COOPERATIVE TEXT AND LINE-ART EXTRACTION FROM A TOPOGRAPHIC MAP

Luyang Li, Panasonic Information and Networking Laboratories, Princeton, NY, lly@research.panasonic.com

George Nagy , Rensselaer Polytechnic Institute, Troy, NY, nagy@ecse.rpi.edu

Ashok Samal & Sharad Seth, University of Nebraska - Lincoln, Lincoln, NE, (samal,seth)@cse.unl.edu

Yihong Xu, Hewlett Packard Laboratories, Palo Alto, CA, xuy@hpl.hp.com

Abstract

The black layer is digitized from a USGS topographic map digitized at 1000 dpi. The connected components of this layer are analyzed and separated into line art, text, and icons in two passes. The paired street casings are converted to polylines by vectorization and associated with street labels from the character recognition phase. The accuracy of character recognition is shown to improve by taking account of the frequently occurring overlap of line art with street labels. The experiments show that complete vectorization of the black line-layer bitmap is the major remaining problem.

1. Introduction

This paper presents new results obtained since the conclusion of the NIMA's Intelligent Map Understanding project but it draws heavily on methods discussed in detail in [5,7,9,10,13,15].

The novel aspect documented here is to demonstrate the effect of cooperative processing of text and line art. This is necessary because the text and line art are often overlapped and recognition results on one can help recognition of the other. We use information from the street lines to locate and orient label boxes and to remove overlaying line segments, and use the output of the character recognition system to refine the street-line sublayer.

We review briefly relevant sources of information. The definitive authority on text-graphics separation is Kasturi [3,6,14]. There are many papers on map conversion in general, with a particularly valuable early paper by a practitioner, Rhind [12]. Good examples of recent projects are [4,8]. A recent survey on vectorization is [1].

2. Data

The source map for all of the following experiments was WASHINGTON EAST (D.C.) QUADRANGLE. The map has a scale of 1:24,000 and elevation contour intervals of 10 feet.

The USGS 7.5-minute series of topographic maps is typical of the best in traditional cartography and packs an enormous amount of information for diverse uses. The extraction of data from such a map is correspondingly more difficult than from specialized maps like cadastral and road maps, or from map separates (overlays) and requires high-resolution digitization..

The focus here is on street lines and labels in the black layer (Fig. 1). In addition to street casings 0.1 mm wide and separated by 0.5 mm, the black layer graphics include solid blocks showing buildings in non-built-up areas like parks; black outlines for running tracks; thick railroad lines; dashed political boundary lines; and straight grid lines.

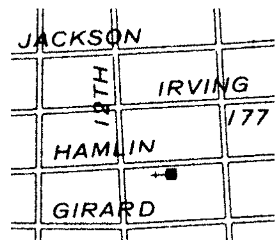

Fig. 1. Street label and street casings

The street labels are slanted sans-serif caps 1.25 mm high or, for principal streets, 1.4 mm high. Text in other fonts denotes districts, monuments, schools, elevation benchmarks, etc. The street labels are typically oriented from left to right for E-W streets, and from bottom to top for N-S streets for viewing from the right. Street labels frequently overlay intersecting street casings, and occasionally touch other icons.

3. Methodology

The interaction between the line art and text processing at several points in the data flow is shown in Fig. 2. The initial processing examines every pixel only once to recover the black layer and extract its connected components. Then the candidate constituents of street names and street casings are identified. The initial classification is, however, error prone and the resulting errors affect subsequent OCR and vectorization. The results can be improved by detailed examination of the vicinity of identified character candidates. The local processing consists of two interlaced steps: character grouping and line structure recovery.

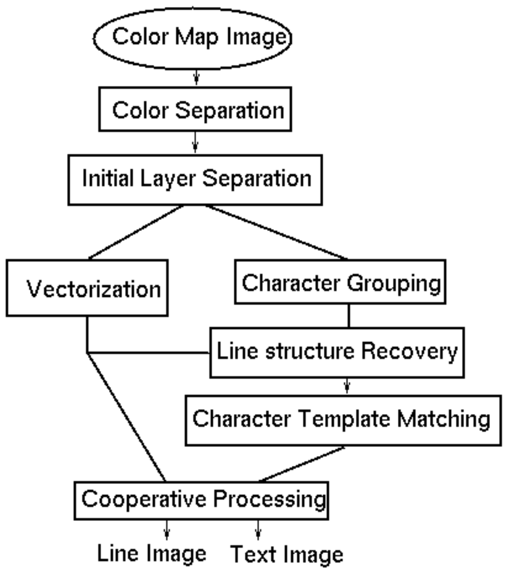

Fig. 2. Sublayer separation

3.1. Character Grouping

Character grouping is based on the character neighborhood connected components. Characters are assembled into word boxes according to the constraint that they lie on a straight line and their centroids are separated by about 1.4 times their average width.

Combining the information from the neighboring street casings and the character neighborhood graph, characters are grouped into word boxes to form the longest possible aligned string. Strings are recursively merged as long as they don't violate orientation and word separation constraints (Fig. 3).

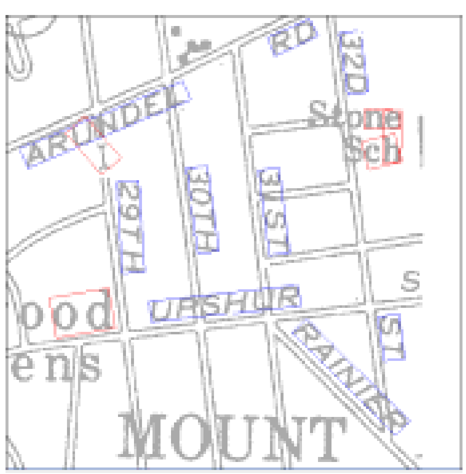

Fig. 3. Example of street-label character grouping

3.2. Line Structure Recovery

The objective of this step is to recover street-line casings that are interrupted because they overlap text. A black pixel in a word-box can belong to a character, to a casing, or to both.

All the vectors in the vicinity of each word box are analyzed. Short vectors, which are not aligned with longer vectors, are typically parts of characters. Long vectors that intersect the word bounding box are extended and connected, and assigned to the street-line layer. The output of the character recognition routine is used to flag the location of text pixels that overlap the street-line layer. The resulting sublayers, segmented text, and segmented street lines, are both subsets of the original black layer. An example is shown in Fig. 4.

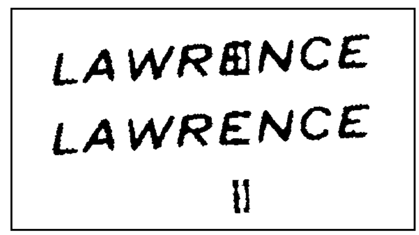

Fig. 4. Line structure recovery based on graphics extrapolation. From top to bottom: Original label box; Segmented text; Segmented street lines.

3.3. Character Template Matching

Our OCR is based on template matching which, in addition to converting the word-box bitmaps to symbolic alphanumeric ASCII labels, allows further improvement of the line art - text segmentation.

Prototypes for each character class are first extracted from bitmaps of operator-labeled training data [11,15]. The extracted prototypes are matched against the word boxes. The classification is based on the best fit for the whole word box, which is found using a level-building algorithm. The matching algorithm can take into account pixels that are flagged as belonging to line art, which may or may not conceal a black character pixel. Thus the recognition result itself can discriminate between these two possibilities, as the best-fitting template should overlay the character pixels. The important point is that the best fitting templates usually form a clean street label, and the residue that is not covered by them (Fig. 5) can be analyzed, as mentioned above, for segmentation.

4. Experiments

The first experiment shows the performance of our custom character-recognition system on unsegmented label boxes. The 32 templates extracted from the 92 street-name boxes used as the training set are shown in

Fig. 5. Label, matching templates, and residue

Fig. 6. Darker shades indicate higher probability of black. A few examples of the 733 automatically-extracted street name boxes used for testing are displayed in Fig. 7.

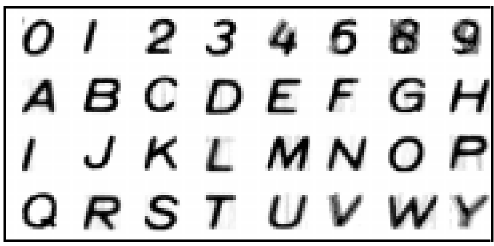

Fig. 6. Templates constructed from the training data

Fig. 7. Examples of test data for street labels, rotated to horizontal

After determining the best-fitting combination of templates that fits a label, the character recognition system may still reject some characters if they are not matched with high enough confidence. Characters that are part of street labels are recognized with an accuracy of 95% at a reject rate of 8%. (The error rate was determined by a string-comparison algorithm.) Most of the graphics fragments are rejected, but some lines perpendicular to the street box are recognized as I or L (here the graphics pixels were not marked). Most alphabetic characters that do not originate from street labels are rejected, but when they are not, they are often misrecognized. The complete results are presented in Table 1.

The second experiment was based on the top half of the map. The classification of the connected components yielded 5745 components that were classified as string orientation, and adjoining an appropriate street-casing pattern. Single components that could not be extended

Table 1. Results of street label recognition

	correct	wrong	rejected
street labels	1285	57	94
other text	151	57	275
Graphics	0	46	573

were discarded. Some curved labels were broken. Altogether 194 label strings were found acceptable for OCR (the criteria were more stringent than those used in Experiment 1), of which 165 were judged by local processing to contain significant graphics overlay.

Of the 194 strings, 69 (36%) had at least one character that was misrecognized. Because most of the boxes considered had significant graphic overlap, this rate is worse than the ~20% *string* error rate that we would expect from the 5% *character* error rate obtained in the first experiment. The reject threshold was not applied: all output labels were accepted, including many corresponding to graphic fragments. Note that correct recognition is assessed only with respect to the extracted label boxes: many of these contain only partial labels. The error count did not include 1-I confusions, as those are identical glyphs on the map and cannot be discriminated without context. The count does include many errors due to the lack of a complete set of templates: for instance, all 5's are recognized as 6's, because there was no 5 in the training data, and therefore no "5" template (Fig. 6).

In order to illustrate the effect of the cooperative processing, we report in further detail the results on the 17 boxes with line-art overlay for which the OCR algorithm produced a different result due to the flagged pixels. Of these cases, 7 strings produced incorrect results in both cases because their font-height was 10-20% larger than the normal street-name font. Only three of the remaining 10 strings were correctly recognized originally. When the flagged pixels were considered, this number jumped to five. Fig. 8 shows an example where the feedback from street recognition helped in correct recognition and another example where it did not.

The street-line layer was evaluated on an 8" x "8" section of the map. Comparison against the Digital Line Graph (DLG) indicated that the vectorization was 97.5% accurate before operator correction, but only 36% of the street lines were extracted. The positional accuracy of the intersections was 12 meters, which is within the National Map Accuracy Standards. An interactive session raised the completeness of the vectorization to 97%, with an accuracy of 8m (see [10] for further details on evaluation).

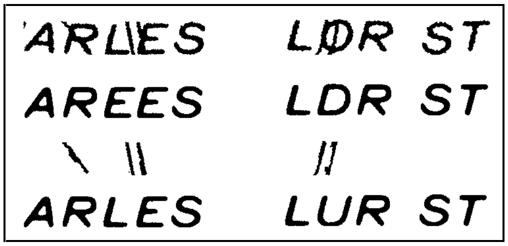

Fig. 8. Example of using flagged pixels. From top to bottom: original label bitmap, old OCR result, flagged street lines, new OCR result.

5. Conclusion

Beginning with the connected components of the black layer of the map, we have shown that almost all street-label boxes can be located and extracted using character alignment and size, and the graphic context of the location of labels adjacent to street lines. Most boxes are correctly oriented, but sizing these boxes precisely is more difficult. Even after extensive processing, many boxes contain some graphics fragments, and some text characters are missing.

The percentage of correctly extracted street vectors, and of correctly extracted and recognized labels, is evidently far too low to be useable without extensive operator interaction. At the previous stage, we estimated that the automated processing that we had developed reduced the time required to digitize the street network from about 24 hours to 16 hours. With some of the additional improvements mentioned here, especially close interaction between text and graphics processing, it appears feasible to reduce the overall time to about 10 hours.

According to the operator log resulting from the correction of the 8" x 8" chip, the most time-consuming aspects of operator intervention are vectorization, identification of the street label locations, and association of street labels with street lines. Our methods are quite effective for the last two of these three tasks, label box positioning and association. As shown earlier, few labels are missed entirely, and given correct street vectors and street labels, the association is practically foolproof.

It therefore appears desirable to concentrate further effort on more complete vectorization, perhaps even by relaxing the constraints that now guarantee high accuracy.

Wholly automatic conversion of documents of the complexity of topographic maps is still well over the horizon, but tools can be developed for reducing the cost of conversion.

Acknowledgment. This work was supported by a grant from the National Imagery and Mapping Agency as a part of the Intelligent Map Understanding Project. Support was also provided by the Central Research Laboratory, Hitachi, Inc., the Center for Communication and Information Science of the University of Nebraska – Lincoln, and the New York State Center for Automation Technologies (CAT) at Rensselaer Polytechnic Institute.

References

1. D. Doermann, An introduction to vectorization and segmentation, in Graphics Recognition: Algorithms and Systems, K. Tombre and A. Chhabra, editors, Lecture Notes in Computer Science 1389, Springer, pp. 1-8, 1998.

2. Environmental Systems Research Institute, Inc., ARC/INFO Software, Redlands, CA 1997.

3. L. A. Fletcher and R. Kasturi, A robust algorithm for text string separation from mixed text-graphics images, IEEE Trans. PAMI-10, 6, pp. 910-918, 1988.

4. J. den Hartog, A framework for knowledge-based map interpretation, Doctoral Dissertation, Technical University of Delft, 1995.

5. K. Kalafala, Identification of street lines and intersections in a scanned urban topographic map, MS Thesis, Rensselaer Polytechnic Institute, Troy, 1997.

6. R. Kasturi and L. Alemany, Information extraction from images of paper-based maps, IEEE. Trans. SE-14, 5, pp. 671-675, 1988.

7. L. Li, Adaptive text/line separation in document images based on vectorization and OCR, MS Thesis, University of Nebraska - Lincoln, 1998.

8. G. K. Myers, P. Mulgaonkar, C. Chen, J. DeCurtins, E. Chen, Verification based approach for automated text and feature extraction from raster-scanned maps, Graphics Recognition: Methods and Applications, R. Kasturi and K. Tombre, editors, Lecture Notes in Computer Science 1072, Springer, pp. 190-203, 1996.

9. G. Nagy, A. Samal, S. Seth, T. Fisher, E. Guthmann, K. Kalafala, L. Li, S. Sivasubraniam, Y. Xu, Reading street names from maps - Technical challenges, in Procs. GIS/LIS Conference, Cincinnati, OH, pp. 89-97, 1997.

10. G. Nagy, A. Samal, S. Seth, T. Fisher, E. Guthmann, K. Kalafala, L. Li, S. Sivasubraniam, Y. Xu, A prototype for adaptive association of street names on maps, in Graphics Recognition: Algorithms and Systems, K. Tombre and A. Chhabra, editors, Lecture Notes in Computer Science 1389, Springer, pp. 302-313, 1998.

11. G. Nagy, Y. Xu, Automatic Prototype Extraction for OCR, Procs. Int'l Conference on Document Analysis and Recognition, ICDAR, Ulm, 1997.

12. D. Rhind, An introduction to the digitizing and editing of mapped data, Automation in Cartography, P. Dale, editor, British Cartographic Society, 1974.

13. S. Sivasubramaniam, Street name association in digitized maps, MS Thesis, Rensselaer Polytechnic Institute, Troy, 1998.

14. P. Tofani and R. Kasturi, Segmentation of text from color map images, Technical Report CSE-98-007, Computer Science & Engineering Department, Penn State University, 1998.

15. Y. Xu, Prototype extraction and OCR, PhD Thesis, Rensselaer Polytechnic Institute, Troy, 1998.