ICIC Express Letters
Part B: Applications
Volume **3**, Number **4**, August **2012**

ICIC International ©2012 ISSN 2185-2766
pp. 901–908

SEMANTIC ANALYSIS OF MULTI-MODAL FEATURES IN SCIENTIFIC AND TECHNICAL LITERATURE

RUIJIA WANG AND YAO LIU*

Information Technology Support Center
Institute of Scientific and Technical Information of China
No. 15, Fuxing Road, Haidian District, Beijing 100038, P. R. China
*Corresponding author: liuy@istic.ac.cn

Received February 2012; accepted May 2012

ABSTRACT. *Scientific and technical literature contains images, tables, formulae, audio files and videos besides the common text information, which will help the users to fully understand the knowledge presented in the literature, so we can take the resource of scientific and technical literature as a kind of multi-modal information. In this paper, we use the multi-modal information to analyze the different semantic modality features in the scientific and technical literature. To be specific, we choose scientific and technical literature in PDF format, and recognize tables, formulae and four kinds of images - the flow chart, bar chart, curve chart and interface diagram from the literature, then analyze the multi-modal information in semantic level, and find the relations between different modality features, in order to optimize the semantic representation of the scientific and technical literature.*

Keywords: Multi-modal information, Scientific and technical literature, Semantic representation, Semantic feature

1. **Introduction.** The modern science and technology is developing swiftly, leading to the rapid expanding of new knowledge and information in the scientific and technical literature and diversification of knowledge expression forms; users can acquire knowledge more directly and easily. The resource of scientific and technical literature is a kind of multi-modal data; it contains various modalities of information, such as images, tables, formulae, audio files and videos. The multi-modal information could complement the text, which is the main information in the literature, and help the users to fully understand the knowledge in the scientific and technical literature. This research utilizes multi-modal semantic information features and the complementarity between them to make semantic representation of images, tables and formulae in scientific and technical literature, and finds the semantic correlation of different modality features, in order to optimize the semantic representation of the scientific and technical literature.

2. Concept and Research Status of Multi-modal.

2.1. **Basic concept of multi-modal.** "Multi-modal" is used in contrast to "unimodal" or "single-modality". At present, there is no specific and generalized definition of "multi-modal", and the research of multi-modal is generally focused on solving a certain problem with at least two modalities of information. The earliest research about multi-modal is the multi-modal pattern discrimination test in 1968 [1]; after that, the concept of "bimodal signals" is proposed based on the concept of "unimodal signals" in the literature of the function research of signal detection in 1970, and named "multimodal signal detection" [2]. During the same period, the concept of multi-modal is used in the research fields of medical multi-modal therapeutic methods, multi-modal learning in biological system

and so on. In the late 1990s, the research of multi-modal is increased and the application fields are broader than before.

"Multi-modal" is a new research field, with multi-modal methods we can solve certain problems with different modalities of information, so it should not be limited to one particular field. Research about multi-modal both domestic and overseas involves different fields, such as multi-modal image automatic indexing and retrieval, multi-modal medical image registration and fusion, multi-modal identity recognition, multi-modal video classification and retrieval, multi-modal human-computer interaction system, multi-modal discourse analysis, robot target recognition and multi-modal emotion recognition.

2.2. **Research status of image semantic representation.** There are lots of images in scientific and technical literature, so the extraction and expression of image semantic features are important in the research work. Research about image semantic features both domestic and overseas involves automatic image annotation and image retrieval.

2.2.1. *Automatic image annotation.* Making the computer label the image with keywords and bridging the gap between low-level visual features and high-level semantic features of image are the main purpose of automatic image annotation. There are four kinds of existing methods, classification-based method, machine translation method, probabilistic model method and method based on Internet dataset.

2.2.2. *Image retrieval.* TBIR (Text-Based Image Retrieval) was used for image retrieval in the early days since 1970s, at which time people tagged the images manually because the amount of images is relatively small. However, TBIR cannot fulfill the requirements in large-scale image database with the development of digital photography and the Internet, it was time-consuming and the tagging results were not accurate because different people have different understanding with a certain image. So CBIR (Content-Based Image Retrieval) was proposed in early 1990s to solve this problem.

In February 1992, the US National Science Foundation (NSF) organized a work shop to "identify major research areas that should be addressed by researchers for visual information management systems that would be useful in scientific, industrial, medical, environmental, educational, entertainment, and other applications" [3]. In CBIR task, user poses a query image and the system returns a set of relevant images which have the similar visual features with that extracted from the image the user posed.

People do not use the similarity of low-level visual features of images to judge whether two images are similar or not; they use high-level and perceptive similarity instead of that. Concept of image is established based on the objects and behaviors described in the image and emotions conveyed in the image, which combines lots of experience from daily life. So SBIR (Semantic-Based Image Retrieval) was proposed based on CBIR, which contains human understanding of image and is more reasonable, but it is in the start stage because the computer vision technology and image understanding technology are not well developed.

There are some multi-modal image retrieval systems available, such as QBIC, Virage, RetrievalWare, Photobook and VisualSEEK.

2.3. **Research status of table semantic representation.** Table recognition is an important application field of OCR (Optical Character Recognition); the initial table recognition technology was image-based because scanned image is composed of pixels [4]. Research in this field includes image preprocessing, table recognition and table contents extraction, and plenty of attention was paid on discussing and improvement of the system and key technologies.

2.4. **Research status of formula semantic analysis and representation.** Dr. Anderson addressed the problem of formula recognition for the first time in 1968 [5], and the research was not rapidly developed until 1990s.

Mathematical formula is widely used in the Internet, and people need to obtain relative information by searching the formula. Mathematical formula searching is difficult because the formulae have special structure, and we cannot process them with standard natural language processing methods. There are two categories of methods existing: one is parsing and translating the mathematical expressions into a set of natural language keywords and using normal information search methods; the other is structure-based [6], the XML structures of formula are directly indexed and compared with the XML structures of the queries. Muhammad [7], Michael [8], both have developed search engines for mathematical expression. Muhammad preprocessed the formula both entered by users and saved in the database to discover important patterns of the formula and assigned meaning to them, the final document vector could be constructed by counting their respective number of occurrences. Michael indexed the formula with substitution tree indexing and searched the formula via their structure and meaning.

3. Semantic Analysis Content and Scheme.

3.1. Semantic analysis content.

3.1.1. *Theoretical and method research of multi-modal semantic analysis.* Analyze the main fields, progress and development direction of multi-modal research both domestic and overseas, emphasize the integration of multi-subjects.

3.1.2. *Inherent rule and semantic correlation of multi-modal features.* Discuss the relationship between multi-modal and semantics, find the semantic correlation between features of different modalities, and construct a structured semantic description system which focuses on the entities, relationships and events in scientific and technical literature, in order to generate semantic representation of the literature content.

3.1.3. *Single mode information analysis and feature extraction.* Develop integrate methods to analyze and extract semantic multi-modal information, and generate effective analysis and extraction of image, table, formula and text in scientific and technology literature.

3.1.4. *Multi-modal semantic feature extraction and expression.* Find the statistical relationship between different multi-modal semantic features, then construct co-occurrence matrix of multi-modal features to generate isomorphic subspaces with different data types and reflect the relevance, and finally generate the expression of relationship between semantic features of different modalities.

3.1.5. *Multi-modal fusion and expression model construction based on context.* Develop a algorithm which is suitable for fusion and collaborative analysis of multi-modal information, generate the fusion of multi-modal semantic features and develop theories and the methods to extract features of multi-modal information.

3.2. **Semantic analysis scheme.** In the research we will use various theories and methods, especially the natural language processing methods. With the technology and system of ontology construction based in semi-structured text, we will propose fusion and expression methods of multi-modal semantic features based in context, develop technologies to analyze the texts, images, tables and formulae in scientific and technical literature and extract their features. We will find the inherent rules and complementary of multi-modal features, and construct a structured semantic description system which focuses on the entities, relationships and events in scientific and technical literature, in order to provide

theoretical and technical support for content understanding and knowledge service. Our research scheme and technology route are as follows:

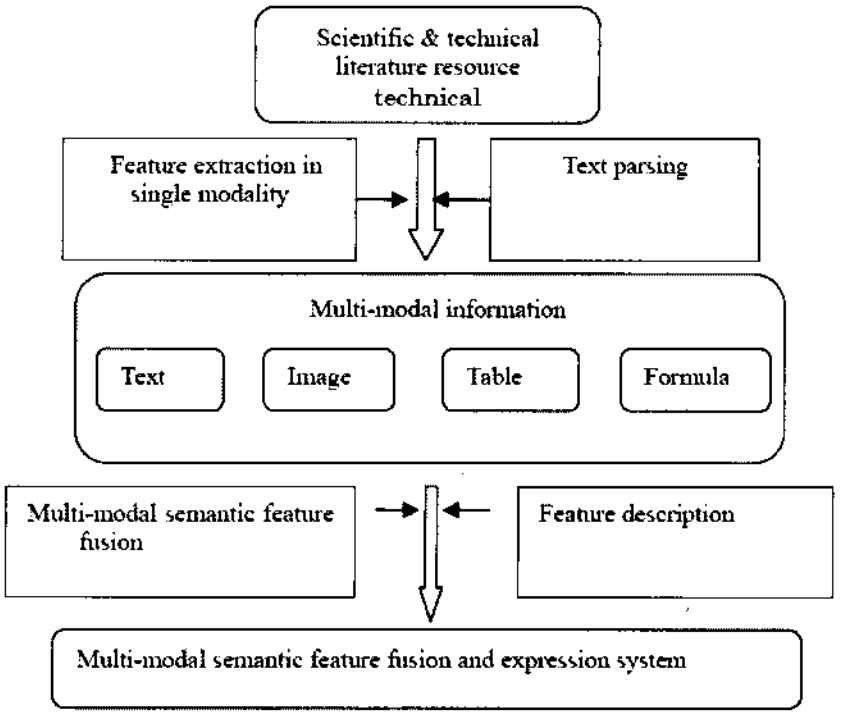

FIGURE 1. Semantic analysis scheme and technology route

4. **Recognition and Semantic Analysis of Multi-modal Features in Scientific and Technical Literature.** Scientific and technical literature has different formats and should be processed with different methods. With PDF files people can browse and exchange electronic documents freely and easily, and do not have to worry about the recognition problem in electronic documents sharing. PDF format has become the open standard of global electronic documents distribution and most scientific and technical literature is saved in this format. Therefore, the research object in this paper is scientific and technical literature in PDF format.

4.1. **Image recognition and semantic analysis.** Visual features include color feature, texture feature and shape feature.

4.1.1. *Color feature of image.* Color is an important visual feature of image and is widely used in image processing. Contrast with geometric features, color feature has better stability and color constancy, and it is less sensitive in size and direction. So color is an effective and the simplest feature to describe an image. Color histogram is the main method to describe color features.

Color histogram is a presentation method of whole image, it reflects the composition and distribution of color, the displayed colors and the probability of a certain color in an image. Color histogram is described as follows:

For image I, the image size is $m \times n$, if $N(C)$ represents the numbers of pixels in color C, the color histogram of the image could be defined as:

$$P(C) = \frac{N(C)}{m \times n}, \quad (C = 0, 1, \ldots, L-1)$$

In this formula, L represents numbers of colors, $P(C)$ represents the occurrence frequency of color C.

4.1.2. *Texture feature of image.* Texture is an internal feature related with surface materials of objects in an image, which contains important information about surface composition and surrounding environment. The change and distribution of texture has repeatability, uniformity and directionality. Structural analysis method, spectrum analysis method and statistical analysis method are three traditional description methods of texture.

4.1.3. *Shape feature of image.* Shape is an important feature to describe the outline and physical structure of objects, which can be used to intuitional distinguish objects and has higher semantics than color feature and texture feature. In two-dimensional images, shape is identified as a region surrounded by closed outline curve. There are two kinds of methods for shape presentation, one is boundary-based method, and the other is region-based method.

4.2. **Table recognition and semantic analysis.** Most scientific and technical literature is presented in PDF format. There are two categories of methods to recognize tables in PDF files and extract the contents, one category of methods converts the files into other formats (like XML, txt and image) and recognizes tables in converted files; and the other category of methods recognizes the tales in original PDF files.

Yildiz, Kaiser and Miksch [9] have developed heuristic-based approaches to recognize and decompose tables in PDF files, and they utilized the absolute coordinates of text elements in PDF files to extract table information. Liu, Mitra and Giles [10] proposed a preprocessing method to improve the table boundary detection performance by considering the sparse-line property of table rows, and they identified that majority lines that belong to the table areas are sparse in terms of the text density. There are also methods to convert the PDF files into images and recognize tables by analyzing the layout of the images [11,12]. Image-based table recognition technology is relatively mature, but it also has limitations; for example, it cannot recognize the text information.

Since different writers may arrange their papers in different ways, tables in scientific and technical literature may exist in image format, or text format. Image-based table recognition technology can be used to analyze the image format tables, specifically, the table image should be preprocessed firstly, extract and merge the table lines after the preprocessing, finally extract the characters and recognize the characters with OCR technology. For tables in text format, graphic elements cannot be used to carry out the analysis, we could rasterize the table region in PDF files, and range the unstructured text information, in order to recognize the table content.

4.3. **Formula recognition and semantic analysis.** Recognizing formula in PDF files received more attention in the recent years. Files in PDF format are special and hard to process, by the influence of image based formula recognition, some researchers convert the PDF files into images and use the methods in image based formula recognition [13]. Because converting the file format would cost the loss of original information, some researchers proposed methods to analyze PDF files, and they believed the characters and layout features obtained with PDF files analysis are richer and more accurate than those got from OCR technology. Josef and Sexton [14] proposed an approach to extract perfect knowledge of the characters used in formula directly from the document, and they analyzed the extracted information into an abstract syntax tree, processed the tree to generate the LaTex output. They improved their approach by exploiting additional font and spacing information available from a PDF file, and the extracted information could be post-processed to produce markup that can be re-inserted into the PDF files [15,16].

Formulae in scientific and technical literature have standard form and could be recognized according to their space features, like row height and line space. In general, lines

with formula in the literature have following space features: the formula line is higher than text lines; the space between formula lines is wider than that between text lines; the formula line is center horizontally; there is formula number at the end of the formula line. Rules according to these space features can be made and formulae can be recognized according to the rules, the characters and symbols in the formulae can be recognized with OCR technology.

5. **Research Progress.** The multi-modal information should be recognized first in order to analyze their semantic features. There are two kinds of images in PDF files, one is XObject, which existed beside the content stream and has its own name; the other is inline image, the image attribute and data are embedded in content stream, this kind of image can present limited images. By processing the PDF files, this paper extracts the images in the files, as shown in the following figures. Figure 2 is a page in PDF files, and Figure 3 is the extracted images after the processing.

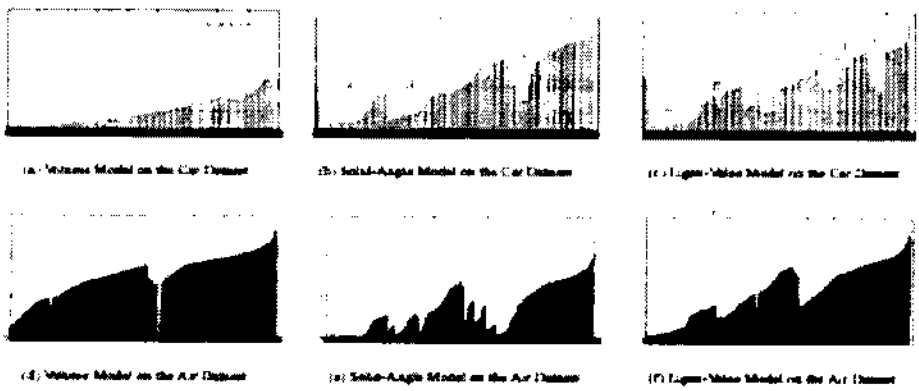

Figure 9: Reachability plots computed by OPTICS for the Car Dataset and the Air Dataset.

Evaluation of the Volume Model. The reachability plots computed by OPTICS using the Volume Model for both the Car Dataset and the Air Dataset are depicted in Figure 9(a) and 9(d). In both cases no clear classification of the objects can be found. None of the groups described in Section 4.1 were distinguished by the clustering algorithm. Although we got satisfying results using k-nn queries (cf. Figure 6), the Volume Model is rather ineffective if applied to the whole data set. This indicates the suitability of clustering to evaluate the quality of similarity models.

Let us note that according to the plot the objects are ranked with increasing reachability values along the ordering. An analysis of representatives shows that the objects are ordered according to their volume. This might be the explanation why the Volume Model is rather ineffective. Objects with related volume are modeled as similar. Most likely this similarity is too simple.

Evaluation of the Solid-Angle Model. The reachability plots computed by OPTICS using the Solid-Angle Model for both the Car Dataset and the Air Dataset are depicted in Figure 9(b) and 9(e).

On the Car Dataset the Solid-Angle Model provides three clusters denoted as *A*, *B*, and *C* in Figure 9(b). We analyzed the resulting clusters by picking samples out of the set of objects grouped in each cluster. The result of this evaluation is presented in Figure 10. As it can be seen, cluster *A* contains mainly of long and thin objects. This might be still inside the intuitive notion of similarity. The same observation can be made for the objects in cluster *C*. But the objects that are grouped together in cluster *B* are no more intuitively similar.

Figure 10: Objects in the clusters A, B, C in Figure 9(b) found by OPTICS.

Evaluating the Solid-Angle Model using the Air Dataset we made similar observations. The reachability plot computed by OPTICS (cf. Figure 9(e)) yields a clustering with a large number of hierarchical classes. But the analysis of the objects within each cluster displays that intuitively dissimilar objects are counted as similar according to the model. A further observation is the following: objects clustered in different groups are intuitively similar.

To sum up, the Solid-Angle Model does not guarantee all clusters for the Car Dataset, whereas for the Air Dataset it yields clusters with dissimilar pairs. This suggests the conclusion that the Solid-Angle Model is also rather unsuitable as a similarity model for our real-world test datasets.

Evaluation of the Eigen Value Model. In contrast to the

FIGURE 2. A page in PDF files

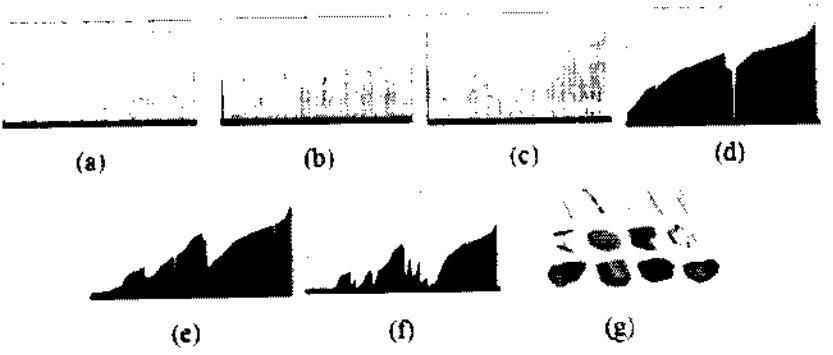

FIGURE 3. The extracted images

Images in scientific and technical literature are mostly grey and cannot be distinguished with color features. Texture features are used to analyze objects with fine texture, such as wood grain, sand and lawn, or objects composed of inerratic elements, such as pattern of cloth and brick. So textures features are not suitable for image analysis in scientific and technical literature either. Therefore, images in scientific and technical literature can be analyzed according to their shape features.

There are various kinds of images in scientific and technical literature, such as bar chart, curve chart, pie chart, flow chart, model diagram and system interface diagram. Each kind of image has its own structure; we can analyze the structure to obtain semantic representation of the image. In this paper, we choose bar chart, flow chart, curve chart and interface diagram as the semantic analysis objects.

6. **Conclusions.** This research recognizes tables, formulae and four kinds of images from the scientific and technical literature in PDF format, and uses the multi-modal information to analyze the different semantic modality features [17,18], in order to optimize the semantic representation of the scientific and technical literature.

Based in the current work, we will recognize four kinds of images, the flow chart, bar chart, curve chart and interface diagram according to the shape features, and carry out the semantic representation; recognize the tables and formulae according to the space features, analyze them in the semantic level; take advantage of text semantic analysis technology, find the semantic relation between information of different modalities, finally construct multi-modal semantic features representation system in scientific and technical literature to improve the semantic analysis and understanding of scientific and technical literature.

Acknowledgment. This work is partially supported by National Key Project of Scientific and Technical Supporting Programs No. 2011BAH10B04, National Social Science Fund No. 12BTQ006, Pre-research found of Institute of Scientific and Technical Information of China No. YY-201125. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] D. Cappon, R. Banks and C. Ramsey, Improvement of recognition on a multi-modal pattern discrimination test, *Perceptual and Motor Skills*, vol.26, no.2, pp.431-441, 1968.

[2] S. Fidell, Sensory function in multimodal signal detection, *Journal of the Acoustical Society of America*, vol.47, pp.1009-1015, 1970.

[3] A. W. M. Smeulders, M. Worring, S. Santini et al., Content-based image retrieval at the end of the early years, *IEEE Trans. in Pattern Analysis and Machine Intelligence*, vol.22, no.12, pp.1349-1380, 2000.

[4] S. Mori, C. Y. Suen and K. Yamamoto, Historical review of OCR research and development, *Proc. of the IEEE*, vol.80, no.7, pp.1029-1058, 1992.

[5] R. H. Anderson, Syntax-directed recognition of hand-printed two-dimensional mathematics, in *Interactive Systems for Experimental Applied Mathematics*, Academic Press, 1968.

[6] K. Yokoi and A. Aizawa, An approach to similarity search for mathematical expressions using MathML, *The 2nd Workshop Towards a Digital Mathematics Library*, pp.27-35, 2009.

[7] M. Adeel, H. Cheung and S. H. Khiyal, Math go! Prototype of a content based mathematical formula search engine, *Journal of Theoretical and Applied Information Technology*, vol.4, no.10, pp.1002-1012, 2008.

[8] M. Kohlhase and L. A. Sucan, A search engine for mathematical formulae, *Computer Science*, vol.4120, pp.241-253, 2006.

[9] B. Yildiz, K. Kaiser and S. Miksch, Pdf2table: A method to extract table information from PDF files, *Proc. of the 2nd Indian International Conference on Artificial Intelligence*, 2005.

[10] Y. Liu, P. Mitra and C. L. Giles, Identifying table boundaries in digital documents via sparse line detection, *Proc. of the 17th ACM Conference on Information and Knowledge Management*, 2008.

[11] H. Wasserman, K. Yukawa, B. Sy, K. Kwok and I. T. Phillips, A theoretical foundation and a method for document table structure extraction and decomposition, *The 5th IAPR International Workship on Document Analysis System*, 2002.

[12] D. W. Embley, D. Lopresti and G. Nagy, Notes on contemporary table recognition, *Proc. of the 7th International Workshop on Document Analysis Systems*, 2006.

[13] M. Suzuki, F. Tamari, R. Fukuda, S. Uchida and T. Kanahori, Infty - An integrated OCR system for mathematical documents, *Proc. of ACM Symposium on Document Engineering*, pp.95-104, 2003.

[14] J. Baker, A. Sexton and V. Sorge, A linear grammar approach to mathematical formula recognition from PDF, *Proc. of Intelligent Computer Mathematics*, 2009.

[15] J. Baker, A. P. Sexton and V. Sorge, Using fonts within PDF files to improve formula recognition, *E-Inclusion in Mathematics and Science*, 2009.

[16] J. B. Baker, A. P. Sexton and V. Sorge, Faithful mathematical formula recognition from PDF documents, *Proc. of the 9th IAPR International Workshop on Document Analysis Systems*, 2010.

[17] S. Wang, F. Xia, Y. Cao, Y. Pei and C. Cao, Mining commonsensical semantic relations from noun-noun phrases, *International Journal of Knowledge and Language Processing*, vol.3, no.1, pp.1-14, 2012.

[18] Y. Liu, Z. Sui, Q. Zhao, Y. Hu and R. Wang, On automatic construction of medical ontology concept's description architecture, *International Journal of Innovative Computing, Information and Control*, vol.8, no.5(B), pp.3601-3616, 2012.