Efficient and Robust Graphics Recognition from Historical Maps

Yao-Yi Chiang,[1] Stefan Leyk,[2] and Craig A. Knoblock[3]
[1] Information Sciences Institute and Spatial Sciences Institute,
University of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292, USA.
yaoyichi@isi.edu
[2] Department of Geography, University of Colorado, UCB260, Boulder, CO 80309, USA.
stefan.leyk@colorado.edu
[3] Department of Computer Science and Information Sciences Institute,
University of Southern California, 4676 Admiralty Way, Marina del Rey, CA 90292, USA.
knoblock@isi.edu

Abstract. Historical maps contain rich cartographic information, such as road networks, but this information is "locked" in images and inaccessible to a geographic information system (GIS). Manual map digitization requires intensive user effort and cannot handle a large number of maps. Previous approaches for automatic map processing generally require expert knowledge in order to fine-tune parameters of the applied graphics recognition techniques and thus are not readily usable for non-expert users. This paper presents an efficient and effective graphics recognition technique that employs interactive user intervention procedures for processing historical raster maps with limited graphical quality. The interactive procedures are performed on color-segmented preprocessing results and are based on straightforward user training processes, which minimize the required user effort for map digitization. This graphics recognition technique eliminates the need for expert users in digitizing map images and provides opportunities to derive unique data for spatiotemporal research by facilitating time-consuming map digitization efforts. The described technique generated accurate road vector data from a historical map image and reduced the time for manual map digitization by 38%.

Keywords: Color image segmentation, road vectorization, historical raster maps, image cleaning

1 Introduction

Maps contain valuable cartographic information, such as locations of historical places, contour lines, building footprints, and hydrography. Extracting such cartographic information from maps (i.e., creating spatial layers that can be processed in a GIS) would support multiple applications and research fields. For example, there are numerous cases in which historical maps have been used to carry out research in land-cover change and biogeography [Kozak et al., 2007; Petit and Lambin, 2002], and urban-area development [Dietzel et al., 2005].

Today, thousands of such maps and map series are available in scanned raster format (i.e., digital map images) in a variety of digital archives. Previous work on extracting cartographic information from raster maps typically requires intensive user intervention for training and parameter tuning, in particular, when processing historical maps of poor graphical quality [Gamba and Mecocci, 1999; Leyk and Boesch, 2010].

Consequently, most studies that utilize cartographic information from historical maps for their analysis are based on time-consuming manual map digitization, which introduces subjectivity because of the missing reproducibility and the limited number of maps that can be digitized. More advanced semi- or fully- automated procedures for cartographic information extraction from historical maps would allow the advancement of such studies by including historical spatial data that are derived from a variety of maps, underlie repeatable procedures with reproducible results, and cover large areas.

In this paper, we present an interactive graphics recognition technique based on straightforward user training processes to minimize the required user effort for processing raster maps, especially for the raster maps that have limited graphical quality. We demonstrate this approach using a historical U.S. Geological Survey (USGS) topographic map in a road vectorization example, which builds on our previous work on color image segmentation (CIS) [Leyk, 2010] and road vectorization [Chiang et al., 2008; Chiang and Knoblock, 2011]. The USGS topographic map suffers from poor graphical quality and thus represents a particularly challenging research object.

The remainder of this paper is organized as follows. Section 2 discusses related work. Section 3 presents our graphics recognition technique using an example of road vectorization from a historical USGS topographic map. Section 4 reports on our experimental results. Section 5 presents the conclusions and future work.

2 Recent Work

Here we provide a brief review of recent map processing research; a more detailed review can be found in [Chiang, 2010; Chiang and Knoblock, 2011]. Extracting cartographic information from raster maps is challenging due to poor image quality caused by scanning or image compression processes, as well as the aging of the archived paper material, which often causes effects of false coloring, blurring or bleaching [Chen et al., 2006; Dhar and Chanda, 2006; Gamba and Mecocci, 1999]. The complexity of raster map contents increases if there are overlapping map layers of geographic features, such as roads, contour lines, and labels in different or similar colors. Color image segmentation has been investigated as a preprocessing step to separate individual map color layers [Chen et al., 2006; Chiang, 2010; Leyk, 2010; Leyk and Boesch, 2010]. However, there are still limitations when processing poor quality images of historical maps [Leyk, 2010].

Much research has been devoted to extracting geographic features from particular map series. For example, Itonaga et al. [2003] describe a road-line vectorization approach from computer-generated maps that cannot be applied to scanned maps. Dhar and Chanda [2006] extract geographic features from Indian survey maps based on user-specified image processing filters that exploit the geometric properties of these features. Another exemplary approach that is highly customized to a specific map series is the work by Raveaux et al. [2008] on extracting quartiers from historical French cadastral maps. As with many other recent approaches, the described extraction procedures from Itonaga et al. [2003], Chanda [2006], and Raveaux et al. [2008] have limited applicability to other map series or types and require expert knowledge for parameter tuning.

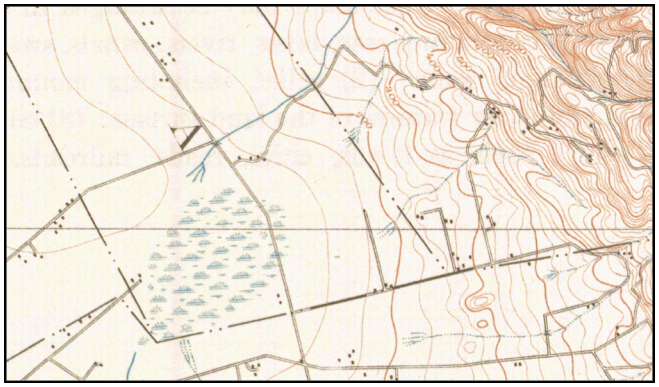


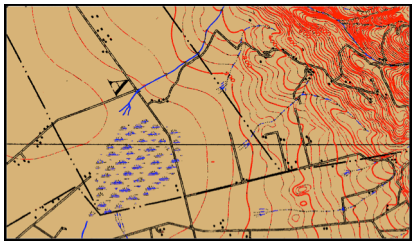
Fig. 1. A sample USGS historical topographic map

3 Methods: Road Vectorization from a Historical USGS Topographic Map

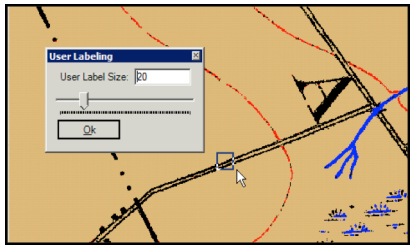
We demonstrate our approach using a road vectorization example. Figures 1 shows a sample USGS historical topographic map. Typical problems existed in old maps (and can be seen here too) that limit the graphical image quality are bleaching and blurring effects as well as mixed or false coloring as consequences of archiving paper material and scanning procedures. Figure 2 illustrates the different steps and the user interface of the described technique. Our graphic recognition approach consists of three major steps: (i) separation of homogeneous thematic map layers using color image segmentation (CIS) [Leyk, 2010], (ii) interactive extraction and cleaning of these separated map layers, and (iii) subsequent raster-to-vector conversion of the cleaned map layers [Chiang et al., 2008; Chiang and Knoblock, 2011].

3.1 Color Image Segmentation

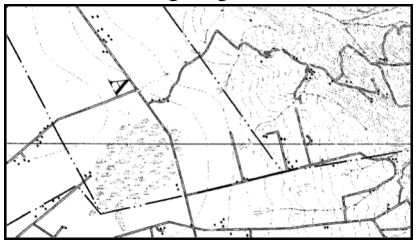
As an important preprocessing step, color image segmentation (CIS) separates thematic homogeneous colored map layers. CIS is of critical importance since the outcome directly determines the image processing methods to be applied in all subsequent stages of data extraction. In our previous work, a hierarchical CIS approach based on homogeneity computation, color space clustering, and iterative global/local color prototype matching has been implemented and tested on historical USGS topographic maps (Figures 1 shows a sample map) [Leyk, 2010]. The only input parameters for this CIS approach are the number and types of color layers. The approach has been improved for this paper to overcome some reported limitations in the final region-growing step, such as merging of nearby elevation or road line features. Constraining the final segments to maximum widths and enforcing connectivity between homogeneous portions was incorporated in order to avoid such merging effects and thus improve the final CIS output.



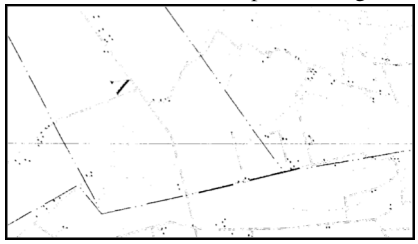
(a) The color image segmentation result



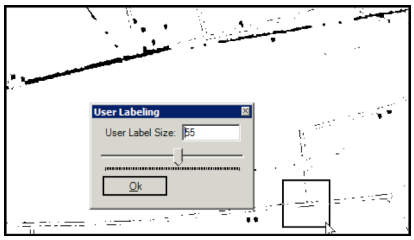
(b) A user label of an exmple road segment



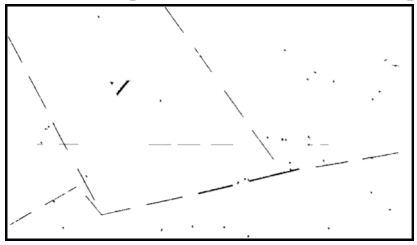
(c) The extracted road layer with noise pixels



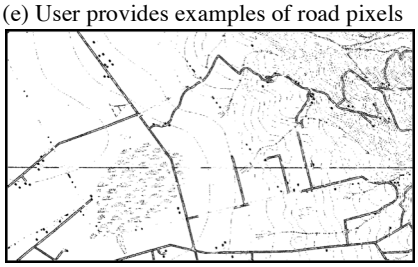
(d) Erosion operator to remove most road pixels



(e) User provides examples of road pixels



(f) Remaining large noise objects



(g) Large noise objects are removed

(h) User provides examples of small noise objects

(i) Noise objects are removed

(j) Raw road vectorization results

Fig. 2. Road layer cleaning with minimal user labeling

Figure 2(a) shows the result of CIS of the original USGS topographic map portion (Figure 1). While the performance of the segmentation is visually satisfying (i.e., individual map layers are shown in unique colors), there are some remaining merging and mixing effects at places with dense elevation contours and road lines as well as gaps in road lines where they intersect with elevation contours; these problems are caused by issues of graphical quality as described earlier. Although this CIS procedure could be further tuned to provide better results, the raw and unrepaired segmentation outcome is used to test the robustness of the subsequent cleaning and road vectorization steps when using a general non-post-processed (and thus sub-optimal) CIS outcome.

3.2 Road Layer Extraction and Cleaning by Example

Road Layer Extraction by Example. To extract the road layer (the black layer in the CIS result) from the segmented map, first one sample area (approximately) centered on a road line has to be identified by a user. Figure 2(b) shows the user interface for the road-vectorization approach. Using this interface a user labels a rectangular area of 20-by-20 pixels to provide a "road sample." The thematic map layer that has the same color as the identified road lines (i.e., the road-like features) are automatically identified in the entire map image and the road type (e.g., parallel lines or single lines) and road width are determined [Chiang and Knoblock, 2011]. This step is carried out in a rotation-invariant way. Figure 2(c) shows the identified road layer. In this example, the majority of the detected linear features have a width of one pixel.

Road Layer Cleaning By Example. The identified road layer might contain non-road pixels depending on the quality of the input maps and the CIS result. Therefore, prior to generating the road vector data, the road layer has to be cleaned. Commercial products for raster-to-vector conversion, such as R2V[1] and Vextractor,[2] generally include image-processing tools, such as morphological or structural operators that can be selected and adjusted by the user to manually clean the raster image. However, this manual process requires expert knowledge and is very time consuming. In order to overcome such limitations, in this paper, we present an interactive technique, which incorporates simple user training processes for removing the undesired (non-road) pixels. This technique exploits user provided "noise samples" in order to identify appropriate image-processing filters and to generate parameter sets. These parameterized image-processing filters are then applied to efficiently clean the map image and remove existing noise.

In this road vectorization example, the first step of our cleaning process removes large non-road objects (noise objects that are thicker than road lines); the second step removes small non-road objects. To identify large non-road objects in the road-layer-extraction result (Figure 2(c)), first, an image (Figure 2(d)) is created in which the majority of road lines are eliminated by applying the erosion operator and thus non-road objects remained. The number of iterations for erosion is determined by the detected road width (one pixel according to the one collected sample) [Chiang and Knoblock, 2011].

---

[1] http://www.ablesw.com/r2v/

[2] http://www.vextrasoft.com/vextractor.htm

The identified large noise objects (Figure 2(d)) contain some road pixels that are not eliminated by the erosion operator. This is because some of the road lines are thicker than the identified road width. To remove these road pixels from the identified large noise objects, the user provides an example containing road pixels that are not removed by the erosion operator (Figure 2(e)). The connected components in Figure 2(d) that have a similar length, width, and number of pixels as the ones in the new user label (Figure 2(e)) are removed automatically; they represent parts of road features and will be preserved. Figure 2(f) shows the remaining large noise objects that exceed the detected (sampled) typical road width. Removing the objects in Figure 2(f) from Figure 2(c) results in the image shown in Figure 2(g) which contains road pixels and small noise objects.

To remove the remaining small noise objects in Figure 2(g), the user again provides local areas containing samples of the objects in question (Figure 2(h)). Connected components that show similar sizes as the ones in the samples are then removed from Figure 2(g) in order to generate the final cleaning result (Figure 2(i)).

3.3 Road Layer Vectorization

Once the road layer has been cleaned, we employ our previously described technique for automatic generation of road geometry [Chiang et al., 2008] and subsequently convert the road geometry to road vector data automatically [Chiang and Knoblock, 2011]. Figure 2(j) shows the road vectorization results (without manual post-processing) consisting of 1-pixel wide road centerlines.

In this road vectorization example, the horizontal map grid line is not removed during the cleaning process and the vectorization result thus contains a portion of these grid lines. Additional operators would be needed or manual post-processing (manually edit the road vector data) would have to be done to remove such elements. Also some broken road lines can be observed, which indicates the need to refine the procedure or to manually edit the final data layer. Such post-processing steps are generally needed for both semi- and fully- automatic approaches to process maps of limited image quality.

4 Experiments

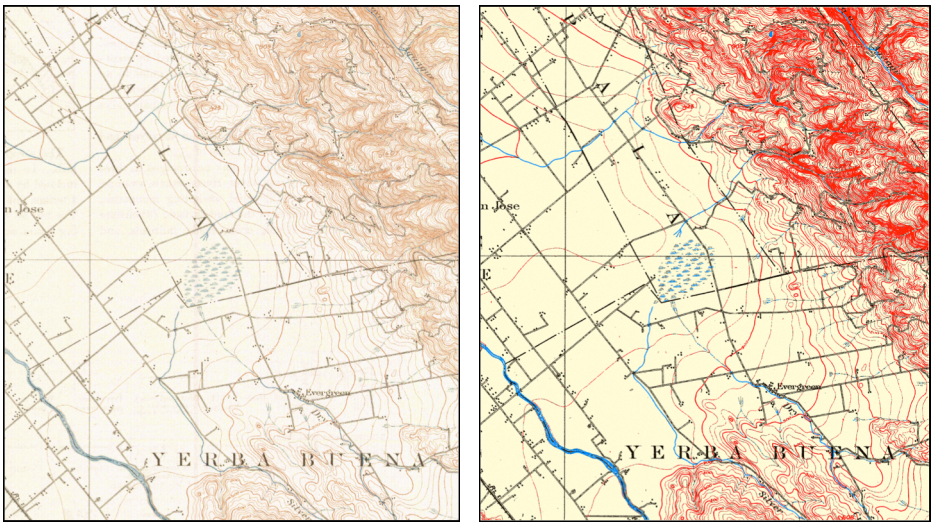
In this section our experiments are described in which the proposed approach is tested for extracting road vector data from a historical USGS topographic map that covers the city of St. Jose, California.[3,4] Figure 3(a) shows the original map tile (2283 × 2608 pixels). We compared the interactive recognition technique with a completely manual map digitization process by (i) the required manual processing time and (ii) the accuracy of extracted road vector data. The required manual processing time for our approach includes the time for user labeling and manual post-processing. We use Esri

[3] USGS Topographic Maps: Map page, San Jose, California (San Jose Quadrangle); edition of 1899 (reprint 1926), engraved in July 1897 by USGS; scale: 1/62,500

[4] The detailed information for obtaining the test map and the ground truth can be found on: http://www.isi.edu/integration/data/maps/prj_map_extract_data.html

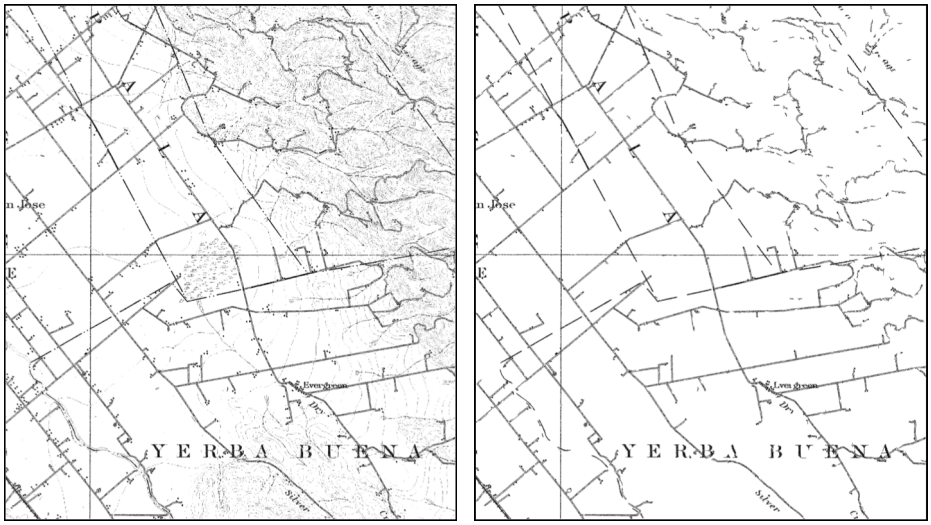
ArcMap[5] for both our manual post-processing step and the compared manual digitization work.

Given the test map tile, a user first specified the number of desired map layers. Then the segmented map with separated color layer was generated based on CIS (Figure 3(b)). The user-specified number of desired map layers was 4 in this case:



(a) The test map tile (b) The CIS result

Fig. 3. The test map tile and intermediate results



(a) The extracted road layer (b) The cleaned road layer

Fig. 4. The test map tile and intermediate results

[5] http://www.esri.com/

Table 1. Comparison of the required user time for road vectorization from the test map

Methods	Task	Manual Work Time (Hours: Minutes: Seconds)
Our Interactive Approach	User labeling	0:0:33
	Manual post-processing	0:49:52
Completely Manual Approach	Manual Digitization	01:21:24

Table 2. Accuracy of the extracted road vector data from our interactive approach using the manual digitization results as the ground truth

Road Buffer	Completeness	Correctness	Redundancy	RMS
1 pixel	99.9%	100%	0.054%	0.152 pixels
2 pixels	100%	100%	0.14%	0.152 pixels

black, red, blue and white (background). Within the segmented map, the user selected 1 sample area of roads and 2 sample areas of noise objects to generate a cleaned road layer. Figure 4(a) shows the extracted road layer with some of the remaining non-road pixels, and Figure 4(b) shows the cleaned result after removing these noise pixels. Some of the broken road lines were also eliminated, accidentally. The cleaned road layer was then automatically converted to a set of road vector line features. Finally, the user manually edited the road vector data (e.g., recover the broken lines) to achieve a complete road vectorization result.

Table 1 shows the required user time for our approach and the completely manual digitization. Our approach required a total of 50 minutes and 33 seconds to vectorize the road network from the test map tile. The computing time for road-layer-extraction was 3 seconds, for road-layer-cleaning 52 seconds, and for road vectorization 56 seconds. In contrast, complete manual digitization required 1 hour, 21 minutes, and 24 seconds. Our approach reduced the digitization time by ***38%***. The manual post-processing time in our approach was mainly spent with creating missing lines in the top-right corner of the map tile and connecting broken lines. The completely manual digitization required more than 1 hour because manual tracing of the exact road centerlines requires precise mouse movement on the pixel level.

We also compared the extracted road vector data from the two approaches based on the completeness, correctness, quality, redundancy, and the root-mean-square (RMS) difference (the average distance between the extracted lines and the ground truth) [Heipke et al., 1997; Chiang and Knoblock, 2011]. We use the manual digitization results as ground truth to compute the described metrics from the road vector data generated by our interactive approach.

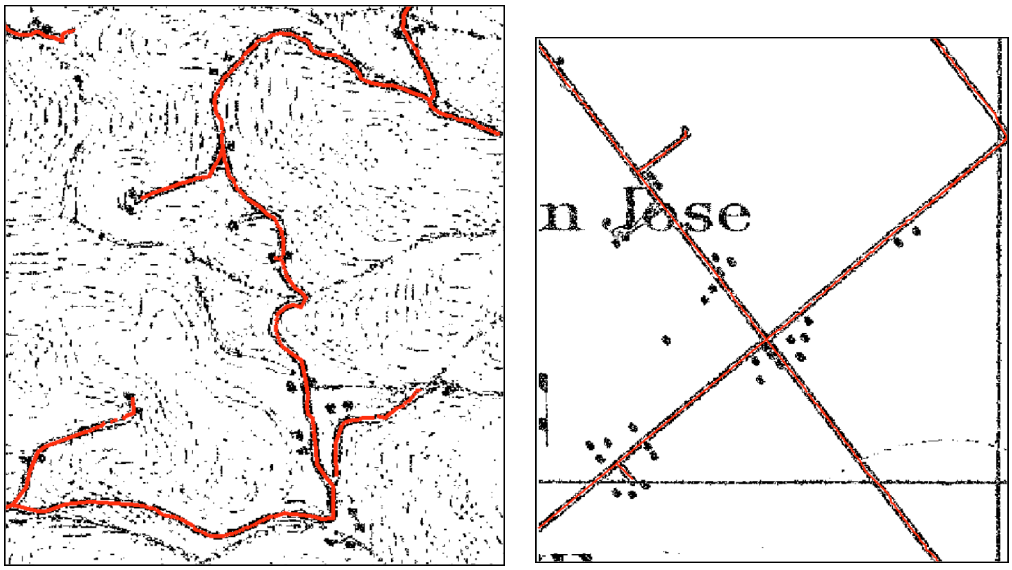


Fig. 5. Samples of our road vectorization results

The completeness is the length of true positives divided by the sum of the lengths of true positives and false negatives, and the optimum is 100%. The correctness is the length of true positives divided by the sum of the lengths of true positives and false positives, and the optimum is 100%. The redundancy is the length of matched extraction minus the length of matched reference. The redundancy shows the percentage of the matched ground truth that is redundant (i.e., more than one true positive line matched to one ground-truth line), and the optimum is 0. The RMS difference is the average distance between the extracted lines and the ground truth, which represents the geometrical accuracy of the extracted road vector data. To identify the length of the true positives, false negatives, and matched ground truth, we used buffer widths of 1 and 2 pixels. This means that a correctly extracted line is no farther than 1 or 2 pixels from the ground truth.

Table 2 shows the accuracy of our extracted road vector data using the manual digitization results as the ground truth. Our approach generated accurate results (high completeness and correctness) at both 1-pixel and 2-pixel buffer sizes. In addition, our results contain very few redundant road lines and are very close to the ground truth regarding their geometry (i.e., low RMS). Figure 5 shows details of portions of the final road vector data from our approach on top of the extracted raster black layer. Note that the resultant road vector data are very close to the original road centerlines.

6 Conclusions and Future Work

We presented an efficient and robust approach for graphics recognition from historical raster maps. We showed that this approach minimizes the required user intervention and reduces the manual digitization time for road vectorization from a historical USGS topographic map by 38%. The presented technique shows high potential for robust

extraction of cartographic information from historical maps of low graphical quality and opens unique opportunities for "spatio-historical" research in various fields.

We plan to improve our graphics recognition technique by further refining and constraining the final region-growing step in the CIS to reduce the number of noise objects in the CIS result. Moreover, we plan to incorporate connectivity and additional geometry constraints in the image-cleaning-by-example step to further reduce the required manual processing time.

References

Chen, Y., Wang, R., Qian, J. (2006) Extracting contour lines from common-conditioned topographic maps. *IEEE Transaction Geoscience and Remote Sensing*, 44(4), 1048–1057.

Chiang, Y.-Y. (2010). Harvesting Geographic Features from Heterogeneous Raster Maps. *Ph.D. thesis*, University of Southern California.

Chiang, Y.-Y., Knoblock, C. A., Shahabi, C., and Chen, C.-C. (2008). Automatic and accurate extraction of road intersections from raster maps. *GeoInformatica*, 13(2):121-157.

Chiang, Y.-Y and Knoblock, C. A. (2011). General Approach for Extracting Road Vector Data from Raster Maps. *International Journal on Document Analysis and Recognition*, 2011.

Dhar, D. B. and Chanda, B. (2006). Extraction and recognition of geographical features from paper maps. *International Journal on Document Analysis and Recognition*, 8(4): 232-245.

Dietzel, C., Herold, M., Hemphill, J.J. and Clarke, K.C. (2005). Spatio-temporal dynamics in California's Central Valley: Empirical links to urban theory. *International Journal of Geographical Information Science*. 19(2):175-195.

Heipke, C., Mayer, H., Wiedemann, C., and Jamet, O. (1997). Evaluation of automatic road extraction. *International Archives of Photogrammetry and Remote Sensing*, 32: 47–56.

Gamba P. and Mecocci A., (1999). Perceptual Grouping for Symbol Chain Tracking in Digitized Topographic Maps, *Pattern Recognition Letters,* 20(4): 355-365.

Itonaga, W., Matsuda, I., Yoneyama, N., and Ito, S. (2003). Automatic extraction of road networks from map images. *Electronics and Communications in Japan*, 86(4):62-72.

Knoblock, C. A., Chen, C., Chiang, Y.-Y., Goel, A., Michelson, M., and Shahabi, C. (2010). A General Approach to Discovering, Registering, and Extracting Features from Raster Maps. In Proceedings of the *Document Recognition and Retrieval XVII of SPIE-IS&T Electronic Imaging*, vol 7534.

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 *GREC. LNCS 6020*, pages 231–241.

Leyk S. and Boesch R. (2010). Colors of the Past: Color Image Segmentation in Historical Topographic Maps Based on Homogeneity. *GeoInformatica* 14(1): 1-21.

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

Petit, C.C. and Lambin, E.F.. (2002): Impact of data integration technique on historical land-use/land-cover change: Comparing historical maps with remote sensing data in the Belgian Ardennes. *Landscape Ecology* 17(2), 117-132.

Raveaux, R., Burie, J.-C., and Ogier, J.-M. (2008). Object extraction from colour cadastral maps. In *Proceedings of the IAPR Document Analysis Systems*, pages. 506-514.