DEVELOPMENT OF A BLUNDER DETECTION APPROACH FOR AUTOMATED POINT MATCHING DURING VECTOR TO IMAGE DATA INTEGRATION

Lawrence V. Stanislawski, Sr. GIS Developer, Science Applications International Corporation
Michael P. Finn, Scientist, U.S. Geological Survey
E. Lynn Usery, Chief Scientist, U.S. Geological Survey
Paul M. Robinette, Computer Engineer, U.S. Geological Survey
Center of Excellence for Geospatial Information Science
1400 Independence Road
Rolla, MO 65401
lstan@usgs.gov
mfinn@usgs.gov
usery@usgs.gov
probinette@usgs.gov

ABSTRACT

Recent (2007) developments in automated integration of vector geospatial data with image data combine techniques for extracting point features from image data with algorithms that match point features between data layers. The precision and accuracy of an image-extracted point feature depend on various factors related to the feature extraction technique and to the quality of the image in the area of extraction. Furthermore, the precision and accuracy of image-extracted points affect the reliability of the subsequent process for matching points between layers, and consequently, the overall adequacy of the data integration approach. Several approaches for detecting and removing improperly matched points are available. The USGS is investigating the use of a weighted affine transformation to filter point matches during automated integration of vector roads with images. The transformation is applied to a local area of match points to detect probable blunders and remove them from the rubber-sheeting algorithm. Aside from blunder detection capabilities, advantages of this approach include the ability to weight control coordinates relative to estimated precisions of extracted point features, and the ability to estimate the precision of the integrated vector layer through error propagation.

INTRODUCTION

During recent years, the U.S. Geological Survey (USGS) has been remodeling the way it provides geospatial data to the nation through *The National Map* program (USGS, 2006). The vision of *The National Map* is to ensure that “current, complete, consistent, and accurate” geographic base information is readily available through a system of web-based interfaces (USGS, 2006). To the extent possible, geospatial data will be aligned to its true geographic position, thereby eliminating any cartographic offsets inherent with some source products. *The National Map* data will be derived from various sources by a consortium of data stewards. As explained by Usery and others (2005), one primary complexity of this vision is the “integration of the various resolutions and accuracies of data in both horizontal and vertical directions,” which is “one of massive proportion” when considering national implementation.

In 2004, USGS scientists in Rolla, Missouri, began research to address the data integration issue for *The National Map*. Thus far, the research agenda has included empirical exploratory analysis to qualitatively and quantitatively assess the extent of the data integration problem, development of a hypothesis for data integration based on resolution and accuracy, and development and testing of automated systems and algorithms to shift features from one dataset into alignment with another to achieve integration (Finn and others, 2004; Usery and others, 2005).

Through a cooperative agreement with the University of Southern California (USC), a system for integrating vector roads with orthoimage data that was developed at USC (Chen and others, 2003a, 2003b) has been emulated and evaluated by USGS researchers (Usery and others, 2005). The system applies automated algorithms that identify road intersection points on orthoimages that match intersection points on a vector road layer. Matching points are filtered to eliminate undesired pairs. Subsequently, a piecewise linear transformation (Saalfeld, 1985; White and Griffin, 1985; Saalfeld, 1993) is applied to fit the vector roads to image data.

Currently (2007), the best approach identified by the USGS for filtering undesired matching pairs applies a vector median filter that eliminates 50 percent of the matching pairs (Usery and others, 2005). In this paper, we hope to take advantage of earlier work (Krarup and others, 1980; Stanislawski and others, 1996; Alsabti and others, 1998; Stanislawski, 2000) and implement a more rigorous statistical filtering technique by identifying possible outlying observations, or blunders, within localized subsets of the project area. We propose and test the use of the k-means clustering technique to define localized subsets of matching point pairs. The objective is to assess an alternative filtering strategy that implements a series of weighted transformations to identify and remove suspected blunders from localized areas.

METHODS

Test Data

In this preliminary analysis, test data were limited to the same sources used for the original study (Usery and others, 2005). Small subsets (about 1 kilometer x 1 kilometer) of image data and associated vector roads were used to develop and test the new filtering technique. The subsets were in urban areas around St. Louis, Missouri. The image source is color orthophotography with approximate 0.33-meter (1-foot) resolution, which was collected for the 133 priority cities of the Homeland Security Infrastructure Program (Vernon, 2004). Vector road data overlaying the images was extracted from the Missouri Department of Transportation (MODOT) layer, which provides one of the most accurate sources for this area.

A sample of the image and vector road layers is shown in figure 1. Notice that in most areas the vector roads are within about one or two road widths of the image roads. From visual inspection, it appears nearly all vector road intersections are within about 50 meters of corresponding image intersections, and the majority are within about 10 meters. A more thorough quantitative summary of preprocessed vector-to-image discrepancies is provided later.

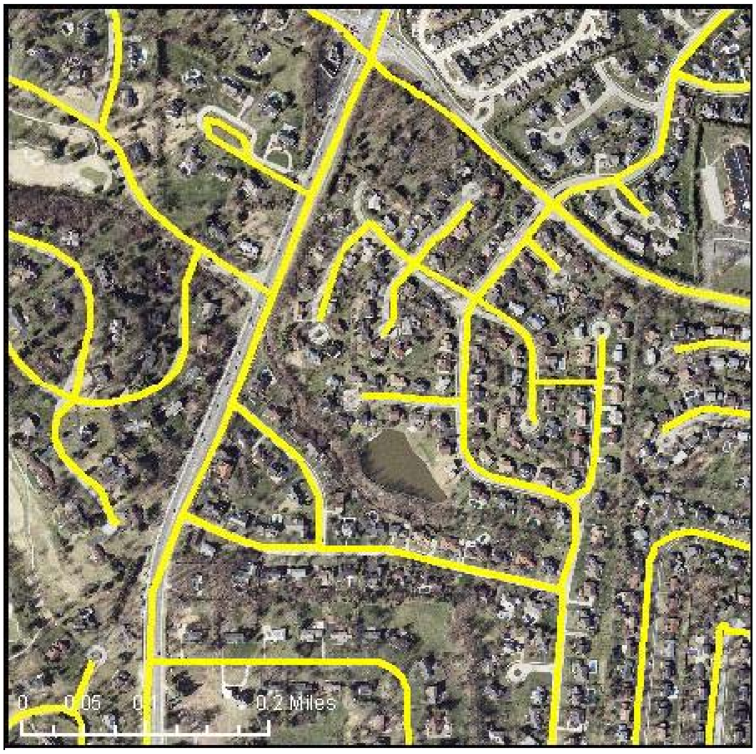

Figure 1. MODOT transportation (yellow) overlaid on an orthographic image.

Processing Steps

In this research, the process to integrate a section of a vector road layer to an associated image is not intended to register one layer's coordinate system to the other. Input layers should already be registered to the same geographic coordinate system and geometrically integrated to a certain extent. This work quantifies the required level of prior integration between the two datasets and the level of improved geometric alignment that is achieved by this process. The steps involved in the process to integrate a section of a vector road layer to an associated orthoimage are as follow:

1. Classify image pixels as road or non-road based on hue, saturation, and value.
2. Locate road intersection points in the vector data.
3. Buffer vector intersection points and create a vector image template (VIT) of road segments within each point buffer that reflects road widths as defined by the vector feature type or associated attributes (figure 2).
4. For each VIT, find the image pattern within a local area of the classified image that best matches the VIT. If a match is found, store the matching image pattern centroid coordinates and the associated vector intersection coordinates.
5. Using the k-means approach, cluster image match points into groups that are larger than 30 and less than or equal to 40.
6. Check if any blunders are detected in each cluster by using the weighted affine transformation to fit the vector intersection points to the matching image points.
7. If any blunders are detected, remove them from the set of match points and repeat steps 5 and 6 until no blunders are detected in any cluster.
8. Perform rubber-sheeting transformation to correct vector roads (Saalfeld, 1985).

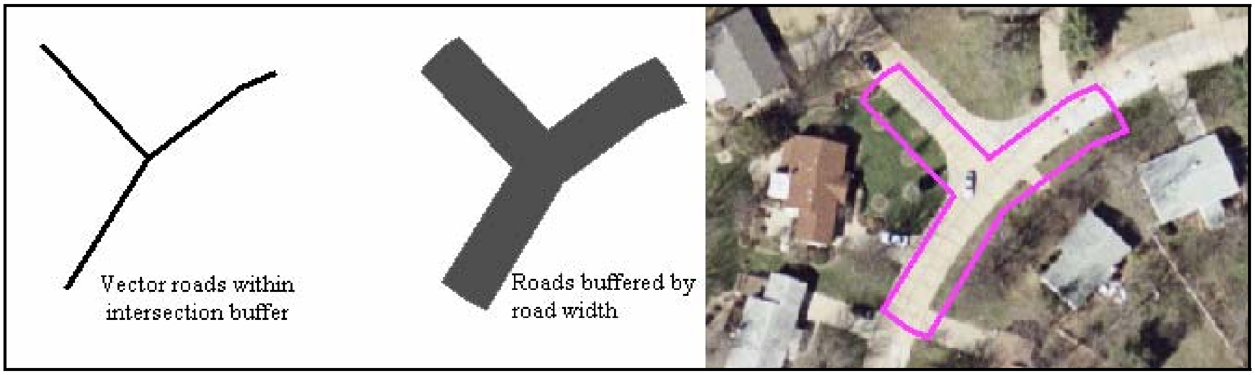
Vector roads within intersection buffer
Roads buffered by road width

Figure 2. Sample vector image template (VIT) generation showing section of roads within intersection buffer, road segments buffered by road width, and template in original vector position overlaid on orthoimage.

All the steps, except step six, have been automated through C programs. The weighted affine transformation is coded in Pascal. ERDAS Imagine software was used to generate training sets for the road and non-road image classes, which are entered into the Bayesian classifier of step one. Pattern matching in step four is completed through a raster-to-raster correlation computation by moving a window of a user-specified size over a user-specified distance. Window size and search distance depend on image characteristics and how well the datasets are integrated before processing. Limitations for these values are yet to be determined. However, we have had good match results for our test data using a 50 meter by 50 meter window within a 70 meter by 70 meter search area around the vector point. For additional details see Usery and others (2005).

K-means

Clustering is any automated process of classifying data into groups, and it is a common technique in exploratory data analysis (Jain and others, 1999). The k-means clustering approach has been effectively implemented in various practical applications (Alsabti and others, 1998). In our integration process, we would like to detect and eliminate any matching point pair that does not conform to an affine fit for a local area; therefore, the k-means clustering algorithm is used to identify localized groups of matching point pairs that are subsequently tested for blunders through the weighted affine transformation. Earlier research suggests that the weighted affine model requires at least 12 points to detect blunder observations (Stanislawski, 2000). Thus, we initially required clusters between 30 and 40 points, with multiple use of some points in adjacent clusters to achieve the minimum size criteria.

Although the k-means algorithm can cluster by proximity in n-dimensions, our implementation groups points by proximity in two-dimensional space. K-means does not limit the size of the cluster, so some post processing must be performed to achieve desired results. The steps of our k-means clustering algorithm follow:

1. Calculate approximate number of clusters by dividing the number of points by the maximum number of points in a cluster.
2. Assign each point to a random cluster.
3. Calculate centroids of each cluster.
4. Assign the nearest point in a cluster as the cluster centroid.
5. Aside from the cluster centroids, assign each point to its nearest centroid.
6. Recompute each cluster centroid.
7. Repeat steps 4-6 until the centroids stop moving based on Euclidean distance, or until a maximum number of iterations is reached (10 times the number of points in the dataset).
8. Divide clusters that are too large into smaller clusters using steps 1 through 7.
9. If any cluster is too small, add nearest points until minimum cluster size is achieved.

To decrease the number of iterations required for clustering, our variation of the k-means algorithm forces each cluster centroid to be a data point.

Weighted Affine Transformation

During step 6 of the integration process (the weighted affine transformation step), coordinates of the intersection points on the vector, or “target” layer, are modeled to fit the matching intersection points on the image, or “control” layer, through a general least squares adjustment. This process is repeated for each cluster of matching point pairs. The general least squares adjustment incorporates weights for control and target coordinates, and minimizes the sum of squares of the weighted residuals (Mikhail and Gracie, 1981). Weights are computed from coordinate variance estimates. Before adjustment, variance estimates for control coordinates are estimated as the width of one or two pixels, but this is subject to further testing. *A priori (initial)* variance estimates for target coordinates should be larger than control variances, which are approximated as ten times the variances of control coordinates. During adjustment, *initial* target coordinate variances are modified using the Danish method to remove associated effects of suspected outlying observations (Krarup and others, 1980). *A posteriori (adjusted)* target coordinate variances are computed as the outlier-corrected variances multiplied by the variance of unit weight after convergence (Stanislawski, 2000).

The method being used to modify the weights of suspected outlying observations was developed by the Geodetic Institute of Denmark and is "especially designed to eliminate gross errors" (Krarup and others, 1980). The diagonal element of the weight matrix for outlying data layer coordinates will be modified as follows:

$$w_{i+1} = w_i \; f(r_i)$$

where w_i represents a diagonal element of the weight matrix for the *i*th iteration, and

$$f(r) = 1 \quad if \quad \frac{|\, r \,| \sqrt{w}}{\hat{\sigma}} < c \text{ , } or$$

$$f(r) = e^{-\left(\frac{|r| \sqrt{w}}{c \, \hat{\sigma}} \right)}$$

where *r* is the residual of the observation, *w* is the corresponding diagonal element of the weight matrix, *c* is a constant usually set to 3, and $\hat{\sigma}$ is the *a posteriori* estimate of the reference variance (Kubik and others, 1986).

Basically, this process iteratively increases the variance of an observation if its standardized weighted residual is larger than the constant. Modification of the variance of a point's coordinates effectively flags it as a blunder, which results in filtering the associated pair as unacceptable for use in the subsequent rubber-sheeting transformation.

STATUS

As of February 2007, all component programs have been written or acquired from previous researchers, and a sample dataset from previous studies has been acquired. We currently are testing the programs with sample data from the St. Louis area. The sample MODOT vector road data has about 100 road intersection points that are possible candidates for image pattern matching. Upon automatically selecting a set of matching road intersection pairs via the pattern matching program (step 4 of processing steps), we will filter these points through the vector median filter and the LAT filter and compare results. Subsequently, we will refine the LAT filter process, or complete the rubber-sheeting transformations using the two sets of filtered points and compare results. If the LAT filter process provides improved rubber-sheet results, we will test the process on a larger area.

Upon completing several tests with the localized affine transformation (LAT) filter, we hope to answer the following questions regarding the tested datasets:

1. Does the LAT approach actually filter any points?
2. If the LAT approach filters any points, does it filter different points than those removed with the vector median filter? Also, are these data integrated better using the LAT filter based on Euclidean distance rather than using the vector median filter?
3. Does the LAT filter take substantially more time than the vector median filter?
4. Can additional information regarding the accuracy of the integration process be derived from the LAT models through error propagation?
5. What are the limitations of the LAT filter approach? How close must the datasets be integrated before implementation? What is the minimum number of points required for the process to be successful? Can the integration system be implemented in rural areas?

REFERENCES

Alsabti, K., S. Ranka, and V. Singh (1998). An efficient k-means clustering algorithm. Information Technology Lab, Hitachi America. Accessed Feb. 13, 2007 at URL: http://www.ece.northwestern.edu/~peters/references/KMeans98.pdf

Chen, C.-C., C.A. Knoblock, C. Shahabi, and S. Thakkar (2003a). Automatically and accurately conflating satellite imagery and maps. Proceedings of the International Workshop on Next Generation Geospatial Information. Cambridge, Massachusetts.

Chen, C.-C., C. Shahabi, and C.A. Knoblock (2003b). Automatically conflating road vector data with high resolution orthoimagey. Report to U. S. Geological Survey on Grant No. 03CRSA0631. University of Southern California, Los Angeles.

Finn, M.P, E.L. Usery, M. Starbuck, B. Weaver, and G.M. Jaromack (2004). Integration of *The National Map*. Abstract presented at the XXth Congress of the International Society of Photogrammetry and Remote Sensing: Istanbul, Turkey, July 2004. Accessed Feb. 12, 2007 at URL: http://carto-research.er.usgs.gov/data_integration/pdf/integrationsISPRS.pdf

Jain, A.K., M.N. Murty, and P.J. Flynn (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264-323.

Krarup, T., J. Juhl, and K. Kubik (1980). Gotterdammerung over least squares adjustment. 14th Congress of the International Society of Photogrammetry, pp. 369-378.

Kubik, K., D. Merchant, and T. Schenk (1986). Grosserrors and robust estimation. pp.250-255 *In* Technical Papers 1986 ACSM-ASPRS Annual Convention, Vol. 4., Washington D.C.

Mikhail, E.M., and G. Gracie. (1981). *Analysis and adjustment of survey measurements*. Van Nostrand Reinhold Company, New York, NY.

Saalfeld, A. (1985). A fast rubber-sheeting transformation using simplicial coordinates. *The American Cartographer*, 12(2):169-173.

Saalfeld, A. (1993). *Conflation: Automated Map Compilation*. Ph.D. Dissertation, Computer Vision Laboratory, Center for Automation Research, University of Maryland.

Stanislawski, L.V. (2000). Mapping positional accuracy of geographic information system data layers through error propagation. pp.597-600 *In* Proceedings of 4th International Symposium On Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Amsterdam, Netherlands. 772pp.

Stanislawski, L.V., B.A. Dewitt, and R. Shrestha (1996). Estimating positional accuracy of data layers within a GIS through error propagation. *Photogrammetric Engineering and Remote Sensing*, 62(4):429-433.

U. S. Geological Survey (2006). *The National Map*. Accessed Feb.10, 2007 at URL: http://nationalmap.gov/index.html

Usery, E.L., M.P. Finn, and M. Starbuck (2005). Integrating data layers to support *The National Map* of the United States. Proceedings International Cartographic Conference, A Coruña, Spain, July 2005. Accessed Feb. 12, 2007 at URL: http://carto-research.er.usgs.gov/data_integration/pdf/data-integration-icc.pdf

Vernon, Jr., D.E. (2004). Geospatial technologies in homeland security. *EOM, Earth Observation Magazine*, 13(1):12-14.

White, M.S. and Griffin, P. (1985). Piecewise linear rubber-sheet map transformation. *The American Cartographer*, 12(2):123-131.