Text, Image and Vector Graphics Based Appraisal of Contemporary Documents

Sang-Chul Lee[2], William McFadden[1] and Peter Bajcsy[1]
[1]*National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign*
[2]*Department of Computer and Information Engineering, Inha University, Incheon, Korea*
sclee@inha.ac.kr, wmcfadd@ncsa.uiuc.edu, pbajcsy@ncsa.uiuc.edu

Abstract

We have designed a framework for content based appraisal of documents. Our motivation is to provide computer assisted support for answering several appraisal criteria according to the general appraisal guidelines in the National Archives and Record Administration (NARA) 1441 directive. The appraisal criteria led us to investigations related to (a) finding groups of PDF documents with similar content, (b) ranking documents according to their creation/ modification time and digital volume, and (c) detecting inconsistency between ranking and content within a group of related documents. The novelty of our work is in designing a methodology and a mathematical framework for document appraisals, and prototyping the framework working with text, image and vector graphics components of PDF documents. We present example results of grouping, ranking and integrity verification for groups of scientific documents about medical topics.

1. Introduction

The objective of our work is to design a methodology, algorithms and a framework for document appraisal by (a) enabling exploratory document analyses and integrity/authenticity verification, (b) supporting automation of some analyses and (c) evaluating computational and storage requirements for archival purposes. In order to address the aforementioned criteria, our approach has been to decompose the series of appraisal criteria[1] into a set of focused analyses, such as (a) find groups of records with similar content, (b) rank records according to their creation/last modification time and digital volume, (c) detect inconsistency between ranking and content within a group of records, and (d) compare sampling strategies for preservation of records.

In this work, we had chosen a specific class of electronic documents that (a) correspond to information content found in scientific publications about medical topics, (b) have an incremental nature of their content in time, and (c) contain the types of information representation that are prevalent in contemporary medical environments. Specifically, we narrowed our focus to those electronic documents that contain primarily text, raster and vector graphics as found in typical medical records in office document file formats. Among the file formats, MS Word can be considered as the most widely used file format for creating documents, while Adobe Portable Document Format (PDF) and Ghostscript could be described as the most widely used for exchanging documents. We selected to work with PDF documents since PDF is an open file format, and the open nature of the file format is critical for automated electronic document appraisal and long term preservation.

Our work is related to the past work of authors in the area of digital libraries [1], content-based image retrieval [2] and appraisal studies [3]. For example, the authors of [4] analyze PDF document by examining the appearance and geometric position of text and image blocks distributed over an entire document. However, they did not use the actual image and vector graphics information in their analyses. Similarly, the focus in [5] is only on the logical structure of PDF documents but not the content. The work in [6] and [7] is based on analyses of vector graphics objects only since it is focused on diagrams represented by a set of statistics, e.g., the number of horizontal lines and vertical lines. Other authors also focused only on chart images using a model-based approach [8]. There is currently no method that would provide a comprehensive content-based PDF comparison and appraisal strategy according to our knowledge.

In order to address the appraisal criteria, we adopted some of the text comparison metrics used in [1], image comparison metrics used in [2] and lessons learnt stated in

[1] http://www.archives.gov/oig/reports/september-2005.html#challenges
POC: Peter Bajcsy, pbajcsy@ncsa.uiuc.edu, 217-265-5387.

[3]. Then, we designed a new methodology for grouping electronic documents based on their content similarity (text, image and vector graphics), and prototyped a solution supporting grouping, ranking and integrity verification of any PDF files and HTML files. First, text based, vector based and multi-image based comparisons are performed separately. Multiple images in each document are grouped first and then groups of images across documents are compared to arrive to an image-based similarity score. The current prototype is based on color histogram comparison, line count in vector graphics and word frequency comparison. The image colors and word/ integers/ floating numbers can be analyzed visually to support exploratory analyses. Subsets of the undesirable text and image primitives could be filtered out from document comparisons (e.g., omitting conjunctions, or background colors). The results of text, image and vector based comparisons are fused to create a pair-wise document similarity score. The matrix of pair-wise document similarity scores are used for grouping. The other appraisal criteria are approached by ranking documents within a group of documents based either on time stamps or on file name indicating the version number. The inconsistency between ranking and content within a group of records is based on frequency tracking, where the frequency of text, image and vector primitives is monitored over the time/version dimension of the grouped documents.

Currently, we hypothesized that the correct temporal ranking correlates with the content (images, vector and text) in such a way that the content is being modified without sharp discontinuities. Sharp content discontinuities are perceived as significant changes of document descriptors that would correspond, for instance, to large text/image deletions followed by large text/image additions or large text/image additions followed by large text/image deletions. We have experimented with real PDF documents of journal papers about medical topics to validate the above hypothesis.

The novelty of our work is in designing a methodology for computer-assisted appraisal of a class of electronic medical records labeled as scientific papers, and in developing a mathematical framework for automation of appraisals based on image, vector graphics and text types of information representation. Furthermore, our contribution is in prototyping a computer assisted appraisal system and demonstrating its performance on sets of PDF documents.

2. Methods

This section presents the methodology and theoretical framework for addressing grouping, ranking and integrity verification problems.

2.1. Methodology

The designed methodology consists of the following main steps: (1) Extract components and properties stored in PDF files/containers. (2) Define text, image and vector graphics primitives, and extract their characteristic features. (3) Group images within each document into clusters based on a pair-wise similarity of image primitives and a clustering similarity threshold. (4) Compute a pair-wise similarity of image clusters across two documents based on their corresponding features. (5) Compute a pair-wise similarity of text & vector graphics primitives across two documents. (6) Calculate fusion coefficients per document to weight the contribution of text-based, image-based and vector-based similarities to the final pair-wise document similarity score. (7) Repeat steps (4-6) for all pairs of documents. (8) Group documents into clusters based on the pair-wise document similarity score and a selected similarity threshold. (9) Assign ranks to all documents based on their time stamps and storage file size. (10) Calculate the second difference of the document characteristic features over time and file size dimensions. Report those documents for which the second difference exceeds a given threshold defining allowed discontinuities in content.

2.2. Theoretical Framework

Clustering statistical features. After extracting and defining the components and properties from a PDF document in step (1) and (2) described in section 2.1, we group the components in the PDF document based on the following similarity analysis. Given a set of document $\{D_i\};\ i = 1, 2, \cdots, N$ compute pair-wise similarity of documents $sim(D_i, D_j)$ and aggregate them into clusters based on the similarity values for further ranking within each cluster.

The similarity of documents is understood as the combined similarity of document components. In our case, it would be the similarity of text, vector and raster (image) graphics components. The three components are decomposed into multiple images I_{ik} and their image primitives e_m^{IMAGE}, vector graphics and their image primitives e_m^{VECTOR}, and text primitives e_m^{TEXT} in textual portions $T_{ik} = T_i$ of a document D_i. The similarity for each component type is derived either directly using the features of its primitives (the case of text) or average features of multiple components of the same type and their primitives (the case of images and vector graphics).

The text feature for the word primitives is the frequency of occurrence of each unique word. The image feature for the color primitive is the frequency of occurrence of each unique color (also denoted a one-dimensional color histogram). The vector graphics feature is the frequency of occurrence of lines forming each vector graphics. The frequency of occurrence provides a statistical estimate of the probability distribution of primitives.

Calculation of Document Similarity. Based on the features computed for each category (text, raster, vector), we calculate the document similarity per category (steps 4-6 in section 2.1). These similarities are then fused in the following integration framework. Given two PDF documents D_i, D_j the similarity is defined as a linear combination of the similarities of the document components. In our case, the formula contains only the text and raster graphics components.

$$\begin{aligned} sim(D_i, D_j) &= w_{TEXT} \cdot sim(T_i, T_j) + \\ & w_{RASTER} \cdot sim(\{I_{ik}\}_{k=1}^{K}, \{I_{jl}\}_{l=1}^{L}) + \\ & w_{VECTOR} \cdot sim(V_i, V_j) \end{aligned} \qquad (1)$$

where the $w_{TEXT}, w_{RASTER}, w_{VECTOR}$ are the weighting coefficients.

We have derived the weighting coefficients from the spatial coverage ratio of images, vector graphics and text in two compared documents. The weight assignment could be viewed as the relevance (the weight) of each PDF component according to the amount of space it occupies in a document. The motivation is based on a typical construction of documents where the space devoted to a textual description or an illustration reflects its importance and hence should be considered in the similarity calculation. Thus, the weighting coefficients are calculated as

$$W_{IMAGE}(D_i, D_j) = \frac{R_{IMAGE}(D_i) + R_{IMAGE}(D_j)}{2},$$

$$W_{IMAGE}(D_i, D_j) + W_{VECTOR}(D_i, D_j) + W_{TEXT}(D_i, D_j) = 1$$

where $R_{IMAGE}(D) = \frac{Area_{IMAGE}(D)}{Area_{IMAGE}(D) + Area_{VECTOR}(D) + Area_{TEXT}(D)}$,

$$R_{IMAGE}(D) + R_{VECTOR}(D) + R_{TEXT}(D) = 1$$

Calculation of Text Similarity: The similarity of text components from two documents $sim(T_i, T_j)$ is computed using the text features and similarity metric defined according to [1]. The equation is provided below.

$$sim(T_i, T_j) = \sum_{k1,k2} \omega_{i,k1} \omega_{j,k2} \qquad (2)$$

where $k1, k2$ are those indices of text primitives that occur in both documents (in other words, there exist $e_{i,k1} = e_{j,k2}; \quad e_{i,k1} \in T_i, e_{j,k2} \in T_j$) . The ω terms are the weights of text primitives computed according to the equation below.

$$\omega_{ik} = \frac{f_{ik} \log(N / n_k)}{\sqrt{\sum_{l=1}^{L} (f_{il})^2 (\log(N / n_l))^2}} \qquad (3)$$

where f_{ik} is the frequency of occurrence of a word e_k in D_i , N is the number of documents being evaluated, *L* is the number of all unique text primitives (words) in both documents, and n_k is the number of documents in that contain the word e_k $(n_k = 1 \; or \; 2)$.

Calculation of Raster Graphics (Image) Similarity: In contrary to text that is viewed as one whole component, there are multiple instances of raster graphics components (images) in one document. Thus, the similarity of image components in two documents is really a similarity of two sets of images.

Due to the fact that many documents contain images that are sub-areas or slightly enhanced versions of other images in the same document, we have observed biases in image-based document similarity if all possible image pairs from two documents are evaluated individually and then the average similarity would be computed. The bias is introduced due to large similarity values of one master image in one document with multiple derived images in another document, although many other images would not have any match.

In order to avoid such biases, we approached the similarity calculation by first computing a pair-wise similarity of all images within each document and clustering them. Next, the pair-wise similarity of clusters of images from each document is computed using the average features of clusters.

A. Intra-document image similarity: The similarity of two raster graphics (image) components from one document $sim(I_{ik} \in D_i, I_{il} \in D_i)$ is computed using the one-dimensional color histogram feature and the same similarity metric as defined before for text according to [1]. The equation is provided below.

$$sim(I_{ik} \in D_i, I_{il} \in D_j) = \sum_{k1,k2} \omega_{i,k1} \omega_{i,k2} \qquad (4)$$

where $k1, k2$ are those colors that occur in both images (in other words, there exist $e_{i,k1} = e_{j,k2}; \quad e_{i,k1} \in I_{ik}, e_{i,k2} \in I_{il}$. The ω terms are the weights computed the same way as before.

B. Inter-document image similarity: The similarity of two sets of raster graphics (image) components, one from each document, $sim(I_{ik} \in D_i, I_{jl} \in D_j)$ is computed using the average one dimensional color histogram feature of all images in a set and the same similarity metric as defined before for text according to [1]. The equation is provided below.

$$sim(\{I_{ik}\} \in D_i, \{I_{jl}\} \in D_j) = \sum_{k1,k2} \omega_{i,k1}\omega_{j,k2} \qquad (5)$$

Calculation of Vector Graphics Similarity: The calculation is computed similarly for vector graphics elements as it was for text. The only difference is in changing the word frequency to line count frequency. The justification for this approach follows from the manner in which vector graphics are generated by the PDF document. The structure of each page of a PDF document is comprised of a series of objects placed at various points on the surface of the page. When the object is a string of text or a raster image, the information of the object is logged within the PDF document. However, since any vector graphics object is really a set of primitive curves, when a PDF document is created any composite vector graphic image is rendered in the PDF document as individual paths. Although the graphic may be interpreted by the reader as a single unit, there will be no way of determining this from the document itself. The method therefore relies on the fact than simple arcs often share endpoints in the construction of splines and polygons. These connected paths are then defined by the number of lines or curves composing them, and the comparison is done based upon the occurrence of paths of matching degree. This way of characterizing vector graphics has advantages in the statistical comparison of documents because documents containing significant volumes of vector graphics will likely contain many unique graphics objects, which will make comparison between two versions of one document difficult were one to attempt to reconstruct the graphic precisely. For example, one finds that similar documents will contain graphs composed of the same number of gridlines although the actual graph data may change from one document to the next.

Preliminary Version Order Ranking. For the initial ordering of the document groups, we first group documents into document clusters based on the pair-wise document similarity threshold (step 8 in section 2.1), and then assign ranks to all documents based on their time stamps and storage file size (step 9 in section 2.1). Assuming the internal timestamp of the PDF document is considered accurate, version order integrity verification takes place to ensure tampering with the timestamp did not occur.

Modification Integrity Check. After the documents are ordered by time stamp, finally certain criteria are checked to ensure modification probably occurred in the order indicated by the current ordering. The criteria currently include (1) appearance or disappearance of document images, (2) appearance and disappearance of dates appearing in documents, (3) file size, (4) image count, (5) number of sentences, and (6) average value of dates found in document.

3. Experimental Results

Using the presented framework, we appraised sets of PDF documents generated during scientific medical journal preparations. The document sample set consisted of 10 documents 4 of which were modified versions of one article and 6 were of a different though related article. The pair-wise comparisons of the features are presented in the following graphs.

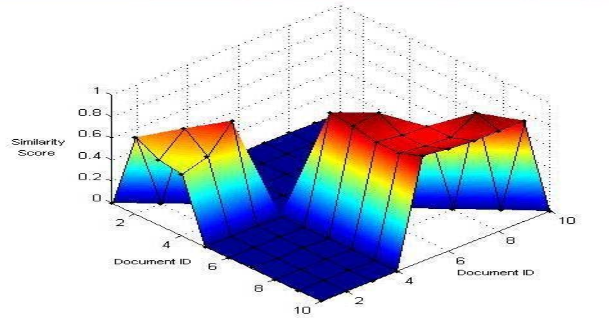


Figure 1. Word Similarity Comparison

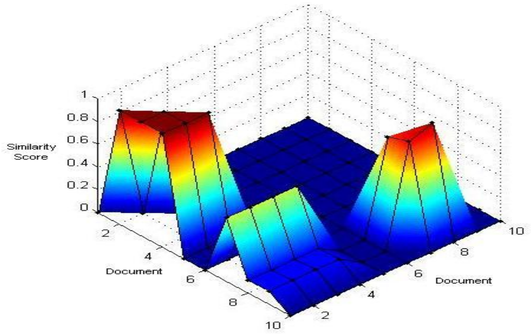


Figure 2. Vector Graphics Similarity Comparison

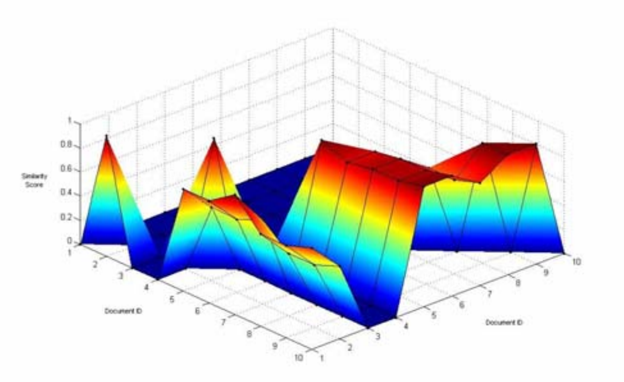


Figure 3. Raster Image Similarity Comparison

The z values of the graphs in Figures 1-3 represent the similarities between the documents identified with the numbers by the x and y axes. The word similarity comparison clearly shows that the documents 5 through 10 score highly in comparison with each other while they score poorly in comparison to the documents 1 through 4. Likewise, documents 1 through 4 show higher scores for comparison within the group. The vector graphics of documents 1 through 4 are nearly identical while in the second subgroup only documents 7 through 9 are conclusively linked. The raster images within the two subgroups of the documents shows high similarities for the documents 5 through 10 score but the rest are not obvious how they are related. The combination of vector graphics comparison along with word comparison results in a clear consensus about which documents belong together as shown in Figure 4.

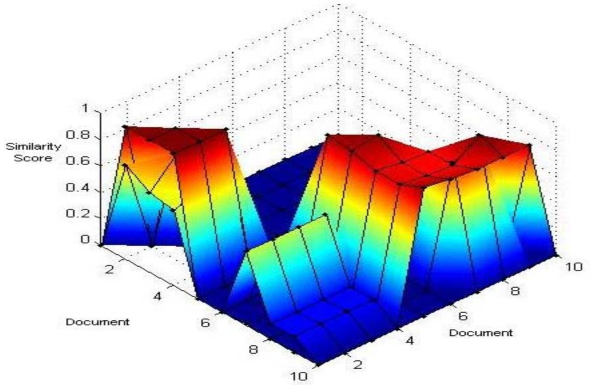


Figure 4. Vector Graphics Similarity and Word Similarity Combined

Throughout the documents the relative apportionment of visible space of the three document features varies. Figure 5 shows the fraction of each document covered by words, images and vector graphics.

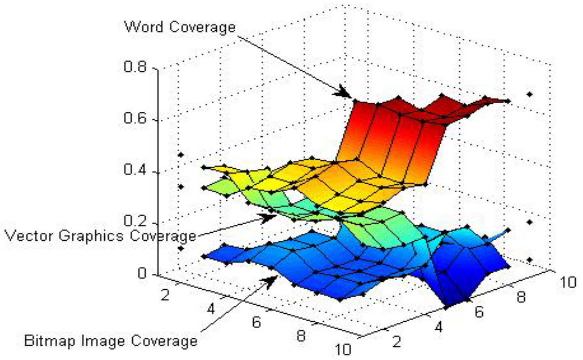


Figure 5. Portion of Document Surface Allotted to Each Document Feature

Combining the three comparison techniques with weights allotted by the proportion of coverage of the feature represented by that comparison allows for a final similarity score to be established. Figure 6 displays the final comparison matrix, which clearly distinguishes the two subgroups of the original document set.

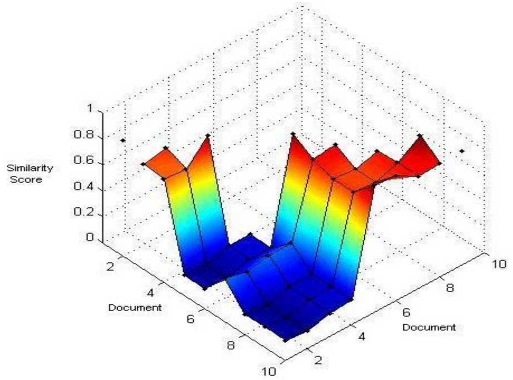
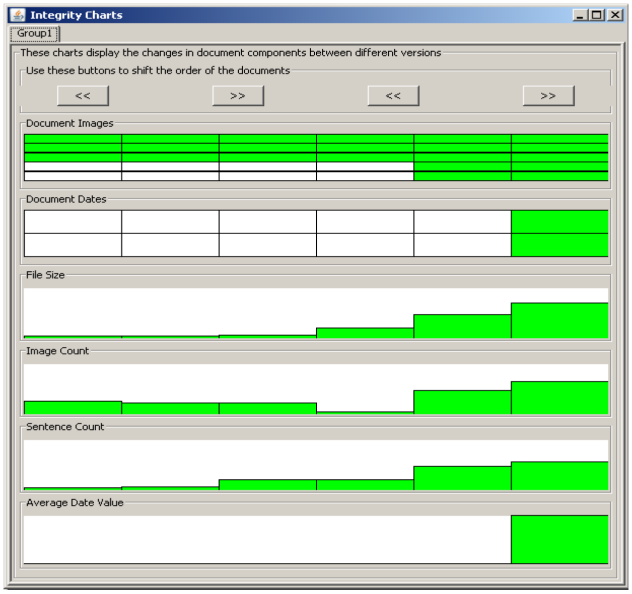


Figure 6. Comparison Using Combination of Document Features in Proportion to Coverage

With the documents adequately grouped and ordered by PDF timestamp, the verification process looks for conspicuous editing habits from one document to the next. Figures 7 and 8 show a visualization of passed and failed tests, where documents are aligned horizontally from earliest (left) to latest (right). The integrity tests are aligned vertically from top to bottom to refer to: (1) appearance or disappearance of document images, (2) appearance and disappearance of dates appearing in documents, (3) file size, (4) image count, (5) number of sentence, and (6) average value of dates found in a document. Green/red color indicates pass/failure of the integrity test.

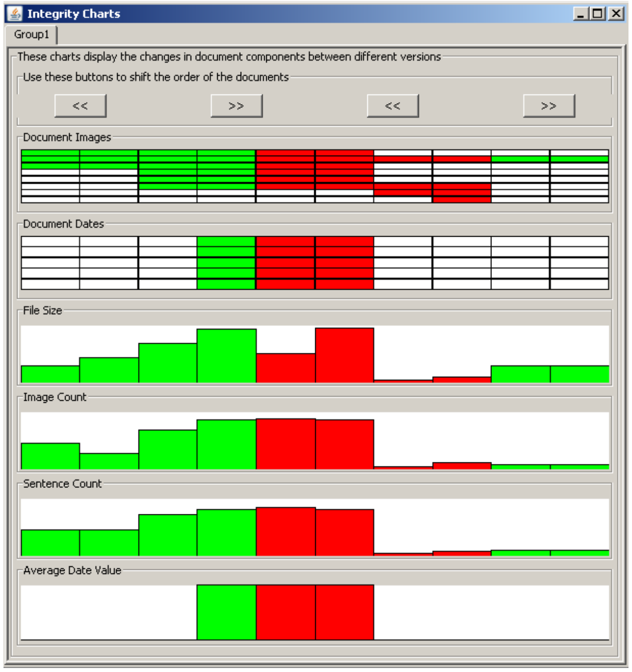

Integrity Charts

Group1

These charts display the changes in document components between different versions

Use these buttons to shift the order of the documents

Figure 7. Verification of the document ordering based upon the time stamps of PDF documents.



These charts display the changes in document components between different versions

Use these buttons to shift the order of the documents

Figure 8. Failed verification of the document ordering based upon the time stamps of PDF documents. Green bars indicate reasonable changes to documents while red bars indicate suspicious document editing behavior such as drastic deletions.

4. Summary

We have designed and prototyped a framework for addressing the appraisal criteria. The framework consists of a comprehensive content-based grouping of documents, ranking based on temporal or file size attributes and verification of document integrity. Although we selected to work with documents in PDF format, the framework is applicable to any file format as long as the information can be loaded from any proprietary file format. In future, we will be exploring other hypotheses to increase the likelihood of detecting inconsistencies and understanding the high-performance computing requirements on computer-assisted appraisal of electronic records.

5. Acknowledgement

This research was partially supported by a National Archive and Records Administration (NARA) supplement to NSF PACI cooperative agreement CA #SCI-9619019. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the National Science Foundation, the National Archive and Records Administration, or the U.S. government.

6. References

[1] G. Salton, J. Allan, and C. Buckley, "Automatic structuring and retrieval of large text files," *in Communication of the ACM*, vol. 37, pp. 97-108, 1994.

[2] D. M. Squire, W. Muller, H. Muller, and T. Pun, "Content-Based query of image databases: inspirations from text retrieval," *Pattern Recognition Letters*, vol. 21, pp. 1193-1198, 2000.

[3] J. A. Marshall, "Accounting For Disposition: A Comparative Case Study of Appraisal Documentation at the NARA in the US, Library and Archives Canada, and the NAA," in *Dep. of Library and Information Science*. vol. PhD: Univ. of Pittsburg, 2006.

[4] W. S. LOVEGROVE and D. F. BRAILSFORD, " Document analysis of PDF files: methods, results and implications," *ELECTRONIC PUBLISHING*, vol. 8, pp. 207-220, JUNE & SEPTEMBER 1995 1995.

[5] A. Anjewierden, "AIDAS: incremental logical structure discovery in PDF documents," in *International Conference on Document Analysis and Recognition*, 2001.

[6] R. P. Futrelle, M. Shao, C. Cieslik, and A. E. Grimes, "Extraction,layout analysis and classification of diagrams in PDF documents," in *International Conference on Document Analysis and Recognition*, 2003, p. 1007.

[7] M. Shao and R. P. Futrelle, "Recognition and Classification of Figures in PDF Documents," in *Lecture Notes in Computer Science*. vol. 3926: Springer Berlin / Heidelberg 2006.

[8] W. Huang, C. L. Tan, and W. K. Leow, "Model-based chart image recognition," in *Fifth IAPR International Workshop on Graphics Recognition Computer Vision Center* Barcelona, Catalonia, Spain, 2003, pp. 87-99.