METHODS AND ALGORITHMS FOR VECTORIZATION OF RASTER IMAGES

[1]Shaxislam Joldasov, [2]Saida Beknazarova
Tashkent University of Information Technologies named after Muhammad Al- Khwarizmi, Tashkent, Uzbekistan[1,2]

ABSTRACT

Technologies for converting graphic documents obtained by digitizing traditional sources – paper media into vector form are currently in high demand. A vector shape is understood as a data model (structure), which is an ordered set of layers of objects that are modeled by points, polylines or polygons located on a plane or sphere with a given coordinate system. This is due to the development of computer-aided design systems that are able to interpret the vector format, as well as the significant advantages of the vector form compared to the raster representation: ease of manipulation and management, the need for less memory, etc. The process of converting images from raster to vector form is called vectorization.

An engineer-operator can simply process design documentation, drawings, drawings with the input of the received graphic data into CAD, however, computer systems initially do not have knowledge of the subject area and human abilities for analysis, thinking, representation, which leads to a number of problems. The analysis of the modern software market in the field under consideration showed the absence of fully automatic vectorization methods. The analysis also showed their insufficient effectiveness in terms of the accuracy of the result obtained, the preservation of important topological characteristics, such as inter-object connectivity, localization of segments of articulation of adjacent objects, etc.

Methodology. Methods of machine graphics, computational geometry, image processing and recognition, graph theory, theory of algorithm analysis, mathematical statistics, and data mining were used in the performance of the dissertation.

The whole variety of algorithms directly involved in the transformation of raster images into vector form is divided into three levels. As a rule, the following three-stage scheme is used to implement an automatic or semi-automatic vectorization algorithm: at the first stage, the original image is divided into several fragments, each of which includes image objects of only one class (for example, text, lines, areal objects). This stage is commonly called image segmentation. Further, at the second stage, each fragment undergoes a primary vectorization procedure, during which a raster model is obtained, consisting of points connected to each other, modeling the lines and boundaries of the area objects of the image. And finally, at the third stage, post-processing of the results is performed in order to improve the quality of the final vector representation.

The general task of segmentation is divided into several subtasks, each of which is solved by specific methods. These are the tasks of segmentation of text and graphics, segmentation of areal objects and lines, segmentation of thin and thick lines, and separation of the raster by a set of base colors.

The problem of primary vectorization of linear (non-areal objects) binary images is usually solved in this way: first of all, the axial lines of objects are highlighted in the form of chains of raster pixels, after which a linear approximation of the found axial lines is performed. As a result, an image model is formed in the form of sets of polylines that meet at their end points. For this scheme, the methods for selecting centerlines can be divided into five groups: (1) based on thinning lines, (2) based on contour matching, (3) based on graphs of object strokes, (4) based on splitting the image with a regular grid, (5) based on sparse viewing of the raster. There are also other schemes for constructing vector models: based on the Hough transform, based on the approximation of raster objects by areal geometric shapes.

Methods of linear approximation. The result of primary vectorization is a set of pixel chains describing the average axes of linear objects and the boundaries of areal objects in the image. This representation is already vector at the current stage, however, it is redundant due to the presence of "extra" description points that are approximately on a straight line between two neighboring pixels in the chain. To reduce the degree of redundancy of the vector shape, while maintaining the necessary informativeness, these points can be removed from the set, thereby forming an approximated description of the image. In order for the vector representation to remain accurate, the approximation procedure usually requires setting the threshold e, which is the criterion for choosing between saving and removing a point from the set. If the distance from the point in question to a straight line passing through two adjacent points is greater than e, then the point is critical and remains in the set, otherwise it is eliminated. The lower the threshold e, the more accurate the approximated form will be, but the redundancy of this one will be higher.

Post-processing of the vector model. This stage is necessary to transform the resulting vector model at the previous stages of vectorization to a more convenient form that meets the user's expectations. Most vectorization methods use sets of simple topological and geometric rules to calculate the attributes of objects, which often leads to a result other than expected.

The bitmap image at the end of the operation of most vectorizer systems can be set in a parametric form, which is characterized by representation in graphs. Indeed, graph models are naturally able to reflect the structure of a graphic document, where each primitive corresponds to a vertex of the graph, and the relationships between primitives are characterized by connections. Also, the use of graph models makes it possible to develop an image recognition system by implementing an isomorphism search between input and reference graphs (graphs in the database).

So, the vector image model is reduced to a set of oriented graphs. All vector objects can be divided into types: line or areal object. Accordingly, there can be three types of relationships between types: "line area object", "area object line", "line line".

REFERENCE

[1] Novikov Yu.L. Research of methods of vectorization of raster images and their implementation in a geoinformation system. - Tomsk: Tomsk. state un-T. Faculty of Computer Science, 1999.-177 p.

[2] Jimenez J., Navalon J. Some Experiments in Image Vectorization IBM Journal of Research and Development.: http://www.research.ibm.com/journal/rd/266/ibmrd2606J.pdf

[3] Kostyuk Yu.L., Novikov Yu.L. Graph models of high-resolution color bitmap images // Bulletin of TSU. 2002. No. 275, April. pp.153–160.