RECOVERING SPATIAL INFORMATION FROM PURELY GRAPHICAL PRODUCTS

Sally E. Goldin*[a] and Kurt T. Rudahl[b]
[a] Faculty, Department of Computer Engineering, King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Road, Bangkok 10140, Thailand; Tel:+66-2-470-9088;
E-mail: seg@goldin-rudahl.com

[b] Faculty, Department of Computer Engineering, King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Road, Bangkok 10140, Thailand; Tel:+66-2-470-9088;
E-mail: kurt@cpe.kmutt.ac.th

KEY WORDS: Doppler Weather Radar, Information Extraction, GIS, Image Processing, Optical Character Recognition

ABSTRACT: Modern geographic information systems require data to be represented in digital form. Although spatial data are increasingly available in digital formats, there are still situations where the only source of the necessary information is non-digital. Examples include both historical maps and graphical products generated by proprietary systems that do not provide access to underlying raw data.

This paper presents a software framework for recovering geographically-referenced information from such products. We describe our implementation of this framework for a specific case, namely Doppler weather radar Plan Position Indicator (PPI) plots gathered by the Thai Bureau of Royal Rainmaking and Agricultural Aviation (BRR). Our system uses image processing and optical character recognition to extract radar return amplitude values from the GIF images output by BRR radar stations. It then georeferences those values and stores them in a database. We demonstrate the utility of these data by synthesizing new display products such as integrated Constant Altitude Plan Position Indicator (CAPPI) plots to support country wide precipitation monitoring.

The results suggest that our general approach can be applied to other cases where the information needed to support geographical decision making is locked away in formats intended for display rather than analysis.

1. INTRODUCTION

Computer-based geographic information systems (GIS) provide tremendous value. They support critical planning, monitoring, design and decision-making activities by allowing users to transform, integrate, and evaluate spatial information represented in flexible digital formats. However, even in today's digital world, there are cases where a digital source of required information is not available. Examples include historical maps and documents as well as graphical products generated by proprietary systems that do not provide access to the underlying raw data.

This paper presents a general framework for recovering geographically-referenced information in such cases, and then describes a specific implementation of this framework in a situation where raw data are urgently needed but are not available. Our example is a relatively simple case, where the input is well-documented and contains little noise. We discuss extensions to the method in order to handle more challenging situations.

2. BACKGROUND

Although an extensive literature exists on extracting vectors and text from scanned paper maps (e.g. Luo et al., 1995; Tofani and Kasturi, 1998), the results are typically not georeferenced and integrated into a spatial data base. A variety of projects have attempted to build searchable databases from scanned historical maps, but in most cases these efforts rely on humans to annotate, index and geocode the map data (Simon et al., 2009). We are not aware of any prior work that automatically extracts data from raster products, georeferences that two or three dimensional spatial information, and then stores it in a form that can be used by other applications.

Automated extraction of textual information from historic documents is an active research area (e.g. Moghaddam and Cheriet, 2009; Tangwongsan and Sumetphong, 2008). Although the goals of these studies are quite different from ours, some of the same principles can be applied, e.g. the suppression of noise, the identification of critical features, and the use of contextual and application knowledge. Text extraction research is also relevant because a significant amount of potentially useful geographic data exists only in the form of printed tables in government reports. Our framework could be extended to extract and store information from such sources.

3. INFORMATION EXTRACTION FRAMEWORK

Figure 1 presents an overview of our framework.

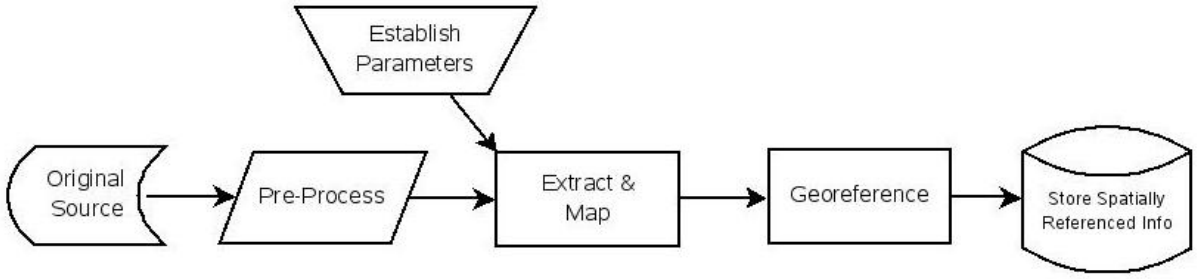

Figure 1. Framework for information extraction

The framework indicates that there are three fundamental activities: preprocessing the information in the original source, extracting the information guided by some application-specific parameters, and transforming the information into a useful geographic coordinate system. Although the figure suggests that these activities are sequential processing steps, that is not necessarily the case. In particular, extraction and georeferencing can sometimes be handled in parallel.

The activities have very general labels because the details can vary a great deal depending on the situation. Pre-processing may involve scanning from a hardcopy source or file format conversion as well as removal of noise or redundant information. The "extract & map" activity may involve anything from a simple color-to-data-transformation, as in our work, to complex feature extraction or automated classification. The "georeference" activity uses some base information from the parameter set to assign X, Y and possibly Z coordinates to the spatial attributes identified in the previous activity. The results are saved in a data base customized for the type of data and its intended use. In many cases, this may be the primary spatial data store of a general purpose GIS package.

Given the generic nature of this framework, one might ask why it is useful. First, the framework unifies a range of different problems, putting them in a common perspective. This allows us to consider whether specific processing techniques applied in one situation might be transferred to new situations. Second, the framework emphasizes the fact that unless extracted information is referenced to a geographic coordinate system and stored in a searchable form, it has little practical value.

The next sections make this framework more concrete by applying it to a specific, real world problem. After describing our method and results, we conclude by considering other potential applications.

4. EXTRACTING RAW MEASUREMENTS FROM WEATHER RADAR PLOTS

4.1 Problem

Continuous measurement and monitoring of precipitation is very important for Thailand. Accurate, timely information on rainfall location and intensity is necessary to predict and respond to droughts and floods, both of which are serious problems for the country. These data are also potentially useful for landslide prediction and can be used to guide cloud seeding activities.

Thailand has approximately 20 weather radar installations in various locations throughout the country, operated by the Bureau of Royal Rainmaking and Agricultural Aviation (BRR) and the Thai Meteorological Department (TMD). Each radar site can capture information on precipitation every 5 to 15 minutes in a radius of 240 km.

BRR evaluates and disseminates the data gathered by these stations. It maintains a website which presents frequently updated sample plots of data gathered from various stations as either Plan Position Indicator (PPI) or Constant Altitude Plan Position Indicator (CAPPI) plots, which are conventional data products in this domain. Figure 2 shows a sample PPI image generated at the radar station at Don Muang airport, just north of Bangkok.

Although BRR and TMD apply these radar images for weather prediction, the utility of these data is limited in several ways:

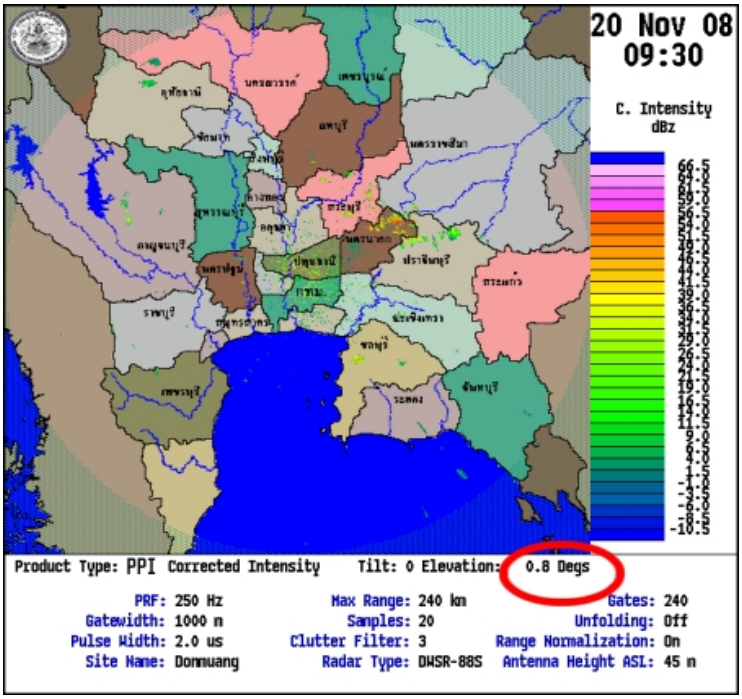

Figure 2. Example PPI image

- It is difficult to integrate regional information into a national perspective. Each radar installation produces plots for a radius of only 240 km around the station. In addition, different stations produce plots of different types and with different geographic and measurement scales. Thus there is no easy way to assess precipitation on a country-wide basis.
- The software associated with the stations produces only processed graphical results. The raw measurements are not currently available.
- Since radar data are available in graphical form only, researchers do not have the numerical, historical data they need to create more sophisticated models and decision support systems. Apparently the company which sells the radar hardware and software can supply a module that outputs raw data, but BRR informed us that this software component is prohibitively expensive.

We undertook this project to provide an affordable solution to the raw data problem as well as to provide integrated visualizations of precipitation that combine information from multiple radar sites. We have designed and implemented a software system that can automatically extract numerical precipitation readings from the graphical products currently available, store these in a database, and then use the stored information to synthesize new precipitation maps that display data from multiple radar stations for any selected region in Thailand. This system can potentially provide near-real-time access to country-wide precipitation information.

4.2 Method

The Precipitation Radar Analysis System (PRAS) software was developed at the Geoinformatics Laboratory at KMUTT (GLaK). The software is written in Gnu C and in Perl, with a small amount of JavaScript used in the web-based user interface. It uses a variety of open source components including the PostgreSQL data base management system (*http://www.postgresql.org*), the Proj4 geographic projection library (*http://trac.osgeo.org/proj/*), the Gocr optical character recognition application (*http://jocr.sourceforge.net/*), Shapelib, a library for reading and writing ESRI Shape files (*http://shapelib.maptools.org/*), ImageMagick, an image transformation application (*http://www.imagemagick.org/*), and Gd, a graphics library for creating images in common formats (*http://www.libgd.org/*).

PRAS involves three core processes, which operate asynchronously.

The first process, *data acquisition*, involves querying the Bureau of Royal Rainmaking website to retrieve the most current PPI images from each station that is on-line. The retrieved images are converted from GIF to RGB format, stored on our server, and queued for further processing. This operation corresponds to the **Preprocess** activity in our framework.

The second process, *data extraction*, uses image analysis techniques to locate pixels in the image which represent radar returns and records data values for those locations. Each radar image includes a legend, which indicates the colors associated with different radar values. The data extraction module analyzes the legend for each image in order to determine the correspondence between pixel color and value, which differs across different radar sites.

This process also uses parameters for each station to assign three-dimensional geographic coordinates to each radar return point. The X and Y coordinates are based on the known location of the radar station and known resolution of the image. The Z coordinate is calculated based on the known elevation of the radar station plus the beam angle, which is recorded on each image and which we extract using optical character recognition. The resulting values are stored in a data base with a format very similar to recently proposed standards for LIDAR data (Ferede et al., 2009).

Thus, the data extraction processes corresponds to both the **Extract & Map** and the **Georeference** activities in the framework. A number of input parameters are required for each station: its geographic location and elevation above sea level, the resolution of the PPI images it produces, the location and size of the legend on those images, and the location of the beam angle string on those images. These values do not change once they are determined; however, we would have to add a new set of parameter values if we extended the software to handle a new radar site.

The third process, *image synthesis,* is outside the framework but demonstrates the use of the extracted information. This process generates integrated CAPPI images based on user requests from a web page. The user selects a rectangular region of the country to view (graphically or using lat/long coordinates), a date and time, and an elevation. The process queries the database to locate radar readings that satisfy these criteria. It then assembles a synthetic CAPPI image, adding reference data such as country and provincial boundaries and water features, and displays this image in the web page.

The database serves both as the repository for collected radar data and as a communication mechanism for coordinating these three main processes. The data acquisition process records the availability of new PPI images in the database. The data extraction process examines the database to discover which PPI images have been uploaded but not yet analyzed. Data extraction analyzes these images and adds their radar readings to the database. After an image has been processed, the data extraction module updates the database to indicate that fact. Finally, the image synthesis process accesses radar readings as well as contextual geographic features stored in the database, and uses the query results to create an integrated image for a region of interest.

4.3 Results

We have successfully implemented and deployed PRAS on our university server. Figure 3 shows an example of an integrated precipitation image generated by the Precipitation Radar Analysis System.

The prototype version of PRAS is publicly available at *http://www.cpe.kmutt.ac.th/glak/radarproject/demo.html* **.** This URL provides instructions plus a link to the application. The prototype allows the user to request integrated images for anywhere in Thailand. However, the software for gathering new images (data acquisition and data extraction) has been turned off because we do not have adequate disk space to handle the large amount of data involved. The database available holds readings from only two dates.

One outcome of this research was the realization that an operational data extraction process for Thai weather radar images would generate a significant volume of data. Processing one image per hour from each radar site produces approximately 260 images per day. (This varies because sites are not always functional.) A single image can generate as many as 250,000 data points, although most contribute far fewer. The average number of points contributed by each image in our sample data base was about 13,000. Using these figures, we estimate that an operational system would add about three million readings, or about 150 MB, of data per day, or about 60 GB per year. If every captured image were processed, the annual storage requirements would approach a terabyte.

Budget limitations have so far prevented the BRR from moving our prototype system into a production mode, but as the price of storage devices continues to drop, these constraints may be reduced. In the meantime, the potential utility of our work is very clear. We have liberated a wealth of information previously locked away in the display-oriented PPI images and made it available for query, analysis, and integrated visualization.

5. DISCUSSION AND CONCLUSIONS

PRAS provides an instructive example of our information extraction framework. We have designed and implemented a system to calculate, georeference and store individual radar return readings from graphical products intended only for viewing. This information can be used for simulation and modeling as well as for near real time or historical visualizations of country-wide precipitation.

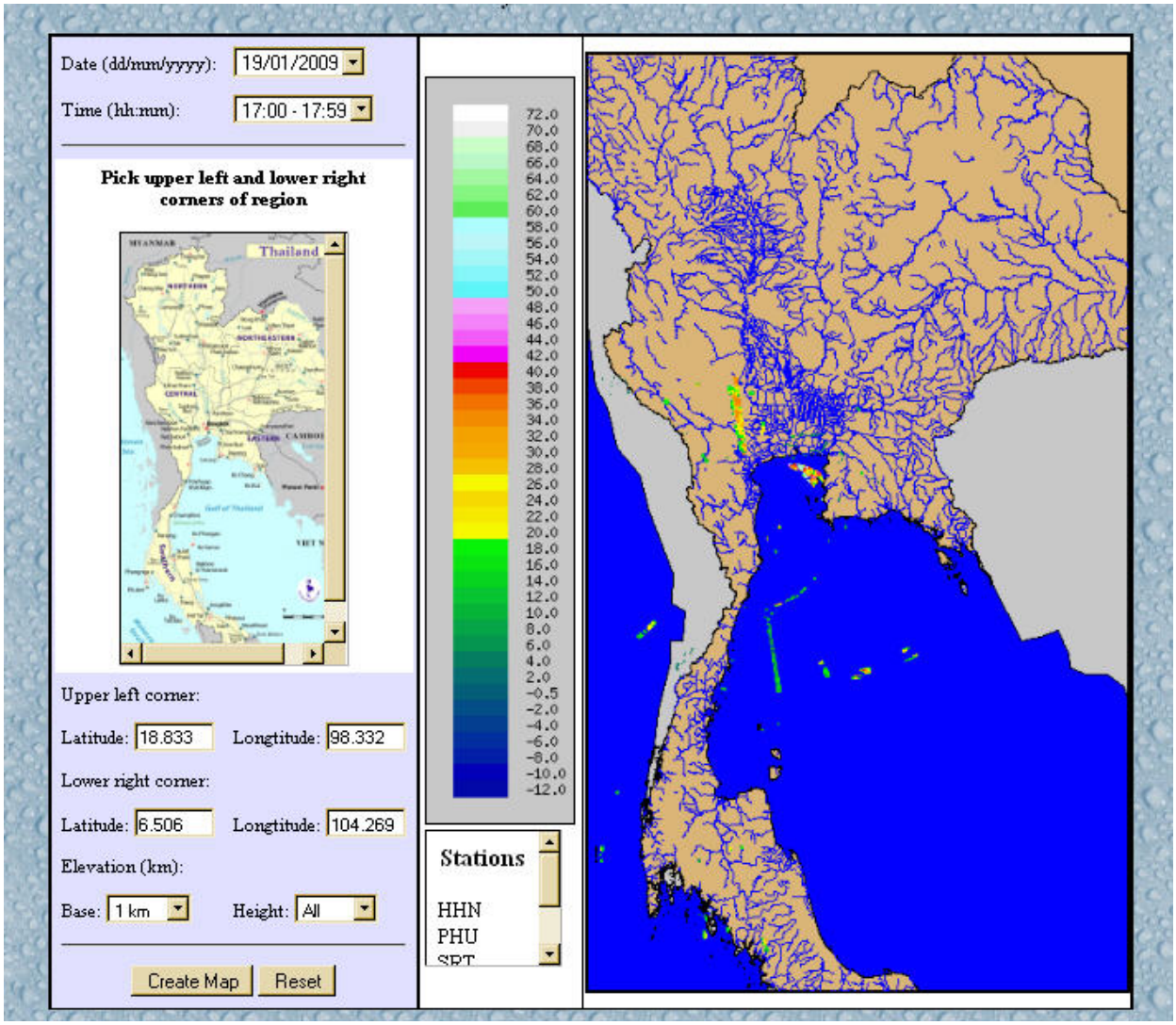

Figure 3. Integrated precipitation image created by PRAS

Our experience with this research has highlighted several important points.

1. Effective information extraction will likely require knowledge about the domain and the process that produced the source documents. In the case of the radar data, we had to determine both geographic and non-geographic parameters related to each station. We visited one radar site in order to gain a detailed understanding of the radar acquisition process. Establishing the correct parameters was actually one of the more time consuming aspects of the project.

2. The detailed processing required to extract and georeference data depends a great deal on the particular sources involved. In our case, we could rely on the color coding of the image since the source product was originally generated in digital form. When sources must be scanned, as in Tofani and Kasturi (1998), color information is likely to be far less uniform and reliable. Preprocessing such as clustering, histogram analysis, or slicing of the scanned image into layers may help improve this situation.

3. Some sort of geographical reference is required in order for this sort of processing to be successful. In our case, we could use the location of the station receiving the data. Paper maps created within the last century should provide fairly reliable coordinate annotation which can be used to bootstrap georeferencing. For older historical maps, however, precise location information may be difficult to obtain, and thus the accuracy of the resulting data may be reduced.

4. Given the amount of work involved in implementing an instantiation of the framework, this will be most worthwhile when there is a significant number of comparable source documents from which information is to be extracted.

We are particularly interested in applying our framework to tabular geographic data. Our experience suggests that in Thailand, at least, large amounts of useful information are collected on a per-subdistrict or per-village basis, but never entered into a GIS. Instead, these data remain trapped in paper reports. We believe that we can use image processing to segment tables into regions or columns, then apply OCR to turn text into values which can be stored as attributes to polygons representing districts or other political units.

GIS offers tremendous potential as a tool for planning, monitoring and decision-making. The more information we can incorporate into our spatial data stores, the more effective our results will be. Our framework suggests a general approach for augmenting the wealth of digital data currently being produced with critical information that is currently available but inaccessible in non-digital sources.

6. ACKNOWLEDGMENTS

This work was supported by a research grant from the Software Industry Promotion Agency via the Thai National Grid Center. We also want to thank Dr. Prasert Aungsuratana from the Bureau of Royal Rainmaking and Agricultural Aviation for his cooperation and assistance.

7. REFERENCES

Ferede, H., Sarkani, S., and Mazzuchi, T.A.., 2009. Multi-dimensional data discovery. Proceedings of ASPRS/MAPPS 2009 Fall Conference, San Antonio, TX. American Society for Photogrammetry and Remote Sensing.

Luo, H., Agam, G., and Dinstein, I. 1995. Directional mathematical morphology approach for line thinning and extraction of character strings from maps and line drawings. Proceedings of ICDAR 1995, Third International Conference on Document Analysis and Recognition, Volume 1, Montreal, Canada, pp. 257-260. IEEE Computer Society.

Moghaddam, R.F. and Cheriet, M., 2009. Application of multi-level classifiers and clustering for automatic word spotting in historical document images. 10th International Conference on Document Analysis and Recognition. IEEE, ICDAR_2009.104.

Simon, R., Korb, J. Sadilek, C. and Schmidt, R., 2009. Collaborative map annotation in the context of historical GIS. IEEE eScience 2009 Workshops.

Tangwongsan, S. and Sumetphong, C., 2008. Optical character recognition techniques for restoration of Thai historical documents. 2008 Conference on Computer and Electrical Engineering. IEEE, ICCEE.2008.142.

Tofani, P. and Kasturi, R., 1998. Segmentation of text from color map images. Proceedings of the 14th International Conference on Pattern Recognition, Volume 1, pp. 945-947. IEEE.