Geometric modeling of raster images of documents with weakly formalized description of objects

D Yu Vasin[1], **V P Gromov**[1] **and S I Rotkov**[2]

[1]National Research Lobachevsky State University of Nizhny Novgorod, Gagarin Ave., 23, Nizhny Novgorod, Russia, 603950
[2]Federal State Budgetary Educational Institution of Higher Education Nizhny Novgorod State University of Architecture and Civil Engineering», Ilyinskaya St., 65, Nizhny Novgorod, Russia, 603950

e-mail: dm04@list.ru, rotkovs@mail.ru

Abstract. In this paper, we analyze graphic documents with a weakly formalized description of objects (WFGD) and reveal their main features that influence the choice of models, methods and algorithms for processing such documents. In the framework of the development of the combinatorial-geometric approach, a geometric model for describing WFGDs with a pronounced orientation of linear objects is proposed. We also propose a technology for vectorization of raster images of WFGDs in the presence of noise in the source data. The effectiveness of an extended class of vector models (linear and segment-node models) used for describing WFGDs with a distinctive linear orientation of objects is shown, which was revealed during practical experiments on real WFGDs.

1. Introduction

In recent years, increasing importance has been given in video information analysis to the development of mathematical methods for constructing a formalized structured description of input video information. The tasks related to obtaining formal descriptions are addressed by studying the internal structure, and content of elements or objects of a simpler nature (non-derivative elements, objects identified in the images being processed at various processing levels, etc.) [1-3].

In this paper, only the images of large-format, semantically rich graphic documents with a complex structure (LFGD) are considered as source information. In this case, the source data, as a rule, are graphic images (GI) on paper, while digital documents must be produced in the terms of the respective problem area. Such documents contain symbols of four classes of objects: point (discrete), linear, two-dimensional (areal), and symbol images. Taken together, these symbols make up spatially-distributed data (SDD) [4 - 9].

The analysis of LFGDs shows that a significant proportion of such documents was produced in a manner inconsistent with the rules for nomenclature description of objects. We shall distinguish such LFGDs as a subclass of documents with a weakly formalized description of objects (WFGD). This subclass includes: engineering drawings, diagrams, floor plans for buildings, topographic maps and nautical charts, data on the Earth’s surface obtained from satellites, etc. The main features of WFGDs, which influence the choice of representation models and methods for their processing, are considered in [4–9]. Taking these features into account places greater demands on the geometric modeling of this class of graphic documents (GD).

2. The problems of creating automatic technologies and systems for WFGD processing

In order to automate WFGD input, various information technologies have been proposed in recent years. They are based on heuristic procedural methods, as well as on recognition methods with learning, and are effective for a limited set of objects with strict limits on their size and orientation. It should be noted that the technology of automatic analysis of WFGD is a complex multi-stage process that involves a large number of processing methods and algorithms: filtering, compression, storage and search, analysis and decision making. For effective operation of this “pipeline”, it is very important to ensure that all mathematical models, methods, algorithms and data representation structures are interconnected and mutually effective: it is obvious that even the highest efficiency at some particular stage of processing can be offset by low performance at other stages.

Huge information redundancy of raster images of WFGD (RIWFGD) certainly places greater demands on automatic processing algorithms. The problem becomes even more serious, if we take into account that automatic processing of RIWFGD at lower levels of the hierarchy should be carried out in real time and with limited memory resources, and the models and methods being developed should be integrated into existing technologies and systems.

Consequently, the models and methods for processing RISFGD must be technologically advanced and must meet the general efficiency requirements for graphic image analysis as a whole [4, 5, 9]:

- technological effectiveness;
- high efficiency in terms of speed and memory;
- natural integrability in the general processing scheme.

It should be noted that vector models of WFGDs obtained by means of automatic procedures do not always produce objects that correspond to their reference description and that are not always specific, both in terms of their composition and the methods for setting them. Besides, the practice of processing WFGDs, especially WFGDs taken from archives, has revealed some new serious problems associated with the transition from the lower (pixel) level of representation to the vector level. This results in a further dramatic increase of complexity of the procedures for automatic recognition of objects in WFGDs and inevitably reduces the time efficiency of the entire processing technology for this class of documents due to the requirement of mandatory interactive control and editing of possible errors.

To avoid multiple duplication in the development of systems for solving various tasks of WFGD processing, it is useful to have a basic system that can be considered as a tool for solving two main tasks: on the one hand, it would be the basis for developing systems specialized in a specific subject area, and on the other hand, it would be an automated workplace for the development and research of algorithms for WFGD processing.

The combinatorial-geometric approach (CGA) proposed in the 1980s for processing raster images of graphic documents (RIGD) in spatially distributed data [4] can become the core of such a system. This approach is based on the hierarchy of mathematical models of image description, the hierarchy of data representation structures, a set of fast and memory-efficient algorithms for solving problems of computational geometry, as well as specialized algorithms for processing video data. The essence of the approach is as follows: a contour image model (CIM) or a linear-contour image model (LCIM) is constructed from the initial RIGD, i.e. the initial RIGD is assigned a set of points, polygons and broken lines. Based on this representation, a hierarchy of interrelated mathematical models of description, structures for representation and decision making is built, where objects are also considered as points, polygons, broken lines and their collections. The objects of the hierarchy of image models are built using a system of logical – geometric predicates (decision rules) that calculate the characteristics and relationships between objects: dimensions, distances, nesting, junction, intersection, and other types of mutual arrangement relationship, characteristics of objects and their parts.

As a result, the entire complex set of tasks related to the analysis of video data is considered from a unified point of view of building a hierarchy of interrelated mathematical models for description, representation and decision-making structures. At the lower level of this hierarchy, raster information from the original source of visual data is processed, while the upper level corresponds to the description of graphic data at the level of content in the user's terms [4].

Therefore, it is important to further increase the intelligence of information technologies for automatic processing of RIWFGD, and, consequently, to further develop the CGA and the hierarchy of description models and appropriate RIWFGD processing methods and algorithms.

3. Effective models for WFGD description

In the framework of the CGA, the mathematical model of the image is understood as a triplet of the form $M_v^\alpha = \left\{E_v^\alpha, C_v^\alpha, R_v^\alpha\right\}$, where: $E_v^\alpha = \left\{e_1^\alpha, e_2^\alpha, ..., e_s^\alpha\right\}$ is the set of non-derivative elements of the rank α model; $C_v^\alpha = \left\{c_1^\alpha, c_2^\alpha, ..., c_n^\alpha\right\}$ is the set of permissible relations between the non-derivative elements of the rank α model; $R_v^\alpha = \left\{r_1^\alpha, r_2^\alpha, ..., r_n^\alpha\right\}$ is the set of characteristics of non-derivative elements of the rank α model; α = 1,2,3, ... *N* is the rank (level) of the model [4].

An image model will be considered invariant with respect to a certain set of classes of GD images, if:

- the model can be built for any document from this set;
- the original image can be restored using a model with known accuracy.

Initially, to automate LFGD processing, mathematical models were proposed for the lower levels of the hierarchy to describe images. In these models, graphic images were presented in the form of a raster (set of pixels), a collection of lines, contours and points (CIM, LCIM). At the same time, LCIM was considered as the base model for all subsequent levels of models. Models of higher levels were built on the basis of CIM or LCIM. For these models, effective computational geometry methods (the general-to-specific method on the basis of hierarchical structures of vector data representation), methods for graphic objects recognition (the correlation extremal method), etc., were developed [4-9].

The practice of automated WFGD processing required the extension of the existing class of models for their description. The original extended classes of raster and vector models for describing RIGD are discussed in detail in [6–9]. In particular, at the raster level, original models of raster simple objects (RSO) and raster composite objects (RCO) were proposed. RSO are divided into raster linear objects (RLO) and raster areal objects (RAO). The class of vector models was extended by introducing linear (LIM) and segment–node (SNIM) models. In this case, CIM is in one-to-one correspondence with RIWFGD, it can be used as an independent model, but it can also be considered as a preparatory stage for producing a linear-contour model, and LIM can be considered as a degenerate case of LCIM. SNIM is a vector model that describes the entire image as a group of connected sets.

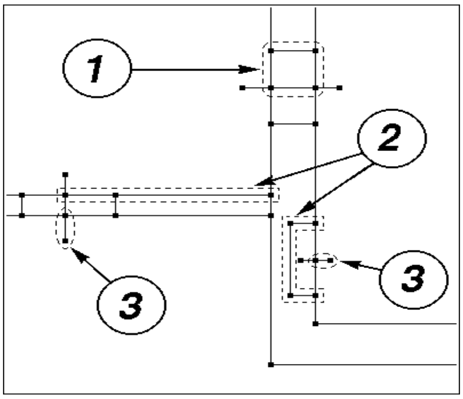

Figure 1. An example of a WFGD represented in the form of SNIM.

Figure1 shows a fragment of the WFGD represented in the form of SNIM. The segment is a broken line, but in particular cases it is a leg between intersections of this line with other lines in the image (Figure 1, positions 1, 2). In this case, the segments are often the boundary between two contour objects. For a section of the line limited by the intersection with another line only on one side, the concept of a segment with a free end was introduced (Figure 1, position 3). In general, Figure 1 shows a fragment of a segment-node image model of a WFGD containing the image of the elements described above that are characteristic of the vector representation of the WFGD.

With a view to a more rigorous description, it is advisable to associate a formalized representation of the SNIM with graph theory [10, 11]. If we impose the restriction that the SNIM segment is exclusively a straight line space bounded on both sides by nodes, then the term “segment” can be considered synonymous with the term “link”. Thus, the SNIM graph is a set ***G*** = ***{U, S}*** consisting of two subsets. The subset ***U*** consists of elements of the “node” type: **U = *{U_1, U_2, …, U_p}***, and the subset ***S***, of the elements of the “link” type: ***S* = *{S_1, S_2, …, S_q}***. When each vertex of the graph ***G*** is assigned its identifying number, it will be possible to refer this graph to the class of labeled graphs.

Another feature of the SNIM graph is its original characteristic for each of its nodes/vertices: the coordinates *(X, Y)* of the node's location on the plane. This feature makes it possible to uniquely map the graph ***G*** on the plane, which distinguishes it from most other graphs of arbitrary nature, for which the specific form of the schematic image (graph diagram) is often not essential. Complete certainty when mapping the SNIM graph on a plane allows us to speak of this model as an image model.

The use of SNIM is especially effective when processing WFGDs with a distinct topological load on linear geometric elements of the image [7–9]. In this case, the topological model is determined by the presence and storage of sets of interrelations, such as interconnected arcs at intersections, an ordered set of segments forming the boundary of each contour, etc. The topological properties of the figures do not change with any deformations that occur without breaks or connections. The topological feature of the WFGD is the presence of a large number of raster rectilinear objects forming mutual intersections. The image of the graph chart has similar features, and due to this the application of SNIM, which is also described in terms of graph theory, is particularly effective precisely for this kind of documents.

Unlike object vector WFGD models, SNIM does not contain objects in the usual sense, such as contours, vectorized lines, segments, etc. The transition from SNIM to the level of object vector models is a separate task. CIM, LCIM and LIM are associated with different vectorization algorithms and provide geometric interpretation of images in the tasks of scene analysis and recognition, as well as the metric description of the information components of the raster [4, 5, 7-9].

4. Problem statement

Vectorization is a basic operation in most processing and analysis systems for graphic images. If a RIWFGD consists mainly of interacting linear extended objects (diagrams, technical drawings and plans, hydrographic maps and plans) with a clearly defined direction of their orientation, then in the case of significant amounts of initial raster data, the dashed form of representation allows reducing them and building simple and reliable algorithms for vectorization and, if necessary, geometric segmentation into linear and area raster objects. It means that effective geometric modeling of RIWFGD can be achieved, and in the case of small volumes of source raster data it is possible to build sufficiently time-effective algorithms directly from RIWFGD in pixel form.

If the RIWFGD of vectorized objects do not contain distortions and noise, then the existing local vectorization algorithms can cope with this task quite well, although it must be borne in mind that the vectorized objects that are being obtained require additional smoothing or approximation. However, it is not always possible to meet the requirements of the metric accuracy of the approximation and the geometric accuracy of a vectorized object at the same time [4, 5, 7–9].

In the framework of this research, the source data for the proposed vectorization technology is a binary raster image (BRIWFGD) with geometric dimensions N x M, which is the description of the RIWFGD source document as a two-layer pixel object:

$$R_{ij} = \begin{cases} 1, \text{if the pixel belongs to the sign layer;} \\ 0, \text{if the pixel belongs to the background layer;} \end{cases} \quad i = 1, \ldots, N; \quad j = 1, \ldots, M$$

We will look for the sought-after geometric model as a collection of sets of non-derivative geometric elements: topological nodes U, segments S, and contours K (Figure 2):

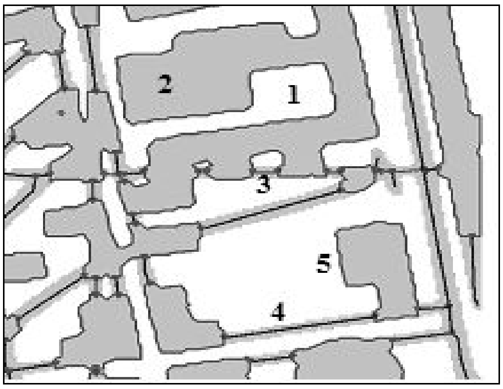

Figure 2. Non-derivative elements of the geometric model of the image.

$U = \{U_i\}$, i=1, 2, 3, …, N_u;
$S = \{S_i\}$, i=1, 2, 3, …, N_s;
$K = \{K_i\}$, i=1, 2, 3, …, N_k).

Figure 2 shows these elements against the background of RIWFGD, where:

- areas 1 define the background raster layer;
- areas 2 define the sign raster layer;
- lines and nodes 3 define topological nodes and short segments;
- lines 4 define segments of centerlines;
- lines 5 define contours of polygon objects.

5. Original procedures and operations for producing a geometric model of WFGD

We propose a set of improved and original algorithms for WFGD vectorization based on a low-level model for describing RIWFGD.

The composition of the sets of non-derivative geometric elements is described below.

A topological node is described by a set of numerical characteristics *U{x, y, swT, spN},* where: *x, y* are the raster coordinates of the node, with the node pixel always belonging to the sign layer; *swT ={0, 1, …, v}* is the coefficient of topological connectivity, which determines the number of segments originating from the node; *spN* $=\{N_1, N_2, …, N_s\}$ is the list of segment numbers originating from this node.

The segment $S=\{x_1, y_1, …, x_t, y_t\}$ is an inter-node fragment of the centerline. It consists of a collection of connected pixels of the sign layer belonging to the centerlines of the RLO, where *t* is the number of linear approximation nodes of the set of connected pixels of the inter-node intervals of the centerline.

The contour K=$\{x_1,y_1,…, x_k,y_k\}$ is a description of the RAO boundaries by a collection of connected boundary pixels of the sign layer, where k is the number of nodes of the linear approximation of the set of boundary RAO pixels.

Using a two-layer model of raster pixel data and the chosen vector description model, it is easy to extend existing methods and algorithms for a vector description of a set of raster composite objects (RCO) of multicolor documents, including full-color raster images of Earth remote sensing, as well as raster hyperspectral images (HSI) that have undergone color (spectral) layering (clustering).

Structural analysis of the sign layer of BRIWFGD shows that this layer contains the following RCOs:

- noise objects of quite small geometric dimensions (“snow”);
- small-sized objects, which are images of elements of a set of discrete signs;
- large-sized objects, which are isolated linear and areal signs, or a conglomerate of the results of the superposition (merger, tangency) of linear, discrete and areal signs.

Within the framework of CGA, the original hierarchical model proposed by the authors for the representation of raster and vector SDD and as our contribution to the development of existing methods and algorithms, we propose the following sequence of original procedures and operations for constructing a WFGD geometric model based on the collection of raster composite objects (RCO) of BRIWFGD, providing a high level of confidence during the semantic interpretation of the document.

Stage 1 Construction of a RCO model from the sign layer of BRIWFGD.

1.1 Compiling a list of sign pixels in the sign layer *Z* of the raster R_1: ListZP = $\{Pix_i=1\}$(i=1, 2,…, Nz), where: Pix_i is the sign pixel with the coordinates x_i,y_i; Nz is the number of sign pixels of the *Z* layer.

1.2 While the ListZP is not empty, the next, not yet clustered pixel is extracted from it, and, starting from it, the RCO is formed by clustering along the 8 - connected neighborhood.

1.3 After completing the clustering procedure, the set of pixels constituting the next RCO are converted into a vector contour (stage 2) and are deleted from the ListZP list.

1.4 If the ListZP is empty, then the stage of construction of the RCO model is considered complete.

Stage 2 Contour analysis of the RCO model.

2.1 Selecting a set of boundary pixels G from the raster sign layer for each RCO.

2.2 Construction of the CIM by following (based on pixels G) the boundaries of the RCO that make up the layer *Z*.

2.3 Parametrization of all contours constructed, by measuring for each contour the following geometrical characteristics: p_1 is the contour length along its perimeter; p_2, p_3 are the width and height of the minimum area rectangle circumscribing the contour, with its sides parallel to coordinate axes; p_4 is the contour area; $p_5 = p_4 \big/ p_1^2$ is the external aspect of the contour.

2.4 Segmentation of contours depending on parameters $p_1 - p_5$ into: noise contours, which are subsequently removed from the initial layer *Z*; discrete signs; isolated areal objects; small-length linear segments and producing from them the set **V** with the subsequent removal of the corresponding RCOs from the initial layer *Z*.

Thus, the output result of the stage is a CIM containing vectorized elements of the *Z* layer and a modified raster $\widetilde{R}_1$ which contains no RCOs that describe vectorized objects.

Stage 3 Splitting of large-sized RCOs of the sign layer into areal and linear signs and their description by a set of geometric elements such as nodes, segments and contours.

3.1 Splitting the modified sign layer of pixels $\widetilde{R}_1$ using its skeleton model [12] into the following types: “linear” pixels belonging to the centerlines of the RLO; "areal" pixels belonging to RAO; “nodal” pixels belonging to the nodal points of the RLO and RAO intersection.

The operation is performed by the method of sequential, parametrically controlled D - multiple morphological pixel operations of diffusion and dilation [12], where D is the half-thickness of the RLO.

The classification of pixels into the types listed above is carried out on the basis of the original recognition mask filters, which make it possible to identify, at a pixel level, certain types of graphic situations with a high degree of confidence.

3.2 For “nodal” pixels, their coordinates are measured and they are entered into the set **V** of vector data as a vector description of the topological nodes of the desired geometric model.

3.3 By using “linear” pixels, the inter-node gaps of the center line are tracked, it is linearly approximated, and the tracking results are entered into the **V** set as LRO segments.

3.4 For “areal” pixels, the procedure of D-multiple “spraying” and the subsequent construction of the contour of the areal object are performed, followed by linear approximation and entering the approximated contour into the set **V**.

3.5 For nodal pixels, their topological characteristics are calculated, based on the constructed metric of nodes, segments and contours: the connectivity coefficients K_{sw} and the list of numbers of segments originating from the nodes.

Thus, after completing all operations, the set of vector data **V** will be formed, consisting of vector elements of the following types: node points, contours, segments that uniquely define the desired geometric model of the original WFGD.

The targeted parametric control of the input parameter D for the whole procedure (LRO half-thickness) can be obtained from the skeleton of the RIWFGD.

Figure 2 shows an adequate geometric model where objects are split into contour (areal) and linearly extended ones, with topological characteristics of their interaction (tangency or intersection).

To support the whole variety of algorithms for classification of signs in graphic documents, the set of vector elements **V** thus obtained is supplemented for the current document with a description of its sign pixel layer **Z** in a dashed format, which allows constructing character recognition algorithms based on their combined and consistent synchronous description (vector and pixel). This feature distinguishes the proposed methods and algorithms for describing RIWFGD from the existing ones.

Figure 3 shows a fragment of a geometric model of a WFGD in a terrain map with a large number of linear objects. The model was obtained by applying the proposed technology.

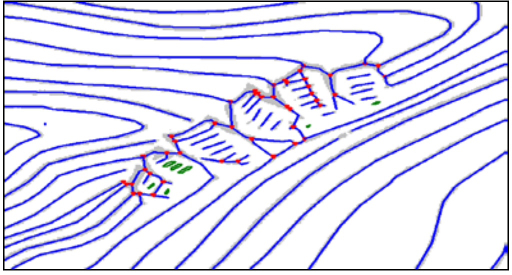

Figure 3. A fragment of a geometric model of a WFGD in a terrain map with a large number of linear signs.

6. Conclusion

In the course of our research, documents of the WFGD type were analyzed, and their main features that determine the choice of models, methods and algorithms for their processing were identified. The effectiveness of the extended class of vector models (LIM and SNIM) for the description of WFGDs with a distinctive linear orientation of objects was confirmed during practical experiments on real WFGDs.

7. References

[1] Vasin Yu G and Yasakov Yu V 2016 Distributed database management system for integrated processing of spatial data in a GIS *Computer Optics* **40(6)** 919-928 DOI: 10.18287/2412-6179-2016-40-6-919-928

[2] Khafizov R G, Okhotnikov S A and Yaranceva T V 2016 Models of the image of object contours with geometrical distortions *Computer Optics* **40(3)** 404-409 DOI: 10.18287/2412-6179-2016-40-3-404-409

[3] Belim S V and Kutlunin P E 2015 Boundary extraction in images using a clustering algorithm *Computer Optics* **39(1)** 119-124 DOI: 10.18287/0134-2452-2015-39-1-119-124

[4] Vasin Yu G, Bashkirov O A and Chudinovich B M 1987 Combinotary geometric approach in complex graphic data analysis tasks *Automation of complex graphic information processing: Inter-university collection* (Gorky: Gorky State University) (in Russian)

[5] Vasin Yu G and Bashkirov O A 1984 Mathematical models for structured description of graphic images *Automation of complex graphic information processing: Inter-university collection* (Gorky: Gorky State University) 92-117 (in Russian)

[6] Vasin D Yu 2015 Automation of characters input based on low-level models of graphic images description *Privolzhsky Science Magazine* **3(35)** 109-115 (in Russian)

[7] Vasin Yu G, Vasin D Yu, Gromov V P and Rotkov S I 2018 Robust vectorization of graphic documents with distinctive orientation of linear objects *Proc. of International Scientific Conference in Computing for Physics and Technology* (Tsargrad, Moscow region) 313-317 (in Russian)

[8] Vasin D Yu, Gromov V P and Rotkov S I 2018 Formation of segment and nodal model of graphic documents with distinctive orientation of linear objects *Proc. of International Scientific Conference in Computing for Physics and Technology* (Tsargrad, Moscow region) 265-280 (in Russian)

[9] Vasin Y, Vasin D, Utesheva T, Lebedev L and Kustov E 2017 Increasing the effectiveness of intelligent information technology for producing digital graphic documents with weakly formalized description of objects *Procedia Engineering Proc.* **201** 341-352 DOI: 10.1016/j.proeng.2017.09.642

[10] Harary F 1996 *Theory of graphs* (London: Addison-Wesley) p 274

[11] Christofides N 1986 *Graph theory: an algorithmic approach* 4th (London: Academic Press) p 400

[12] Gonzalez R and Woods R 2018 *Digital image processing* (New York: Pearson) p 1019

Acknowledgements
This work was supported by the Russian Foundation for Basic Research, RFBR grants No. 17–07–00543 and No. 18–07–00715.