Finding Captions in PDF-Documents for Semantic Annotations of Images

Gerd Maderlechner, Jiri Panyr, and Peter Suda

Corporate Technology
Siemens AG,
D-81730 München, Germany

Abstract. The Portable Document Format (PDF) is widely-used in the Web and searchable by search engines, but only for the text content. The goal of this work is the extraction and annotation of images in PDF-documents, to make them searchable and to perform semantic image annotation. The first step is the extraction and conversion of the images into a standard format like jpeg, and the recognition of corresponding image captions using the layout structure and geometric relationships. The second step uses linguistic-semantic analysis of the image caption text in the context of the document domain. The result on a PDF-document collection with about 3300 pages with 6500 images has a precision of 95.5% and a recall of 88.8% for the correct image captions.

1 Introduction

This work is motivated by the following five facts: (1) in the world wide web nearly all searchable documents are in HTML-format. The next important format is the PDF-format (portable document format), with a proportion of about 3% of the searchable web (experimental result with Google and Yahoo). Since most PDF documents are larger than HTML-pages we estimate that about 10% of the searchable information is in PDF format [17]. The remaining document formats are below 1%. (2) PDF is a standard format for archiving all types of documents in libraries, government or companies. PDF has an open published specification. PDF/A is an ISO standard for archiving. (3) PDF is popular for electronic publishing because it is a page description format which preserves even complex layouts consisting of text, graphics and images on all output devices. (4). The existing image search engines like Google, Yahoo, Picsearch etc. do not consider images in PDF-documents. (5) Present text based image search engines use keywords and not semantic annotations of the images.

From this we conclude that it is worthwhile to consider the PDF format for image search. Furthermore the image search quality can be improved by semantic annotations of the image captions. This will allow image searching not only by keywords but using questions like "Show me the player who scored the goal 1:0 in the match Mexico-Costa Rica on August 17th, 2005". The semantic annotation can be applied also to image captions in HTML-pages. An example will be presented below (Fig. 5).

Semantic and index information for image understanding and searching may be obtained from the image content using image processing methods [1], or from some text describing the image [2]. Some approaches use a combination of either information [3].

D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109, pp. 422 – 430, 2006.

This paper describes a new approach that does not use the image content analysis but relies on the recognition of existing image captions using layout analysis. We concentrate on the image caption recognition and do not go into details of the linguistic-semantic methods.

2 Related Work

2.1 PDF-Document Analysis

There exist many commercial and public domain programs for processing of PDF-documents [4, 5]. But the result of our investigation for the purpose of image capture recognition was disappointing. We found some papers [6, 7] on PDF-document analysis using the open source library xpdf and the programs pdftotext, pdfimages and pdf2html. These tools are helpful but we had to add considerable functions to the existing programs which is described in chapter 3.

Lovegrove et.al.[8] analyze PDF files using Adobe SDK with the goal to perform logic labeling of the layout objects, which also contains image captions with only few examples and no quantitative evaluation.

Chao and Lin [9] developed a proprietary system for PDF-layout analysis with a different purpose.

2.2 Image Caption Recognition

For *HTML* web pages there are many research and commercial systems available which use also image captions, e.g. Google image search: "Google analyzes the text on the page adjacent to the image, the image caption and dozens of other factors to determine the image content. Google also uses sophisticated algorithms to remove duplicates and ensures that the highest quality images are presented first in your results." [15].

Rohini Srihari [3] applied natural language processing to figure captions in newspapers in combination with face detection in the corresponding image. This work does not locate but takes image captions as granted. Her focus is on natural language processing and segmentation of faces in the images to associate them with person names in captions.

Rowe et.al. [10] stress the importance of captions for indexing of images. But they do not use layout (geometry) but only neighborhood in ASCII text representation. They determine statistically relevant presence or absence of particular keywords in the potential caption sentences. (MARIE-4 system).

Paek et.al.[11] classify photographs with corresponding captions into indoors and outdoors. They compare text based and image content based methods for classification. The text based method achieved 83% accuracy on a test set size of 1339.

3 Approach

The proposed approach consists of two steps: First Recognition of the image captions and second semantic annotation of the image based on the caption text (see Figure 1).

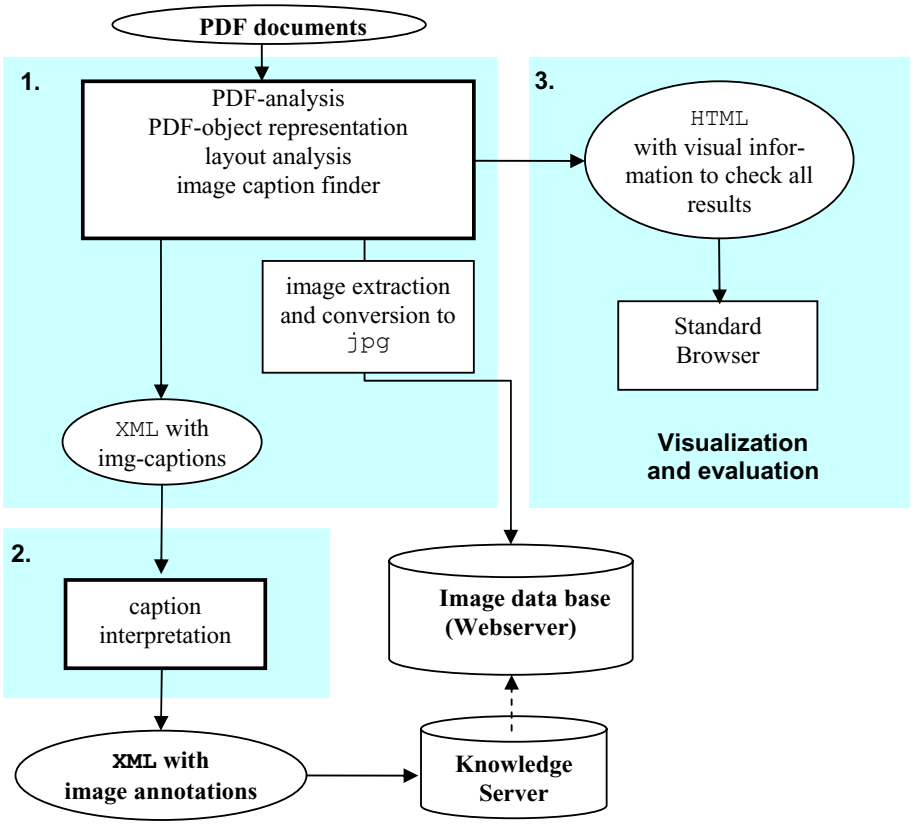

Fig. 1. Two step approach for image annotation in PDF documents: 1.Recognition of image captions and storage of extracted images in an image data base. 2. Interpretation of the captions and storage as semantic annotations of the extracted images in a knowledge server with references to the image data base. 3. The results of layout and caption analysis may be visualized for evaluation using a standard HTML-browser.

3.1 Recognition of the Image Captions

In contrast to the methods mentioned above (chapter 2) this approach tries to locate the existing *image captions* using the layout structure of the document, i.e. positions and sizes of the images and text objects and their geometric arrangement on each page in complete PDF-documents.

An *image caption* is defined as a *text block* which is intentionally placed (by the author resp. publisher) below or above the image to describe the semantics of the image. Left and right positions of captions are neglected currently because of their rare occurrence. A *text block* is defined as a visually separable unit of one or more text lines with homogeneous layout features.

The generation of text blocks applies a bottom-up process starting from the glyphs to build words, text lines, text blocks, and columns. This is similar to document image

analysis methods [12, 13], but more accurate, because there are no distortions due to scanning noise or page skew. Instead of the pixel image we use the PDF objects and streams for text and image layout analysis. The images are located and saved in separate files after conversion to jpg format if necessary.

Fig. 2. Global resolution of Top-Bottom caption conflict using font attributes from the whole PDF-document. The bottom text block is recognized (cyan) as caption of image 1_11. The rectangles display the dimensions of the text lines and text blocks from the original PDF-page.

The main problem is to decide which text block is an image caption resp. which image has a caption. This problem is solved as a constraint satisfaction problem using generic layout rules. The rules are derived from the standard publishing rules for allocating text blocks and image captions using font, line, and block attributes.

The recognition process has three phases: (1) Neighborhood analysis starting from the images with local conflict resolution, (2) Neighborhood analysis starting from the text blocks and local conflict resolution, (3) Global resolution of the remaining ambiguities. The last step tries to determine dominant layout attributes in the whole document, like font type, font size and font style to discriminate the captions from other text blocks (see an example in Fig. 2).

3.2 Semantic Annotation of Caption Text

The semantic annotation of the caption text is performed in cooperation with the DFKI (German Center for Artificial Intelligence). It is based on linguistic-semantic analysis of the caption text and the whole document text using the SPROUT tool. A detailed description is given in [14].

4 Results and Discussion

The first test set is a document collection (corpus) of 290 PDF documents downloaded from the FIFA web site *http://fifaworldcup.yahoo.com/06/de/index.html* containing 3323 pages with a total of 6507 images. This corpus was chosen because this work is part of a larger project called SmartWeb [16], which has the goal to allow natural language questions to be answered automatically by semantic web technologies. A first use case is the soccer domain in the context of the FIFA WorldCup 2006 in Germany. The results are summarized in Table 1 and Table 2.

Table 1 shows the confusion matrix between the recognized image captions and the Ground Truth data of the test set. The diagonal entries show the correct results. For images without captions small images (below a size threshold) are separately shown, and all of them are correctly recognized (true negatives TN). From the remaining 2716 images without captions 3+8=11 images erroneously got an image caption (false positives FP). This proves the intended high specificity (TN/(TN+FP) = 99.83%) of our approach.

In total 76 image captions were not found (false negatives FN), from which 6 captions were left/right captions that are not yet considered in our approach.

The majority of captions are located below the images (caption type Bottom). There are 9 + 6 = 15 captions associated with the wrong images (Top/Bottom confusion), which we also count as false positives (FP) in Table 2.

Table 1. Confusion matrix between recognized types of caption and the ground truth (GT) for the test set of 290 PDF documents with a total number of 3323 pages and 6507 images

Ground-Truth: / **Recognized:**	Small images	Without captions	Top	Bottom	Left	Right	**Sum of GT images**
Small images (no caption)	3145	0	0	0	0	0	3145
Img without caption	0	2705	3	8	0	0	2716
Img with Top caption	0	49	135	9	0	0	193
Img with Bottom caption	0	21	6	420	0	0	447
Img with left caption	0	4	0	0	0	0	4
Img with right caption	0	2	0	0	0	0	2
Sum of images	3145	2781	144	437	0	0	6507

In Table 2 the results for top and bottom captions are summarized. In terms of the common quality measures of *precision* and *recall* the result is as follows:

Precision = TP / (TP + FP) = 555/(555+26) = 95.5% and Recall = TP / (TP + FN) = 555/(555+70) = 88.8%, whereas FP consists of 3+8=11 non-captions and 9 + 6 = 15 Top/Bottom confusion errors. The 5850 true negatives consist of 3145 small images and 2705 images recognized without caption.

The average processing time per document is 0.74 sec and per page about 0.06 sec on a standard PC with a 2.7 GHz Pentium 4 processor.

Table 2. Recognition results for image captions over the whole test set of 290 documents with 3323 pages and 6507 images. This results in a precision of 95.5% and recall of 88.8% for both top and bottom caption recognition.

Caption type	True positives	False positives	True negatives	False negatives
Top	135	9 (=6+3)		49
Bottom	420	17 (=9+8)		21
Both (sum)	*555*	*26*	*5850*	*70*

Fig. 3. Result of image caption recognition on a complex PDF page with several background images and images without captions. The only image caption (no. 4) of image 1_4 (cyan) with the text "Prof. Wahlster; Ministerpräsident Müller; Prof. Seibert, FORGIS" was correctly recognized.

In Figure 3 we show the result on a complex PDF-document with a lot of background images and many images without captions.

The second test was performed with a small set of PDF-documents which were converted from HTML to PDF using Adobes PDFprinter. The purpose of this test was to check the quality of the resulting HTML-files (Fig. 1, No. 3) by comparing it with the original HTML.

These PDF-documents may contain a large number of images per page consisting of small graphical objects in gif format because HTML does not support graphics format. The results were comparable to the first test set, but some new problems occurred: Some image captions belonged to the text of buttons, which was sometimes misleading. Figure 4 shows an image caption which was correctly located but does

Fig. 4. The image caption containing the text "Foto vergrössern, Fotogalerie" was correctly recognized, but does not describe the image content. This PDF file was generated from an HTML page. In the original HTML format the "image caption" is a button that has to be clicked by the user to display the image together with the actual caption from data base.

Jared Borgetti ist der neue Rekordtorschütze der mexikanischen National-mannschaft

Name:	Jared Borgetti
Team:	Mexico
Match:	Mexico - Costa Rica
Location:	Mexico City
Date:	2005-08-17
Scorer:	1:0
Minute:	62

Fig. 5. The Semantic Annotation of the figure caption (middle) recognized the name "Jared Borgetti" of the player. Using the semantic annotation of the whole document further data of the soccer event can be determined (right).

not describe the image content. We did not observe such image captions in original PDF documents. This weakness of the layout based caption recognition is obvious, but can be remedied by the following linguistic and semantic post processing.

The semantic annotation is not in the main focus of this paper (see [14]). An example of the results is given in Figure 5. Detailed results and discussion will be presented in a future paper.

5 Conclusion

This paper describes a new layout based approach to find image captions in PDF-documents. The application of layout rules to discriminate image caption text blocks

from other text blocks is successful. This work complements existing systems for image indexing like Google image search, which do not support PDF-documents.

The method was tested on a test set of 290 PDF-documents containing about 3000 pages with about 6500 images. The precision and recall of correct image captions is 95.5% resp. 88.8%. Only 0.17% of images without captions are erroneously associated with a caption.

The subsequent semantic annotation of the image captions using the SPROUT tool is promising. This technique is applicable also to HTML-pages.

Applications of this work are not limited to searching in the web but also suitable for analysis of existing electronic archives of legacy documents in PDF format.

Acknowledgements

We would like to thank Paul Buitelaar and his colleagues from the German Research Center for Artificial Intelligence (DFKI) for providing the corpus of PDF documents and the linguistic-semantic annotation tools.

This work was supported in part by the German Federal Ministry of Education and Research under grant no. BMBF FKZ 01IMD01K.

References

1. Flickner, M., et. al.: Query by image and video content: the QBIC system. IEEE Computer 28 (9), 1995, 23--32
2. Sable, Carl L., Hatzivassiloglou, Vasileios: Text-based approaches for non-topical image categorization. Int. J. Digital Libraries (2000) 261–275
3. Srihari, Rohini K.: Automatic Indexing and Content-Based Retrieval of Captioned Images. IEEE Computer, September, (1995) 49-56
4. www.adobe.com
5. www.xpdf.com, www.foolabs.com
6. Kou, Zhenzhen, Cohen, W.W., Wang, R., Murphy, R.F.: Extracting information from text ands images for location proteomics. Proceedings of the 3rd ACM SIGKDD Int. Workshop on Data Mining in Bioinformatics, Washington DC, USA, Aug. (2003) 2-9
7. W.W. Cohen, R. Wang, R.F. Murphy: Understanding Captions in Biomedical Publications. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington DC, USA, Aug. (2003) 499-504
8. William S. Lovegrove and Davis F. Brailsford, Document analysis of PDF files: methods, results and implications, Electronic Publishing, Vol. 8, 1995, 207 - 220
9. H. Chao and X. Lin, Capturing the Layout of Electronic Documents for the Reuse in Variable Data Printing, Proc. 7th International Conference on Document Analysis and Recognition, Seoul, Korea, August 2005, 940 - 944
10. Neil C. Rowe and Brian Frew, Automatic Caption Localization for Photographs on World Wide WebPages, Information Processing and Management, Volume 34, 1998, 95 - 107
11. S. Paek, C. L. Sable, and V. Hatzivassiloglou, Integration of Visual and Text Based Approaches for the Content Labeling and Classification of Photographs, ACM SIGIR Workshop on Multimedia Indexing and Retrieval, 1999
12. S. Mao, A. Rosenfeld, T. Kanungo, Document structure analysis algorithms: a literature survey Proc. SPIE Electronic Imaging, January 2003 SPIE Vol. 5010, 197-207

13. Gerd Maderlechner, Peter Suda, and Thomas Brückner, Classification of documents by form and content, Pattern Recognition Letters 18 (11-13), 1997, 1225-1231
14. M. Becker, W. Drozdzynski, H.-U. Krieger, J. Piskorski, U. Schäfer, F. Xu, SProUT, Shallow Processing with Unification and Typed Feature Structures, Proceedings of the International Conference on NLP (ICON 2002). December 18-21, Mumbai, India, 2002
15. www.google.com/help/faq_images.html (at 10.02.2006)
16. http://SmartWeb.dfki.de
17. Philipp Mayr, Das Dateiformat PDF im Web - eine statistische Erhebung, Informationswissenschaft & Praxis, Vol. 53, 2002, 475 - 481