatibility matrix. By the constant updating of compatibility of candidate matching objects, the compatibility matrix is converged and multi-scale planar entity matching is achieved.

Walter and Fritsch [21] utilized a matching method based on probability statistics, which first selects the candidate matching set and then utilizes regional statistics to determine the threshold value. Finally, it applies the merit function to finalize the matching results. Tong [22] studied the multi-characteristic matching method based on probability, and discovered that calculating the matching probability of the entity to determine the matching entities avoids the selection of the exact threshold for the matching index.

The error-based matching method is applicable to the matching of multi-scale geographic data under strict cartographic specifications. Safra [23] proposed a spatial data matching method based on location. This method compares the distance between spatial objects with the tolerance error of the map to determine the matching relationship. To solve the difficulty of determining the threshold and many-to-many relationships of multi-scale spatial data matching, Liu [24] suggested a planar entity matching algorithm based on mean square error and adjacent entity relationships.

In recent years, to improve the automation of spatial entity matching, researchers have proposed a matching method based on pattern recognition. Zhang [25] utilized a multi-scale residential areas matching method based on pattern classification, and Wang [26] proposed a multi-represented feature matching method based on a back-propagation neural network.

Similarity-based matching methods coincide with human spatial cognition and are the most widely used method types in multi-scale residential areas matching studies. These methods present difficulties, including the following: (1) owing to different scales, production units, and production

times, the residential areas in a multi-scale spatial database are large in number, great in disparity, and complicated in matching relationships, making it difficult to conduct similarity measurement; and (2) it is difficult to determine the similar weights and threshold values, and manual intervention is largely required. For the first problem, we aim to improve the characteristic similarity measurement method for multi-scale residential areas. The study provides a solution for similarity measurements of matching objects that are not one-to-one. For the second problem, machine learning is a more effective method to solve the problem of threshold and weight; however, the existing machine learning-based matching methods require a large number of labeled training samples, and it is difficult to identify multiple matching relationships [25,26]. The study presents a matching method that is based on the RVM algorithm and active learning, which avoid manually setting characteristic weights and matching thresholds in the match methods. Moreover, the work associated with sample labeling can be reduced. In Section 2, we describe our methodology, while experimental results are analyzed and discussed in Section 3. The final section includes conclusions and future prospects.

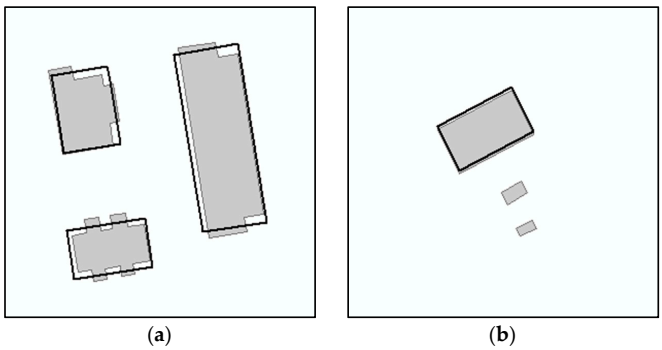
2. Methods

2.1. Multi-Scale Residential Areas Matching Relationship

Matching of multi-scale geospatial data is difficult because of the comprehensive impact of cartography, errors yielded during data production, and the alteration of geographic entities themselves. Following is an introduction to the matching relations of different scale spatial entities after cartographic generalization. To begin with, we assume that the multi-scale spatial entities have the same spatial coordinates, and the small-scale maps are the generalizations from the large-scale maps. Thus, there are differences in the spatial expression of different scaled entities. For planar residential areas, the matching relations of large and small scales can be classified as:

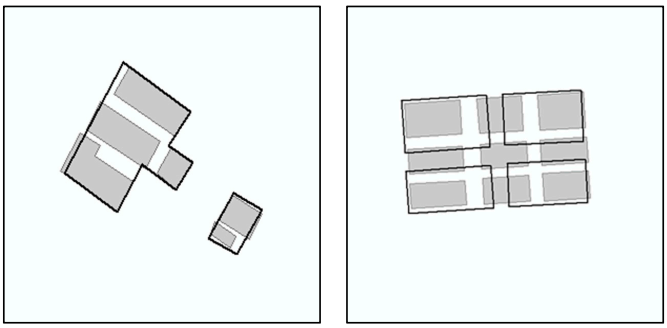
- 1:1 (Figure 1a)—in both large and small scales the entities with the same name have a 1:1 matching relationship.
- 1:0 (Figure 1b)—where some entities occur in large-scale maps, but are invisible in small-scales maps because during map generalization some small entities are omitted.
- *m*:1 (Figure 1c)—a many-to-one relationship for entities between large-scale maps and small-scale maps, where in the process of map generalization large-scale objects are combined to form small-scale objects.
- *m*:*n* ($m > n$) (Figure 1d)—a many-to-many relationship for entities between large-scale maps and small-scale maps, where during the map generalization process a stylization operation is conducted to reflect the shapes and spatial distribution features of residential areas.

In addition to the differences reflected in number, operations such as shape simplification and displacement are also conducted during map generalization. In this way, different entities with the same name will differ in shape and position in different scale maps.



(**a**) (**b**)

Figure 1. *Cont.*



(**c**) (**d**)

Figure 1. Matching of multi-scale data exemplified (larger and smaller scale objects are in gray and white, respectively). (**a**) 1:1; (**b**) 1:0; (**c**) *m*:1; (**d**) *m*:*n*.

2.2. Overall Design

The present study aims to constitute a multi-scale residential areas matching method compatible with data characteristics, by introducing the concept of categorization from pattern recognition. In addition, the study aims to place selected samples into classification models by machine learning designed for application to object matching within identical scenarios. The overall framework is shown in Figure 2, and is as follows:

(i) Selecting training samples of both matched and unmatched objects via human-machine cooperation.
(ii) For candidate matching objects, converting matching relations that do not correspond one-to-one into one-to-one relations for the convenience of similarity computation by data processing.
(iii) Computing characteristic similarities of the sample data.
(iv) Applying a relevance vector machine (RVM) algorithm to characteristic similarities and matching results to generate classifiers.
(v) Inputting residential data at various scales after data processing into (iv) classifiers to yield classification results.
(vi) Multi-matching data classified as matched to obtain the final matching results.

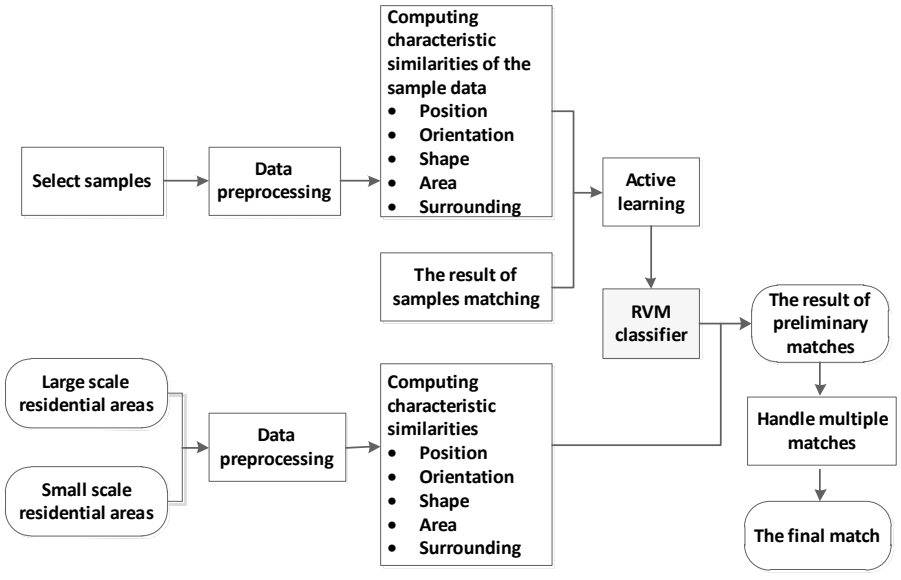


Figure 2. Overall framework.

2.3. Object Merging

First, we apply the buffer to search for candidate matching objects. To contain the many-to-many matching relationship, large-scale and small-scale objects must undergo a forward and reverse bidirectional iterative search [16,25]. Object merging is the most efficient way to convert one-to-many and many-to-many relations into one-to-one matching relations. Considering the complexity of the reduction and one-to-one matching relations that do not need to be converted, the present study conducts the data processing in a way that merges without simplifying.

Our aim is to maintain the outer contour of residential areas during merging because several merging methods are feasible for objects with large-scales owing to the variety of cartographic generalization methods. Residential areas that meet each other are merged by removing the joining edges, while residential areas discrete from each other are merged by generating and processing Delaunay triangulation [27], as shown in Figure 3. The exact approach is as follows:

(i) To perform node encryption by inserting nodes into the contour of residential areas elements, and to construct Delaunay triangulation, thus categorizing triangles outside and inside residential areas elements into external and internal triangles.

(ii) To generate convex hull from pre-merging residential areas and remove the following three types of external triangles that have joining edges with a convex hull: (1) all three vertices are located inside an identical residential areas; (2) vertices are located in two residential areas with one interior angle measuring more than θ (θ is obtuse) and one edge sharing the same edge with either of the two residential areas; and (3) vertices are located in two residential areas with an edge overlapping with the contour of residential areas and the altitude of this edge is larger than the threshold. This rule is applied in the measurement of the distance between residential areas.

(iii) To apply a recursive algorithm to search for other external triangles that have joining edges with removed triangles and applying rule (2) and (3) in the previous step to remove suitable triangles.

(iv) To merge triangles remaining by removing joining edges.

After the merging processing, the many-to-one and many-to-many relations in residential areas of various scales are converted into one-to-one relations to be matched.

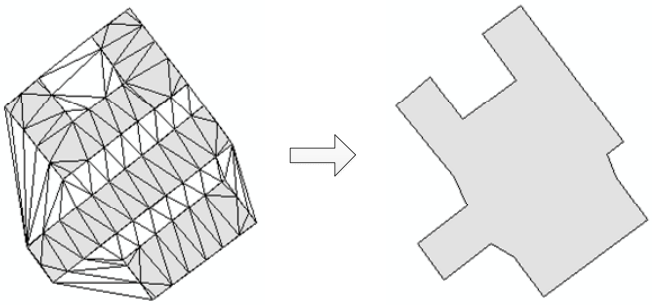


Figure 3. Applying Delaunay triangulation to merge residential areas.

2.4. Similarity Computation

Numerous one-to-many and many-to-many relations exist within residential areas of various scales that require data processing to convert them into one-to-one relations suitable for characteristic similarity computation. In the present study, the method described in Section 2.3 is used to transfer these relationships into one-to-one relations to facilitate the calculation of feature similarity. Based on the features of residential areas with human spatial cognition, five characteristics, i.e., position, area, shape, orientation, and surroundings are utilized to evaluate the similarity for matching relations. The value of each characteristic similarity index is between 0 and 1. Although there are problems related to incomplete attribute information and inconsistent criteria of data, this study adopts a method

independent of semantic information by selecting spatial characteristics, which are less affected by multi-scale representation and map generalization than characteristics such as perimeter and overlapping areas are.

2.4.1. Position Similarity Index

The closeness of spatial entities indicates a high similarity of position. For a geographical area entity, the centroid might best reflect the characteristics of its location. The present study applies Equation (1) below to measure position similarity by calculating the ratio of the Euclidean distance of the centroid of the two residential groups and the maximum distance D. In Equation (1), (x_1, y_1) and (x_2, y_2) are the centroid coordinates of the two entities to be matched. The maximum distance D of the matched entities is determined by statistical analysis of the centroid distance of positive and negative matching samples. In this study, the value of D is double the centroid distance of matching samples.

$$S_{position} = 1 - \frac{\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}}{D} \tag{1}$$

2.4.2. Area Similarity Index

Area is an important characteristic that reveals the size of a geographical entity. Although differences in area exist in area entities of various scales because of map generalizations and other factors, maintaining the characteristics of size of a geographic entity is one principle of map generalization. Residential areas that have matching relations would have similarities in area. This study applies Equation (2) below to measure the area similarity of residential areas by calculating the ratio of areas of the residential areas to be matched, where A and B refer to the residential areas to be matched.

$$S_{area} = \frac{MIN(Area(A), Area(B))}{MAX(Area(A), Area(B))} \tag{2}$$

2.4.3. Shape Similarity Index

The quantitative description of shape is an enigma in the field of GIS and computers [28]. Taking into account the shapes of residential entities, the present study uses the shape index (compactness) proposed by Peter to measure it [29]. Compactness is affected by the size and boundary of the object [30] and is calculated as shown in Equation (3), where p represents the area entity. The calculation method of shape similarity is proposed in Equation (4), where A and B represent objects to be matched in large and small scales, respectively.

$$Compact(p) = \frac{perimeter(p)}{2\sqrt{\pi * Area(p)}} \tag{3}$$

$$S_{shape} = \frac{|Compact(A) - Compact(B)|}{MAX(Compact(A), Compact(B))} \tag{4}$$

2.4.4. Orientation Similarity Index

The orientation similarity of residential areas refers to the overall extension direction. Commonly used methods include the long side method, the wall-based statistical method, and the smallest minimum bounding rectangle (SMBR) method [30]. The SMBR method uses the long axis direction of the SMBR of the entities to be matched as the orientation of residential areas. The angle difference of the long axis direction is the angle difference of the two area entities. This method cannot recognize the orientation of two area entities that rotate their orientations 180 degrees. The present study proposes an improved method for the SMBR method. If the residential areas to be matched are A and B, the orientation similarity of the residential areas is calculated according to Equation (5), where θ_A and θ_B

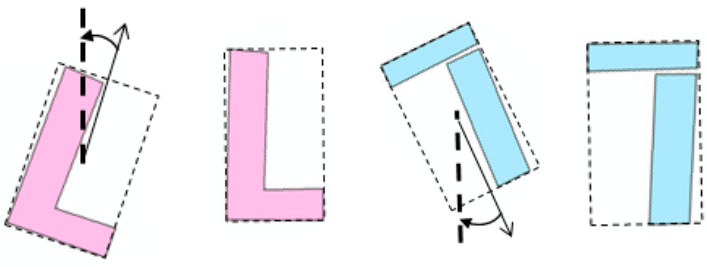
are the angles of the SMBR long axes of A and B and the y-axis, respectively; the value intervals are $[0, \pi/2]$; and F is the Boolean function that determines whether the residential areas rotate 180 degrees.

$$S_{orientation} = F \times (1 - \frac{|\theta_A - \theta_B|}{\pi/2}) \quad (5)$$

If the shape similarity index of the two objects is low, then the F in Equation (5) is 0. If the shape similarity index is high, then it needs to be further processed using the following method: the minimum angles a_A and a_B are calculated by counterclockwise rotation of the to-be-matched object A and B and their SMBR when the long axes of the SMBR and the y-axis are parallel. When $|a_A - a_B| > \pi/2$, object A and B are counterclockwise rotated for a_A and a_B, and the new objects A′ and B′ are obtained. When $|a_A - a_B| > \pi/2$, object A is clockwise rotated for a_A to obtain A′, and object B is counterclockwise rotated for $\pi - a_B$ to obtain B′. As Figure 4 shows, after rotating Figure 4a, Figure 4b is obtained, and after rotating Figure 4c, Figure 4d is achieved. The bottom right corner of the SMBRs of object A′ and B′ are used as the points of origin. A coordinate system is established with the short axis as direction x and the long axis as direction y. The SMBRs of A′ and B′ are equally divided to m rectangles along the long side, and n rectangles along the short side. The ratio between the intersection area of each rectangle and object, and the area of corresponding rectangle is calculated, the value interval of which is [0,1]. Two histograms are generated using each ratio in the x-axis positive direction and the y-axis positive direction. Figure 5a,b are area histograms of Figures 4b and 5c,d are area histograms of Figure 4d. The horizontal axis of the histogram is a serial number of rectangles, and the vertical axis is the ratio between the intersection area of the rectangle and the object and the area of the corresponding rectangle.

The procedures for recognizing direction via a histogram are as follows: (1) Smooth the histograms by using Equation (6) with an interpolation method, where x in Equation (6) is the horizontal coordinate value of the histogram, f(x) is the vertical coordinate value, step is the step length, and Z is the value of histograms after smoothing. The x-axis direction of the area histogram after smoothing has h units, and the y-axis direction of the area histogram has j units. (2) Compare the average values of the histograms and each unit value, excluding the histogram that has the smallest unit value difference. The histogram represents a rectangle that does not need to be compared. The F value is 1. (3) Compare the ith unit of a histogram with the (h-i)th unit of another histogram. When their difference is smaller than the given threshold value, these two values are regarded as the same. After comparing h groups, the number of the same units reaches u (here u/h > 0.9), which means the two histograms are opposite. (4) When there is the opposite corresponding histogram, the F value in Equation (5) is 0; otherwise it is 1.

$$Z = \frac{\sum_{i=0}^{i=step-1} f(x+i)}{step} \quad (6)$$



(**a**) (**b**) (**c**) (**d**)

Figure 4. Rotation of object: (**a**) is rotated counterclockwise to obtain (**b**), and (**c**) is rotated counterclockwise to obtain (**d**).

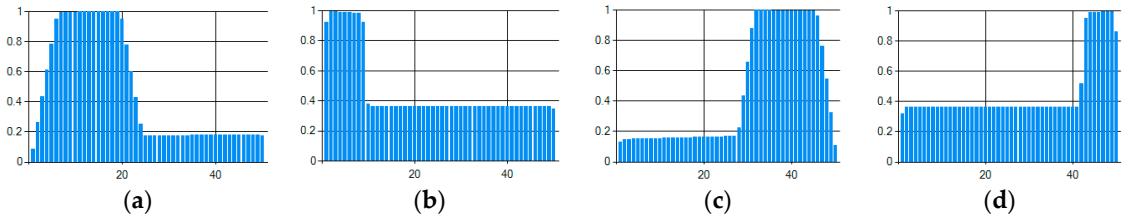


Figure 5. Projective area histograms: (**a**) Figure 4b area projective histogram as per *x*-axis; (**b**) Figure 4b area projective histogram as per *y*-axis; (**c**) Figure 4d area projective histogram as per *x*-axis, and (**d**) Figure 4d area projective histogram as per *y*-axis.

2.4.5. Surroundings Similarity Index

Commonly on a large-scale map, constructions in the same area have the same shape, with mistakes being made if only the position and geometric characteristic are used for matching. According to the spatial cognition habit, artificial matching of the information in surrounding areas is frequently combined to identify entities. We determine the measurement of surroundings similarity by measuring the characteristics of surrounding entities in residential areas, as shown in Figure 6, the mass center of the entity to be matched is used as a center point to construct a 2×2 square grid, which is parallel to the coordinate axis. The side length of the grid is set at twice the length of the long side of the SMBR element of the small-scale residential areas to be matched. G_1, G_2, G_3, and G_4 represent the upper left, upper right, bottom left, and bottom right of the grid, respectively, as shown in Figure 6. The surroundings similarity of each grid area is calculated according to Equation (7), where $Area(SM_i)$ and Area(LA_i) are the surrounding residential areas located in the area of the grid in the small-scale and large-scale data, respectively. When the value of $Area(SM_i)$ and the value of $Area(LA_i)$ are not 0, the value of $Sim(\mathrm{G_i})$ will be the area ratio. When the value of $Area(SM_i)$ and the value of $Area(LA_i)$ are 0, the value of $S_{grid}(G_i)$ will be 1. With different scales, the representations of geographic entities also show differences, so some very small residential areas in small-scale that would be presented in large-scale data might not be observed. When the value of $Area(SM_i)$ is 0 and the value of $Area(LA_i)$ is a small number (less than the threshold ε), the value of $S_{grid}(G_i)$ will be 1. When the value of $Area(SM_i)$ is not 0 and the value of $Area(LA_i)$ is 0, the value of $S_{grid}(G_i)$ will be 0. The total surroundings similarity is calculated according to Equation (8).

$$S_{grid}(G_i) = \begin{cases} \frac{Min(Area(SM_i), Area(LA_i))}{Max(Area(SM_i), Area(SM_i))} & IF\ Area(SM_i) \neq 0\ AND\ Area(LA_i) \neq 0 \\ 1 & IF\ Area(SM_i) = 0\ AND\ Area(LA_i) \neq 0\ AND\ Area(LA_i) < \varepsilon \\ 0 & IF\ Area(SM_i) \neq 0\ AND\ Area(LA_i) = 0 \\ 1 & IF\ Area(SM_i) = 0\ AND\ Area(LA_i) = 0 \end{cases} \tag{7}$$

$$S_{surrounding} = \sum_{i=1}^{4} S_{grid}(Gi)/4 \tag{8}$$

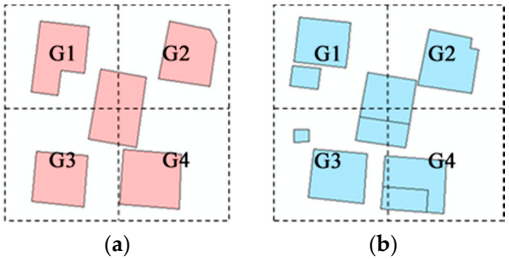


Figure 6. Grid-based surroundings similarity: (**a**) grid constructed by small-scale data; and (**b**) grid constructed by large-scale data.

2.5. Matching Approach

2.5.1. Relevance Vector Machine

After calculating each feature similarity of the matching candidates, the standard matching approach is to obtain comprehensive similarity by weighing characteristic similarity and to select matching results using thresholds [13,15]. The weighing and threshold setting processes within this approach require manual intervention, which makes it cumbersome for adoption in different data fields. In the present study, a machine learning approach is designed to realize the matching of spatial entities.

The RVM [31] is a new type of machine learning approach that has been developed in recent years. It is similar to the support vector machine (SVM), as it is especially suitable for binary classification of small samples.

In the present study, the input vector of the RVM is defined as being five-dimensional, including the similarities of five characteristics (i.e., position, area, shape, orientation, and surroundings). Classification is defined as "match" or "mismatch". The output of the RVM can be utilized as an assessment of the reliability of the classification results. The output function of RVM is shown in Equation (9) [32], where $y \in [0, 1]$.

$$y = \sigma(z) = 1/(1 + e^{-z}) \tag{9}$$

The value of z in Equation (9) is calculated as shown in Equation (10), where $Q(x, x_n)$ is the kernel function and, w_n is the weight of the model.

$$z = f(x; W) = \sum_{n=1}^{N} w_n Q(x, x_n) + w_0 \tag{10}$$

The estimates of the dataset obtained from the likelihood estimator is shown as Equation (11), where $t = (t_1 \cdots t_N)^T$ and $W = (w_0 \cdots w_N)^T$.

$$p(t|W) = \prod_{n=1}^{N} \sigma\{f(x_n; W)\}^{t_n} [1 - \sigma\{f(x_n; W)\}]^{1-t_n} \tag{11}$$

In the Bayesian framework, the weights W in Equation (11) can be obtained with the maximum likelihood estimation method. However, to avoid over-learning, RVM defines a Gaussian prior probability distribution for each weight to constrain the parameters (Equation (12)), where α in Equation (12) is a N + 1-dimensional hyper parameter. Although the posterior probability of the weights cannot be calculated, it can be approximated by the Laplacian theory. The maximum possible weight W_{MP} is calculated for the currently fixed α value. Because $p(w|t, \alpha) \propto p(t|w)p(w|\alpha)$, it can be translated into the maximum of Equation (13).

$$p(w|\alpha) = \prod_{i=0}^{n} N(w_i \Big| 0, \alpha_i^{-1}) \tag{12}$$

$$\log\{p(t|w)p(w|\alpha)\} = \sum_{n=1}^{N} [t_n \log y_n + (1 - t_n) \log(1 - y_n)] - \frac{1}{2} w^T A w \tag{13}$$

$\frac{1}{2} w^T A w$ is a constant when the maximum possible W_{MP} is obtained. When the relationship between two objects is matched, the value of y tends to 1 so that the result of Equation (13) is maximum. When the relationship between two objects is mismatched, the value of y tends to 0 so that the result of Equation (13) is maximum. Therefore, the reliability of matching between objects is higher when the output values are closer to 1, and the reliability of mismatching between objects is higher when the output values are closer to 0.

2.5.2. Active Learning

When using a machine learning process, sample selection is time consuming. Selecting matched samples from residential data requires artificial identification. To decrease the number of training samples and improve the efficiency of matching, the present study adopts the active learning approach [33]. The main idea behind this approach is that, by multiple iteration sampling, samples that are beneficial to improve classification performance are selected and, by small-scale labeled sample training, learning performance that can be acquired by large-scale labeled samples is achieved. The learning procedures are as follows:

(i) *N* samples from candidate sample set U are selected; their classifications are manually labeled to form initial training sample set D. Each classification needs to have at least one sample in D.
(ii) Samples in D are trained, and initial classifier F is established.
(iii) Classifier F is used to classify unlabeled samples, and classification results of low confidence coefficient are added after they are manually labeled.
(iv) Classifier is trained for another time until it meets the criterion of finishing classifier training. The criterion is that the number of loops reaches the predetermined value or the number of labeled samples reaches the expected value.

RVM and active learning to construct the classifier are shown in Figure 7.

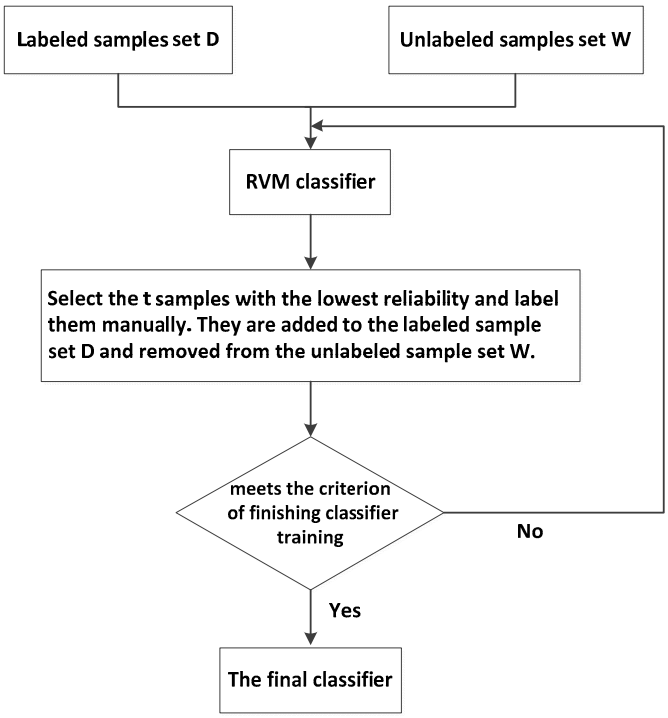


Figure 7. RVM and active learning to construct the classifier.

2.5.3. Matching Strategy

The following are some of the important steps in the matching strategy.

- Data preprocessing: The first step of data preprocessing is to identify the candidate elements for matching. A buffer area is generated from small-scale residential areas, and the large-scale features that intersect with this buffer area are identified as candidate features. Second, the candidate elements are determined by recognizing pairs of multiple match relations and using a bidirectional search. Third, after obtaining the candidate features, since there are many 1:*n* or *m*:*n* match

relations in multi-scale residential areas, permutation and combination are used to generate a combination of candidate matching objects to recognize the match relation. Since the combination of candidate-matching objects is determined according to the number of elements, we set as mismatches the objects that are impossible to be matched based on the matching types of multi-scale residential areas (e.g., the object whose quantitative relationship between the large-scale object and small-scale object is 1:n or m:n ($m < n$)) and deleted these objects from the candidate-matching objects. For ease of similarity measuring, we use the method of entity merging described in Section 2.3 to combine the multiple entities into a single entity.

- Sample selecting: For the training samples, we adopt the method of human-computer cooperation, generating a buffer area with the use of source objects to search for the candidate-matching objects. Manual work is used for identifying and labeling matches or mismatches between candidate elements and source elements. Unlabeled samples are obtained by searching for candidate-matching objects in the buffer area generated from source elements.
- Multiple matching relation processing: There might be some cases of multiple matching based on the output of the classifier, as Figure 8 shows. Entities A and B and A and BC are all classified as matches. However, determining the final match relation depends on the reliability output of the RVM. To be specific, we determine the matching pairs containing the same elements in the category of the match, after being selected by the classifier. Utilizing Equation (14), the set with maximum reliability is selected as the final match result.

$$R = Max(r_1, r_2, r_3, ..., r_n) \tag{14}$$

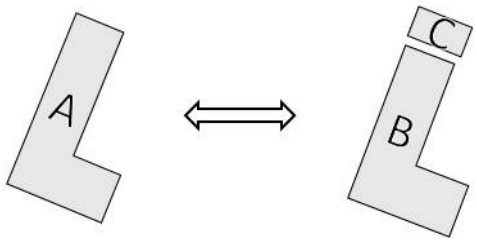


Figure 8. Example of multiple matching.

2.6. Experimental Design

To verify the effectiveness of the proposed method, we selected the residential areas of 1:5000 and 1:25,000 in the Tianhe District, Guangzhou, for the matching experiment (Figure 9). The 1:5000 and 1:25,000 datasets contain 4375 and 1023 entities, respectively. In this experiment, ArcGIS Engine10.0 was secondary developed by Visual Studio 2010 to obtain the characteristic similarity of the spatial entities, together with RVM_Matlab toolbox for the classifications.

In the experiment, a buffer area is generated with the small-scale residential area as the source element and a dataset consisting of 503 recordings is constructed by program automatic selection, which accounts for 70% of the training set, and 30% of the test set. From this we manually selected 76 labeled classification samples for constructing an initial classifier. In the labeled samples, 21 pairs are of 1:1 matching relations, 28 pairs are of 1:m matching relations, seven pairs are of m:n matching relations, and 20 pairs are mismatched. The active learning approach is adopted to continuously optimize the classifier, and the iteration number of the active learning is set at 10. Table 1 shows the initial training samples, where SOURCEID and TARGETID are the element serial identifier code of the small-scale residential area and large-scale residential areas, respectively. Column headings of LOCAL, ORIEN, AREA, SHAPE, and SUR represent the five characteristic similarities, i.e., position, orientation, area, shape, and surroundings, respectively. The column heading RESULT shows the classification

result of manual recognition, in which 1 represents match and 0 represents mismatch. The uncertainty interval of classification confidence is set at [0.1, 0.9]. This uncertainty interval demonstrates that the results of classification are largely in doubt and the rate of wrong classification increases for values in an interval. When the result of the classification training output of unlabeled samples is within the doubting interval, 10 samples with the lowest reliability are selected, and it is manually determined as either a "match" or "mismatch". The manually judged classification result is added to the dataset for retraining and a new classifier is formed. A test is then implemented with the test set. This procedure is repeated until the classification result is convergent, and the final classifier is obtained. Following this, the characteristic similarity of candidate matching pair selected in the buffer is inputted into the classifier to obtain the output of binary classification result. Eventually, the final matching object is determined according to the classification reliability.

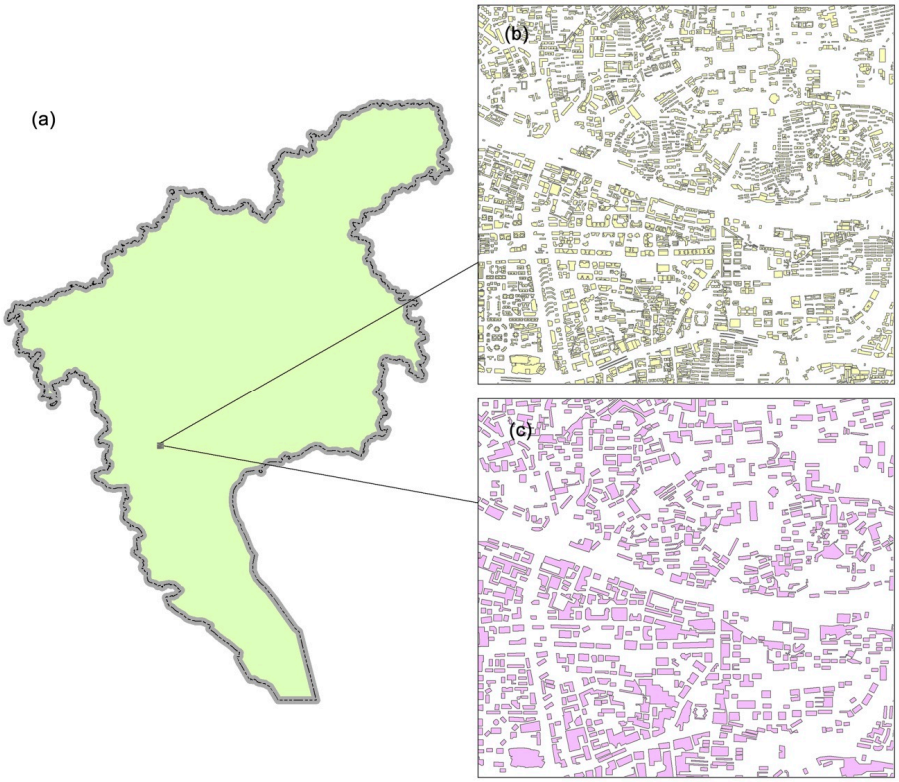


Figure 9. Experimental data: (**a**) Guangzhou map; (**b**) 1:5000 residential areas, and (**c**) 1:25,000 residential areas.

Table 1. Example of initial training samples.

Source ID	Target ID	Local	Orien.	Area	Shape	Sur.	Result
73	352, 353	0.97	1	0.96	0.96	0.75	1
170	528, 529, 530	0.91	0.99	0.92	0.79	0.95	1
195	685, 687	0.94	0.92	0.95	0.76	0.63	1
803	1307	0.86	0.95	0.98	0.81	0.83	1
...	...	...	...	...	...	...	...
161	334, 335	0.35	0.49	0.57	0.60	0.63	0
463	725	0.31	0.96	0.66	0.69	0.51	0
532	2406, 2407, 2409	0.52	0.92	0.65	0.62	0.37	0
697	2908, 2910	0.38	0.11	0.53	0.59	0.40	0

The evaluation of the experimental results is performed by comparing the results of manual matching by professional cartographers with the automatically matched results. The objects that are

labeled as matched by both manual matching and automatic matching are *TP*; those that are matched by manual matching while not recognized by automatic matching are *NP*; and those that are identified as matched by automatic matching but unrecognized by manual matching are *FP*. The indicators of evaluation are precision and the recall rate. Equation (15) is utilized to calculate precision, and Equation (16) is used to obtain the recall rate. The parameter *F*1 value was introduced to measure the harmonic mean of precision and recall, with Equation (17) used to calculate F1.

$$Precision = \frac{TP}{TP + FP} \tag{15}$$

$$Recall = \frac{TP}{TP + NP} \tag{16}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{17}$$

3. Results and Discussion

3.1. Merging of Residential Areas

Merging tests of selected matching samples are carried out based on the merging approach proposed in the present study and the convex hull-based merging method [26]. The results are shown in Table 2 and Figures 10–13. The results reveal: (1) The proposed method obtains the higher average similarity than the convex hull-based merging method; and (2) the position similarity values, area similarity values, and shape similarity values are flatter in the proposed method. Thus, compared with the traditional convex hull-based merging method, the proposed method that transfers the one-to-many and many-to-many relations into one-to-one relations is more suitable for adopting to the similarity measure, with the feature similarity and measurements selected being more applicable for the matching of multi-scale candidates.

Table 2. Average similarity calculated by different merging methods.

	Mean Position Similarity	Mean Orientation Similarity	Mean Area Similarity	Mean Shape Similarity
Convex hull-based	0.79	0.99	0.66	0.51
The proposed method	0.95	0.99	0.94	0.90

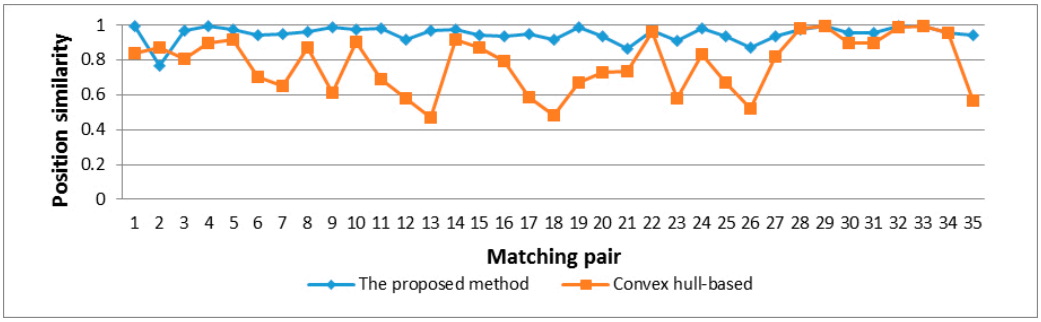


Figure 10. Position similarity values between the matching pair using the two merging approaches.

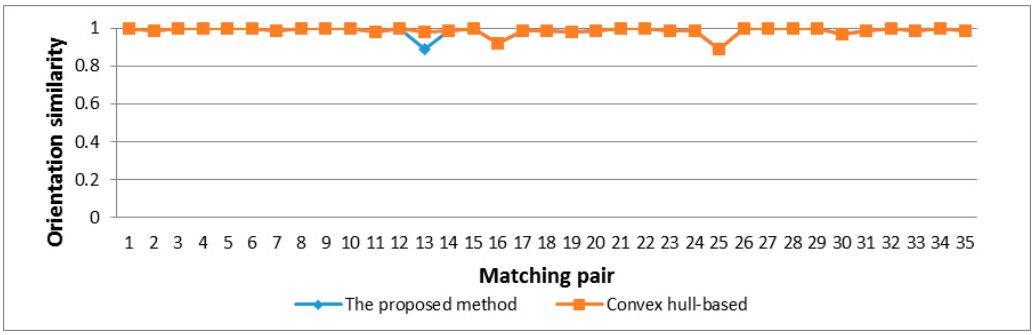


Figure 11. Orientation similarity values between the matching pair using the two merging approaches.

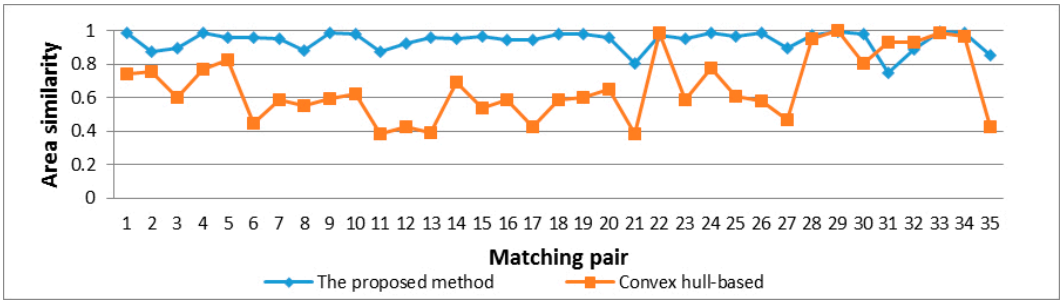


Figure 12. Area similarity values between the matching pair using the two merging approaches.

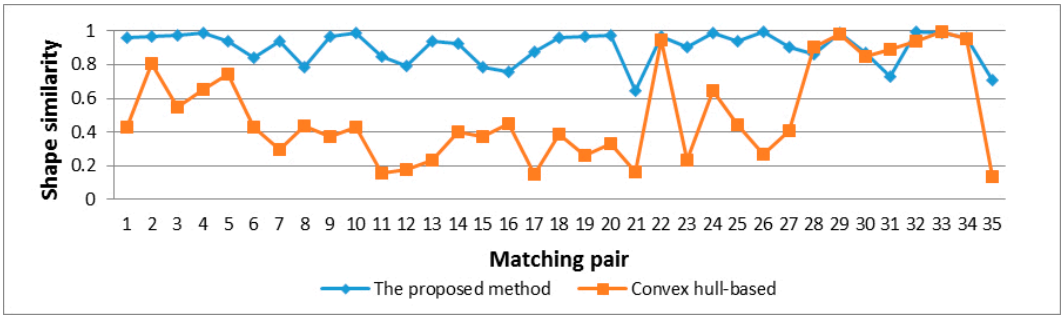


Figure 13. Shape similarity values between the matching pair using the two merging approaches.

Figure 14 displays examples of merging tests of features. Figure 14a,b are the residential areas to be matched at different scales. They are matched visually, but by using different programs to make automatic identification and using the convex hull-based method to merge Figure 14a into Figure 14c, we observe that they are quite different geometrically. However, using the proposed method to merge Figure 14a and obtain Figure 14d, we discover that the geometric similarity between Figure 14d and 14a is higher than that between Figure 14c and 14a.

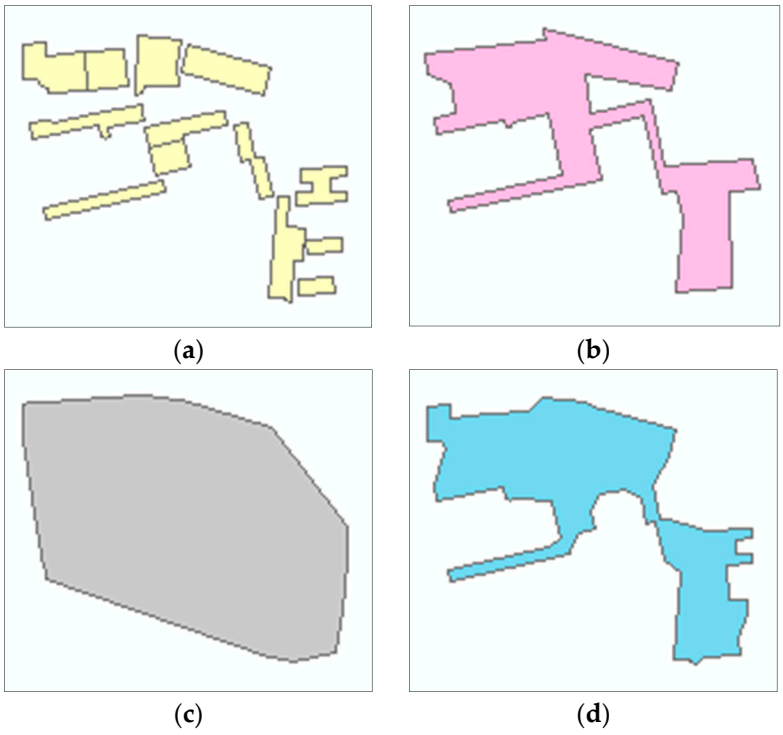


Figure 14. Examples of feature merging: (**a**) large-scale residential areas; (**b**) small-scale residential areas; (**c**) merging effect of the convex hull-based method; and (**d**) merging effect of the proposed method.

3.2. Feature Similarity Measure

In terms of the measurement of residential areas' similarity (Figure 15), by calculating the long axis direction of the SMBR of the entities, the direction of the merged entity of No. 2625 and No. 2626 is recognized as being the same as the direction of No. 2629 by the former SMBR measuring method [26], which is similar to the direction of residential area No. 729. However, the results from the proposed measurement method reveal that the direction similarity is 0, which means that No. 2629 and No. 729 are in opposite directions, and the similarity between the merged entity of No. 2625, No. 2626, and No. 729 is 1, which is consistent with the artificial perception.

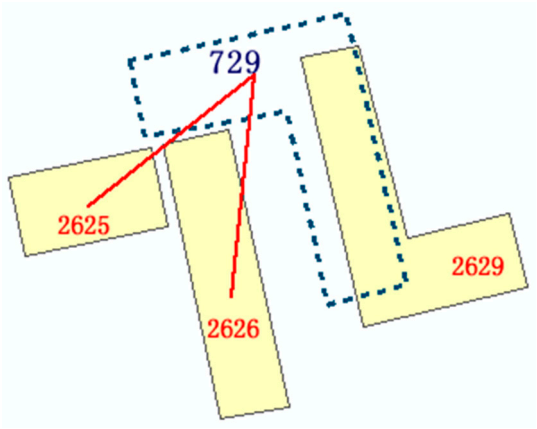


Figure 15. An example of the metric of direction similarity index.

3.3. Result Comparison of the Matching Methods

In the process of constructing the classifier using the RVM and active learning, we counted the correct number of classification results in 151 test samples. As shown in Figure 16, with an increase in the number of iterations, the correct number of test samples is gradually increased. After eight iterations, the classification results are stable. Using the same number of labeled samples for passive learning (samples are randomly selected from the sample set), the accuracy of the classification of the test sample is lower than the active learning method. Figure 16 shows that active learning can achieve better classification results with fewer labeled samples.

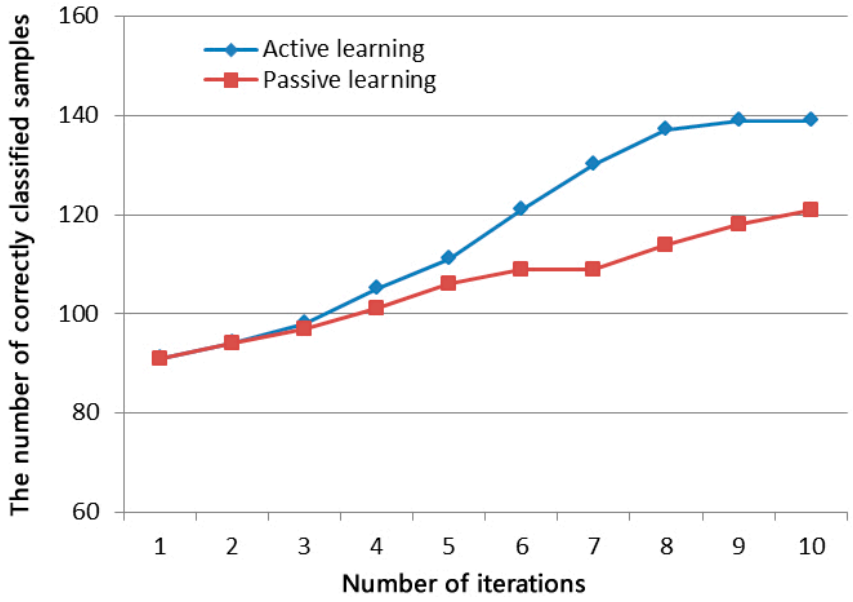


Figure 16. The statistics of the correct number of categories in test samples.

The results of the experiments are shown in Table 3, which shows that by taking the proposed RVM to process the selected features, the accuracy of matching is 92.1% and the recall rate is 91.8%. Compared with other methods, the RVM shows a distinct advantage in successful matching. The overlay method [8] for determining the buffer area is commonly used in matching data of the same scale, because in multi-scale data the shifting, merging, simplifying, and other operations during the cartographic generalization results in a low overlapping rate of buffer areas with different scales. In addition, the overlapping threshold is difficult to identify. Therefore, the accuracy of this method is relatively low. When taking the proposed selected characteristics to process, among the weighting feature similarity [15], SVM [25] and the mentioned RVM methods, the accuracy of matching is higher than that when using the selected characteristics proposed in Zhang [25]. For the weighted matching method [15] based on characteristics' similarities, the characteristics' weight and matching thresholds have a great impact on accuracy, and it is difficult to identify them manually; thus, the rate of successful matching is not high. Using the SVM [25] algorithm can avoid the manual setting of characteristic weights and match thresholds. In addition, this method is suitable for classifying two-type problems, yet when multiple matches commonly exist in multi-scale objects, it is unable to further identify best match pair and may judge some mismatched objects as being matched.

Table 3. The statistical evaluation of the proposed method and other methods.

Matching Method	Characteristics	TP	FP	NP	Precision (%)	Recall (%)	F1 (%)
Buffer area of overlap [8]	-	2351	1205	1112	66.1	67.9	67.0
Characteristics of the weighted similarity [15]	The characteristics of the literature [25]	2715	689	748	79.8	78.4	79.1
Characteristics of the weighted similarity [15]	The proposed characteristics	2790	631	673	81.6	80.6	81.1
SVM [25]	The characteristics of the literature [25]	3071	509	392	85.8	88.7	87.2
SVM [25]	The proposed characteristics	3111	457	352	87.2	89.8	88.5
The proposed method	The characteristics of the literature [25]	3119	338	344	90.2	90.0	90.1
The proposed method	The proposed characteristics	3177	271	286	92.1	91.8	91.9

To ensure high classification accuracy, the method of active learning adopted in the present study can reduce the workload in labeling samples manually. Moreover, the output of RVM can be used to further recognize the multiple matching relations. Therefore, the proposed method displays a higher rate of successful matching and a smaller manual intervention workload than other methods. The results of matching are shown in Figure 17. Figure 17c demonstrates the matching effects of simple objects, while Figure 17b illustrates the matching effects for complicated objects. In this figure, the entity with gray solid line is illustrating large-scale data, the entity with blue gray line is showing the corresponding small-scale data, and the red solid line refers to the matching relations.

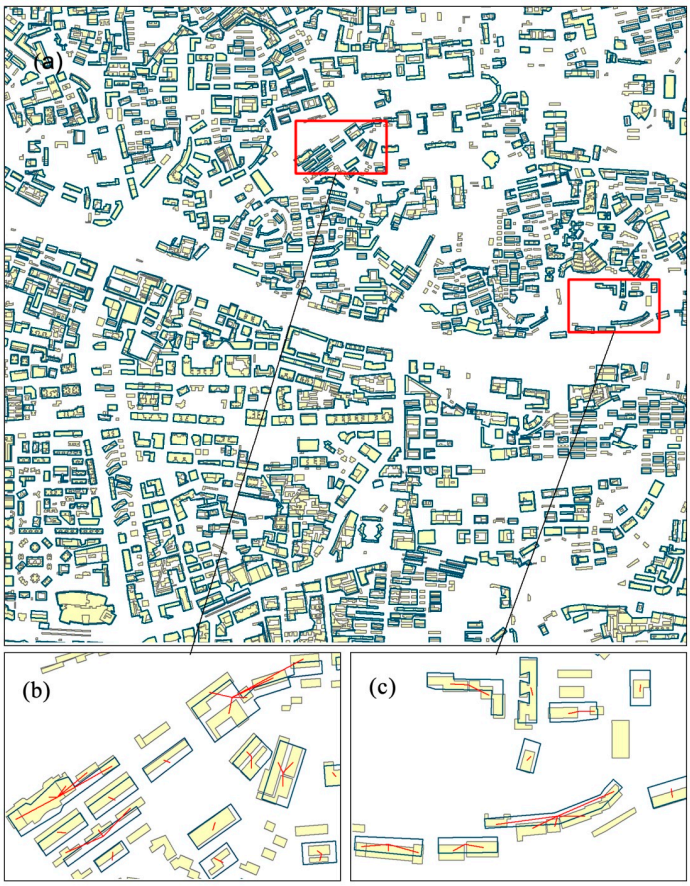


Figure 17. Display of matching effect: (**a**) global display; (**b**) matching effect of complicated objects; (**c**) matching effect of simple objects.

Figure 18 is an example of multiple matching, which shows that with the use of a buffer, No. 509 small-scale residential area (the blue dashed border) can search the three large-scale candidate elements Nos. 1782, 1783, and 1784 (the gray solid borders). If there are three combinations as judged by the classifier, then they are all matched. The method of SVM usually misjudges No. 1784 as one of the matching objects of No. 509. As shown in Table 4, the proposed method in the present study compares the value of classifying output and selects the combinations with the highest reliability, Nos. 1782 and 1783, as the final matching objects, which is consistent with the manual recognition result.

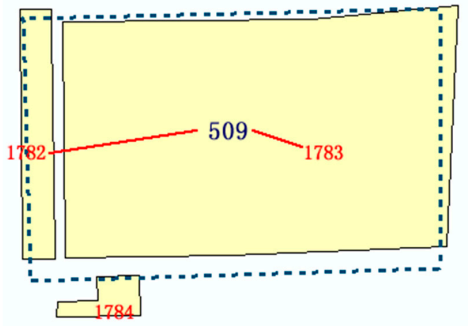


Figure 18. An example of the multiple matching.

Table 4. The values of the matching reliability.

Source ID	Target ID	Results	Output Value
509	1783	matched	0.923
509	1782, 1783	matched	0.987
509	1782, 1783, 1784	matched	0.916

The experimental results from the proposed method are calculated using three matching types: 1:1, 1:*m*, and *m*:*n*. As shown in Figure 19, the matching precision of 1:1 is highest, reaching 95.5%, with a recall rate of 96%, and *F*1 (the harmonic mean of precision and recall) value is 95.7%. The matching of 1:*m* needs several merging operations and the method of merging could bring some complicated shapes such as cavity, which makes it more difficult to measure the similarity. The accuracy of 1:*m* is lower than 1:1 (i.e., precision = 91.9%, recall rate = 91.4% and F1 = 91.6%). The quantity of *m*:*n* type is small and is more complicated. The large number of candidate matching entities might result in partial mistakes in the selection of candidate matching objects, and, at the same time, the measurement may be influenced by complicated shapes. Its matching precision is 82.2%, the recall rate is 83.3%, and the *F*1 value is 82.7%.

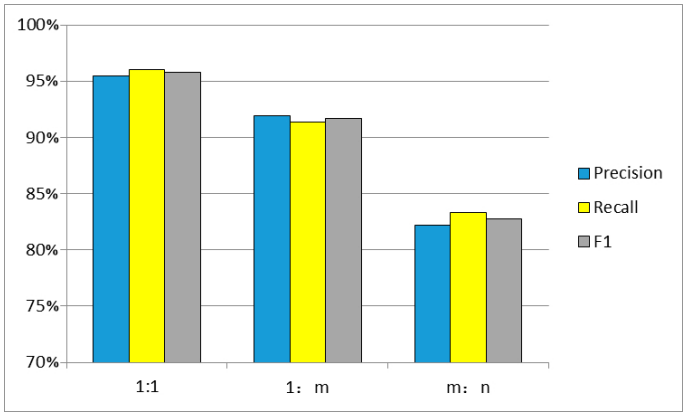


Figure 19. Calculation of accuracy of various matching types with the proposed method.

4. Conclusions

Multi-scale object matching is the key technology used in cascading updates and fusion of multi-scale spatial data. This study presents a type of multi-scale residential areas matching method based on RVM algorithm and active learning. It proposes the rule to merge and not to simplify the method, using Delaunay triangulation, which converts the one-to-many or many-to-many relationships into one-to-one relationships in the matching of residential areas, thereby facilitating the measurement of geometric similarity. According to the characteristics of multi-scale area objects, the five characteristics of position, area, shape, orientation, and surroundings are selected to achieve similarity measurements. Improvement of the orientation similarity measurement using the histogram of area projection is achieved, and a grid-based method for measuring surroundings similarity is designed. The classifying method of the RVM can avoid manual work for determining weights and threshold values. The active learning strategy achieves reasonable classification results with a small number of labeled samples, which can reduce the work of marking samples manually. This work enhances the automation and intellectualization of multi-scale spatial entities matching.

By means of the matching experiment that utilized 1:5000 scale residential areas and 1:25,000 scale residential areas, it is shown that the proposed method has obvious advantages in entity merging, similarity measuring, and matching compared with other methods. The overall precision of matching exceeds 90%, with the accuracy of 1:1 being highest and the other two (1:*m* and *m*:*n*) also having high matching precision. However, this method still needs further improvements: (1) the measurement of shape similarity needs further development to be suitable for area entities that have extremely complex shapes (e.g., an area entity with many cavities); and (2) when the values of *m* and *n* are relatively large in matching relations of 1:*m* and *m*:*n*, they will generate more candidates to be matched, which requires a longer processing time. Therefore, the selection process of matching groups needs further enhancement to increase efficiency.

Acknowledgments: The work described in this article was supported by National Natural Science Foundation of China (grant nos. 41431178 and 41671453) and the Natural Science Foundation of Guangdong Province, China (grant no. 2016A030311016).

Author Contributions: Xinchang Zhang and Guowei Luo developed the framework and wrote the manuscript. Guangjing He and Liyan Chen took part in the experiment.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript

GIS	Geographic Information System
RVM	relevance vector machine
SVM	support vector machine
SMBR	smallest minimum bounding rectangle

References

1. Cooper, A. The Concepts of Incremental Updating and Versioning. In Proceedings of the 21st International Cartographic Conference, Durban, South Africa, 10–16 August 2003.
2. Zhang, X.; Guo, T.; Huang, J.; Xin, Q. Propagating Updates of Residential Areas in Multi-Representation Databases Using Constrained Delaunay Triangulations. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 80. [CrossRef]
3. Haunert, J.H.; Sester, M. Propagating updates between linked datasets of different scales. In Proceedings of the XXII International Cartographic Conference, A Coruña, Spain, 9–16 July 2005.
4. Qi, H.B.; Li, Z.L.; Chen, J. Automated change detection for updating settlements at smaller-scale maps from updated larger-scale maps. *J. Spat. Sci.* **2010**, *55*, 133–146. [CrossRef]
5. Ying, S.; Wen, W.; Wan, Y.; Duan, X. Modelling the spatial evolution of map objects by map agents. *Geocarto Int.* **2016**, *31*, 408–427. [CrossRef]

6. Samal, A.; Seth, S.; Cueto1, K. A feature-based approach to conflation of geospatial sources. *Int. J. Geogr. Inf. Sci.* **2004**, *18*, 459–489. [CrossRef]
7. Vauglin, F.; Ali, A.B.H. Geometric matching of polygonal surfaces in GISs. In Proceedings of the ASPRS Annual Meeting, Tampa, FL, USA, 30 March–3 April 1998.
8. Goesseln, G.; Sester, M. Change Detection and Integration of Topographic Updates from ATKIS to Geoscientific Data Sets. In *Next Generation Geospatial Information*; CRC Press: Boca Raton, FL, USA, 2005; pp. 85–100.
9. Min, D.; Zhilin, L.; Xiaoyong, C. Extended Hausdorff distance for spatial objects in GIS. *Int. J. Geogr. Inf. Sci.* **2007**, *21*, 459–475. [CrossRef]
10. Tong, X.; Liang, D.; Jin, Y. A linear road object matching method for conflation based on optimization and logistic regression. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 824–846. [CrossRef]
11. Ai, T.; Cheng, X.; Liu, P.; Yang, M. A shape analysis and template matching of building features by the Fourier transform method. *Comput. Environ. Urban Syst.* **2013**, *41*, 219–233. [CrossRef]
12. An, X.; Sun, Q.; Xiao, Q.; Yan, W. A shape multilevel Description method and application in measuring geometry similarity of multi-scale spatial data. *Acta Geod. Cartogr. Sin.* **2011**, *40*, 495–501.
13. Fu, Z.; Lu, Y. Establishment of the comprehensive model for similarity of polygon entity by using the bending radius complex fuction. *Acta Geod. Cartogr. Sin.* **2013**, *42*, 145–151.
14. Hastings, J.T. Automated conflation of digital gazetteer data. *Int. J. Geogr. Inf. Sci.* **2008**, *22*, 1109–1127. [CrossRef]
15. Shao, S. *Researehes and Applications on Polygon Entity Matching for Multi-Scale Vector Data Based on Geometric Features*; Wuhan University: Wuhan, China, 2011.
16. Luo, G.; Zhang, X.; Qi, L.; Guo, T. The fast positioning and optimal combination matching method of change vector object. *Acta Geod. Cartogr. Sin.* **2014**, *43*, 1285–1291.
17. Huh, Y.; Yu, K.; Heo, J. Detecting conjugate-point pairs for map alignment between two polygon datasets. *Comput. Environ. Urban Syst.* **2011**, *35*, 250–262. [CrossRef]
18. Kieler, B.; Huang, W.; Haunert, J.H.; Jiang, J. Matching River Datasets of Different Scales. In *Advances in GIScience*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 135–154.
19. Kim, J.O.; Yu, K.; Heo, J.; Lee, W.H. A new method for matching objects in two different geospatial datasets based on the geographic context. *Comput. Geosci.* **2010**, *36*, 1115–1122. [CrossRef]
20. Zhang, X.; Ai, T.; Stoter, J.; Zhao, X. Data matching of building polygons at multiple map scales improved by contextual information and relaxation. *ISPRS J. Photogramm. Remote Sens.* **2014**, *92*, 147–163. [CrossRef]
21. Walter, V.; Fritsch, D. Matching spatial data sets: A statistical approach. *Int. J. Geogr. Inf. Sci.* **1999**, *13*, 445–473. [CrossRef]
22. Tong, X.; Shi, W.; Deng, S. A probability-based multi-measure feature matching method in map conflation. *Int. J. Remote Sens.* **2009**, *30*, 5453–5472. [CrossRef]
23. Safra, E.; Kanza, Y.; Sagiv, Y.; Beeri, C.; Doytsher, Y. Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets. *Int. J. Geogr. Inf. Sci.* **2010**, *24*, 69–106. [CrossRef]
24. Liu, P.; Zhang, Y.; Gong, J. Root mean square error and neighbouring relation matching approach for multi-scale areal feature. *Acta Geod. Cartogr. Sin.* **2014**, *43*, 419–425.
25. Zhang, X.; Zhao, X.; Molenaar, M.; Stoter, J.; Kraak, M.J.; Ai, T. Pattern classification approaches to matching building polygons at multiple scales. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*; XXII ISPRS Congress; International Society for Photogrammetry and Remote Sensing: Tempe, AZ, USA, 2012; Volume I-2, pp. 19–24.
26. Wang, Y.; Chen, D.; Zhao, Z.; Ren, F.; Du, Q. A Back-Propagation Neural Network-Based Approach for Multi–Represented Feature Matching in Update Propagation. *Trans. GIS* **2015**, *19*, 964–993. [CrossRef]
27. Fisher, J. Visualizing the connection among convex hull, Voronoi diagram and Delaunay triangulation. In Proceedings of the 37th Midwest Instruction and Computing Symposium, Minneapolis, MN, USA, 16–17 April 2004.
28. Li, W.; Goodchild, M.F.; Church, R. An efficient measure of compactness for two-dimensional shapes and its application in regionalization problems. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 1227–1250. [CrossRef]
29. Peter, B.; Weibel, R. Using vector and raster-based techniques in categorical map generalization. In Proceedings of the Third ICA Workshop on Progress in Automated Map Generalization, Ottawa, ON, Canada, 12–14 August 1999.

30. Zhang, X.; Xiao, P.; Song, X.; She, J. Boundary-constrained multi-scale segmentation method for remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2013**, *78*, 15–25. [CrossRef]
31. Revell, P.; Antoine, B. Automated matching of building features of differing levels of detail: A case study. In Proceedings of the 24th International Cartographic Conference, Santiago, Chile, 15–19 November 2009.
32. Tipping, M.E. Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* **2001**, *1*, 211–244.
33. Wu, Y.; Kozintsev, I.; Bouguet, J.Y.; Dulong, C. Sampling strategies for active learning in personal photo retrieval. In Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, Toronto, ON, Canada, 9–12 July 2006; pp. 529–532.