LINGUISTIC SUPPORT OF THE KNOWLEDGE BASE FOR IMAGE ANALYSIS AND UNDERSTANDING SYSTEM

Yulia Trusova, Igor Gurevich, Victor Beloozerov and Dmitri Murashov

Dorodnicyn Computing Center, Russian Academy of Sciences
40 Vavilov str.,119991 Moscow, Russian Federation

Keywords: Image analysis, pattern recognition, knowledge bases, domain thesauri, ontologies, knowledge portals.

Abstract: The problem of lexical and semantic support of the knowledge base for the system for automation of scientific research in image processing, analysis and understanding is discussed. The main contribution is the image analysis thesaurus which has been developed as a main tool for solving this problem. A structure of the thesaurus and functional characteristics of the basic version of the thesaurus are described. Lexical categories of terms and relationships between terms in the domain of image processing, analysis and recognition are considered. The thesaurus was implemented as an autonomous program module. The description of the thesaurus module and its use are provided. The developed thesaurus was applied for automation of early diagnosis of hematological diseases on the base of cytological specimens.

1 INTRODUCTION

The problem of scientific research automation in the subject domain of image processing, analysis and understanding is one of the fundamental problems of computer science. The paper is devoted to the description of the image analysis thesaurus, which: 1) allows systematization of poorly structured and changing terminology in the domain of image processing, analysis and understanding; 2) provides automation of an information retrieval in knowledge bases for image processing, analysis and understanding; 3) is a stand-alone reference book, which helps to navigate in the subject domain.

At present, universal systems designed for image processing, analysis, and understanding, which are not related to a specific subject domain, attract a lot of interest. A knowledge base is the most important component of such systems. It contains knowledge on the image processing, on the classes of scenes analysed, and on the available computational methods (Bertino et al., 2001). For efficient information retrieval in the knowledge base it is necessary to have a tool for semantic interpretation and matching of textual object descriptions and user queries. In practice a domain thesaurus can be used to solve the problem. The relationships between terms fixed in the thesaurus help to specify and

extend the user query for more successful information retrieval.

For several recent years the authors have been developing the knowledge base for the "Black Square. Versions 1.0, 1.1, 1.2 software system for the automation of scientific research in image processing, analysis, recognition, and understanding" (KBBS 1.0) (Gurevich et al., 1999; 2006).

KBBS 1.0 is aimed at the support and automation of solving problems of image analysis, estimation, understanding, and recognition. The automation depends extremely on solving the following main problems: 1) automation of image analysis algorithm retrieval; 2) automation of algorithm development and combination; 3) algorithm matching on the basis of its comparative effectiveness, accuracy, and ability to be programmed. To solve these problems it is necessary to use a thesaurus on image processing, analysis and understanding.

Analysis of the literature testifies that till the present time no thesauri were developed for the domain of image processing, analysis, and recognition. "Image Analysis Thesaurus. Version 1.0" (IAT 1.0) compensates the lack. The thesaurus presented is being used to solve the following tasks:

- classification of algorithms and tasks of image processing, analysis, understanding and recognition;
- generation of the descriptions of algorithms and tasks of image processing, analysis, understanding and recognition;
- automation of information retrieval;
- classification and retrieval of bibliographic and reference data.

One of the IAT 1.0 distinctive features is that it can be used not only as a part of KBBS 1.0, but also as a separate linguistic resource. IAT 1.0. is a bilingual and contains terms and their definitions in two languages (Russian and English).

The IAT 1.0 was applied for automation of early diagnosis of hematological diseases on the base of cytological specimens. The application confirmed its efficiency. Its details will be described in future.

2 THE USE OF IAT 1.0

In general, a thesaurus is a controlled vocabulary of terms and relationships between them. The thesaurus structure, its lexical content and program implementation depend on subject domain specificity and tasks to be solved (Aitchison et al., 2002).

IAT 1.0 can be used as a stand-alone reference book on image processing, analysis and recognition. It contains definitions of terms and references. IAT 1.0 can be recommended to both professional and non-professional users. In particular, it will help those users who are not specialists in the subject domain to use efficiently KBBS 1.0.

The basic version of IAT 1.0 contains 1538 terms, including 230 terms in "Image" section, 634 terms in "Image Processing" section, 464 terms in "Image Analysis" section, and 210 terms in "Pattern Recognition" section. The maximum number of hierarchy levels is 6.

Below we consider the main functional characteristics of the IAT 1.0 in the framework of its use in KBBS 1.0.

2.1 Descriptions of Algorithms

Textual description of an algorithm in KBBS 1.0 consists of a name of a task (goal), a name of an algorithm, description of input and output data, context and references. For that terms of the following functional categories are included in IAT 1.0:

a) "Objects", which includes:

- names of image types (e.g., aspect image, range image, 2D image, quantized image, etc.);
- names of image elements (e.g., contour, region, pixel, etc);

b) "Tasks", which includes:
- names of classes of image processing tasks (e.g., image enhancement, image restoration, image quantization, etc.);
- names of classes of image analysis tasks (e.g., image segmentation, texture analysis, etc.);
- names of classes of pattern recognition problems, including names of image recognition tasks (e.g., feature selection, error estimation, etc.);

c) "Instruments", which includes:
- names of classes of image processing instruments (methods, algorithms, techniques, operations, functions, operators, transformations) (e.g., median filtering, Hough transform, etc.);
- names of classes of image analysis instruments (methods) (e.g., contour-based shape descriptor, region growing method, etc.);
- names of classes of pattern recognition methods, including names of classes of image recognition techniques (e.g., maximum likelihood decision rule, cluster assignment function, etc.);

d) "Properties", which includes:
- names of instrument properties (e.g., hexagonal sampling grid, structuring element, convolution kernel, etc.);
- names of image description elements (e.g., brightness, color model, contrast difference, etc.).

The example of algorithm description is as follows.

1. Task name: *median filtering*.

2. Task goal: *noise removing*.

3. Input data: *gray-level image* (image depth: 8 bpp; image width: 1024 pixels; image height: – 1024 pixels).

4. Result: *gray-level image* (image depth: 8 bpp; image width: 1024 pixels; image height: – 1024 pixels).

5. Operator name: mediana.

Each element of the description is, in turn, an object characterized by a set of properties. The latter for such objects can be described by IAT 1.0 descriptors.

2.2 Classifications of Algorithms and Tasks

For automation of image processing, analysis and recognition the uniform descriptions of standard and solved tasks were included in the KBBS 1.0.

General classification of tasks of image processing, analysis and recognition is based on representation of tasks in the form of a sequence of operations. In this sequence each operation corresponds to a task of image processing, analysis and recognition.

Task classification in the KBBS 1.0 is developed on the basis of the IAT 1.0. The classification is based on the functional hierarchical classification of algorithms for basic operations of image processing, analysis and recognition.

The following lists are examples of hierarchical classification of thesaurus terms related to image processing operations (a) and image processing tasks (b):

a) image processing operation
    geometric image processing operation
    linear image processing operation
    mathematics-based image processing operation
       arithmetic-based image processing operation
         image addition
         image blending
         image division
         image multiplication
         image subtraction
    morphology-based image processing operation
    neighborhood image processing operation
    non-linear image processing operation
    point image processing operation
    smoothing image processing operation
     ...
b) image processing task
    image compression
    image enhancement
      contrast enhancement
       histogram equalization
       ...
      edge enhancement
      image sharpening
      image smoothing
      noise suppression
    image preprocessing
    image restoration
...

2.3 Planning and Control of Problem Solving

The main objective of the KBBS 1.0 is the support of planning and control of problem solving. To this end, the logical and pragmatic relationships between terms representing task descriptions and solution techniques should be defined and included into the thesaurus.

According to the specificity of the domain IAT 1.0 contains the following basic relationships between descriptors:

- image type – image description element (e.g., video-image - aspect ratio);
- process – applied instrument (e.g., edge detection – Hueckel edge operator);
- image transformation – result (e.g., thresholding - binary image);
- applied instrument – instrument characteristic (e.g., morphologic dilation operator - structuring element);
- applied instrument – result (e.g., edge detector – edge map);
- image type – image acquisition technique;
- image type – image transformation.

2.4 Applied Terminology of IAT 1.0

The IAT 1.0 was experimentally tested on the problems connected with automation of cytological image analysis. The applied part of the thesaurus includes the following hematological terms:

- names of blood cells classes;
- names of cells parts and organs;
- names of cells morphological characteristics;
- morphological characteristics values;
- names of physiological processes in blood;
- diagnostic terms.

As a source of hematological terminology we used atlases of blood cells and tumors of lymphatic system (Vorob'ev, 2001). The number of term records is more than 350.

The relationships between hematological descriptors are defined by standard for thesauri relationships (ISO-5964, 1985) – hierarchical generic and whole-part relationships. Other specific relations between terms are defined by associative relationships (e.g., relations between names of morphological characteristics and characteristic values, relations between characteristics and names of blood cells classes, and relations between blood cells classes and diagnostic terms).

3 PROGRAM IMPLEMENTATION

The IAT 1.0 was implemented in Visual FoxPro 7.0. The IAT 1.0 program module requires the following resources:

- Intel Pentium 200 processor and higher;

- 64Mb main memory;
- 50Mb for programs;
- operating system - Windows 98/ME/2000/XP.

Detailed description of the IAT 1.0 module is presented in (Beloozerov et al., 2003).

The thesaurus module provides the following functions:

- visualizating and editing of the hierarchical structure of terms;
- adding and editing of terms, descriptors and records;
- adding and editing of relationships between terms;
- context searching in the database of a thesaurus.

The IAT 1.0 module consists of the database of a thesaurus, software tools for database control, a user interface, a database interface, a LAN interface, and an Internet interface.

The database of the IAT 1.0 contains descriptor records and the following main tables:

- the table of descriptors;
- the table of terms;
- the table of definitions;
- the table of relations between terms;
- the table of the types of relations;
- the table of languages;
- the table of user interfaces,

and index files.

The user interface is employed to present a system of definitions, to visualize the hierarchical structure of terms and other types of relations between the terms, to add and edit the records of a thesaurus, to formulate queries, and to represent the search results.

The user interface consists of graphic forms and user menus:

- window menu;
- "Thesaurus display" form (Figure1);
- menu for editing thesaurus structure;
- "Adding and editing of terms" form (Figure 2);
- "Adding and editing of descriptor relationships" form;
- "Search" form.

4 CONCLUSIONS

At present we are developing the Internet reference-providing information resource on image processing, analysis and recognition based on the presented thesaurus on image analysis.

The Internet is the main and highly available source of information. So, the data reflecting achievements and problems in the domain of image processing, analysis and understanding should be presented on the Internet.

The available tools for information representation and retrieval on the Internet do not provide the effective exploitation of Internet resources for automated image processing and analysis. The reasons are insufficient volume of data, insufficient data systematization and disadvantages of retrieval tools searching for the formal coincidence of terms in user query and information source, and invalidity of data.

For successful information retrieval it is necessary to have an ontology of the domain of interest. The practice of development of information retrieval systems shows that domain ontology can be adequately represented with the help of information retrieval thesaurus, where objects are represented by descriptors and semantic relations of objects are represented by the formal relations between descriptors reflecting semantic content of notion, logical and pragmatic relations between notions.

The Internet resource on image processing, analysis and recognition will contain:

1) a reference book in the field of image processing, analysis, and recognition in the form of a thesaurus;

2) a bibliographic database of descriptions of papers and monographs, and web links to the electronic publications in the given domain;

3) tools for relevant information retrieval on the Internet;

4) a catalogue of Internet resources on image processing, analysis, and recognition including (a) web links to electronic libraries, (b) web links to bibliographic databases, (c) a list of the websites of institutions, scientific centres, laboratories and IT companies involved in research and development in the field of image processing and analysis, (d) a list of the websites of publishing houses, and (e) a regularly updating list of relevant conferences with their websites.

The Internet reference-providing information resource on image processing, analysis and recognition will provide integration of existing information sources and will support intelligent information retrieval in the domain of image processing, analysis, recognition and understanding.
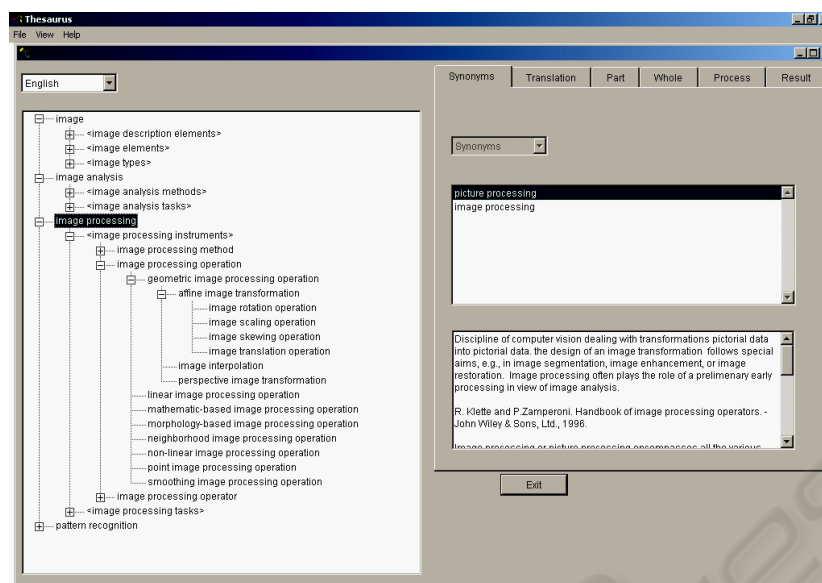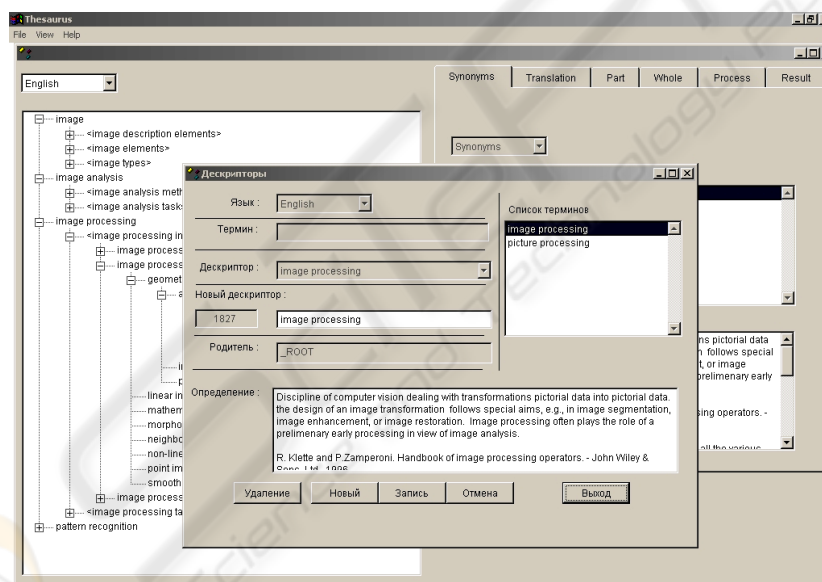
Figure 1: "Thesaurus display" form.

Figure 2: "Adding and editing of terms" form.

ACKNOWLEDGEMENTS

REFERENCES

Aitchison, J., Gilchrist, A., Bawden, D., 2002. *Thesaurus construction and use: a practical manual.* Aslib. London, 4th edition.

Beloozerov, V.N., Gurevich, I.B., Gurevich, N.G., Murashov, D.M., Trusova, Yu.O., 2003. Thesaurus for Image Analysis: Basic Version. In *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, 13 (4), 556-569.* MAIK "Nauka/Interperiodica", Moscow.

Bertino, E., Catania, B., Zarri, G.P., 2001. *Intelligent Database Systems.* ACM Press.

Gurevich, I.B., Khilkov, A.V., Koryabkina, I.V., Murashov, D.M., Trusova, Yu.O., 2006. An Open General-Purposes Research System for Automating the Development and Application of Information Technologies in the Area of Image Processing, Analysis, and Evaluation. In *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, 16 (4), 530-563.* MAIK "Nauka/Interperiodica", Moscow.

Gurevich, I.B., Murashov, D.M., Zhuravlev, Yu.I. et al., 1999. Knowledge-Based System for Automatization of Scientific Research in Image Analysis and Understanding. Part 1. In *Optoelectronics, Instrumentation and Data Processing (Avtometria), 6, 18 –36.*

ISO-5964: 1985. Documentation - Guidelines for the establishment and development of multilingual thesauri.

Vorob'ev, A.I. (ed.), 2001. *Atlas "Tumors of lymphatic system".* Hematological Scientific Center of the Russian Academy of Medical Sciences.