Integrating Color Image Segmentation and User Labeling for 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

Keywords- color image segmentation, road vectorization, historical raster maps, image cleaning

I. 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 [10, 14], or urban-area development [6].

Today, thousands of such maps and map series are available in 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 [7, 13]. Consequently, most studies on analyzing and using cartographic information from historical maps are based on time-consuming manual map digitization, which introduces subjectivity because of the limited numbers 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 cover large areas, are derived from a variety of maps, and underlie reproducible procedures.

In this paper, we describe a first demonstration of an interactive approach for cartographic information extraction from raster maps that have limited graphical quality and contain thematic color layers. In particular, this approach integrates an efficient and effective image-cleaning procedure that is based on simple user training processes with recently described techniques of Color Image Segmentation (CIS) [11] and road vectorization [3, 4] for processing historical maps. We demonstrate this approach on the extraction of road features from a historical topographic map, which suffers from poor graphical quality and thus represents a particularly challenging research object.

II. RECENT WORK

Here we provide a brief review of recent map processing research; a more detailed review can be found in [2]. Extracting cartographic information from raster maps is challenging due to poor image quality that can be 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 [1, 5, 7]. 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 different map color layers [1, 2, 11, 12], but there are still limitations when processing poor quality images of historical maps [11].

Much research has been devoted to extracting geographic features from particular map series. For example, Itonaga et al. [8] describe a road-line vectorization approach from computer-generated maps that cannot be applied to scanned maps. Dhar and Chanda [5] extract geographic features from Indian survey maps based on user-specified filters that exploit the geometric properties of these features. As with many other approaches their extraction procedure has limited applicability to other map series. Another exemplary approach that is highly customized to a specific map series is the work by Raveaux et al. [15] on extracting quartiers from historical French cadastral maps.

III. METHOD OVERVIEW

Our interactive approach for cartographic information extraction, which is described here using the example of road vectorization, contains two major steps: (i) the separation of homogeneous thematic map layers using Color Image Segmentation (CIS) [11]; (ii) the cleaning of these separated map layers (i.e., identifying and removing noise pixels from the CIS results) and the subsequent raster-to-vector conversion of the cleaned map layers [3, 4]. These steps allow the generation of spatial data in vector format, which then can be registered in time and space according to the information on the map and used in a GIS for spatiotemporal analysis.

A. Color Image Segmentation

As an important preprocessing step, Color Image Segmentation (CIS) separates thematic homogeneous color 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 (Figure 1) [11]. The only input parameters for this CIS approach are the number and type of color layers. The approach has been improved for this case study to overcome some limitations in the final region-growing step, which led to problems in the CIS output e.g., merging of nearby features such as elevation contours or roads. Constraining the final segments to a maximum dimensionality and connectivity tests prevented a large proportion of such merging effects.

B. Road Layer Extraction, Cleaning, and Vectorization

To extract the road layer from the segmented map, first a sample area (approximately) centered on a road line has to be labeled (Figure 2(a)). The thematic map layer that has the same color as the sampled 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 (Figure 2(b)) [4]. This step is carried out in a rotation-invariant way.

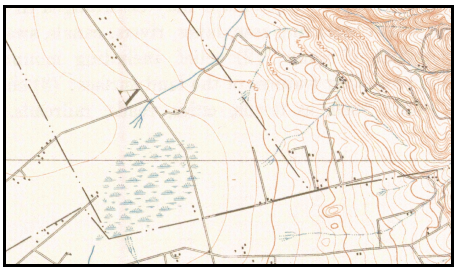
The identified road layer might contain non-road pixels depending on the quality of the input maps. Therefore, prior to generating the road vector data, the road layer has to be cleaned. Commercial products for raster-to-vector conversion, such as R2V[1] and Vextractor,[2] include image-processing tools, 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 highly time demanding. In order to overcome such limitations, in this case study, we use 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 tools and generate parameter sets. These parameterized tools are then applied to achieve acceptable cleaning results. The user training process is demonstrated for an example of road layer extraction in the next section.

Once the road layer has been cleaned, we exploit our previously described technique for automatic generation of the road geometry [3] and subsequently convert the road geometry to road vector data [4].

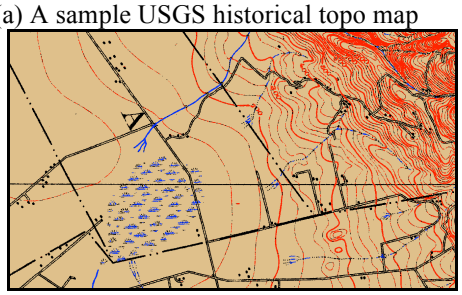
IV. CASE STUDY: ROAD VECTORIZATION FROM A HISTORICAL USGS TOPOGRAPHIC MAP

We demonstrate our approach using a road vectorization example. Figures 1-3 illustrate the different steps and the user interface of the described technique.

Color Image Segmentation: Figure 1(b) shows the result of CIS of the original USGS topographic map (Figure 1(a)). As can be seen, the segmentation successfully separates the color layers, but there are still some remaining merging effects of dense elevation contours and road lines as well as missing road pixels when road pixels intersect with elevation contours – these are caused by issues of graphical quality as described earlier. Although this procedure could be tuned to provide better results, the raw and unrepaired segmentation outcome was used to test the robustness of the subsequent cleaning and the road vectorization steps when using a general non-post-processed (and thus sub-optimal) CIS outcome.

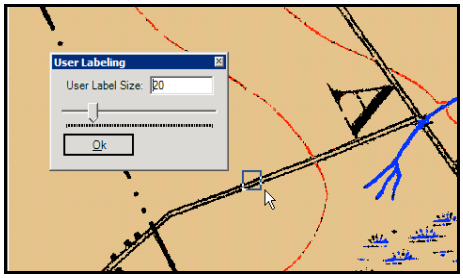


(a) A sample USGS historical topo map



(b) The color image segmentation result

Figure 1. Color Image Segmentation



(a) A user label of an exmple road segment

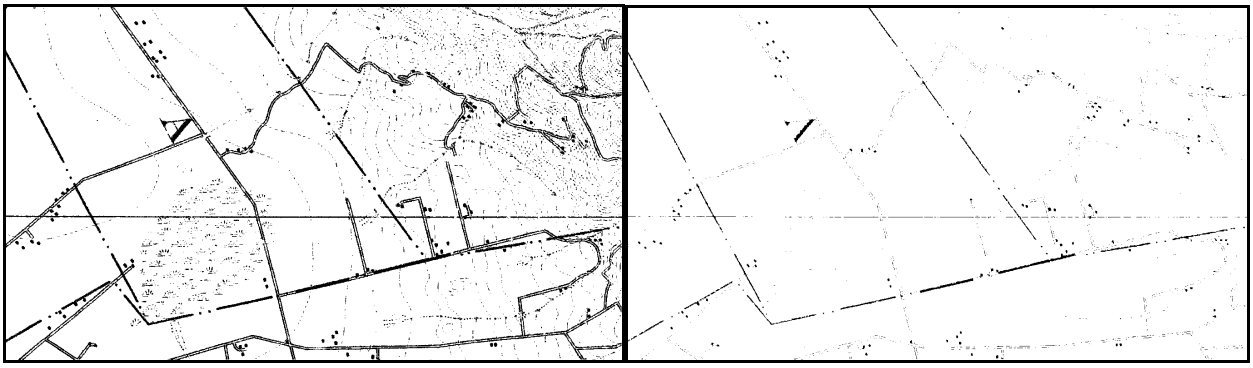
(b)The extracted road layer with noise pixels

Figure 2. Supervised road layer extraction

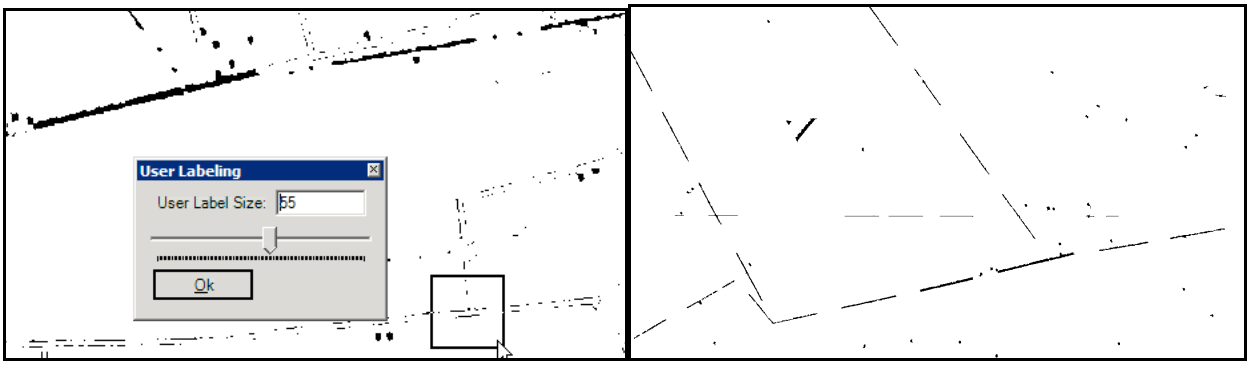
Road Layer Extraction: Figure 2(a) shows the user interface for the road-vectorization approach. Using this interface a user labels a rectangular area of 20 pixels to provide a "road sample". Figure 2(b) shows the identified road layer. In this example, the majority of the detected linear features have a width of one pixel.

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

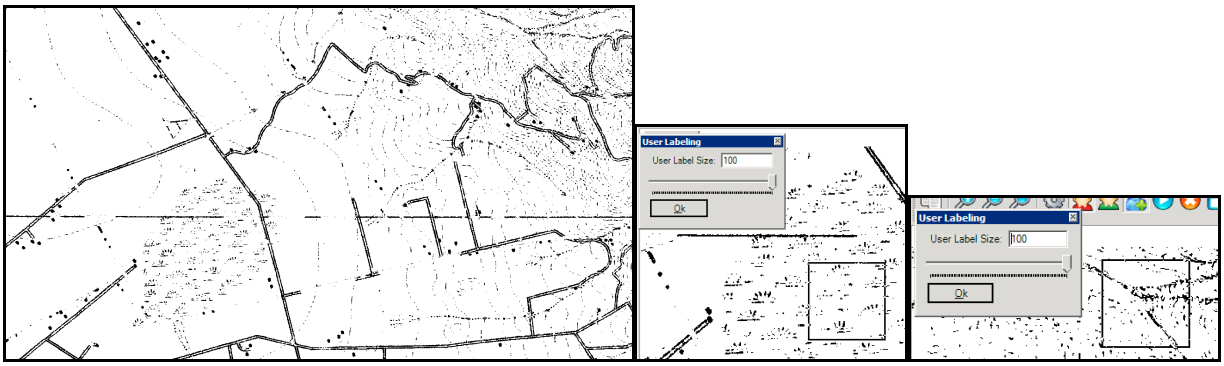
[2] http://www.vextrasoft.com/vextractor.htm



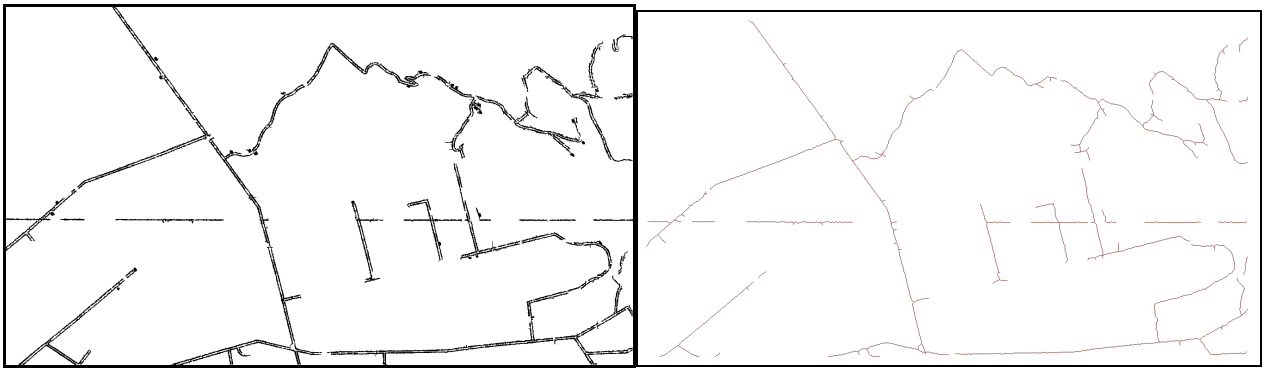
(a) Input for the cleaning process (b) Erosion operator to remove most road pixels



(c) User provides examples of remaining road pixels (d) Large noise objects i.e., thicker than road lines



(e) Large noise objects are removed (f) User provides examples of small noise objects



(g) Cleaning result: noise objects are removed (h) Raw road vectorization results

Figure 3. Road layer cleaning with minimal user labeling

Road Layer Cleaning: The first step of our cleaning process removes large non-road objects; the second step removes small non-road objects. To remove large non-road objects from the road-layer-extraction result (Figure 3(a)), first, an image (Figure 3(b)) was created in which the majority of road lines were eliminated and thus non-road objects remained. The majority of road lines were eliminated by applying the erosion operator. The number of iterations for erosion was determined by the detected road width (one pixel according to the one collected sample shown). Next, the user provides an example containing road pixels that are not removed by the erosion operator (Figure 3(c)). The connected non-road components in Figure 3(b) that have a similar length, width, and number of pixels as the ones in the new user label (Figure 3(c)) are also eliminated automatically; they represent parts of road features and will be preserved. Figure 3(d) shows the remaining large noise objects that exceed the detected (sampled) typical road width. Removing the objects in Figure 3(d) from Figure 3(a) results in the image shown in Figure 3(e).

To remove the remaining small noisy objects in Figure 3(e), the user provides local areas containing samples of the objects in question (Figure 3(f)). Connected components that show similar sizes as the ones in the samples were then removed from Figure 3(e) and the final cleaning result was processed (Figure 3(g)).

Road Layer Vectorization: The final road vector data layer (Figure 3(h)) was generated using the vectorization technique described in [4] with the cleaning result (Figure 3(g)) as input. Note that since the horizontal map grid line was not removed in the cleaning process, the vectorization result contains a portion of these grid lines. Additional operators would be needed or manual post-processing (e.g., manually edit the road vector data) would have to be carried out to remove such elements. Also some broken road lines can be observed, which indicates a need to refine the procedure or to manually edit the final data layer. Such post-processing steps are commonly required even for conventional approaches, which are applied for common-conditioned maps and could not handle maps of such limited image quality.

V. Discussion and Conclusion

We present the integration of a Color Image Segmentation (CIS) step with an interactive road-layer extraction process that consists of an image cleaning and a vectorization step. We describe a case study to demonstrate the performance of this integrated approach, which minimizes user effort for generating road vector data from historical raster maps. Our final vectorization is based on an efficient user intervention strategy: the CIS requires only the input of the numbers of thematic layers in the historical map; the interactive extraction technique takes only 4 user labels as shown in Figures 2 and 3.

The described shortcomings in the CIS will be improved by further refining and constraining the final region-growing step; the road extraction process will incorporate connectivity constraints of the road lines and additional image processing operators to increase the robustness of the final result. 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.

References

[1] Chen, Y., Wang, R., Qian, J. (2006) Extracting contour lines from common-conditioned topographic maps. IEEE Trans. Geosci. Rem. Sens. 44(4), 1048–1057.

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

[3] 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.

[4] Chiang, Y.-Y and Knoblock, C. A. (2010). Extracting Road Vector Data from Raster Maps. In *GREC. LNCS 6020*, pp. 93–105.

[5] Dhar, D. B. and Chanda, B. (2006). Extraction and recognition of geographical features from paper maps. *IJDAR*, 8(4): 232-245.

[6] 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.

[7] Gamba P. and Mecocci A., Perceptual Grouping for Symbol Chain Tracking in Digitized Topographic Maps, Pattern Recognition Lett. 20 (1999) 355-365.

[8] 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.

[9] 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.

[10] 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.

[11] Leyk S. (2010). Segmentation of Colour Layers in Historical Maps based on Hierarchical Colour Sampling. In *GREC. LNCS 6020*, pp. 231–241.

[12] 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.

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

[14] 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.

[15] Raveaux, R., Burie, J.-C., and Ogier, J.-M. (2008). Object extraction from colour cadastral maps. In *Proceedings of the IAPR DAS*, pp. 506-514.