WEB DOCUMENT IMAGE RETRIEVAL SYSTEM BASED ON WORD SPOTTING

K. Zagoris, N. Papamarkos and C. Chamzas*

*Image Processing and Multimedia Laboratory
Department of Electrical & Computer Engineering
Democritus University of Thrace
67100 Xanthi, Greece, papamark@ee.duth.gr

ABSTRACT

Nowadays, the huge non-indexing quantities of image archives (especially document images) require the development of intelligent tools for their retrieval with convenience comparable of the texts search engines. The proposed technique addresses the document retrieval problem by a word matching procedure. It performs matching directly in the images bypassing OCR and using word-images as queries. It is constituted of two different parts: The offline and the online operation. In the offline operation, the archive of document images is examined and the results are stored in a database. The online operation consists of the web interface, the creation of the word's image and finally, the matching stage. The proposed matching process it can be described shortly as a two-threshold rating system. Finally, the proposed system has been build and it can be found in at the web address: http://orpheus.ee.duth.gr/irs2.

Index Terms—Internet, Document image processing, Information retrieval, Feature extraction

1. INTRODUCTION

In the last years, the world has experience a phenomenal growth of the size of multimedia data and especially documents images, which has been sparked by the easiness to create such images from the today scanners or digital photocopiers. Also, the problems which are associated of maintaining huge quantities of printed documents have created large images archives without any indexing information. All these have caused the need to store, retrieve and transmit such images with convenience comparable of the texts search engines.

In the last years, an alternative approach has emerged. These systems, which are called Document Image Retrieval, are performing the matching directly in the image data bypassing OCR and using word-images as queries [1]-[7].

This work is co-funded by the project "ΠΕΝΕΔ 2003-03ΕΔ679".

Their goal is not to recognize exact all the text but to provide an answer if the word that the user is searching are situated inside the document.

In this paper, we propose a Web Document Image Retrieval System (WDIRS) based on exact word matching. Our system has the ability to search words in the documents images. We assume that a page layout analysis technique has found the uniform text regions which then fed to our document image retrieval system.

2. WEB DIR SYSTEM STRUCTURE

Figure 1, depicts the overall structure of the proposed system. It is consist of two different parts: The OFFLINE and the ONLINE operation.

In the offline operation the archive of document images are examined and the results are stored in a database. This digital "scanning" consists of three stages. At first stage the document passes the preprocessing stage which includes a binarization with the Otsu [8] method, a mean filter and a skeletonization operation [9].

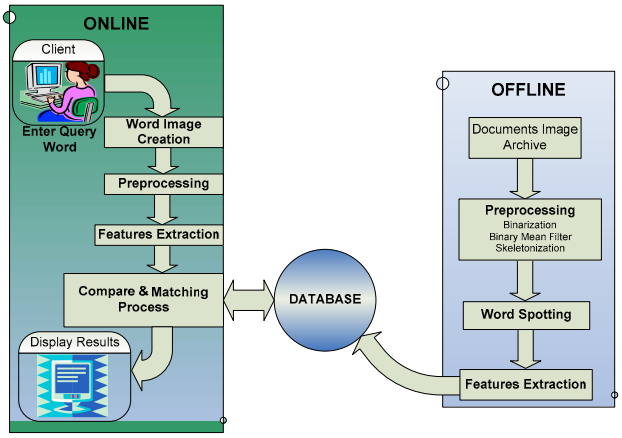

Figure 1. The overall structure of the Web Document Image Retrieval System

The word segmentation stage is following the preprocessing. Its primary goal is to detect the word blocks.

This is achieved with the continuously use of vertical and horizontal projections (Recursive X-Y Cuts) [10].

Finally, the small words (~2 characters) are rejected for the reasons that no one searching for them ("to", "in" etc). So each time a word-box has smaller width than a threshold which is called Minimum Width Box (MWB), it is rejected. In practice, it is calculated that a threshold equal to 6% of the total document's width is enough. Figure 2 presents the final extracted word blocks after applying the MWB threshold.

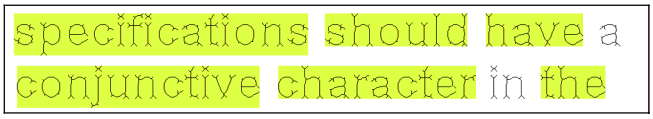

Figure 2. The result of the segmentation process with Minimum Width Box: 6%

In the final stage of the offline operation the features of each word are calculated and stored in the database. In Section 3 these features are described analytically.

The online operation of the proposed WDIRS is the visible part from the user perspective. It consists of the web interface from which the user can manipulate the system (enter the query word, see the results), the creation of the word's image, the preprocessing and features extraction stages which are the same with that in the offline operation, and finally, the matching process of the query word's features with them in the database. The matching process is described analytically in the Section 4.

3. FEATURES

The proposed WDIRS use nine (9) distinct features which are employ to describe the word's image. These are:

- *Width to Height Ratio*: The width to height ratio of the query word constitutes important shape information.
- *End Points*: An end point is a point of the skeleton word which has only one black pixel in its 8-neighborhood.

Figure 3. A visual representation of the Top-Bottom Shape Projections: (a) Original Image. (b) Top Shape Projection. (c) Bottom Shape Projection

- *Cross Points*: Cross points are the points where two or more lines of the skeleton word cross over.
- *Image Area*: This feature represents the percentage of the black pixels included in the word-box area.
- *Center of Gravity*: It represents the Euclidean distance from the word's center of gravity to the upper left corner of the word-box.
- *Horizontal – Vertical Projections*: This feature consists of a 40 elements vector. The first ten elements correspond to the first ten coefficients of the Discrete Cosine Transform (DCT) of the horizontal projection and the rest thirty elements are equal to the first thirty DCT coefficients of the vertical projection.
- *Top–Bottom Shape Projections*: As depicts the Figure 3, the *Top–Bottom Shape Projections* can be considered as signatures of the word shape. These signatures lead to a 50 elements feature vector, where the first 25 values are the first 25 coefficients of the *Top Shape Projection* DCT (Figure 3(b)) and the rest 25 values are equal to the first 25 coefficients of the *Bottom Shape Projection* DCT (Figure 3(c)).
- *Number of Characters*: This feature represents the number of characters included in the word image. This number is calculated by using a fast and effective character separation technique proposed by Papamarkos and Koutalianos [11].

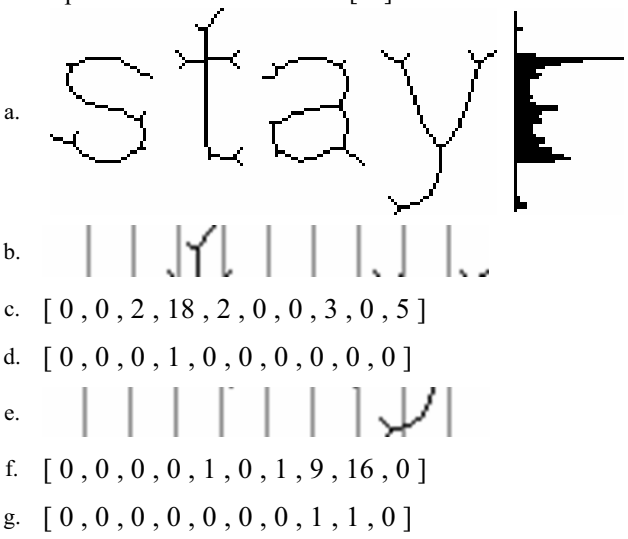

Figure 4. A visual representation of the Upper Grid and Down Grid Information feature extraction procedures: (a) The original word-box image and its horizontal projection. (b) The extracted upper part and its separation in 10 parts. (c) The number of black pixels in each part. (d) The final vector of Upper Grid feature with threshold = 7 (The half of the height of the extracted upper part image). (e) The extracted lower part and its separation in 10 parts. (f) The number of black pixels in each part. (g) The final vector of Down Grid feature with threshold = 8 (The half of the height of the extracted lower part image).

- *Upper Grid and Down Grid features*: As it is described in Figure 4, these features are extracted from the upper and lower parts of the word images and lead to two vectors with 10 binary values each (10 for the Upper Grid and 10 for the Down Grid).

4. COMPARE AND MATCHING PROCESS

The matching process it can be described shortly as a two-threshold rating system. Figure 5 illustrates the structure of the matching method.

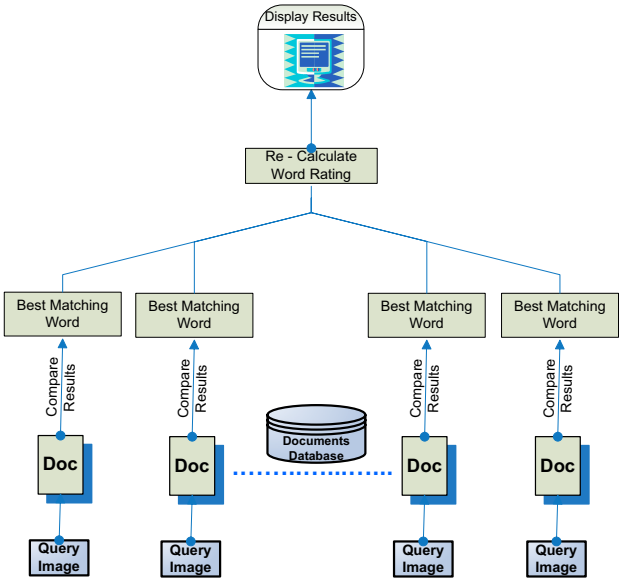

Figure 5. The overall structure of the matching process

When the user enters the query word, the system constructs a word image for each document stored in the database. The font height of the word's query image is the same as the average height of the word-boxes which have been founded in the processing document. The font type of the query image is "Arial".

First, the Euclidean distance between the features of the query image and the word included in the processing document is calculated:

$$CR(k,i)=\sqrt{\sum_{\ell=1}^{n}\left(QF(k_\ell)-WF(k_\ell,i)\right)^2}$$

where k is the feature which is being compared, QF is the feature of the query word image, WF is the feature of the document word, i is the number of the document compared word and n is the number of elements of the feature vector. So finally, there is a set $CR(k,i)$ which consists of the Euclidean distances between each document word and the query word for each of the features which have been described in Section 3.

Next, all the words of the document which doesn't satisfy the following criterions are excluded:

1st: $CR_{CN} \le T_{CN}$,

2nd: $CR_{UGIF} \le T_{UGIF}$,

3rd: $CR_{DGIF} \le T_{DGIF}$

CR_{CN}, CR_{UGIF}, CR_{DGIF} are the Euclidean distances corresponding to the "*Number of Character*", "*Upper Grid*" and "*Down Grid*" features, respectively. Also, the T_{CN}, T_{UGIF}, T_{DGIF} thresholds which define the similarity between the query and the document words. In the implementation of the proposed system, the values of the above thresholds are: $T_{CN}=0$, $T_{UGIF}=1$ and $T_{DGIF}=1.5$.

In the next stage, the rating of the remaining words is computed. The rating is a normalized value from 0 to 100 which depicts how similar are the remaining words with the query word. The rating for each word, which included in a document, is defined as:

$$R_i = 100\frac{\sum_k p(i,k)}{3}$$

where R_i is the rating value of the word i and k the feature. Also, $p(i,k)$ is a function which is defined as:

$$p(i,k)=\begin{cases}3, \text{ if } D(i,k)=0\\ 2, \text{ if } D(i,k)\le T_1(k)\\ 1, \text{ if } D(i,k)>T_1(k) \text{ and } D(i,k)\le T_2(k)\\ 0, \text{ if } D(i,k)>T_2(k)\end{cases}$$

Vector $D(i,k)$ is the difference between the Euclidean distance of the feature k of the word i and the minimum value of the feature k of the entire document words and it defined as:

$$D(i,k)=CR(i,k)-\min\{CR(k)\}$$

Table 1. The values of T_1 and T_2 thresholds

Feature	T_1	T_2
Width to Height Ratio	0.5	1
End Points	3	6
Cross Points	3	6
Image Area	1	2
Center of Gravity	4	8
Horizontal – Vertical Projections	20	40
Top-Bottom Shape Projections	40	80
Down Grid Feature	0.75	1.5
Upper Grid Feature	0.5	1

Finally, T_1 and T_2 are thresholds values which they define how similar are the features of the query image with those of the document words. Table 1 shows the values of T_1 and T_2 which are used in the implementation of the proposed WDIRS. These values have been calculated by

experimental means. So, the word with the maximum rating is nominated as the best matching word for the particular document. The same steps are repeated for each document stored in the database. Finally, a new $CR(k,i)$ has been created which consist from the best matching words of each document that has been examined.

The next and final stage consists of re-calculating the words rating of the new set. The element which changes is the minimum values of the features $min\{CR(k)\}$. At the end, the system presents the documents which contain the words with the corresponding rating in descending order. In the implementation of the proposed system, the documents which have been presented to the user are those that have a rating above 40.

5. IMPLEMENTATION AND EVALUATION

The proposed WDIRS is implemented with the help of the C#/.NET Framework. It can be found at the http:// orpheus.ee.duth.gr/irs2. The database of the documents has been created automatically from various digital text documents. This gives the advantage to implement a text search engine parallel to the WDIRS system so it's easier to verify and evaluate the search results of the WDIRS system.

In order to have an evaluation of the entire system, in this paper thirty searches have been made with random words and the Precision and Recall coefficients results that have been computed are presented in the Figure 6.

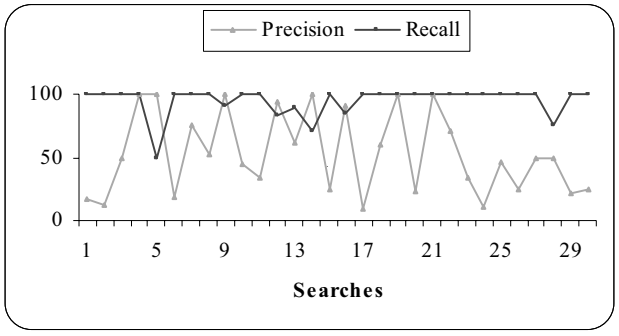

Figure 6. The variation of the Precision and Recall coefficients of the proposed WDIRS for 30 searches. The Mean Precision is 53.43 and the Mean Recall is 94.78.

Also, at the http://orpheus.ee.duth.gr/irs1demo a system has been offered which search for words included in a document and uses the proposed compare and matching process.

6. CONCLUSION

In this paper a new method for document image retrieval is proposed which is based on word spotting by using a set of powerful features and a two-threshold rating compare and matching scheme. Specifically, the proposed technique addresses the document retrieval problem by a word matching procedure avoiding OCR and using only word-images as queries. Also, an experimental platform has implemented and offered in the web. A number of test search results have shown a lot of potential. Specifically, the experiments revealed high Recall and good Precision rates proportional to the ratings thresholds predefined to the system.

The proposed system can be easily combined with page layout analysis techniques to develop a general document retrieval system.

REFERENCES

[1] D. Doermann, "The Indexing and Retrieval of Document Images: A Survey", *Computer Vision and image Understanding*, Vol.70, no. 3, pp. 287-298, 1998.

[2] J. DeCurtins and E. Chen, "Keyword Spotting via Word Shape Recognition", *Vincent, L. M., Baird, H. S. (eds.) Proceedings of SPIE*, Document Recognition II, Vol.2422, San Jose, California pp. 270–277, 1995.

[3] F. R. Chen, L. D Wilcox and D. S. Bloomberg, "Word Spotting in Scanned Images Using Hidden Markov Models", *Proc. of the International Conference on Acoustics, Speech, and Signal Processing*, Vol.5, pp. 1–4, 1993.

[4] F. R. Chen, L. D Wilcox and D. S. Bloomberg, "Detecting and Locating Partially Specified Keywords in Scanned Images Using Hidden Markov Models", *Proc. of the International Conference on Document Analysis and Recognition*, pp. 133–138, 1993.

[5] Yue Lu, Li Zhang and Chew Lim Tan, "Retrieving Imaged Documents in Digital Libraries Based on Word Image Coding", *Proceedings of the First International Workshop on Document Image Analysis for Libraries (DIAL '04)*, pp. 174, 2004.

[6] Y. Lu, C. L. Tan, W. Huang and L. Fan, "An Approach to Word Image Matching Based on Weighted Hausdorff Distance", *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, Seattle, USA , pp 921–925, 2001.

[7] S. Kuo and O. F. Agazzi, "Keyword Spotting in Poorly Printed Documents Using Pseudo 2-D Hidden Markov Models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.16, no.8, pp. 842–848, 1994.

[8] N. Otsu, "A threshold selection method from gray-level histograms", *IEEE Trans. Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62-66, 1979.

[9] B. K. Jang and R. T. Chin, "Analysis of Thinning Algorithms Using Mathematical Morphology", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v.12 n.6, pp. 541-551, June 1990.

[10] G. Nagy and S. Seth, "Hierarchical representation of optically scanned documents", *Proc. 7th Int. Conf. Pattern Recognition,* pp. 347-349, 1984.

[11] N Papamarkos, T Koutalianos, "Separation of overlapping characters", *Proceedings of ICECS '99*, vol.2, pp. 867-870, 1999.