An Approach to Mining Picture Objects Based on Textual Cues

Adeoye I. Adegorite, Otman A. Basir, Mohamed S. Kamel & Khaled B. Shaban

Pattern Analysis and Machine Intelligence Lab
Department of Electrical and Computer Engineering
University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada
{aiadegor, obasir, mkamel, kshaban} @uwaterloo.ca

Abstract. The task of extracting knowledge from text is an important research problem for information processing and document understanding. Approaches to capture the semantics of picture objects in documents constitute subjects of great interest in the domain of document mining recently. In this paper, we present an approach to extracting information about picture objects in a document using cues from the text written about them. The goal of this work is to mine a document and understand the content of picture objects in the document based on meaning inferred from the texts written about such objects. We apply some Natural Language Processing techniques to extract semantic information about picture objects in a document and process texts written about them. The mining algorithms were developed and implemented as a working system and gone through testing and experimentations. Results and future extensions of the work are discussed in this paper.

1 Introduction

The number of electronic documents that contain rich picture objects (PO) has grown enormously in all kinds of information repository, e.g. the World Wide Web (WWW). This growth can be attributed to the increasing use of scanners, digital cameras, and camera-phones in this modern era. Most of these documents contain pictures and texts. Often times, these texts have some cues regarding the contents of the pictures in the document. In the context of this paper, the definition of a PO includes images of different kinds, such as; figures, tables, diagrams, charts, pictures, and graphics. These kinds of picture objects are found in documents databases of medical images, satellite images and digital photographs [1]. Consequently, the option of manually seeking information about POs in a document is highly tedious, particularly when one is dealing with large databases. Thus, there is a need for an efficient mining system that can automatically extract semantically meaningful information about the picture objects from these large document repositories.

There has been some research work focused on either documents mining or image mining separately, in order to extract information from documents [2-5]. The problem addressed in this paper is that of being able to extract information about PO in a document without necessarily carrying out a detailed low-level pixel image mining

processes on the PO. The output of this mining system is in form of statements about the PO that are indicative of their contents.

Dixon in [6], has defined document mining as the process of finding interesting or useful patterns in a corpus of textual information. However, image mining deals mainly with the extraction of implicit knowledge, image data relationship, and/or other patterns not directly obvious in the image. Despite the development of many algorithms in these individual research fields, research in image mining is still evolving and at an experimental state [2].

The proposed approach in this paper is to mine the contents of images without performing any low-level pixel/vector based processing. Rather we take advantage of text in the document that reveals some useful information about them.

2 Related Works

There are few related research works that are concerned with the type of problems we are dealing with.

2.1 Mining Multimedia Data

Zaiane *et al.* [7-9], have implemented a prototype for mining high-level multimedia information and knowledge from large multimedia databases. For each image collected, the database would have some descriptive information composed of feature and layout descriptors. The original image is not directly stored in the database; only its feature descriptors are. The descriptive information encompasses fields such as: image file name, image URL, image and video type (i.e. gif, jpeg, bmp, avi, mpeg,), a list of all known web pages referring to the image (i.e. parent URLs), a list of keywords, and a thumbnail used by the user interface for image and video browsing. The image information extractor uses image contextual information, like HTML tags in web pages, to derive keywords. The set of all keywords collected this way, is reduced by eliminating stop-words (e.g., the, a, this) or common verbs (e.g., is, do, have, was), or aggregating words from the same canonical form (e.g., clearing, cleared, clears, clear) as presented in [10].

A drawback of this method is the fact that it is structure-dependent. It relies only on the HTML tags to locate the images in a document. For input files with plain texts without HTML tags, it will be difficult to successfully apply this method.

2.2 Summarization of Diagrams in Documents

Futrelle [11], presented a research work on summarization that attempts to mine information from many diagrams and generates a representative figure that captures all

diagrams. The research focused on diagrams, which are line drawings such as data plots or block diagrams. The overall process is to analysis and to develop structural descriptions. These descriptions are aggregated to produce an all-encompassing structure that summary diagram. One major constraint in this work is the fact that the diagram of interest must be vector-based, as contrasted with normal images, which requires detailed image processing and segmentation in order to analyze them.

Moreover, lexical resources or text accompanying figures were not exploited to guide summarization processes, rather the diagram itself was analyzed by visual parsing. The visual parsing is the only phase that has been reported to be achieved.

3 Text-Based PO Mining

In the following sub-section, we describe the steps involved in mining the contents of PO using the text written about them.

3.1 Systems Procedure & Description

The strategy of information extraction utilized in this project focuses on the PO in a document. Our mining algorithm can be summarized into the following steps:

Step 1: Identify and locate the captions/labels or image tags for each of the PO,taking into consideration the structure of the document. For example, image file name, ALT field in the IMG tag for HTML files can be used to pick the label for any PO.
Step 2: Use the labels or tags obtained in step 1 to derive keywords to search through the document to identify where they appear again in the whole document.
Step 3: Capture the relevant sentences in which the captions/labels already located in step 2 are utilized to further describe or explain the PO.
Step 4: Combine all the sentences captured in step 3 for each PO.
Step 5: Output the statements for each PO.

To better illustrate how these implementation steps were achieved in this project, figure 1 shows a further breakdown of specific processes carried out on each input file. Each of these modules are further explained below:

Input File – The input file to our mining system is the text. The actual PO are not included in the input file, only their captions/labels/image tags and all the text in the document are included.

Pre-processing: The pre-processing tasks done here is tokenization and Part-Of-Speech tagging. The tokenization converts the input file to tokens or units of alphabetic, numeric, punctuation, and white-space charaters. Part-Of-Speech(POS) tagging is done according to the grammatical context of the words in the sentences.

Syntactic Analysis: The purpose of syntactic analysis is to determine the structure of the input text. Here, a tree structure is formed with one node for each phrase. This structure consists of a hierarchy of *phrases*, the smallest of which are the *basic symbols* and the largest of which is the *sentence*.

Semantic Analysis: This is the method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the aggregation of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other.

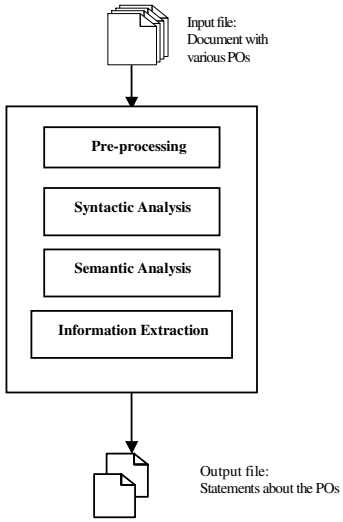


Fig. 1. Text-Based PO Mining

Information Extraction: Our contribution to this research area is in the strategy utilized at the stage of information extraction. Mining contents of PO is carried out by using captions, labels, image file name and ALT tags to search for relevant information about the PO. Sub-section 3.2 gives summarized algorithmic steps of the implementation.

Output File*:* The output of our system is in form of statements about each PO.

3.2 Mining Algorithmic Steps

Let X = $\{x_1, x_2,\ldots x_n\}$ denote a set of PO in a document D = $\{s_1,s_{2\ldots}s_m\}$, where $s_{1\ldots}s_m$ are sentences in the document.
Let Y = $\{y_1, y_2,\ldots y_n\}$ denote the caption of the PO and is a subset of D, where $y_i = (l_i, f_i, c_i)$ and l_i = ALT<label>, f_i = image file name and c_i = label or title.

Let Z = $\{z_{y1}, z_{y2}...z_{yn}\}$ denote set of relevant sentences that contain captions (y_1, $y_2...y_n$) of each PO.

Algorithm 1: Extraction of Information from text about PO

```
1:  Input D ← New Document
2:    Extract PO tags, captions to X
3:        for each x_i ∈ X in D do
4:          y_i ← extract the caption
5:           Add y_i to Y (List of image-file name and captions for PO)
6:        Z ←Empty List (List of relevant
7:                              sentences about each PO)
8:           for each y_i in Y do
8:                    search for sentences s_i ∈ D that
                      contain y_i
9:         If y_i is a word or phrase in s_i ∈ D , then
10:          Add s_i to z_yi ∈ Z
11:       else
12:          Discard s_i (irrelevant sentence to PO)
13:       end if
14:     end for
15:   end for
16:   Output the content of Y and Z
```

3.3 Systems Architecture

The multi-pass architecture used to implement the proposed algorithm in Visual Text[1] is illustrated in Figure 2.

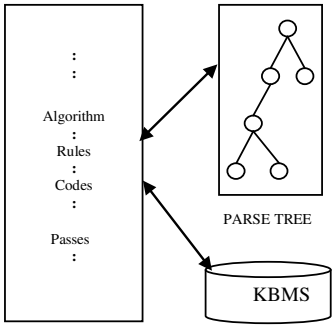


Fig. 2. Text Analyzer Multi-Pass Architecture

Visual Text is an Integrated Development Environment (IDE) that was developed with NLP++ programming language. The passes of the multi-pass architecture are

[1] *Visual Text is a trademark of Text Analysis International*

constrained to share a single parse tree. Each pass receives the cumulative parse tree, elaborates it further, and then hands it to a subsequent pass. In addition to managing a unique parse tree, the passes may also update and access a knowledge base, as well as general program data structures. The Visual-Text IDE [14] uses a hierarchical knowledge base management system (KBMS), for mapping knowledge in a more natural fashion than a relational database. The structure of the programming language is a key component of the architecture; that enables the NLP to manage the passes, parse tree and associated knowledge base. By splitting the NLP system into multiple passes, each pass can be constrained to operate on particular contexts. Passes within the architecture can also dynamically create and execute new passes in the processes of tuning the system to get an optimal result.

3.4 Utilized NLP Techniques

In the following sub-sections, we discuss the details of the NLP techniques used in the implementation of this work.

3.4.1 Syntactic Analysis - Parsing

Parsing is the process of linking the part-of-speech tags into a tree structure that indicates the grammatical structure of the sentence. The interior nodes representing phrases, links represent the application of grammatical rules and leaf nodes represent words [13]. Two major types of parsing that are relevant to our system are discussed below:

3.4.1.1 Parsing PCFG

A Probabilistic Context-Free Grammar (PCFG) is a context-free grammar which has a probability associated with each rule normalized so that the probabilities of the rules associated with a particular non-terminal sum to 1. Disambiguation is achieved by selecting the parse tree with the highest probability. The probability of a parse tree, π for a sentence S is given by

$$P(S,\pi) = \prod_{n \in \pi} (P(r(n))$$

where *p(r(n))* is probability of the rule *r,* that has been applied to expand non-terminal *n* in the parse tree.

3.4.1.2 Lexicalized Parsing

Lexicalized parsers collect two sorts of statistics. Firstly, the probability *P(r|h),* of which rule *r*, should be applied given the head *h,* of the phrase *c* to be expanded. Secondly, the probability that a sub-phrase *q,* has head *h*, given the head of the phrase being expanded *m* (for "mother"). The total probability for a parse π, of a sentence S is then

$$\rho(S,\pi) = \prod_{c \in \pi} (P(h(c) \mid m(c))P(r(c) \mid h(c))$$

One way of thinking about lexicalized parsers is to imagine them as CFGs (Context-Free Grammar) with a profusion of rules, one for each word in the vocabulary.

3.5 Level of Outputs

We built a multi-pass text analyzer. The analyzer is tuned in terms of number of passes in order to achieve a better result. We identified and established four different levels of Output. Figure 3 depicts as: Text level, Caption Level, Description level and the Semantic level. The details of these levels are explained as follows:

- Text Level:
  This is the state of the input text file at the early part of entry into the Text analyzer. Various passes that have worked on the input file has processed the files into paragraphs and paragraphs into sentences.
- Caption Level:
  Here, the system has identified and captured the captions and/or labels of the PO.
- Description level:
  At this stage the mining system has passed through the input file many times, identified the locations of where the figures were referred to in the document and then captured the sentences written about the MMO.
- Semantic level:
  This is the highest and final level of output. The semantic passes utilize the parse tree and data schemas within the knowledge base. It builds concepts in the knowledge base for sentences, events, and objects in the text that it is processing.

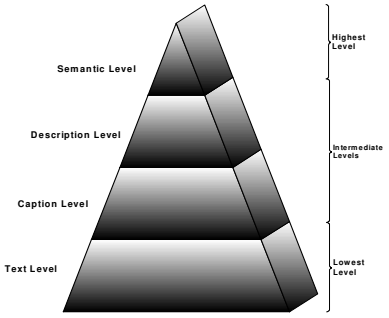


Fig. 3. Levels of Output

The four information levels can be further generalized to three layers: the text level is the lowest level, while the caption and description level forms the intermediate level and the semantic level is the highest level of output. Figure 4.6 shows these levels.

4 Experimental Results

As indicated in the earlier sub-sections, we built a multi-pass text analyzer. The analyzer is tuned in terms of number of passes in order to achieve an optimal result. After building the text analyzer, many samples of documents were presented to the analyzer. All these documents can be generally categorized into two types. There are some documents with PO that has captions with them, while some documents do not have captions clearly written with the PO. In the following sub-sections, we present representatives of these two categories and also discuss their results.

4.1 Sample Type 1 – Documents containing PO with Captions

This is a general case of documents with PO that has captions or labels directly written with the PO. In this case, the text analyzer can easily pick the labels and use this to search through the document to extract sentences that contains these labels. One sample of such documents is shown in figure 4. This document contains many PO, but we have only shown page one, which has four POs. The POs that have captions are written below their corresponding objects. For instance, *"Fig 3: Staff of LWF"* is shown under the particular PO concerned.

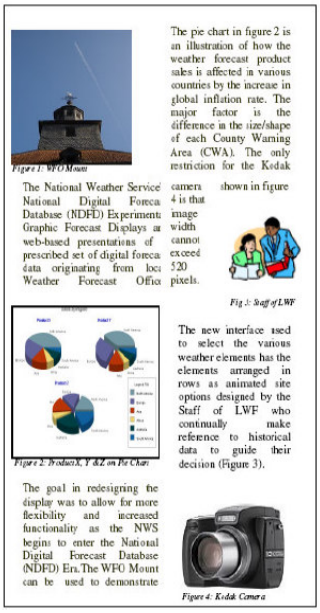
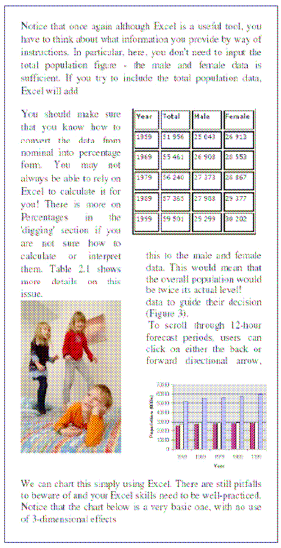


Fig. 4. Sample Document with labeled POs **Fig. 5.** Levels of Output

The result obtained from this example is as shown in figure 5 with the respective levels of output.

4.2 Sample Type 2 – Documents containing PO without Captions

This is a general case of documents in which the captions or labels are not found directly written with the PO. In this case, the text analyzer relies on the ALT<label> tab or image file name to pick the labels and use this to search through the document to extract sentences that contains this labels. A sample document and output is as shown in figure 6 and Figure 7 respectively.



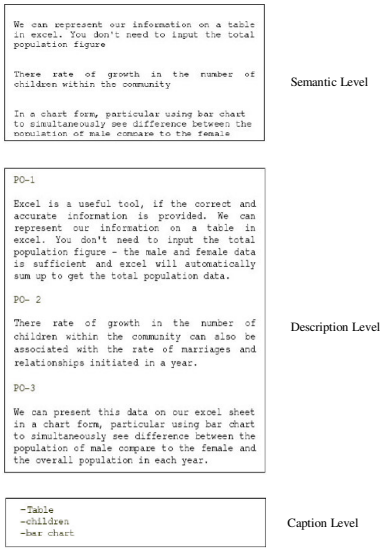


Fig. 6. Sample Document without Labeled POs

Fig. 7. Level of Outputs

5 Conclusion and Future Works

This paper presents a research focused on text-based image mining. It is text-based in the sense that, information is extracted from text in the document regardless of the position of the actual wordings in the whole structure. The main advantage of this system is eliminating the detailed image processing that is often suggested by other techniques.

We also established four-level hierarchical structure of output for describing the PO. The major contribution of this work is the mining algorithm formulated and implemented to produce a text analyzer that can take in an input file in form of text and output statements about the contents of the PO in the document concerned. Comparing our work with some of the related work in the literature [7, 8, 9], the approach we presented is not structure dependent. Another advantage comes in the fact that our approach does not depend on the HTML tags to identify the location of the images. Also, unlike the work done in [11], where diagrams alone are the major focus, our definition of PO is not limited to diagrams, but includes images such as tables, fig-

ures, pictures and graphics. Some application areas of this research approach include: Digital library, manuals and indexing. The algorithms developed can also be applied to learning object mining, knowledge representation and knowledge sharing.

The definition of our PO, which encompasses representations in documents such as diagrams, tables, pictures, charts and graphics, implies a constraint boundary, which invariably has limited us to considering only images that falls under these categories. An extension of this research for future work would be to investigate ways of adapting our strategy to capture other types of images in a document, such as video clips, audio clips, flash animated objects, dynamic web-contents and other multimedia objects that may be in any document.

Reference:

1. H. WYNNE, L. L. MONG AND J. ZHANG. *"Image Mining: Trends and Developments"*. In Journal of Intelligent Information System (JISS): Special Issue on Multimedia Data Mining, pp 97-106 Kluwer Academic, 2002.
2. A. POPESCUL, L.H. UNGAR, S. LAWRENCE AND D. M. PENNOCK. *"Statistical relational learning for document mining"*. Third IEEE International Conference on Data Mining, ICDM 2003, pp275 – 282, November 19-22, 2003.
3. S. QIN-BAO, L. NAI-QIAN, S. JUN-YI AND C. LI-MING. *"Web documents mining"*. In Proceedings of 2002 International Conference on Machine Learning and Cybernetics, Volume: 2, pp791 – 795 November 4-5, 2002.
4. U. FAYYAD, G. PIATETSKY-SHAPIRO, AND P. SMYTH. *"The KDD process for extracting useful knowledge from volumes of data"*. Communications of the ACM, *39*(11): pp27-34, November 1996.
5. H. AHONEN, O. HEINONEN, M. KLEMETTINEN AND A. I. VERKAMO. *"Applying Data mining techniques in text analysis"* Report C-1997-23, University of Helsinki, Department of Computer Science, March 1997.
6. M. DIXON: *"An Overview of Document Mining Technology"* A research report Computer Based Learning Unit, University of Leeds, October 1997.
7. S. J. SIMOFF, C. DJERABA AND O. R. ZAIANE. *"MDM/KDD2002: multimedia data mining between promises and problems"* ACM SIGKDD Explorations Newsletter Volume 4, Issue 2. pp 118 – 121, December 2002.
8. O. R. ZAIANE AND S. J. SIMOFF. *"MDM/KDD: multimedia data mining for the second time"* ACM SIGKDD Explorations Newsletter Volume 3 , Issue 2 COLUMN: Reports from KDD-2001 pp65 – 67, January 2002
9. O. R. ZAIANE, J. HAN, Z. LI AND J. HOU "*Mining Multimedia Data*". CASCON'98: Meeting of Minds, pp 83-96, Toronto, Canada, November 1998.
10. O. R. ZAIANE, A. FALL, R. V. DAHL, AND P. TARAU. "*On-line resource discovery using natural language.*" In Proceedings, RIAO'97 Montreal, Canada, pp65-73, June 25-27, 1997.
11. R. P. FUTRELLE, *"Summarization of Diagrams in Documents"* In I. Mani & M. Maybury (Eds.), Advances in Automated Text Summarization. Cambridge, MA pp 61-65. March 1999.
12. J. ZHANG, H. WYNNE AND L. L. MONG. *"Image Mining: Issues, Frameworks and Techniques"* In Second International Workshop on Multimedia Data Mining-MDM/KDD San Francisco, U.S.A., pp 34-42 August 2001,
13. T. K. LANDAUER, P. W. FOLTZ, AND D. LAHAM. "*An Introduction to Latent Semantic Analysis"*. Discourse Processes, (25): pp 259-284, October 1998.
14. A. MEYERS AND D. HILSTER. "*Description of the TexUS System as used for MUC-4"*. Proceedings of MUC-4. DARPA. pp 207-214. March 1992.