

AUTOMATIC VECTORIZING

In this chapter:

- Automatic vectorizing in Easy Trace Pro
- About images for vectorizing
- Autotrace Lines utility
- Outline Contours utility
- Raw Line Filtering utility
- Breakup Joining utility
- Line Form Optimization utility
- Autodetect Swamps utility
- Autodetect Grid utility
- Autodetect Circles utility
- Autodetect Ortho-Objects utility
- Autodetect Lakes utility

This chapter describes an automatic vectorizing technique for cartographical data. Our approach is somewhat uncommon as it includes operations being usually executed in a GIS. It is aimed at maximal acceleration of “turnkey” data preparing. Ideally, the data will not require any additional correction at all.

Here you will find a description of the utilities that form the backbone of the process.

Automatic vectorizing in Easy Trace Pro

The traditional approach to automatic vectorizing implies specifying of numerous parameters, long processing and even longer manual correction.

Only developers of the software and few experts understand deep-laid connections between parameters and results of vectorizing.

This may work for drawings but not for maps. Map variety corresponds to one of the outer world and any attempt to bring it to a set of parameters is doomed to failure.

Is there a way out? Yes, there is if we deviate from traditions. But let us briefly consider objects of vectorizing first.

Objects of vectorizing

We stipulate from the very beginning that a set of meaningless lines mirroring the hard copy is insufficient. We need something more. What exactly?

- *Reconstruction of objects' geometry* – depends on object type. If isolines, they should be continuous, smoothed and without intersections. If areas, there should be correct polygons rather than sets of line fragments that constitute their boundaries. There may be additional requirements. For example, buildings have right angles as a rule and are aligned along centerlines of streets.
- *Object attribution to layers* – is something more complex than simple separation on the base of formal parameters like line width or length of strokes.
- *Topological connectedness of objects* – comprises common boundaries, common vertices or nodes as well as agreed relations between point, linear, and polygonal objects. In general, all the rules of the topological model adopted in your GIS must be obeyed.
- *Forming of derivative objects and features* – polygon assembling, forming of settlement borders, buffer zones, decoding and assignment of attributes, etc.
- *Deletion of irrelevant objects* - defects, noises, hatchings, fragments of inscriptions and topological symbols.

It is hardly possible to imagine a set of vectorizing parameters sufficient to meet all these requirements.

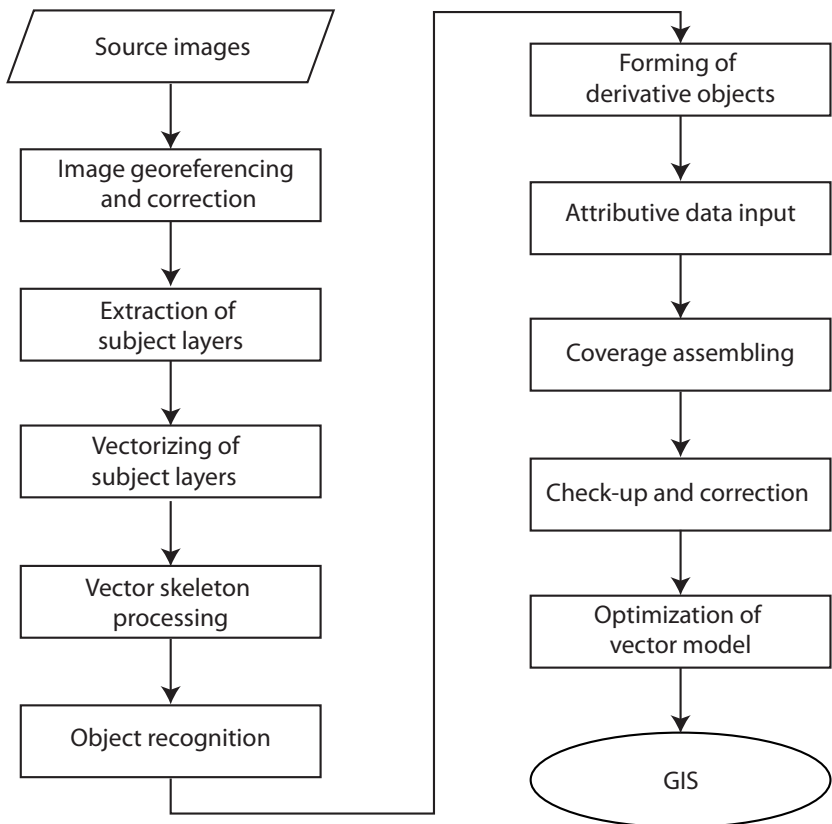
How does Easy Trace do it?

From the very beginning, we do not hope to get by with parameter specifying and one click of a button. We want to get a complete and valuable digital model of an area rather than a nice drawing that resembles the image very much.

That's why the process of vectorizing is divided into several steps. This is the only way to interpret the image in detail. Every step is supported by a specialized utility or a group of utilities. At that, the order of their work is not an invariable tenet - everything depends on the task and your experience. Some steps comprise operator's handwork.

Yes, manual processing is inevitable but utilities are charged with the lion's share of the work. So, the process may be not fully-automatic but it is highly automatized at least. Whatever the name though, it is really fast.

The approach to vectorizing used in ET is best described by the ancient expression "Divide and Conquer". It's time to consider the framework of the process.



1 There are two tasks to be fulfilled at this stage. These are deletion of geometrical distortions in the image and image georeferencing in the coordinate system.

Geometry of the source image is always distorted. At best, these are distortions caused by printing, storing, and scanning of the hard copy.

At worst, one will have to divide the image into fragments and then restore the coverage out of these “scraps”. Hard job indeed but have you ever try to fulfill referencing of an image if the map sheet is torn at several places and “mended” with the Scotch tape?

Another typical problem is low resolution at scanning and merciless compression into JPEG format. In this case, it may be useful to resample the image up doubly. There will be space at least to squeeze pixels into and thus separate neighboring lines. Resampling should be done BEFORE georeferencing and geometrical correction, of course.

Result – the image is corrected and tied to the project field.

2 Vectorizing proper starts with extraction of subject layers (images). Several of them. It is reasonable to make two images even if the source one is black-and-white. One will be used for line tracing and another – for vectorizing of point objects and filled areas.

Advantages of separate processing may be not so evident for black-and-white images but they raise no doubts in the case of color maps. They should be parted into rasters of individual colors.

Consider the nature of your data of course. It is senseless to extract a color layer if it contains only a dozen of lines and several points. But it’s quite another matter if there are hundreds and thousands of objects in the layer.

On the other hand, do not economize on the operation. It takes minutes but save hours and tens of hours.

Remember: we do not alternate the source image but make several copies and process them differently. ET even has a special command - «Duplicate raster layer». The source image always remains inviolable.

The first copy usually serves as a “cover” of the entire image package. It differs from the source image by decreased brightness and contrast. It is easier to control quality of vectorizing against such background.

That’s clear. Color lines are poorly discernible against a bright background and tire operator’s eyes. Besides, it is reasonable to decrease number of colors down to 256 for the “cover”. That’s enough for work and saves memory (disk memory is usually sufficient, but we save RAM).

All next copies of the source image are intended for extraction of subject layers – black-and-white images that contain objects of a certain subject-matter.

Note that vector data may be used at subject layer forming together with raster operations. Usually they help to delete objects from the image. Objects may be subtracted from a raw copy of the source material or from the result of its processing, i.e. a thinned black-and-white subject layer.

Hence, do not try to prepare all subject layers at once. Vector data to be used for object subtraction from images are missing yet...

Result – a package of raster layers: the «cover» plus several black-and-white subject layers.

3 This step is the simplest one. It consists of either automatic line tracing or automatic outlining, depending on the type of your data.

The program does not analyze the nature of raster lines; it just generates their skeletons. These are vector polylines with high density of vertices – about one vertex per one pixel of the image.

Junction of three and more skeleton lines is a node. Dotted lines are represented by rows of individual strokes.

Result – vector skeletons of raster lines.

4 Processing of the skeletons comes to deletion of defects and joining of line fragments across gaps. Both tasks will be mostly done by utilities although some manual editing is inevitable at this stage.

However thoroughly we have extracted subject layers, lines are always discontinued in them. Even lines of the most intact “black” layer are broken by the grid and inscriptions.

Processing consists usually of iterative cycles: Filtering - Breakup Joining –Editing.

Result – improved and mended vector skeletons.

5 Object recognition in skeleton vector data is isolated into an individual step conventionally as it is closely associated with the previous task. We extract objects from the skeleton and thereby make remaining lines available for correction and breakup joining. For example, recognition of water bodies in the skeleton of the “blue” image frees the river net.

On the other hand, recognized objects may be used as barriers that prevent false joining of line segments.

But this is not the whole story. An object is really recognized when its connections with other objects are restored. So, shape correction comprises reconstruction of topological relations. It should be done before shape optimization i.e. substitution of densely packed skeleton lines by the minimal set of vertices that can describe the object with the specified accuracy.

Result – topologically connected and optimized vector objects attributed to different layers.

6 Forming of derivative objects is the next step after total disassembling of the vector skeleton. There should be a polygon in GIS where “arable land” is written in the map. Borders of the polygon already exist at this stage – for example, these may be a forest edge, a river, and several roads.

Another example – elevation marks of geodetic monuments. The monuments themselves are missing in the map but should be included into the vector model.

We apply object generators at this stage. These are utilities that create polygons and buffer zones, tracing tools that can use existing boundaries, and tools – decorators.

Even utilities for topology check-up may be useful. For example, they can find all crossings of highways with railway lines. Or find, mark and classify bridges, highway crossings, and tunnels.

Result - complete set of vector objects represented in the image.

7 Attributive data input finalizes classification of objects. Polyline orientation (directing) and input of isolines’ elevation values also belong to this step.

Some part of attributes may be input at once, applying the Group Editor for example. The rest remain for automatic conveyance and manual input.

There are special means for attributive data control - support of value domains and attribute-based representation of objects as well as generation of attribute-based inscriptions. Use of typical icons instead of text descriptions of qualifier values additionally facilitates the operation.

Sets of attributive values created in advance may be used for standard samples.

Result – complete classification of objects by the specified set of their features.

8 Check of topological relations, polygonal coverage, and line shape concludes processing of an individual map sheet. Errors are inevitable and it is important to search and correct them quickly.

Besides, there are a lot of errors in paper maps, let alone discrepancies at borders of adjacent sheets.

Check and correction of topology is supported by a set of utilities. They mark revealed errors and supply them with “beacons” that allow to navigate quickly to problematic places. Flexibility and quickness are advantages of the utilities.

Topology check-up and correction is in use actually at all stages of vectorizing. It takes only seconds and provides a convenient mean to find places where operator’s intervention is required.

Result – topologically correct vector model of the map sheet.

9 Joining of coverages or junction of neighboring map sheets is a typical task in GIS projects. It is also automatized in Easy Trace of course. Classification of vector layers by line types enables both smooth and linear automatic joining of objects belonging to adjacent sheets.

Unfortunately, typical maps are poorly coordinated at sheet borders. That's why search and marking of disagreements is one of the tasks to be accomplished at this stage. Final decision can be made by the operator only of course.

Result – complete coverage of an area coordinate at borders.

10 Optimization of the assembled vector model is the finality of vectorizing. Superfluous vertices in polylines may originate from many reasons, manual tracing and editing first of all.

Redundant information may constitute up to one third of existing GIS projects. It inevitably affects the speed of work in the system.

This unique utility can shake the rubbish out without violation of topological connectedness of a coverages containing hundreds thousand of objects.

Result – optimized vector data model for a large GIS project.

The scheme may look rather awesome, and one may be dumbfounded at first sight. Quite groundless though. Described in equal detail, coffee making is also a complicated process but it is not a good reason to refuse from this splendid drink.

Associated products

A good working place comprises a lot of things besides a powerful computer and a big screen. These are a comfortable seat, sufficient light, air conditioning, silence, strong tea (coffee, beer...), and so on. Similarly, a good vectorizer does not come to tools and utilities but has a set of additional features that form its shell.

So, what else is provided for comfortable work?

- Means of regular data inspection without “blank spots” and unintentional returns.
- View modes. This is the mean to look at the material from different points of view. Control of data structure and coverages, attributive data input, forming of the elevation model, joint use of maps and photos require different ways of material presentation.
- Special means for selection and marking of any vector objects as well as “delivery line” organizing for their check-up, correction, and attributive data input. Access to objects through tables of attributes.
- Accelerators. These are custom tools adjusted for different object types. These are special buttons for pattern-based attribute data input and the list of previously used values in attributive fields. Finally, these are icons instead of hundreds of topographical symbols’ names.
- Inheritance. Everything adjusted in one project may be used in another. Projects have “the memory of generation” about projects-prototypes.
- Help. The program has a powerful HELP menu together with a detail manual kit. Besides, there are comprehensive instructions at the bottom of every utility’s dialog box. Videos of technological approaches and projects-examples are also very useful.
- Register. This is an integrated system of time accounting together with information on amount and type of vectorized objects. You have control over operators’ per man-hour output and can easily estimate time required for vectorizing of this particular material.
- Compatibility. Results of vectorizing are destined for a GIS, that’s clear. But the vectorizer should be also able to take from the GIS a lot of information. The list of layers, layer color, structure of fillings. Attributive tables. Domains of values for every attributive field. This is the only way to ensure data compatibility.

About images for vectorizing

Resolution (DPI) selection for color and grey-scale images planned for vectorizing is based on the following considerations:

- It is possible to extract subject layers suitable for automatic vectorizing from a high-resolution image even after overcompressing by JPEG format.
- Time of automatic vectorizing is several times (up to 10 and more for a data-rich high-quality relief layer) less than one of image processing with the cleverest tracing tools.
- The cost of resources (GBs of disk and memory, processor frequency, local net, etc.) is falling constantly. On the contrary, the cost of man-hour is increasing.

What scanning resolution is sufficient for automatic vectorizing? One that ensures:

- 50% of pixels in lines that unambiguously have the initial color;
- gaps at least 1.5-2 pixels wide between lines of the same color.

Below is a small “lyrical digression”, which may be skipped.

Offset printing was used for most of topographical maps. It means that inks of different colors were deposited successively on the paper through grooves. The grooves were seldom placed with sufficient accuracy of course, there was always a shift. As a result, lines of different colors overlapped each other rather often. It caused fluctuations of the width of color lines.

Chromotypography is production on the line. Next ink is often deposited before previous one dries up completely. It results in their interpenetration and color gabbling.

The map you are scanning could be printed decades ago. The inks have degraded and the paper has grown yellow. Besides, maps wear out.

Scanner sensors capture not the line only but the surrounding background as well at low-resolution scanning. Even an ideally red line becomes brown in the image if it is drawn on the green background.

When saved in a lossy format (such as JPEG), true colors of pixels are substituted by analogous ones. For example, black becomes dark violet, red becomes purple, etc.

The substitution is almost imperceptible at first glance until you zoom up the image and inspect individual pixels.

In the final analysis, the colors used at map printing are actually missing in the image, especially in a low-resolution one. There are sets of tints that may be more or less reliably attributed to certain colors.

A bit of arithmetic now. Given:

- Scanner resolution is 300 dpi
- Line width in a topographic map is 0.4 mm

Find: what is line width in the image?

Ideally, 300 dots per inch give $300/25.4 = 11.8$ dots per mm. It means that the width of raster line is $11.8 \times 0.4 = 4.72$ dots.

That's quite enough seemingly, but... Line edges in the map are of unexpected colors. Against the white background, they are much lighter than the center zone of the line;

against a color background, they are represented by a color mixture. So, we have to reject outermost pixels. The remainder is 2.72. Minus one dot for discreteness of scanner sensors. The answer is 1.72 dots – a border-line value without any reserves.

A line with 2 pixel wide meaningful parts of horizontal and vertical segments becomes a staircase at 45-degree turns and its width varies from 1 pixel to ZERO.

The same is true for gaps between closely spaced lines; it causes line “agglutination”. There are zones without visible gaps between lines even in images of topographical maps scanned at 440 dpi.

Defects of line structure increase avalanche-like at decrease of scanning resolution. It may be possible to extract something for automatic vectorizing from a data-rich image received at 300 dpi but 200 dpi is hopeless.

Optimal resolution value for color and (strangely enough!) shabby black-and-white materials lies in the range of 400-600 dpi. It allows reliable extraction of line “crests”, i.e. parts that have color best corresponding to the color used at printing.

In case of a black-and-white map scanned in the grey scale, such resolution allows deletion of numerous agglutinations of neighboring lines.

The point is that brightness of thin individual lines is actually equal to one of gaps between closely spaced lines. To save lines-individualist, people usually have to consent to conglutination at bottlenecks.

Sufficient resolution allows applying of the “Unsharp Mask” operation. It evens brightness (or rather darkness) of thin lines and emphasizes (clears) gaps between neighbors. But effective use of the operation requires mask radius to be no less than 3 pixels. And the width of lines should be close to the mask size.

Conclusion:

Source data scanning at 300 or 500 DPI may mean further processing ONE or THREE months long.