

## **SYSTEM APPROACH TO A RASTER-TO-VECTOR CONVERSION: From Research to Commercial System**

Dr. Eugene Bodansky  
ESRI

### **Extended abstract**

Contact between scientists and the developers who create new commercial systems is mutually enriching. Developers use the results of scientific researches, and the scientists receive new tasks from developers.

Presently there are a lot of commercial raster-to-vector conversion systems. Why? Because nothing is perfect. As long as there is a need to capture data and convert raster files into vector format there will be a need to develop new efficient conversion systems.

Since complex systems such as these are more than simply the sum of their parts, the presentation will emphasize analysis of the whole system. So my objective is not a review of isolated methods and algorithms, which draw a rapt attention of scientists. I would like to analyze data automation and the conversion process and to reveal the new important problems that have not attract due attention yet. They have to be solved to develop a new generation of feature extraction and conversion systems. We will show how some of these problems could be resolved.

There is no strict definition for the problem of vectorization, raster-to-vector conversion and feature extraction. But we can obtain some idea about the scope of the tasks and problems that have to be resolved with conversion systems with the help of examples.

Usually the result of raster-to-vector conversion is intended for vector database, which manipulates graphic objects such as points, lines, and polygons. The conversion system has to vectorize raster files or produce such objects from raster file obtained by scanning source "paper" documents.

The resulting vector graphic objects do not necessarily look like the raster objects of the source raster file nor should they. For example, the dashed lines of a raster image should be represented with continuous lines in a vector database.

Sometimes graphic objects have to be labeled with attributes (the depth of a well for a point object, the width of a river, the kind of soil.) So attributing may be also part of conversion. For attributing graphic objects, the operator must extract that information from the source raster files or paper documents. Usually this procedure is time-consuming. There are sets of guidelines for drawing maps and engineering drawings. These rules form graphic languages and allow the reader to understand the drawings. However, it is very difficult to use these guidelines for the automatic interpretation of

drawings and it is not possible to rely entirely on these guidelines, because draftsmen do not always follow them strictly.

Sometimes vectorization involves geometric recognition, recognition of circles, rectangles, arcs, and so on. These tasks should be resolved in spite of the noise caused by different sources: deterioration and deformation of the source paper document, errors of scanning, errors of discretization and binarization, and so on.

The whole document does not always have to be vectorized. In some situations, it is only necessary to vectorize rivers or roads from a source document that contains roads and rivers. So sometimes the users need a conversion system that can do selective vectorization.

Why, if there are so many automatic vectorization methods and algorithms heads-up digitizing still one of the most widespread vectorization methods?

For automated system, tasks have to be formalized. But no one can predict all the tasks that have to be done in data automation and even in vectorization, as the samples discussed above demonstrate.

Nevertheless automatic conversion systems are used because sometimes they are more efficient, measured by the price of the job or the time spent to obtain an acceptable result. But even with automatic vectorization methods there must be a human being at least for verification of the result and correction revealed errors. So the effectiveness of automatic conversion depends not only on the speed of the automatic procedures and the correctness of their results, but on the functionality and effectiveness of the raster and vector editors. If the vector editor has tools for the automatic correction of topological errors, conversion of parcel maps can be done faster. The time to convert engineering drawings could be decreased if the editor has tools for smoothing and geometrical recognition.

Effectiveness of heads-up digitizing decreases when requirement to the geometrical accuracy of centerlines is increase. The more accurate centerlines need to be, the less effective heads-up digitizing becomes. Precise placement of the cursor in the middle of the raster linear object requires additional time. To do it with high precision, zooming may be required. Digitizing free curves, such as contours, takes more time too.

Tradition seen scientists who do research on raster-to-vector conversion attend mainly to the problems of automatic building skeletons and centerlines (raw vectorization). In fact, there are a lot of another important problems for developing effective conversion systems. So hypertrophied attention to automatic vectorization can be excused only by the fact that these methods are prevailing in text recognition, where interactive vectorization can't be used.

Amazingly, it is very difficult to find fundamental research dedicated to interactive vectorization methods like tracing or raster snapping. If you look for the

words “tracing” or “raster snapping” on the Internet you will receive tons of links to numerous commercial companies, and almost nothing to scientific publications.

Some tasks cannot be solved with heads-up digitizing or automatic vectorization. So many users want to use methods that are intermediate between heads-up digitizing and automatic vectorization. These methods let the operator check the vectorization process uninterruptedly and correct errors directly during vectorization. In contrast to heads-up digitizing, they set the operator free from routine and tiresome operations.

Operators who uses tracing should only put the cursor into the starting point and define the direction of the further tracing of the linear object when the cursor stops in an intersection. Of course it is much easier than digitizing each vertex.

If the document contains a lot of straight segments, as parcel maps, and there are a lot of intersections because of noise and text that touches the linear elements, heads-up digitizing with raster snapping becomes very effective.

To develop a good tracing algorithm, it is necessary to solve the problems of the optimal segmentation of the raster image and to suppress edge effects. The tracing has to be done in a real time, and it is not important how long it takes. Therefore some algorithms and methods, labor-intensive and time-consuming for the automatic vectorization, can be used for tracing. Raster snapping algorithms can use implicit information coming from the operator, for example, that the foreground pixel, closest to the point designated by the operator belongs to a linear object.

To build an effective conversion system, it is necessary to divide tasks between the operator and computer appropriately. The operator should do only those actions that are not difficult. The computer should solve only those tasks that can be solved with simple and stable algorithms. “Render unto Caesar the things that are Caesar’s, and to God the things that are God’s.”

Following examples demonstrate how it is possible to increase effectiveness of data capturing with redistribution tasks between an operator and computer.

1. It is well known that the most difficult part of building centerline algorithms is making it automatically recognize what type of intersection it’s dealing with. Some very complex algorithms for solving this task have been developed but there are a lot of cases where the result is unsatisfactory. Operator still needs to test every solution of each intersection and correct the errors manually. It may require a lot of time.  
An operator however can recognize types of intersection easily. It is relatively simple to develop an algorithm that will resolve intersections of a given type. So an effective interactive method of resolving intersections can be suggested: an operator recognizes the type of the intersections and chooses the appropriate template; that intersection can then be resolved automatically.

2. To run many algorithms of automatic vectorization it is necessary to define the maximum width of the linear elements first. To determine this threshold, the operator should measure the local thickness of different lines. Vector editors use a ruler. To measure the width of the line with a ruler, the operator has to put the cursor exactly on opposite sides of the line. If the source document is a big one, it is necessary to zoom an image each time before measuring.  
It is difficult to develop an algorithm that will measure the maximum width of the linear elements of the image, but it is possible to develop an algorithm that will automatically evaluate the local width of the line. Such a tool can save time.
3. At present a lot of methods and programs for text recognition exist. However there are no effective text recognition programs for graphic documents. On maps, engineering drawings, electrical schematics, and other graphic documents, the text may touch linear objects, it can have not a horizontal orientation and even be meandering, it can be short, and it can be written with different fonts and with a different size. To recognize a text of graphical documents it is necessary first to separate it from linear objects, solids, and symbols and to define its orientation. I do not know an algorithm that can solve this task stably. But this can be done interactively. If the operator will draw the line through the text it will be not so difficult to separate it and recognize its orientation.
4. Many conversion systems allow us to clean the raster image of noise before vectorization. To solve this task it is necessary to find speckles and holes as connected components that meet the requirements for their size, area, and sometimes more sophisticated characteristics. But small graphic elements (dashes, dots, and others) can be deleted together with noise. So it is more efficient to automatically select speckles and holes and highlight them first. Then the operator can verify the result of selection and, if it is necessary, to correct it interactively pointing with the cursor to the connected components, which have to be selected or deselected before cleaning.
5. When processing the map of a city it is often difficult to vectorize rectangular buildings. If it is a small-scale map, contours of buildings can have relatively big noise. The program for an automatic recognition of rectangles with big noise is complex and not always gives the good and stable result. Manual drawing of rectangle contours is time-consuming procedure. But an operator can recognize the rectangular buildings easily and there are programs, which approximate connected components with rectangles good and stable. So it is possible to develop an effective interactive procedure of one click vectorization of rectangular buildings.
6. We have already discussed that verification of results is labor-intensive and time-consuming. Is it possible to accelerate this procedure? Usually the result vectors have to be located inside the corresponding raster linear element and

be very close to the centerline. The violation of this condition may be caused by errors of vectorization or by finishing of lines (smoothing, generalization, or compression). It is possible to develop algorithms that will automatically evaluate deviations of the result vectors from centerlines and mark the peculiar places.

One way to increase the effectiveness of conversion systems is by using learning algorithms. Say that the corners between straight line segments and the boundary points of critical curves along lines are called critical points. Often the recognition of critical points is an important component of the conversion process. All the algorithms to solve this task use some thresholds and parameters. Often it is difficult to evaluate them because they depend on so many factors: maximum and minimum curvatures, noise, thickness of lines, and so on. But it is relatively easy to show these points on the screen. The effectiveness of the conversion system will be increased if an algorithm can be developed that can automatically evaluate necessary thresholds and parameters using information about location of some of the critical points.

The new version of the conversion system ARCSCAN, which was developed by ESRI and which is a part of GIS ArcInfo, is a prototype of the new generation of conversion systems. The presentation uses ArcScan to illustrate some of our statements, assumptions, and conclusions.

#### References.

Arvind Ganesan, "Integration of Surveying and Cadastral GIS: From Field-to-Fabric & Land Records-to-Fabric", ESRI, 2002 User Conference Proceedings, <http://gis.esri.com/library/userconf/proc02/abstracts/a0868.html>

Lawrence O'Gorman, "Basic Techniques and Symbol-Level Recognition – An Overview", Graphic Recognition. Methods and Application. LNCS 1072, R.Kasturi, K.Tombre (Eds.) Springer, 1996. pp.1-12.

A.J.Filipsky, R.Flandrena. "Automated Conversion of Engineering Drawing to CAD Form", Proceedings of the IEEE, V.80, #7, 1992, pp.1195-1209.

L.Baotto, V.Consorti, M.Del Buono, S. Di Zenzo, V.Eramo, A.Esposito, F.Melcarne, M.Meucci, A.Morelli, M. Mosciatti, S.Scarci, M.Tucci, "An Interpretation System for Land Register Maps", IEEE Computer, V.25, #7, 1992, pp.25-33.

S. Levachkine, E. Polchkov, "Integrated Technique for Automated Digitization of Raster Maps", Revista Digital Universitaria, Vol. 1, No. 1, Art. 4 (2000). On-line: <http://www.revista.unam.mx/vol.1/art4/>

Eugene Bodansky, Alexander Gribov, "Closing Gaps of Discontinuous Lines: A New Criterion for Choosing the Best Prolongation", LNCS 2423, Document Analysis Systems V, 5th International Workshop, DAS 2002 Princeton, NJ, USA, August 2002

Proceedings. Springer. Daniel Lopresti, Jianying Hu, Ramanujan Kashi (Eds.), 1992, pp. 119-122.

Bodansky Eugene, Pilouk Morakot, "Using Local Deviations of vectorization to enhance the performance of raster-to-vector conversion systems", International Journal on Document Analysis and Recognition", No. 3, 2000, pp. 67-72.

Bodansky Eugene, Gribov Alexander, Pilouk Morakot, "Post-processing of lines obtained by raster-to-vector conversion", Vision (machine Vision Association of SME), Vol.18, #1 ([www.sme.org/mva](http://www.sme.org/mva)), First Quarter 2002.