Fusion of Word Spotting and Spatial Information for Figure Caption Retrieval in Historical Document Images

Khurram Khurshid[1], Claudie Faure[2], Nicole Vincent[1]
[1]*Laboratoire CRIP5 - SIP, Université Paris Descartes -* [2]*UMR CNRS 5141 - Télécom-ParisTech*
{khurram.khurshid ; nicole.vincent} @ mi.parisdescartes.fr - cfaure@telecom-paristech.fr

Abstract

We present a method for figure caption detection by employing a fusion of several information sources. The evaluation is performed on documents gathered from the collection of the historical medical digital library Medic@. A method based on perceptual grouping simultaneously segments the vertical and horizontal text lines in a page. Spatial relationships between the text lines and the graphics are considered to select a set of caption line candidates. A feature-based word-spotting method is proposed to retrieve the occurrences of word images similar to a given query. Word-spotting is applied to detect the label of the captions, a word like 'Fig', 'FIG', 'Figure' ... followed by the figure number. Combining spatial information and word recognition greatly improve the detection of caption lines. Our initial experiments process more than 300 pages from three different books.

1. Introduction

Indexing books to build digital libraries is most often done manually as no automatic method is yet adapted to the needs of the archivists. One of the indexing tasks is to extract images with the associated captions and to save them in a table of figures. Figure 1 shows a screenshot of the historical medical library Medic@[12] with manual indexing of Figure/caption pairs for a book. We propose to facilitate the task of the archivists with an automatic detection of graphics and figure captions. Caption lines are extracted by merging results issued from different systems. Thus, the new system outperforms the individual systems. Here, we demonstrate this feature by involving our word spotting method in the figure caption detection for historical books where character recognition does not work efficiently with commercial OCR systems.

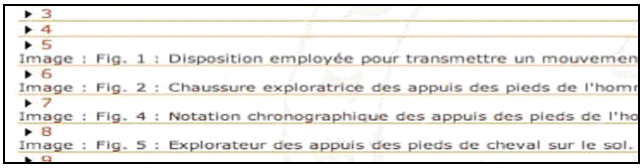


Figure 1. Manual Figure/caption indexing on BIUM [12] web base

The occurrences of the caption labels of a book (e.g. 'Fig.', 'Figure') are searched by word-spotting. The pages are preprocessed to segment graphics and the horizontal and vertical text lines. Text lines are sorted using spatial criteria to select caption line candidates in which the caption label occurrences are searched. The set of text lines explored by word-spotting and the value of the word-similarity parameter are updated during the process according to the word search results.

The next section describes the word spotting system. Section 3 outlines how spatial organization is processed in document images to simultaneously extract vertical and horizontal text lines and to select the caption line candidates. The use of word-spotting in combination with spatial information is presented in the last section.

2. Word Spotting

Lot of work has been done in the field of historical document analysis [13], more specifically, information retrieval using word spotting. Rath and Manmatha [1, 6] introduced an approach which involves grouping word images into clusters of similar words by using word image matching. Four profile features for the word images are matched using different methods [1]. Rothfeder et al. [8] used the corner feature correspondences to rank word images by similarity in historical handwritten manuscripts. Telugu scripts have been characterized by wavelet representations of the words in [7]. But this wavelet representation does not give good results for the Latin letters. Adamek et al. [9]

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.161

introduced the matching of word contours for holistic word recognition. The word contours are extracted and matched using elastic contour matching technique.

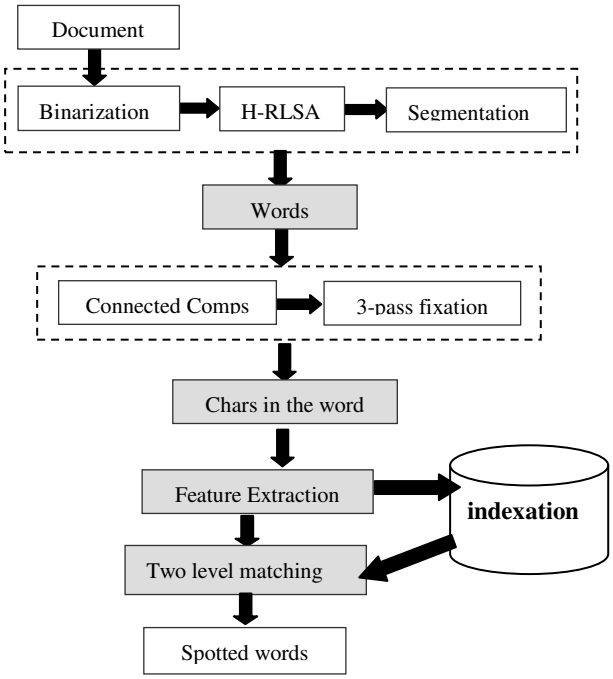


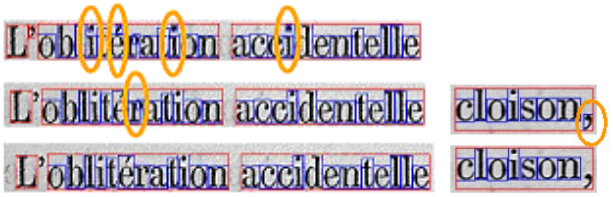
Figure 2. Different stages of word spotting

Our word spotting model is based on feature vectors representing the character images for the purpose of word matching and on the use of two different similarity measures chosen according to the level (character or word) at which comparison is performed. As opposed to [1] where features are extracted from the whole word image, here the words are segmented into characters and the features are extracted from each character of a word, thus giving more precision in word spotting as proved by the results [2]. Figure 2 describes the system. First, the document image is binarized using NICK algorithm presented in [3]. Text is separated from graphics using a size criterion to discriminate the connected components. The words in the document are extracted by applying a horizontal Run Length Smoothing Algorithm (RLSA) [5] on the binarized image and by finding all the connected components in this RLSA image.

The next sub-sections present the character segmentation, the different features extracted and finally the two levels of similarity computation.

2.1. Character segmentation

The connected components (CC) of the binary image do not always correspond to characters, especially in historical documents. A character may be broken into multiple components or multiple characters may form a single component. Character segmentation is obtained after a **3-pass** processing of the CCs included in a word component. In the first pass, components on the top of each other are assigned to the same character (Fig 3a). This is particularly efficient to segment characters such as i, j, é, à ... In the 2nd pass, overlapping components are merged into a single character. This helps to fix the broken characters due to bad quality of the printing (Fig 3a). In the 3rd pass, the punctuation marks (‘,’ ‘.’) are detected using size and location criteria to be removed (Fig 3b). Testing this method on 48 images of 12 different BIUM books having a total of 82,264 characters, we were able to properly extract 98.6% of the characters, while without these passes, we had around 85% proper characters.



(a) (b)

Figure 3. a) Pass 1 & 2 b) Pass 3

2.2. Feature selection

Feature selection is an important part of the process as the indexation of the text parts in the book will be relying on these features. We have defined a set of six feature vectors for a better representation of the characters [2]. We have made use of the four features used at the word level by [1] and introduced two new ones. The length of vector sequence associated with a character is equal to the width of the character bounding box. The features we use are: Vertical projection profile (on the gray level image); Upper character profile position; Lower character profile position; Vertical histogram; Number of ink/non-ink transitions; Middle row transition state. These feature vectors are built for each pixel column of a character. An index file is created for each document where the coordinates of each word, number & position of characters in the word, and the features of each character are stored.

2.3. Word matching

For matching two words, we proceed at two different observation levels: word and character. At the word level we compare the character sequences using the Levenshtein Edit distance [11]. At the character

level, a non-linear elastic matching is more appropriate to compute the character similarity. Some deformations may occur in the characters; therefore, the length of the vector sequences may be different for occurrences of the same character. Elastic matching is able to account for the nonlinear stretch or compression of characters.

A character is represented by a sequence of feature vectors: $X = (x_1 \ldots x_m)$ where x_i is a 6-dimensional vector. To determine the DTW distance between two sequences, X and $Y=(y_1 \ldots y_n)$, D(m,n) is computed as:

$$D(i, j) = \min \begin{Bmatrix} D(i, j-1) \\ D(i-1, j) \\ D(i-1, j-1) \end{Bmatrix} + d(x_i, y_j)$$

$d(x_i, y_j)$, is the Euclidean distance in the feature space, i varies from 1 to m, j from 1 to n. The distance between two characters is equal to D(m,n) divided by the number of steps of the warping path. Two characters are similar if their matching distance is lower than an empirically fixed threshold.

At the word level, the substitution, deletion and insertion costs are derived from the previous character distance. The cost of inserting or deleting a character is the distance between this character and an "empty" character for which the features are set to 0. The edit distance is compared to a threshold, to decide whether the two words are similar or not.

3. Selection of figure caption candidates

The pages in the historical books of Medic@ may contain vertical and horizontal text lines and the caption lines are not always in the direction of the main text. The proposed method detects vertical and horizontal text lines without prior assumption on their direction. It is explained in detail in [4]. A size criterion is used to interpret the large CCs as Graphics; they are labelled CCG and are discarded from the grouping process leading to text lines. The remaining CCs are grouped according to the main properties that enable a human reader to detect symbols alignments (proximity, similarity, direction continuity). Each CC is labelled NNH if its nearest neighbour is found in the horizontal direction or NNV if it is found in the vertical one. The labelled CCs are the input of a rule-based incremental grouping process. The sequences of consecutive NNH or NNV define horizontal and vertical alignments, grouping CCs according to proximity and continuity of direction. These first alignments are expanded in the later steps of the grouping process. An alignment is expanded along its main direction by merging it with its nearest neighbour alignment or by adding the nearest neighbour CC found in the main alignment direction. Typographic conditions are defined to take into account similarity and continuity properties: expanding an alignment by adding a CC or merging two alignments is allowed if the height of the resulting alignment is smaller than 1.5 its height before being expanded and if the distance between the alignment and the CC or between the two alignments is smaller than twice the height of the alignment to be expanded. This stepwise method takes advantage of emerging organisation and is easy to control. Therefore, the spatial information involved in the grouping rules is not reduced to the local information between CCs. After each step of the grouping process, previously detected alignments are reinforced or eliminated. Conflict detection is activated after each grouping step. The main conflict is detected when a CC belongs both to a vertical and a horizontal text line. A voting rule solves this conflict: the vertical (horizontal) line is eliminated if it contains a number of CCs smaller than the number of CCs in the horizontal (vertical) line intersecting it. To be consistent with layout conventions, a text line cannot straddle the borders of the CCG bounding boxes. Therefore, text lines can be detected outside or inside a CCG.

Spatial criteria are defined to sort the detected text lines in order to select caption line candidates. Actually, these criteria are aimed at selecting the line of the captions which is closest to the figure. For each side of a CCG, the nearest text line is found (at most four lines). The confidence of these text lines is increased by one if it is the closest line to the CCG or if the centre of the CCG and the centre of the text line are aligned along a vertical or horizontal direction. Text lines with a positive confidence belong to the set of caption line candidates. With the current system version, text lines included in the CCG bounding box cannot be among this CCG caption line candidates.

4. Fusion of information

The most important phase of this study is the fusion of symbolic information obtained by word spotting and spatial information used to select the caption line candidates. Once the bounding boxes of the caption line candidates are obtained, word spotting is applied in order to increase the confidence of some candidates, to eliminate false positive candidates and also to find caption lines missing in the candidate set. When applied to the extraction of figure captions, the input of the word-spotting process is both a query word and the location and direction of the candidate lines.

The caption line candidates constitute the first set of text lines that are explored by word-spotting. For each

book, a caption label, which can be a word such as 'Fig', 'Figure' or 'FIGURE', is chosen to become the query word. The vertical caption line candidates are rotated to become horizontal in order to apply character segmentation, feature extraction and word matching as previously defined for horizontal lines.

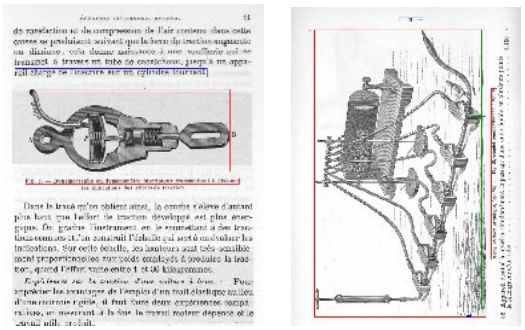


Figure 4. Examples of figure caption candidates along with the figure bounding box

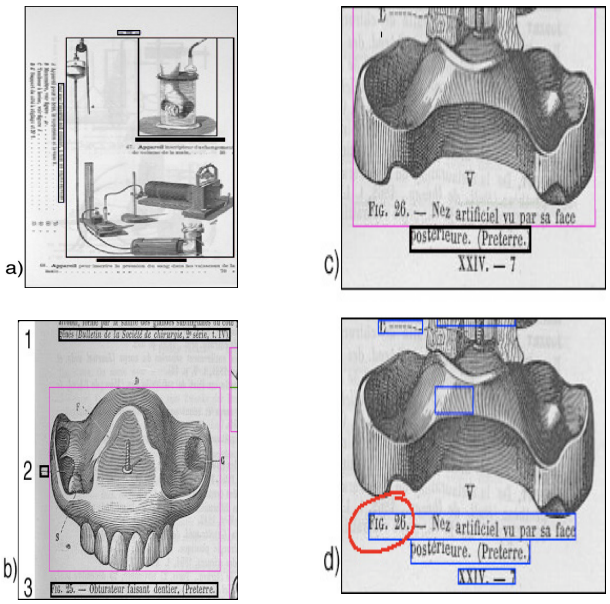


Figure 5. a) Figure and caption candidates, true caption lines are filled in black. b) Three line candidates, after word spotting of "Fig" candidates 1 and 2 are eliminated. c and d) Candidates do not include the first caption line retrieved by word spotting of "Fig."

If no occurrence of the query word is found in the candidate lines associated with a figure, the word similarity threshold is increased. Increasing the tolerance in word-matching is acceptable in this context as the search is limited to the bounding box caption candidate. It enables to detect missing occurrences without increasing false positive detection. Once a caption label is found in a candidate line associated with a figure, its confidence is increased and the other lines are neglected. Fig 5b gives an example where, out of three line candidates, the word 'Fig' is detected only in one of them and is then accepted as the caption line.

If the query word is not found in any caption line candidate associated with a figure, this is interpreted as a missing caption line in the candidate set. Query word is then searched in all the lines of the page. If it is found in the spatial neighborhood of a figure, then the relative position of that line and the actual figure is examined to see whether or not this line may represent a caption. Fig 5c,d shows an example where the first line of the caption was not detected in the caption line candidates (this line is included within the bounding box of the figure itself). Word spotting of the word 'Fig' in this line permits to detect the caption line missing in the candidate set.

The results of information fusion for the detection of caption lines are analyzed to assess the number of confirmed as well as suspected caption candidates.

5. Results

The fusion of text retrieval and spatial information reasoning has been tested on three different historical books printed in the XIXth century. They are available in the digital library Medic@. A total number of 210 figure caption exist in these books. Out of these 210, 204 captions contain a caption label: the word 'Fig', while the other six do not contain any caption label.

Spatial information leads to 449 caption candidates. This number is well above the actual caption number. Out of the ground truth 210 caption lines, 192 text lines have been detected perfectly and selected as caption line candidates, while 18 were either not detected perfectly or were not selected at all in the candidates.

Now to evaluate these caption candidates, word spotting provides information to distinguish between true and false positives. By applying our word spotting algorithm for query word 'Fig' in the candidate lines using a fixed threshold, we were able to detect 160 occurrences of 'Fig', thus confirming 36% candidates. Increasing the threshold enabled to confirm 9 new candidates. By applying word spotting for 'Fig' in all text lines of the document images, 11 caption lines missing in the candidate set are detected. The results are summarized in Table 1.

Results show the efficiency of our system to confirm or eliminate caption line candidates. The use of word spotting leads to associate 180 figures (out of 210) with a single caption line containing the word "Fig.". The six figures for which the caption does not have a label ("Fig." or other) are not in this set. A caption

candidate may be associated to one or several figures and a figure may have one or several caption candidates. Once a label is recognized in a caption candidate, the other candidates (if any) associated to the same figure are eliminated.

Table 1. Results for figure captions

Caption line candidates	
# Figure caption in ground truth	210
#Total caption candidates	449
#Well detected text lines selected to be caption candidates (Recall)	192 (91%)
# False positive caption candidates (Precision)	252 (43%)
Word spotting	
# Captions in ground truth with a label ('Fig')	204
# Caption candidates confirmed by word spotting	160
# Caption candidates confirmed by increasing word similarity threshold	9
# Captions retrieved by word spotting of query word in all text lines	11
# Total captions detected using word spotting (Recall)	180 (88%)
# False positives during word spotting in candidates (Precision)	4 (98%)

One advantage of using word-spotting is that a large number of figure and caption pairs are directly detected based on spatial and visual properties without further analysis. Another is the possibility to detect caption lines which are not in the candidate set. The errors resulting from false positives recognition do not have a strong contribution in the results.

6. Conclusion

We have proposed a two level word spotting process that enables to extract automatically the symbolic information from document images. This capability was applied to the detection of figure and caption pairs in order to take advantage of the symbolic information that is often encountered in captions to label them (the word "Fig." in our data set) and to avoid a complex decision process based on visual and spatial information. Most of the figures were directly associated with a single caption by word-spotting, thus further processing is activated only for a small number of figure and caption pairs, among them the figures with a non-labeled caption.
Our system can bring some help to speed up and semi-automatize the indexing of books and their presentation on the web. The expected final result is the detection of figures associated with caption blocs and also the recognition of the figure numbers following the caption labels (if any). Establishing a link between figure numbers occurring in the text and the associated caption is another goal for the combination of word-spotting and spatial analysis of document images.

7. References

[1] T. M. Rath, R. Manmatha, "Word Spotting for historical documents", IJDAR (2007) 9:139-152

[2] K. Khurshid, C. Faure, N. Vincent, "Feature based word spotting in ancient printed documents", *PRIS* 2008.

[3] K. Khurshid, I. Siddiqi, C. Faure, N. Vincent, "Comparison of Niblack inspired binarization techniques for ancient document images", *16th Int. Conf. DRR*, 2009.

[4] C. Faure., N. Vincent, "Simultaneous detection of vertical and horizontal text lines based on perceptual organization". *16th Int. Conf. DRR*, San Jose, 2009.

[5] K.Y. Wang, R.G. Casey, F.M. Wahl, "Document analysis system", IBM J. Res.Development, pp. 647-656, 1982.

[6] Tony M. Rath, R. Manmatha, "Features for Word Spotting in Historical Manuscripts", *ICDAR,* 2003.

[7] A. K. Pujari, C.D. Naidu, B.C. Jinaga, "An adaptive character recogniser for telugu scripts using multiresolution analysis and associative memory", ICVGIP-2002

[8] J. L. Rothfeder, S. Feng and T. M. Rath, "Using corner feature correspondences to rank word images by similarity", *Conference CVPR*, Madison, USA, 2003, pp. 30-35.

[9] Adamek, T., O'Connor, N. E. and Smeaton, A. F., "Word matching using single closed contours for indexing handwritten historical documents", IJDAR, 2007, 9, 153 – 16

[10] Ar, I., and Karsligli, M.E., "Text Area Detection in Digital Documents Images Using Textural Features," CAIP, LNCS 4673, Springer-Verlag, 555-562, 2007.

[11] R. A. Wagner and M.J. Fischer, "The string-to-string correction Problem", Journal of ACM, v21, pp 168-173,1974

[12] Digital Library of BIUM (Bibliothèque Interuniversitaire de Médecine, Paris), http://www.bium.univ-paris5.fr/histmed/medica.htm.

[13] H. S. Baird, "Difficult and urgent open problems in document image analysis for libraries", *1st International workshop on Document Image Analysis for Libraries,* 2004.

[14] E. Keogh, M. Pazzani, "Derivative Dynamic Time Warping", 1st *SIAM I. Conf. on Data Mining, Chicago,* 2001.