Pattern Based Feature Matching for Geospatial Data Conflation

Weiping Yang, Dan Lee, Nobbir Ahmed
Geoprocessing
Esri, Inc.
Redlands, USA
wyang@esri.com; dlee@esri.com; nahmed@esri.com

***Abstract*—This paper describes a linear feature matching method for overlapping data sources aimed at developing geospatial conflation tools. This method is based on identifying distinguished feature patterns, which are categorized as atomic patterns and composite patterns. It is hoped that these feature level patterns, once identified and modeled, will serve as a signature and as a helpful vehicle to enrich semantic meanings of a dataset. Features from overlapping sources will then be matched, by first matching large or complex structures and then breaking down to individual features of varying cardinalities. This feature matching method has been implemented as a core component for geoprocessing conflation tools to perform spatial adjustment, attribute transfer, and feature change detection. Applying these tools, in workflows, to real world data has produced promising results.**

Keywords-geospatial featture patterns; feature matching; geospatial data conflation; geoprocessing.

I. Introduction

Geospatial data conflation is one of the classic requirements in data integration where attributes from one data source need to be transferred, or geometries adjusted, based on feature correspondences of another data source. With the vastly increasing distributed geospatial data and sharing of these data over the web, high demands on conflation solutions have been seen in recent years. The concept of conflation has also been extended to detect feature changes between updated and existing datasets where new features emerged and previously existing features disappeared.

Due to differences in data capturing purposes, methods, scales, accuracy, or collecting time, corresponding geospatial features have discrepancies which may be spatial and non-spatial. Spatial discrepancies can be topological, geometrical, and metrical. A road feature matching two roads broken at a new T-intersection, for example, is a topological change; a circular cul-de-sac corresponding to a rectangular loop reveals a geometrical difference; and a short extension of a road from an intersection finding its peer prolonged shifted and slightly rotated causes altered metric measures. A non-spatial discrepancy occurs when the road name “Main St.” in one dataset is meant to be “Main Street” in another. Hence, finding relationships between corresponding features must be equipped with capabilities of distinguishing and handling these discrepancies. As a general consensus, the automation of conflating geospatial data requires a multitude of processes and possibly iterations with these processes, among which correctly matching features is central to the success of the automation.

Section II of the paper overviews some published feature matching methods, and the one we have developed for the purpose of creating a suite of conflation tools. It is followed by a discussion in Section III on identifying distinct structural shapes which are the basis of matching features. The matching process, starting from matching structures is presented in Section IV. The result of structural matching is then broken down to match individual features, which is discussed in Section V. The testing of the feature matching method in building conflation tools and the practical uses of the tools in workflow contexts are discussed in Section VI. The paper will end with conclusions and a brief discussion on future work.

II. Overview of Feature Matching Methods

Given two overlapping sets of geospatial features named A and B, the problem of feature matching can be phrased as: for each feature in A, find its most likely counterpart in B if it exists. The catch here is the modifier “most likely” since it implies the application of cognitive knowledge and a conclusion can be elusive in complex configurations. Finding a matching method supported by objective measures has been a challenge to academicians and practicing professionals working on Geographic Information Systems (GIS) over the past 30 years.

Ever since an automated, interactive feature matching system via iterative rubber-sheeting was developed by the U.S. Census Bureau [1], researchers have been looking to use similarity measures against features to enhance matches between datasets. Various definitions of similarities and their measuring methods have been discussed, from considering geometric properties such as orientation, shape, length, etc. to combining themes and semantic attributes [2], [3]. At the operational level, analytical [2], [4], statistical [5], and linear programming models have been proposed [6], [7].

In what follows, a method of matching overlapping linear features is first outlined, leaving elaboration of details in subsequent sections. This method is in line with some of the published approaches [2], [4] that find similarities between corresponding features of different sources, based on topological, geometrical and metric properties of features. The main differences of the proposed method are that it emphasizes on discovering shape patterns from a single

ISBN: 978-1-61208-326-1

dataset first and then cross-referring these patterns to probe and to assemble correspondent patterns in another. The method regards geometric patterns built progressively from low level shapes as objects to compare for similarities.

The concept of identifying a hierarchy of pattern structures in a geospatial dataset has been inspired by a discussion on modeling patterns and structures in maps [8]. In cities and towns, design patterns are generally observed in urban and regional planning [9]. Researchers in GIScience have been detecting and describing street network patterns in geospatial databases out of a mass of seemingly chaotic data collections [10], [11]. It is believed that such a pattern system would reveal “meanings” of geospatial data for intelligent queries and applications.

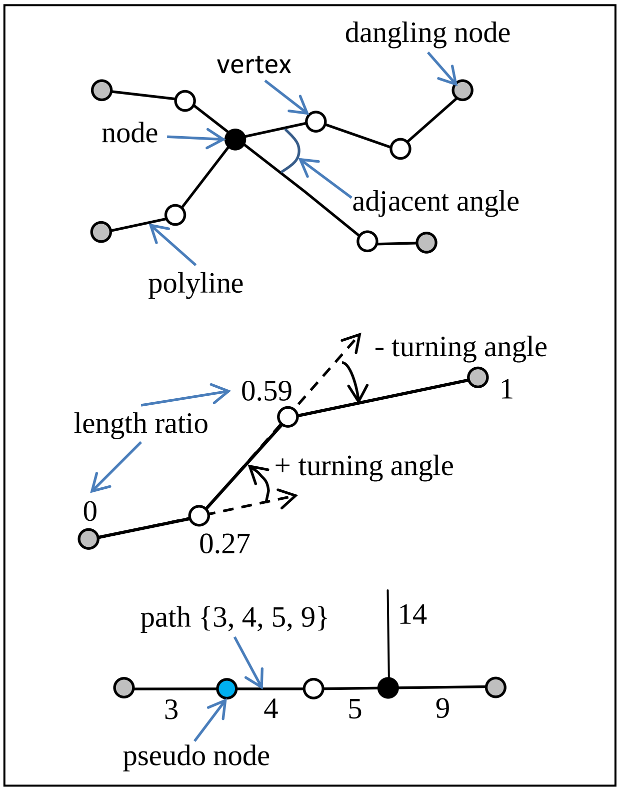

Figure 1. Illustration of terms.

Figure 1 illustrates the terms that are useful to the methods described in this paper. Linear features include roads, rivers, utility lines, or land use boundaries, etc. These features can be represented with polylines extended by a series of vertices in 2D space. A network of nodes and paths are formed from all input polylines. Informally, nodes are located at the two extreme vertices of a polyline. A node can be a connection of one or more polylines when at least one of their extremes coincide or are intersected at the same location. A dangling node does not make a connection to any other polylines while a pseudo node intersects exactly two polylines. A path is formed by one or more polylines connected at or incident to nodes and constrained by metric properties. In the method described in this paper, a path can span a number of nodes some of which may have more than two incident paths. Properties such as orientation, length, normalized length ratios of accumulative segments to the whole length, and turning angles are used to characterize a path. The number of adjacent paths and the adjacent angles formed by two adjacent paths are associated with a node. For clarity, the series of feature IDs composing a path is represented in a curled bracket as ordered list, as shown in Figure 1.

Individual polylines can be short and long, straight, curved, or otherwise forming various shapes. The matching method discussed here uses a combination of top-down and bottom-up approach. The *bottom-up process* constructs progressively larger structures by connecting polylines that are constrained to form descriptive shapes or patterns in one dataset. By obtaining well behaved large patterns such as highways or river networks, global information about the data can be grasped. Models and analysis on top of patterns could be developed to derive semantics about the data.

The *top-down process* takes each of the pattern structures to locate and to match similar patterns composed of a single or a plural set of constructive patterns from the other dataset. The matches of a few well distributed, distinct, and robust patterns could shed light on how the two datasets are shifted and rotated. The information could be further used for constraining proximity searches for matching of smaller or weakly-determined shapes. On the other hand, if the number of mismatches among the pattern structures is high, the matching process could report a strong dissimilarity between the two datasets. After the process of matching structures is completed, matching individual features will be followed by breaking down a pair of matched structures into corresponding feature pairs. In the break-down process, a method of gauging “affiliation” between features using metric and topological properties will be applied.

III. IDENTIFYING STRUCTURES IN A DATASET

Two kinds of structures can be identified in a dataset alone without referring to an overlapping or context layer. The first kind, termed atomic patterns, is formed by a single polyline geometry of a feature. The second, termed composite patterns, consists of a series of polyline geometries of two or more features which themselves may be of atomic patterns. Once identified, all these patterns and supporting polylines can be organized with a dynamic hierarchical structure for easy manipulation.

A. Atomic Patterns

Atomic patterns are captured when shapes of a linear feature class are read in and cached. No searches are involved in this stage. Depending on the nature and purpose of the feature class, there are unlimited ways that a linear feature can be shaped. It is impossible to design all stereotypes to fit all features in an ordinary feature class. Nevertheless, certain shapes stand out to observers’ eyes; others don’t. It is possible to devise well-structured scalable stereotypes to filter and describe them using a few parameters. The research undertaken at Environmental Systems Research Institute (Esri) has focused on developing an increasing set of filters to catch Circular arcs, L-, Spoon-, Door-, U-, Sine-, Z-, Stairs-, Straddle-, and Straight-shapes, etc. All other polyline features not caught by any of these stereotypes fall into the Unknown-shapes. The figures in Figure 2 illustrate the stereotypes that are presently modeled.

The analytical model of the stereotype for an L-shape, for example, is composed of two nearly straight sections connected by a sharp turning point. For Circular arcs, all segments should have their lengths close to an average length and all consecutive turning angles have a same sign and a near average magnitude. A Z-shape stereotype features two consecutive near-90^o turning angles in opposite signs, and all polyline sections are near straight. If consecutive near-90^o turning angles with alternating signs reach a count larger than 4, a Stairs-shape is formed.

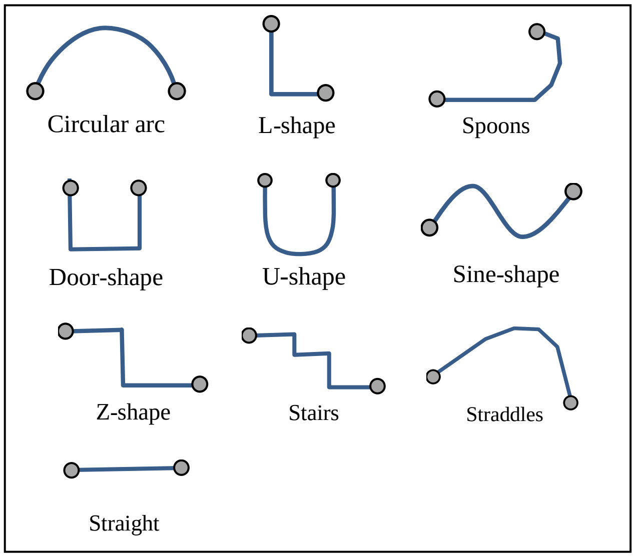

Figure 2. Stereotypes filtering atomic patterns.

Note that real world data of seemingly stereotypic shapes may not all demonstrate expected regularity, stereotype filters need to be able to detect and handle insignificant turbulences in a series of vertices. Techniques such as polyline generalization and statistical deviations can be employed to screen and remove abnormal vertices. Maintaining a robust set of low level pattern filters to hand real data is key to the pattern based feature match method, as the filters will be repeatedly employed in various stages of processing, including to detect non-atomic shapes.

The atomic patterns illustrated in Figure 2 are common shapes that can be observed in most linear geospatial datasets representing urban structures. The set of atomic patterns reflects the maturity of the system in catching geometric shapes using analytic models. New and more complicated patterns would be added as the system evolves.

B. Composite Patterns

Composite patterns can be formed by a number of connected polylines, when together they present some distinct figures. In addition to the shapes shown in Figure 2, Circles, Carriageways, Cul-de-sacs, and Straight can be assembled from atomic patterns, as shown in Figure 3.

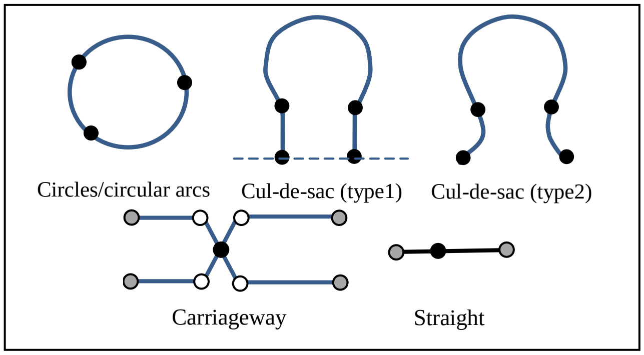

Figure 3. Composite patterns.

A composite Circular arc, for example, can be traced out by looking for consecutive atomic Circular arcs that have a similar curvature and radius. A composite Circle is formed when a Circular arc is closed, else the tracing stops when a currently probed shape does not have similar parameters. A Cul-de-sac can be formed by tracing from both extremes of a Circular arc for a pair of near parallel lines (type1) or a pair of curvature reversed arcs (type2). A Carriageway pattern, generally, involves 4 polylines and is characterized by cross sections joined at a common intersection from which two pairs of near parallel polyline sections are split. Presently, only the above 5 composite structures are assembled in a dataset without referencing structures in a counterpart dataset. Other composite structures will be further formed by reference during the matching process, to be discussed in matching structures.

Forming a composite pattern involves searches at the extending extreme nodes of the current path. It also requires decisions whether a testing atomic shape could be accepted and added into the path. Node topology, path continuity, and pattern compatibility are the factors in the decisions.

C. Pattern Graph

A pattern-path-node graph structure is created to collect and operate on the identified patterns, their associative paths, and their geometries, as shown in Figure 4.

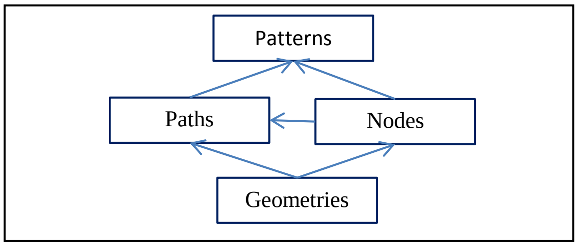

Figure 4. Pattern-path-node graph.

The Geometries are polylines imported from ArcGIS® feature classes, identified by Object ID. Excessive vertices, either duplicated or with distances smaller than a resolution tolerance, are removed through a simplification process. Multipart polylines are consecutively connected as a single shape. Associated with each of the Geometries is a structure holding computed properties to describe the characteristics of a single polyline shape, including the ID of an atomic pattern type. The Nodes are identified as extreme vertices of geometry objects. Associated with each node are IDs of incident Geometries. A path in Paths contains a series of polylines identified by feature IDs that are connected to form a composite structure. A similar data structure to atomic patterns is used to hold characteristic properties of a path. The Patterns is a collection of Paths and associated Nodes that together describe a distinctive structure within a dataset. A pattern object contains a list of main paths that are central to the structure and lists of subordinate paths and nodes that are related to the main paths.

IV. Matching structures

Prior to matching, an inventory of discovered patterns in each graph is established. Furthermore, an indexing of linear features from both graphs is created to facilitate proximity searches. The inventory is used for matching patterns which are listed by pattern types and spatially registered in a coarse grid. The indexing structure is detailed to line segments for refined searches after searches in inventory are firstly attempted. It is convenient to combine segments from both graphs into one indexing structure to pick best matches where multiple candidates are available.

An order of importance will be determined for structure matching, which dictates which pattern type is to be specifically matched first. Our experience reveals that it should proceed from the most complicated to relatively simple structures. This is because simple structures may be part of a larger one. Straight lines will be matched last, just before matching Unknown-shape structures.

For clarity, the first and second datasets involved in matching are named source and target, respectively. The matching process takes a pattern type from source and searches for a counterpart in the target inventory with the closest characteristic values for each of its members. If such a counterpart is found, a match is made. Otherwise, the process will perform proximity searches to identify piecewise shapes from target to fit the larger shape referenced in source. A successful fit will add the composed large structure and modify hierarchical relationships in the target graph. After all known structures in source are exhausted, the process should be repeated for a list of unmatched known shapes from target to match a counterpart from source. It is necessary to repeat matches initiated from either graph, as a large atomic structure existing in one graph may not exist in the other. The reverse process ensures that large known patterns be processed first prior to matching unknown shapes. At the end, structures of both graphs will either find a match or be declared not matched.

Matched pairs may not be exactly of same patterns, but they must be compatible. For example, L- and Spoon-shape patterns are compatible, so are Door- and U-patterns. A straddle may be matched with one of the following compatible shape combinations: a straddle; two spoons; one spoon and one straight; one circular and two straight shapes.

The diagram in Figure 5 illustrates the matching of an atomic straddle shape from the red graph to a number of piecewise polylines in the black graph. The first attempt of finding an atomic straddle counterpart from the black graph is failed. The characteristic sections of the red straddle will be identified, which are a circular arc in the middle part and two near straight sections. Proximity searches then start from the circular section. If a similar circular arc or a spoon shape is returned from the black graph, the rest of the parts will be traced from it. In the example, the search will first find feature 17, which has the closest curvature parameters to the middle section of the red feature 3301. Straight features 12 and 18 are then traced out from 17. Pieced together, the composite straddle in black has properties most similar to that of 3301 in red. A match is done with the red atomic and the black composite straddles.

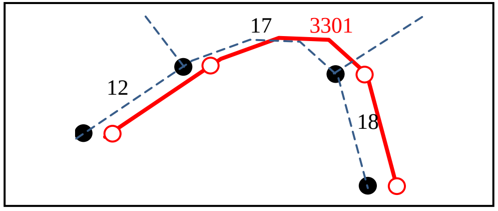

Figure 5. Matching a straddle.

While composing structures during a match, gaps can sometimes exist in one of the graphs, as illustrated in Figure 6. In the example, black circular arcs {1587, 1606, 1608, 1592} form a composite circular path prior to matching. When taking the path to match a counterpart in red graph, two circular paths, {1833} and {1857, 1840} are found, which have similar circular parameters to that of the black. The two red paths together will form a matched composite path, with a gap in between.

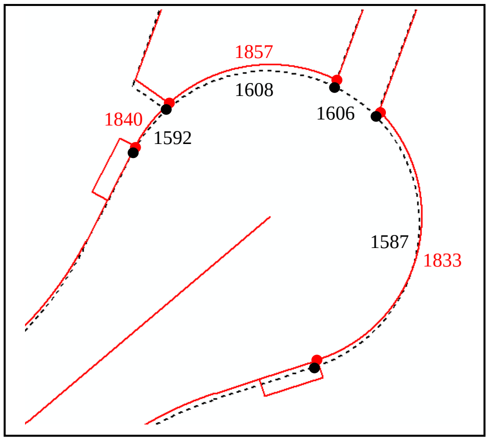

Figure 6. Matching a circular path with gap.

There is also a need to split a previously generated path. This is especially true for large straight paths. Figure 7 illustrates such a case. In the diagram, the black graph shows a straight path composed of shapes {4, 5, 6} and two other straight paths {12} and {9}. The red graph has one red L-shape {64}, and two straight paths {33} and {48}. During the matching process, the red L-shape is matched with {12} and possibly the longer straight path in black. It is necessary to break the long straight into two short ones {4} and {5, 6}.

After breaking, the L-shape is matched with {4, 12} and {5, 6} is added into the black graph structure for further match.

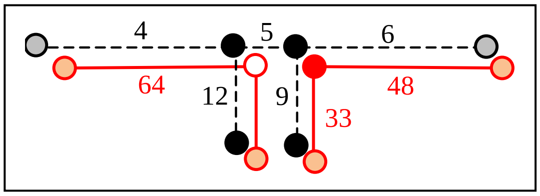

Figure 7. Ilustration of splitting a path.

To match unknown shapes, more comprehensive proximity computations, like buffering, will be needed, since little is known about their characteristics. For each unknown shape, a search for a set of candidates can be constrained by a given or derived search distance, whichever is smaller. The derived search distance can be computed from the result of uniquely matching known structures.

V. MATCHING FEATURES

A final step with the matched structures is to break them down to feature by feature match. Due to changes in real world and differences in data capturing, features from both graphs will not all have one-to-one correspondences. The following cardinality relationships exist in linear features, as shown in Figure 8. The m:n relationship occurs when it would be ambiguous to break the set of features further down to simpler correspondences. The 1:0 and 0:1 relationships are included to indicate no corresponding features could be matched to satisfy similarity measures.

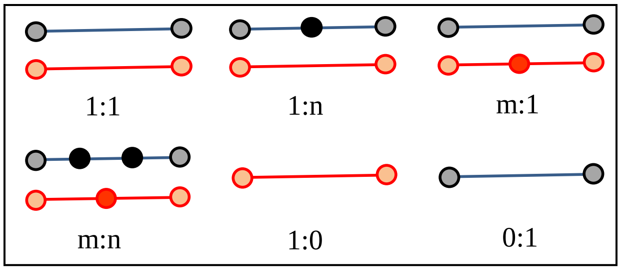

Figure 8. Cardinalities between matched features.

In addition to the length ratio parameter associated with paths, two constraints are considered while matching features from matched structures. The first is topological when other paths incident to nodes and separated by adjacent angles will be analyzed and compared. If there is insufficient topological information, the measure of "orthogonally projected overlapping" between elements of two paths will be applied.

Figure 9 illustrates the process that respects topological measures. Two paths of, source (red) {6, 7, 8} and target (black) {4, 5, 6, 7}, are matched. Features from source will be taken, one at a time, to match features in target. In the diagram, red feature identified as 6 is first matched with black 4. Two determinations will be applied to append black 5 to the match list after 4. First, the length ratio of black 4 is still smaller than that of red 6; second, the extreme node of black 4 is a pseudo node. Adding black 5 to 4 satisfies the length ratio better, furthermore, its front end fits better topologically to the front end of red 6. Other features are matched by applying similar reasoning process using local neighborhood, and considering shape characteristics of incident paths, if necessary.

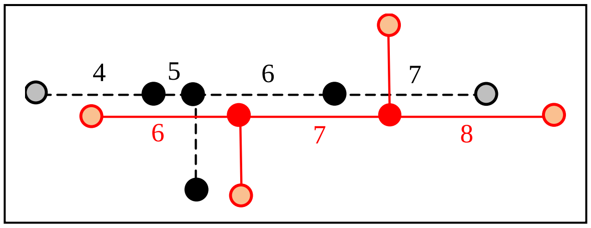

Figure 9. Example of feature matching process.

An orthogonally projected overlapping length is obtained by projecting the extremes of polyline A onto polyline B. If a projection is footed on the extension from an extreme of B, the extreme vertex will be projected back onto polyline A. The projected overlapping length is then calculated with the two points enclosing the orthogonally overlapped section. Figure 10 shows five cases that overlapping lengths are enclosed.

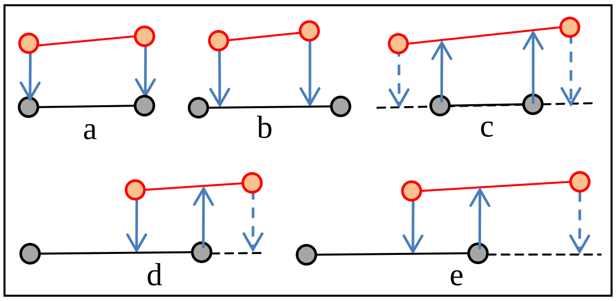

Figure 10. An example of obtaining orthogonally projected length.

The method described in this paper requires that for a pair of features to be considered a match, the projected overlapping length must not be less than half length of the shorter polyline. Case e in Figure 10, for example, does not satisfy the requirement. Using this measure, gaps in the path could correspond to a feature with no match. An example of this case is shown in Figure 6 where the black feature 1606 does not have a sufficient projected overlap with either red 1833 or 1857. It will have a 1:0 match.

VI. APPLICATION AND RESULTS

The method presented in this paper has been designed as a core component to support three geoprocessing tools for ArcGIS, the commercial GIS software produced and marketed by Esri Inc. They are Detect Feature Changes, Generate Rubbersheet Links, and Transfer Attributes, all of which rely on matching features of two datasets from separate sources covering the same geographic areas.

One of the challenges in developing feature matching techniques, and application tools based on them, is to find effective ways to assess results produced by these tools by feeding user data of various complexities. The evaluation is necessary for users to gain confidence on the levels of accuracy that could be expected from using these tools. Another challenge is to find practical workflows to complete conflation tasks with high levels of automation.

Applying the above mentioned conflation tools to real world data, Lee, Yang, and Ahmed [12] have devised workflows to perform data integration tasks such as spatial adjustment, transferring attributes, and generating reports on feature changes by detecting spatial and attribute discrepancies. Furthermore, a set of script tools, written with Python or built by chaining other tools in ArcGIS, have been developed to automatically verify and evaluate outputs produced by the conflation tools. Their assessment shows an achievement of above 90% of feature matching and conflation accuracy in executing the workflows on top of multiple user datasets demonstrating excellent, ordinary, and poor similarities. The successful rates are compatible to those claimed in published papers [4], [6]. Due to unavailability of software developed based on the other published feature matching methods, a cross-comparison under similar conditions on performance, ease of use, and robustness, etc., cannot be reported in this paper.

VII. CONCLUSION AND FUTURE WORK

A method of matching linear features overlapping the same geographic area has been described. By catching atomic and composite feature patterns to construct a pattern graph, better understanding of a dataset is obtained. Patterns of a dataset are recognized through a set of stereotypes as low level constructs, which can be applied to compose large pattern structures within a dataset and with reference to the other dataset during the matching process. Based on the graphs built on source and target data, matching features starts with matching structures, in which locating of paired structures becomes less dependent on coordinates, rotation, and shift, but more on referencing local neighborhood structures. Processes of matching individual features are explained. Consideration factors, determining how matched structures are broken down into matched features, are also elaborated. The method is implemented as a core component which is used for producing conflation geoprocessing tools in ArcGIS. Testing and application of the tools in practical workflows have demonstrated promising results.

While the research reported in this paper has established a framework in developing feature matching based tools, more work is needed to complete the missing parts of the methodology. First, a full analysis on the algorithms in accomplishing matching features is necessary in terms of time and space complexity, from which comparisons to other methods could be made. Evidence of practical uses, and results from applying the conflation tools and workflows in solving real world problems, should be included as an integral part of the method. Meanwhile, it is anticipated that the method and its applications will continue to evolve on the following fronts:

- Maintaining existing pattern recognition stereotypes so that they become more versatile and robust;
- Developing new patterns to reduce the number of unknown-shape elements in graphs;
- Developing reasoning on top of identified geometric patterns to enrich semantic meanings, hence the metadata of a dataset;
- Considering matching patterns coming from datasets with varying generalization scales; and
- Researching on geospatial matching and conflation between vector datasets and raster images.

ACKNOWLEDGMENT

Many members in the Esri Geoprocessing team have contributed to the conflation project through discussions.

Some diagrams contain sections of LA parcel data from the Department of Public Works, Los Angeles County, USA.

REFERENCES

[1] M.P. Lynch and A. Saalfeld, “Conflation: Automated Map Compilation - A Video Game Approach,” AUTOCARTO 7 Proceedings, 1985, pp. 343- 352.

[2] A. Samal, S. Seth, and K. Cueto, “A feature-based approach to conflation of geospatial sources,” International Journal of Geographical Information Science, 18:5, 2004, pp. 459-489.

[3] A. Schwering, “Approaches to Semantic Similarity Measurement for Geo-Spatial Data: A Survey,” Transactions in GIS, 2008, 12(1), pp. 5–29.

[4] M. Zhang and L. Meng, “Delimited stroke oriented algorithm-working principle and implementation for the matching of road network,” Geographic Information Sciences, Hong Kong, June 2008, 14(1), pp. 44-53.

[5] V. Walter and D. Fritsch, “Matching spatial data sets: a statistical approach,” International Journal of Geographical Information Science, 13:5, 1999, pp. 445-473.

[6] L. Li and M. F. Goodchild, “An optimisation model for linear feature matching in geographical data conflation,” International Journal of Image and Data Fusion, vol. 2:4, 2001, pp. 309-328.

[7] G. Touya, A. Coupé, J. L. Jollec, O. Dorie, and F. Fuchs, “Conflation Optimized by Least Squares to Maintain Geographic Shapes,” ISPRS Int. J. Geo-Inf. 2013, 2, pp. 621-644, doi:10.3390/ijgi2030621.

[8] W. Mackaness and G. Edwards, “The importance of modelling pattern and structure in automated map generalisation,” Joint ISPRS/ICA Workshop: Multi-Scale Representation of Spatial Data. Ottawa, Canada, July 7-8, 2002, pp. 1-12.

[9] S. Marshall, “Streets and patterns,” 1st ed., London and New York: Spon Press, 2005, ISBN 0-415-31750-9.

[10] F. Heinzle, K.-H. Anders, and M. Sester, “Automatic detection of patterns in road networks – methods and evaluations,” 2007, Proceedings of Joint Workshop Visualization and Exploration of Geospatial Data, Stuttgart, vol. XXXVI-4/W45 (CD-ROM).

[11] B. Jiang and X. Liu, “Automatic generation of the axial lines of urban environments to capture what we perceive,” International Journal of Geographical Information Science, 2010, 24(4), pp. 545–558.

[12] D. Lee, W. Yang, and N. Ahmed, “Conflation in geoprocessing framework – case studies,” submitted to GEOProcessing, 2014, Barcelona, Spain.