Information Extraction based on the Concept of Geographic Context

Stefan Leyk and Yao-Yi Chiang

ABSTRACT: State-of-the-art graphics recognition technologies for extracting geographic information from scanned map images are very labor intensive and do not scale well to process a large number of maps. Moreover, many historical scanned maps suffer from poor graphical quality due to bleaching of the original paper maps and archiving practices, and are ill-posed for traditional one-time training and recognition efforts. This paper introduces a novel context-based framework for automated recognition of geographic information from cartographic documents including historical scanned map documents of varying graphical quality. This recognition approach makes use of the fact that map content of the same area changes incrementally over time, and such dependencies can be used as geographic context to fully automate the recognition process. If successful, such a framework can be useful for extracting geographic information from whole map series at a national level and over long time periods enabling new forms of landscape research. We demonstrate this framework in a case study on exploiting contextual information about a map area to extract hotel symbols from a scanned paper map, automatically, and discuss the applicability of this framework using geographic layers as context.

KEYWORDS: Pattern recognition, historical maps, massive map archives, context-based information extraction

Introduction

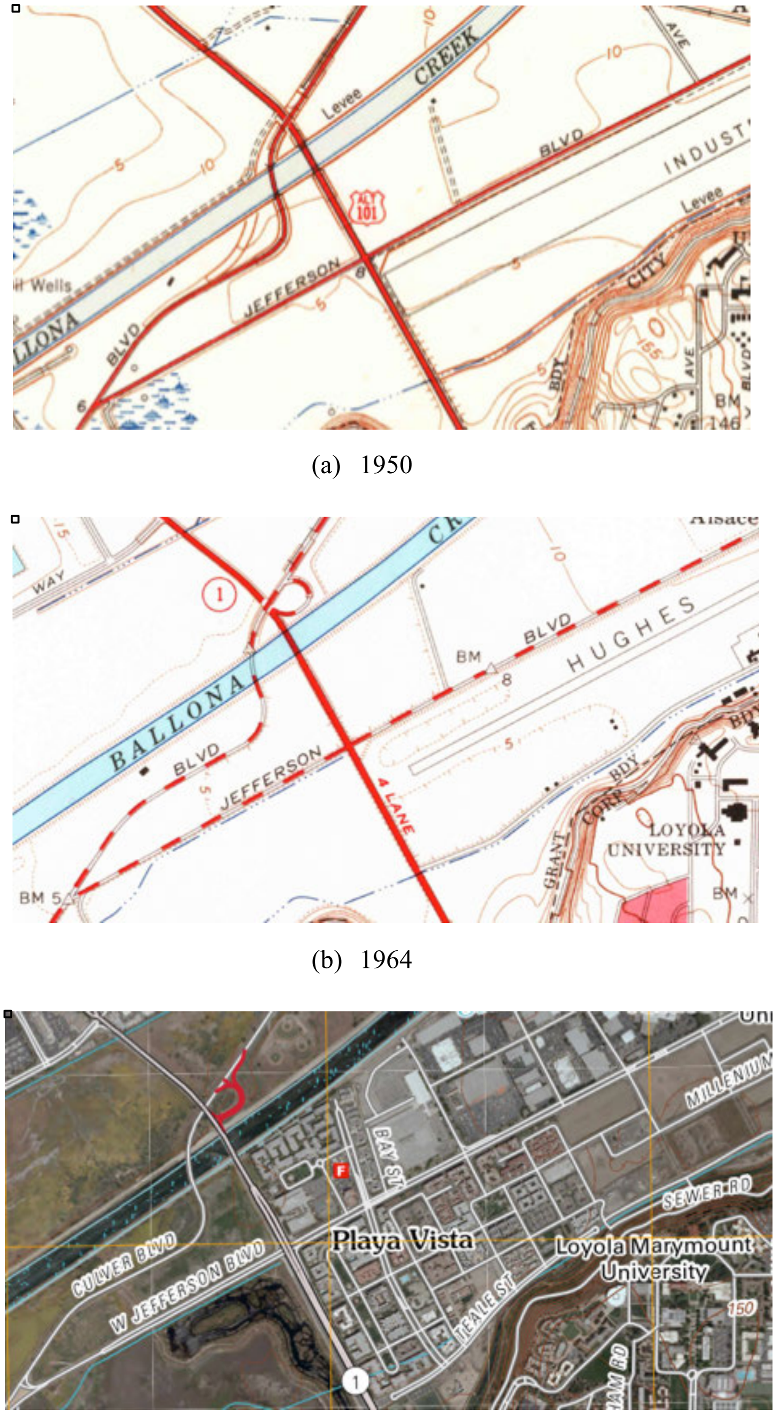
Detailed data on the states of landscapes in the past is essential to understanding the causes and consequences of environmental change. Numerous national mapping agencies have created thousands of maps over long time periods and scanned these paper maps to build digital map archives. The potential of these historical map archives has not been realized because typically, the information is stored as scanned images, and the scanned images have only recently become available. While these scanned images can be easily read and understood by humans, systematic exploration of their contents requires robust, efficient data extraction and conversion into a format that allows meaningful analysis in a geographic information system (GIS) (Chiang et al., 2014a). Extracting this kind of information has the potential to unlock unique research opportunities in the social and environmental sciences related to large geographic areas covering long time periods (e.g., Kozak et al., 2007; Stein et al., 2010).

State-of-the-art graphics recognition technologies for extracting cartographic symbols from scanned map images rely on a user labeling process to generate a set of shape, color, and gradient descriptors of the feature of interest (Frischknecht and Kanani, 1998; Khotanzad and Zink 1996; Chiang et al., 2013; Chiang et al., 2014a; Chiang et al., 2014b). For example, in previous work, we developed an interactive approach for extracting road lines from historical USGS maps (Chiang et al., 2013), which uses manually collected road and non-road (e.g., wetland) samples to reliably extract road vector data and remove noise. While this approach reduced the overall processing time by 38% compared to a manual digitization it still took four labeling steps and a total of 50 minutes (sample labeling plus result curating) to process a map tile of 2283 × 2608 pixels that covers a small portion of the City of San Jose, California. Using a one-pixel buffer for evaluation, the result completeness was 99.9%, the correctness was 100%, and the redundancy was 0.054%. In another study, we presented a semi-automatic *"training-by-example"* approach to extract cartographic symbols from several scanned map images (Chiang et al., 2014b). This approach achieved 96.7% precision and 78.9% recall for extracting 147 symbols from four different maps with varying graphical properties.

While these previous approaches demonstrated the ability to produce accurate results, there is still the need for manual labeling and curating in processing each individual map sheet in different editions which limits their efficiency. These manual processing steps are necessary since most paper maps are printed and archived documents and often suffer from bleaching, blurring, and false coloring (e.g., Khotanzad and Zink, 2003; Leyk et al., 2006; Leyk and Boesch, 2009), properties that are propagated into the digital images during scanning. The significance of these image quality issues can even vary from one area to another within a map (Leyk, 2010), which means that additional sampling would be required to process an entire map sheet, and the trained feature descriptors are not directly applicable to another map.

In this paper, we describe the conceptual idea of geographic context that can support the development of fully automated recognition tools for the extraction of geographic information from individual scanned maps but can also be extended to process large-volume map archives. Geographic context can be derived from external databases such as online gazetteers or dictionaries or from geographic layers directly related to the processed map such as in a map series. Mapping agencies such as the USGS, update maps over time; they do not recreate them from scratch. Thus, maps are visual documents that change incrementally over time and share a great deal of content (i.e., the phenomenon described). This kind of dependence between editions (i.e., release versions) of the same map can be used to establish geographic context for more effective and automated extraction of geographic data from historical multi-edition map archives. Figure 1 shows an example of this dependence between historical 15-minute USGS topographic map sheets and the contemporary USGS National Map covering an area in the city of Marina del Rey, California. The bridge appears in all of the map editions. The major roads stay the same across editions (Highway 1 and Jefferson Blvd) with a minor addition (a ramp). More minor roads appear in the most recent addition, and they are connected to the major roads.

As a preliminary study, we demonstrate the basic idea of exploiting contextual information in an experiment of extracting hotel symbols from a scanned paper map, automatically. While the contextual information used here comes from an online gazetteer database, the results will shed light on future applicability of this approach to develop an effective recognition framework also for large map archives.



(a) 1950

(b) 1964

(c) 2012

Figure 1. Historic 15-minute USGS topographic maps (a, b) & USGS National Map data (c), Marina del Rey, CA.

The concept of geographic context in information extraction

The idea of context-based extraction starts with building generic information that describes position, geometry, topology, and other properties (called geographic context) of the feature of interest found in other data sources. This contextual information has the potential to make the recognition process more efficient, eliminating user intervention by improving localization and identification of features of interest in map documents and thus guiding the extraction process. For example, such geographic context can be used to ensure that graphics sampling and the computation of feature descriptors will be guided (e.g., sampling at the approximate location or nearby in the map) to reliably identify representations of the feature of interest as well as those of other features. The contextual data can also be used to detect extraction conflicts based on which the recognition process can be refined to minimize result curating efforts.

Building contextual information from existing geographic data

In order to describe the contextual information of the geographic phenomenon of interest, a generic semantic model needs to be built. This model may contain the locations of individual phenomena (e.g., x-y coordinates of nodes and vertices), their types (e.g., railroads) and attributes (e.g., road width), geometry types (e.g., polylines), and geometric characteristics (e.g., length of a road segment). Once the semantic model is built, the modeled data can be used to perform reasoning and infer *semantic rules* describing the spatial relationships between geographic phenomena of the same and different types (e.g., proximity or overlap between features). The semantic rules will dictate which contextual situations are possible (e.g., intersecting roads) vs. those that are unlikely (e.g., short isolated road segments) and guide the feature extraction process.

For example, to identify hotel symbols in a map, the contextual data (e.g., map features from a different map or locational descriptions including coordinates and attribute information recorded in a gazetteer) can be consulted knowing that there will be considerable overlap and similarity. This requires that the features in the contextual data are semantically described i.e., a feature is of type hotel, its location (e.g., coordinates) and the semantic rules are defined and implemented (e.g., hotel symbols must be close to roads).

Adaptive graphics sampling using contextual information

To carry out effective recognition the system needs to learn how the feature(s) of interest can be described. The contextual information as described above (e.g., locations of features and their assumed density/proximity) can be used to spatially constrain the area of interest and automatically collect "*nearby*" graphics examples from this area. More specifically, the semantic description "location" in the contextual data can be used to determine the position in the scanned map and define the sample areas (defined by a buffer distance possibly inferred by map knowledge). These sample areas represent the graphics examples used for further processing. Because of the expected overlap between map contents and contextual information, it can be assumed that most of the collected samples (image subsections) represent or contain the feature of interest.

To continue with the example of extracting hotel symbols in the processed map, the system would locate the positions given by the coordinates in the gazetteer data source, use a buffer distance of e.g., 200m (or 40 pixels if one pixel represents 5m) to determine and crop the sampling area assuming that these sampling areas likely contain a hotel symbol.

Computing feature descriptors of graphics samples and building a knowledge base

These image subsections are then input to computing various feature descriptors for shape (e.g., geometry, centroid, or eccentricity), color and texture properties of the graphical contents to build up a knowledge base of descriptors for each feature type. Such descriptors have been used in graphics recognition tasks in scanned maps for template matching (Maderlechner and Mayer, 1995; Reiher et al., 1996; Frischknecht and Kanani, 1998; Yin and Huang, 2001), but usually require a great amount of manual labeling. The built knowledge base can then be used to classify and detect features of interest (e.g., hospital symbols are black and white, 5-pixel wide circular homogeneous objects) and to identify and remove sample outliers based on variability measures.

This generated knowledge base can finally be applied to common recognition tools (e.g., Chiang and Knoblock, 2014; Chiang and Leyk, 2015) in order to create the extracted geographic information layer. The image will be scanned for properties that are similar to the descriptors of the feature(s) of interest through a matching process (e.g., template matching or histogram matching). Wherever template and image are similar enough the system will label a feature of interest (e.g., a hotel symbol).

In general, if the contextual information is meaningful and reliable (i.e., overlaps between map and contextual information), the above described context-based recognition approach has the potential to improve feature extraction from scanned map documents. The constrained graphics sampling process will by design collect subsections that are very likely to contain features of interest rather than other features, which therefore can be identified as statistical outliers. Furthermore, the collected sample can be used to estimate inherent variations in the target symbol thus building up the knowledge base which can help to overcome localized graphical quality issues and graphical differences.

Implementing the concept of geographic concept in an automatic recognition approach in a case study

A case study was conducted to run a first experiment on the implementation of the concept of geographic context in a recognition approach. The goal of this experiment is to develop a better understanding of the challenges of and opportunities in applying geographic context in map recognition tasks for full automation of the process. The results will also inform how this concept could be extended to the extraction of geographic information from massive map archives at high levels of automation and with high accuracy. As described above, we exploit the fact that information about geographic features in a region are not independent between data sources, such as a map and an online gazetteer, that both provide some kind of geographic content to automatically generate training samples and thus enable fully automatic cartographic symbol recognition. For example, the hotel information presented in a map might not be exactly the same (in terms of their presence, completeness and locations) as in an online gazetteer, but their contents should have a significant amount of overlap if both data sources display information from approximately the same time. Based on this assumption, we tested an information extraction approach that takes a scanned map and uses the map extent to query the GeoNames gazetteer[1] to find locations of a particular

[1] http://www.geonames.org/

feature type (e.g., coordinates of hotels). This approach then utilizes the feature locations to automatically identify the positions in question on the map, label and collect graphical samples of the feature symbol in the map, learns a set of feature descriptors from the samples, and finally uses the descriptors to find every symbol of the feature type of interest in the map. Traditionally, this type of extraction task requires a user to provide a sample (or template) of the feature of interest, and then the algorithm learns the descriptions of the feature to extract similar features from the whole map (e.g., Chiang et al., 2014b).

In this case study, we use the hotel symbol recognition as an example to explain our fully automated approach using contextual information from a gazetteer following the conceptual idea described above. Given a scanned map covering Baghdad, Iraq (current edition) (Figure 2), the task at hand is to find all hotel locations in the map without any user intervention for training the underlying feature recognition algorithms, knowing that not all hotels are listed in the gazetteer database but we aim at identifying all hotel locations in the map.

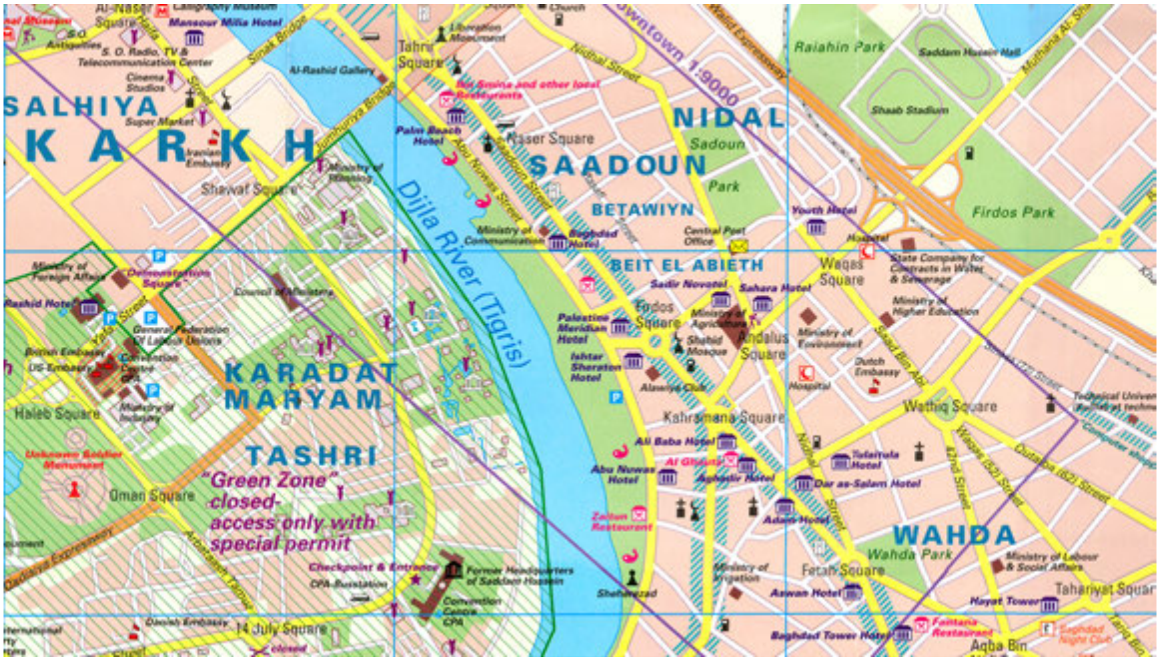


Figure 2. The test map of Baghdad, Iraq (Source: Gecko Maps).

Building contextual information

The system queried the GeoNames gazetteer using the map extent and the keyword “hotel” as filters. The returned query results contained only two hotel locations: Baghdad Hotel (33.31867, 44.41516) and Palestine Hotel (33.31539, 44.41882).

Adaptive graphics sampling

Using the coordinates of the two hotel entries, the system locates these positions in the map and uses a crop distance to crop the areas around these two point locations with the assumption that each of these two cropped areas contain at least one hotel symbol (Figure 3). In the case study, the crop distance was defined considering the map scale and feature type. Note that the second cropped sample contains more than one hotel symbol because the two hotels are close to each other. These subsections represent the graphics examples that will be used for feature descriptor computation.

Computing feature descriptors and recognition

Next, the system computes the SURF (Speeded Up Robust Features) descriptors (Lowe, 1999; Bay et al., 2006) from the two cropped areas and stores these descriptors as a knowledge base. The SURF descriptors are a type of scale-invariant feature descriptors that capture local "interest" points and their properties for image registration and object recognition. These interest points describe the image intensity at certain pixels and the intensity differences between adjacent pixels. Using the SURF descriptors derived from the samples, the system then scans through the entire map to find sub-sections in the map that contain descriptors with similar values (see Lowe (1999) for details of this object recognition procedure). If certain combinations of descriptors are matched the system labels these locations as hotel instances. This matching process using SURF for extracting map symbols is described in detail in our previous work (Chiang et al., 2014b), but this previous work relies on manually selected samples of cartographic symbols (Figure 4). The use of contextual information to automatically identify locations and extract graphics examples to build up a knowledge base of the features of interest represents an important benchmark test to better understand the potential of context-based recognition approaches.

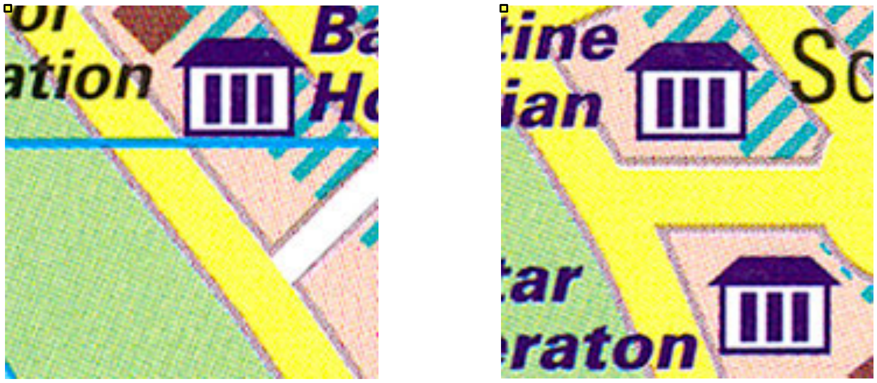


Figure 3. Automatically cropped areas (center of these areas are the locations given in the gazetteer) that contain hotel symbols used as graphics examples to compute descriptors that can be applied to extract all other hotel symbols in the map: Baghdad Hotel (left) and Palestine Hotel (right).

Experimental results

This experiment has been conducted on a scanned map covering Baghdad, Iraq. We purchased the paper map online from Gecko Maps and scanned the map to produce a map image with 600 dot-per-inch resolution. The system extracted 13 hotel locations from the processed map based on two hotel entries found in the GeoNames gazetteer. Out of the 13 extracted hotels, 12 of them were correct (precision 92.3%). The total number of hotels in the map is 17 (i.e., the ground truth), thus the approach missed 5 hotel symbols (recall 70.58%). Figure 5 shows some examples of the extracted symbols. This result shows slightly lower precision and recall than in our previous work which described the same task and test data but involved user intervention to manually label hotel samples in the map (100% and 88.23%, precision and recall, respectively) (Chiang et al., 2014b).

The fact that we were able to develop a fully automatic approach for cartographic symbol recognition using contextual information (here from online data sources) represents a very important step forward in developing more capable recognition systems that scale well for massive data archives. The approach uses existing knowledge of an area (describing some kind of geographic context) to guide the feature sampling and extraction process to eliminate user intervention. The ability to process maps without user intervention is necessary to exploit the large number of existing maps as well as the full richness of large volume digital historical map archives.

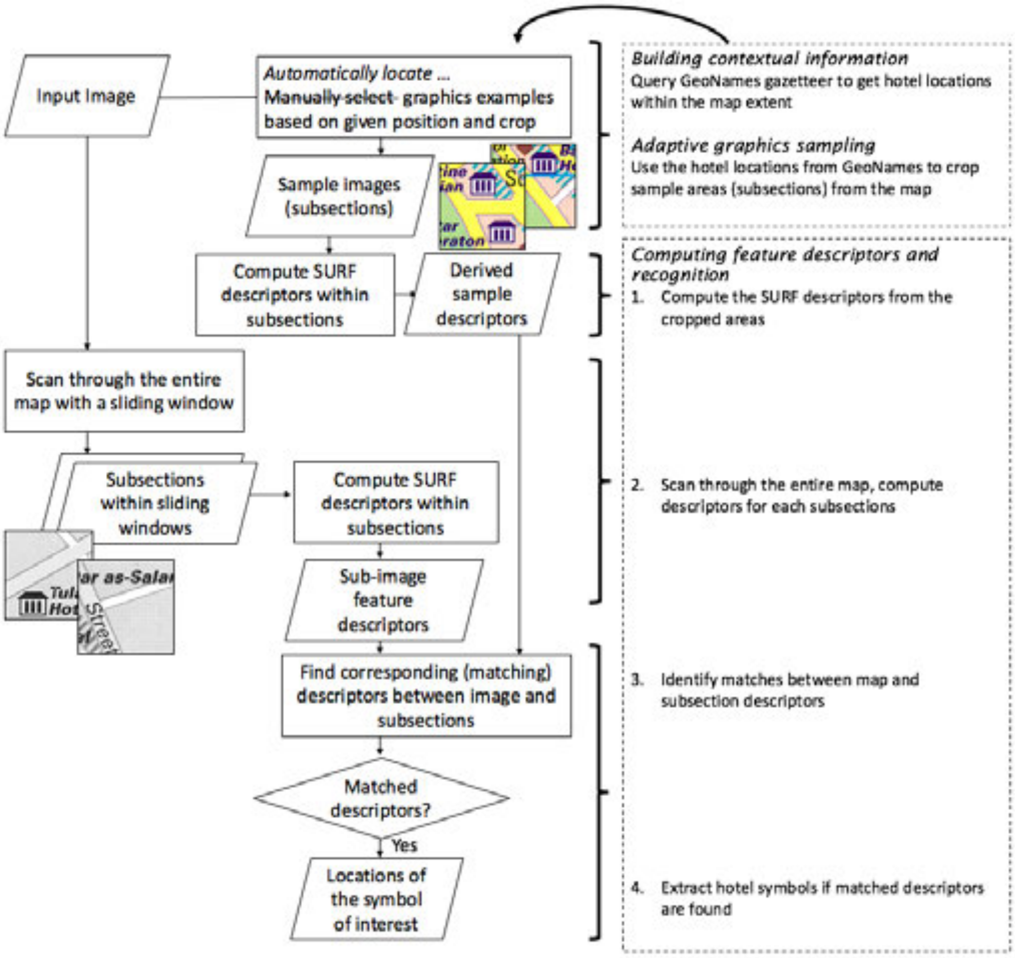
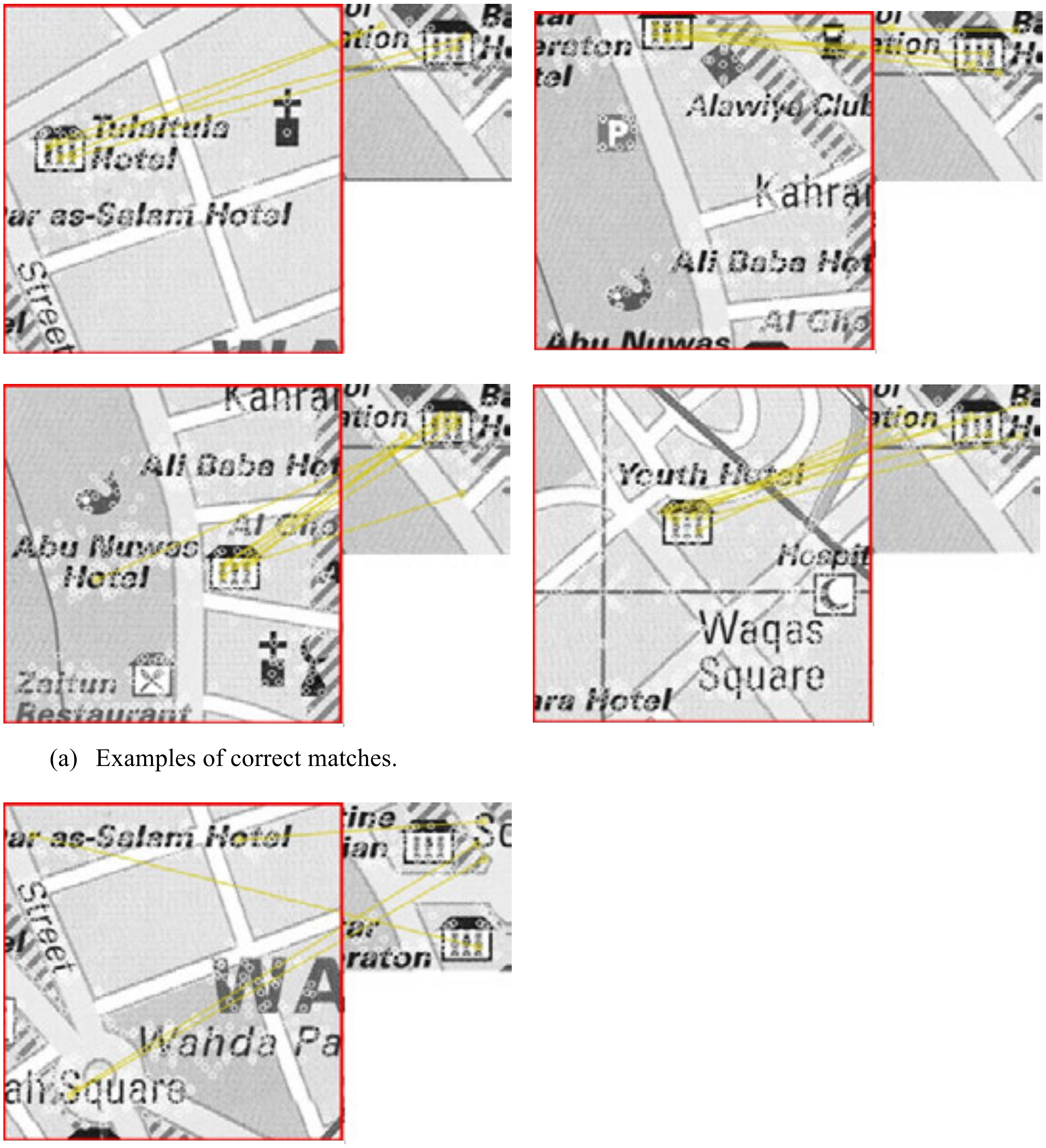

Input Image
Automatically locate ...
~~Manually select~~ graphics examples based on given position and crop
Sample images (subsections)
Compute SURF descriptors within subsections
Derived sample descriptors
Scan through the entire map with a sliding window
Subsections within sliding windows
Compute SURF descriptors within subsections
Sub-image feature descriptors
Find corresponding (matching) descriptors between image and subsections
Matched descriptors?
Yes
Locations of the symbol of interest
Building contextual information
Query GeoNames gazetteer to get hotel locations within the map extent
Adaptive graphics sampling
Use the hotel locations from GeoNames to crop sample areas (subsections) from the map
Computing feature descriptors and recognition
1. Compute the SURF descriptors from the cropped areas
2. Scan through the entire map, compute descriptors for each subsections
3. Identify matches between map and subsection descriptors
4. Extract hotel symbols if matched descriptors are found

Figure 4. The overall process for fully automated extraction of hotel symbols. The process flow on the left shows a traditional approach for feature recognition using the SURF framework (Chiang et al., 2014b). In this case study, we eliminated the need for user-selected samples (model images) and thus fully automated the recognition process.



(a) Examples of correct matches.

(b) The only incorrect match.

Figure 5. Sample extraction results from the hotel recognition case study. The red rectangles indicate the identified map areas with a target symbol. The small image on the right is one of the automatically identified hotel samples. The small white circles on the maps are the locations of the SURF sample. The yellow lines connect matches between the SURF descriptors of the map area and the sample. The samples in (a) show correct identified hotel symbols in four map areas. The sample in (b) shows the only incorrectly identified area.

Discussion

In this study we employed a first experiment to examine how the concept of geographic context can be used for the development of an effective way to fully automate the extraction of geographic information from scanned maps. As the results indicate, this is a promising concept that has the potential to eliminate manual processing, labeling and curating steps while achieving acceptable accuracy. Therefore, the approach of context-based recognition has also potential for processing large-volume map archives but would require a more complex framework of employing geographic context and semantic modeling. There are numerous such map archives (e.g., the Ordinance Survey map archive, The National Library of Scotland) containing millions of scanned maps which cannot be exploited, efficiently, to date.

We will extend these ideas to develop a recognition framework that exploits the fact that map contents are not independent between editions in one map series (e.g., the USGS topographic map series) and change gradually and often cumulatively. We will build generic contextual information (e.g., hydrographic feature types, locations and geometry) using existing contemporary map data (e.g., the USGS National Map layers) and use these contextual data to guide the feature extraction process in the most recent edition of a map similar to what we described in our experiment above including adaptive sampling, computing graphics descriptors and the matching process. The extracted data from the most recent edition can then be used as contextual information for the next (older) map edition in order to carry out the same recognition process. This way, older maps that suffer from low graphical quality can be processed using contextual information from map sheets that are close in time and the results are expected to be more robust than with common recognition methods.

References

Bay, H., Tuytelaars, T. and Gool, L. V., (2006) SURF: Speeded up robust features. In the Proceedings of the 9th ECCV, pages 404–417.

Chiang Y.-Y. and Leyk S. (2015). Exploiting online gazetteer for fully automatic extraction of cartographic symbols. Proceedings of the 27th International Cartographic Conference ICC 2015, Rio, Brazil, August 23-28.

Chiang, Y.-Y, Chioh, P. and Moghaddam, S. (2014b). A Training-by-Example Approach for Symbol Spotting from Raster Maps. In Proceedings of the 8th Geographic Information Science, 2014.

Chiang, Y.-Y. and Knoblock, C.A. (2014). Recognizing text in raster maps. GeoInformatica 19(1):1-27.

Chiang, Y.-Y., Leyk, S. and Knoblock, C. A. (2013). Efficient and robust graphics recognition from historical maps. In Graphics Recognition: Achievements, Challenges, and Evolution, Selected Papers of the 8th International Workshop on Graphics Recognition (GREC), Lecture Notes in Computer Science, 7423, pages 25-35

Chiang, Y.-Y., Leyk, S. and Knoblock, C.A. (2014a). A Survey of Digital Map Processing Techniques. ACM Computing Surveys 47(1): 1-44.

Frischknecht, S. and Kanani, E. (1998). Automatic interpretation of scanned topographic maps: A raster-based approach. In Graphics Recognition Algorithms and Systems, Tombre, K. and Chhabra, A. (eds.). Lecture Notes in Computer Science 1389, pages 207-220.

Khotanzad, A. and Zink, E. (1996). Color paper map segmentation using eigenvector line-fitting. In Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation, pages 190-194.

Khotanzad, A. and Zink, E. (2003). Contour line and geographic feature extraction from USGS color topographical paper maps. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(1):18-31.

Kozak, J., Estreguil, C. and Troll, M. (2007). Forest cover changes in the northern Carpathians in the 20th century: a slow transition. Journal of Land Use Science 2(2):127-146.

Leyk, S. (2010). Segmentation of colour layers in historical maps based on hierarchical colour sampling. In Ogier J.-M., Liu W., and Lladós J. (eds.): Graphics Recognition, GREC 2009, Lecture Notes in Computer Science 6020, pages 231-241.

Leyk, S. and Boesch, R. (2009). Extracting Composite Cartographic Area Features in Low-Quality Maps. Cartography and Geographical Information Science 36(1): 71-79.

Leyk, S., Boesch, R., and Weibel, R. (2006). Saliency and semantic processing: Extracting forest cover from historical topographic maps. Pattern Recognition 39(5): 953 – 968.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In ICCV, vol. 2, pages 1150–1157.

Maderlechner, G. and Mayer, H. (1995). Conversion of high level information from scanned maps into geographic information systems. In Proceedings of the International Conference on Document Analysis and Recognition. Vol. 1, pages 253–256.

Reiher, E., Ying, L., Delle Donne V., Lalonde, M., Hayne, C. and Chona Z. (1996). A system for efficient and robust map symbol recognition. In Proceedings of the 13th International Conference on Pattern Recognition, Vol. 3., pages 783-787.

Stein, E., Dark, S., Longcore, T., Grossinger, R., Hall, N. and Beland, M. (2010). Historical ecology as a tool for assessing landscape change and informing wetland restoration priorities. Wetlands, 30(3):589-601.

Yin, P.-Y. and Huang, Y.-B. (2001). Automating data extraction and identification on Chinese road maps. Optical Engineering 40(5): 663–673.

Stefan Leyk, Department of Geography, University of Colorado at Boulder, Boulder, CO 80309

Yao-Yi Chiang, Spatial Sciences Institute, University of Southern California, Los Angeles, CA 90089