

Geospatial Data Integration

1. Research Team

Project Leader:	Prof. Craig Knoblock, <i>Computer Science</i> Prof. Cyrus Shahabi, <i>Computer Science</i>
Other Faculty:	Prof. Dennis Mcleod, <i>Computer Science</i>
Post Doc(s):	Jose Luis Ambite
Graduate Students:	Ching-Chien Chen, Mohammad-Reza Kolahdouzan, Mehdi Sharifzadeh, Snehal Thakkar
Industrial Partner(s):	IBM, Microsoft, SUN, Navigation Technologies

2. Statement of Project Goals

Recent growth of the geo-spatial information on the web has made it possible to easily access a wide variety of geo-spatial data. Accurately integrating diverse geo-spatial data remains a challenging task, since geospatial data obtained from various data sources may have different projections, different accuracy levels, and different inconsistencies. One of the major applications of integrating these data is to automatically identify objects, such as roads or buildings, in the satellite imagery and annotate these objects with the information from other data sources. GIS and computer vision researchers have worked on identifying objects in the satellite imagery for a long time. However, the resulting algorithms take a large amount of processing time and may produce inaccurate results. The goal of the project is the design and implementation of a novel information integration technique, which utilizes geospatial and textual data available on Internet to automatically and efficiently integrate geo-spatial data and identify objects.

3. Project Role in Support of IMSC Strategic Plan

The main objective of this research sub area is to provide a transparent layer for retrieving and integrating data from several heterogeneous databases and online sources. The unifying framework for integration is the objects' time and space dimensions. At IMSC, several immersive environments, such as *2020Classroom*, require integration of various data (e.g., course-related material or news) from different data sources to provide rich content to the users. These environments have also rich spatial and temporal characteristics (see the *Immersidata* report) and it is very natural to query and access objects based on their space and time dimensions (e.g., user looking to location X, Y, Z at time t, hence, all the objects relevant to that area and time should be accessed and displayed). The techniques designed and developed in this research can be utilized to efficiently integrate distributed spatio-temporal information. In particular for this year's achievements, the results would enable accurate annotation of spatial objects in the imagery using information integration.

4. Discussion of Methodology Used

During the previous year, we have developed techniques to integrate spatio-temporal information. Currently, we are working on integrating various geo-spatial datasets with certain spatial inconsistencies between them. In this report, we will focus on the integration of satellite imagery with geo-spatial vector data. Our proposed mechanism utilizes the *conflation* approach [5] for integration.

Conflation technique compiles two geo-spatial datasets covering the overlapping regions by establishing the correspondence between the matched entities and transforming other objects accordingly. The process can be divided into the following tasks: (1) find a set of conjugate point pairs, termed control point pairs, in both datasets, (2) filter control point pairs, and (3) utilize algorithms, such as triangulation and rubber-sheeting, to align the rest of the points and lines in two datasets using the control point pairs. Traditionally, human input has been essential to find control point pairs and/or filter control points. Instead, we developed different techniques to find control point pairs in both datasets and designed novel filtering techniques to filter inaccurate control points completely automatically.

5. Short Description of Achievements in Previous Years

1. Wrapper building tools that utilize Machine Learning techniques to reduce the amount of user interaction while wrapping different web sources.
2. Information mediator termed Ariadne [2] that provides a uniform interface to multiple heterogeneous sources.
3. The distributed query plans that can query moving objects with pre-defined paths in a distributed environment efficiently [3].
4. The application that performs route planning and K-Nearest Neighbor search based on shortest path or shortest time as well as Euclidean distance, utilizing space embedding techniques to reduce the cost of computation of K-Nearest Neighbor algorithms [4].

5a. Detail of Accomplishments During the Past Year

During the past year, we developed the following techniques:

- **Conflating satellite imagery and vector data automatically:** Finding accurate control point pairs is a very important step in the conflation process as all the other points in both datasets are aligned based on the control point pairs. The traditional conflation technique locates control points on different datasets manually or semi-automatically. We developed two techniques, “generating control points from online data” and “generating control points by localized image processing”, to automatically find the control points. Utilizing these techniques, we can realize automatic conflation.
- **Generating control points from on-line information:** We obtain control points by querying information from online web sources to identify corresponding feature points on both datasets.
- **Generating control points by localized image processing:** This technique relies on the imagery and vector data for actually acquiring control points. We find feature points, such as the road intersection points, from the vector dataset. For each intersection point,

we perform image processing in a small area around the intersection point to find the corresponding point in the satellite image.

- **Filtering control points:** Both of the control points detection techniques mentioned above may detect some inaccurate control point pairs. For example, the approach to identify control point pairs using localized image processing may produce inaccurate control point pairs due to some image noise or linear features that are not roads. For instance, in Figure 2(b), the control point pairs 1 and 2 are inaccurate. Therefore, it is very important to filter out inaccurate control point pairs. The inaccurate control point pairs can be detected by identifying those pairs with significantly different relationship as compared to the other nearby control point pairs. We evaluated two different filters, RANSAC [6] and vector median filter (VMF)[7], to filter out the inaccurate control point pairs.

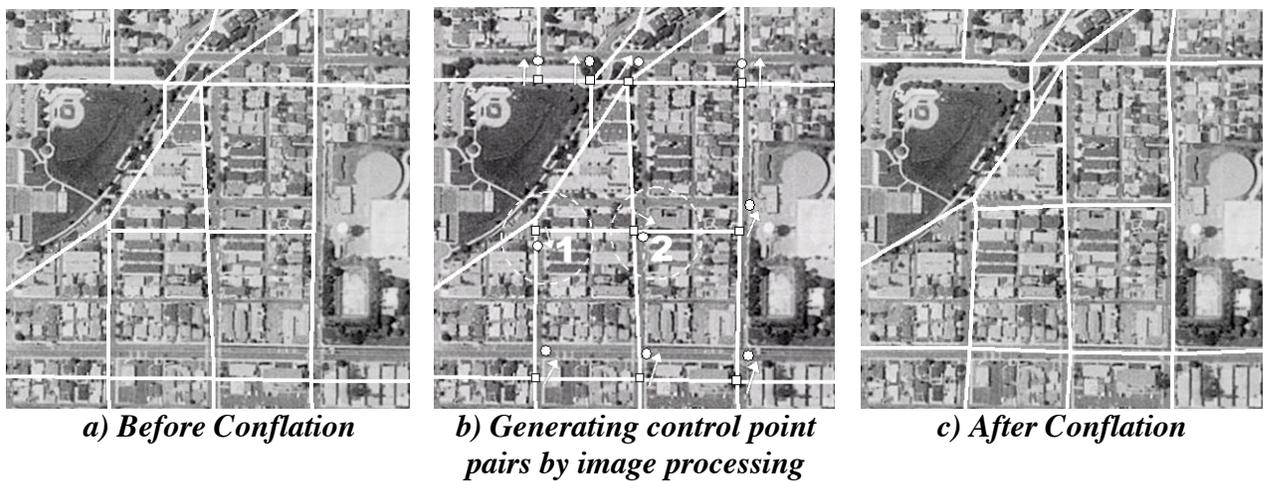


Figure 2 Example images during the conflation process

In [1], we discussed other techniques to improve the accuracy and performance of conflation even further.

6. Other Relevant Work Being Conducted and How this Project is Different

There have already been substantial approaches on integrating geo-spatial data and annotating spatial objects [8, 9, 10]. These approaches utilize different methods for locating the counterpart elements between datasets. Most of these approaches are time-consuming and semi-automatic. Our research is different from the other relevant works as we focus on efficient and autonomous integration of geospatial data, utilizing wide variety of geospatial and textual data available on the Internet.

7. Plan for the Next Year

We plan on continuing our research in the following ways:

1. To design and develop techniques to alleviate the spatial inconsistencies for the area where there is no feature point (such as landmarks or intersection points) on vector data and imagery to perform conflation.
2. To improve the conflation result by an iterative conflation process. The process could work as follows: the vector-image conflation operations, automatic control point pairs generation and vector to imagery alignment, are alternately applied until no further control point pairs are identifiable.
3. To apply our approach to the integration of imagery and vector data collected from National Imagery and Mapping Agency (NIMA) and verify the conflation results.

8. Expected Milestones and Deliverables

1. Building Finder [11]: this web application utilizes the satellite imagery from Microsoft TerraService [12] and the Tigerline vector files [13] from US Census Bureau (as well as some online sources) to annotate buildings on the imagery.
2. Road Finder: a windows application where users can navigate the satellite imagery from Microsoft TerraService with the conflated Tigerline vector data (roads) on it. Moreover, the application applied intuitive UIs to add/remove control points manually or automatically.

9. Member Company Benefits

IBM Informix database server and SUN servers are part of our image, map and vector data repository system. Navigational Technologies provides accurate vector data sets to us and is interested in building next generation route planning applications. We use .NET and Web Services technology as the primary development environments, because of which Microsoft Corporation has been providing us with unrestricted cash gifts since January 2002.

10. References

- [1] Ching-Chien Chen, Snehal Thakkar, Craig A. Knoblock and Cyrus Shahabi (2003). *An Information Integration Approach to Automatically Annotate Spatial Objects in Satellite Imagery*. Submitted for review.
- [2] Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Naveen Ashish, Ion Muslea, Andrew G. Philpot and Sheila Tejada (2001). *The Ariadne Approach to Web-based Information Integration*. International the Journal on Cooperative Information Systems (IJCIS) Special Issue on Intelligent Information Agents: Theory and Applications 10(1-2): 145-169.
- [3] Cyrus Shahabi, Mohammad R. Kolahdouzan, Snehal Thakkar, Jose Luis Ambite and Craig A. Knoblock (2001). *Efficiently Querying Moving Objects with Pre-defined Paths in a Distributed Environment*. The Ninth ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS'01), Atlanta, GA, USA.
- [4] Cyrus Shahabi, Mohammad R. Kolahdouzan and Mehdi Sharifzadeh (2002). *A Road Network Embedding Technique for K-Nearest Neighbor Search in Moving Object Databases*.

The 10th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS'02), McLean, VA.

[5] Alan Saalfeld (1993). *Conflation: Automated Map Compilation*. Computer Vision Laboratory, Center for Automation Research, University of Maryland.

[6] Martin Fischler, Robert Bolles (1981). *Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography*. Communications of the ACM, 1981. 24.

[7] Jaakko Astola, Petri. Haavisto, et al. (1990). *Vector Median Filter*. In Proceedings of the IEEE. 1990.

[8] Albert Baumgartner, Carsten Steger, et al. (1996). *Update of Roads in GIS from Aerial Imagery: Verification and Multi-Resolution Extraction*. IAPRS, 1996. 31.

[9] Sagi Filin and Yerahmiel Doytsher (2000). *A Linear Conflation Approach for the Integration of Photogrammetric Information and GIS Data*. IAPRS, 2000. 33.

[10] Heiner Hild and Dieter Fritsch (1998). *Integration of vector data and satellite imagery for geocoding*. IAPRS, 1998. 32.

[11] <http://apollo.isi.edu/BuildingFinder/WebForm1.aspx>

[12] <http://terraservice.net/>

[13] <http://www.census.gov/geo/www/tiger/>