A Holistic Methodology for Keyword Search in Historical Typewritten Documents

Basilis Gatos, Thomas Konidaris, Ioannis Pratikakis, and Stavros J. Perantonis

Computational Intelligence Laboratory,
Institute of Informatics and Telecommunications,
National Research Center "Demokritos",
153 10 Athens, Greece
{bgat, tkonid, ipratika, sper}@iit.demokritos.gr
http://www.iit.demokritos.gr/cil

Abstract. In this paper, we propose a novel holistic methodology for keyword search in historical typewritten documents combining synthetic data and user's feedback. The holistic approach treats the word as a single entity and entails the recognition of the whole word rather than of individual characters. Our aim is to search for keywords typed by the user in a large collection of digitized typewritten historical documents. The proposed method is based on: (i) creation of synthetic image words; (ii) word segmentation using dynamic parameters; (iii) efficient hybrid feature extraction for each image word and (iv) a retrieval procedure that is optimized by user's feedback. Experimental results prove the efficiency of the proposed approach.

1 Introduction

A robust indexing of historical typewritten documents is essential for quick and efficient content exploitation of the valuable historical collections. In this paper, we deal with historical typewritten Greek documents that date since the period of Renaissance and Enlightenment (1471-1821) and are considered among the first Greek typewritten historical documents. Nevertheless, the proposed methodology is generic having the potential to be applied to other than Greek historical typewritten documents.

Traditional approaches in document indexing usually involve an OCR step [3]. In the case of typewritten historical documents OCR, several factors affect the final performance like low paper quality, paper positioning variations (skew, translations, etc), low print contrast, typesetting imperfections. Usually, typewritten OCR systems involve a character segmentation step followed by a recognition step using pattern classification algorithms. Due to document degradations, OCR systems often fail to support a correct segmentation of the typewritten historical documents into individual characters [1]. In literature, two general approaches can be identified: the segmentation approach and the holistic or segmentation-free approach. The segmentation approach requires that each word has to be segmented into characters while the holistic approach entails the recognition of the whole word. In the segmentation approach, the crucial step is to split a scanned bitmap image of a document into individual

G. Antoniou et al. (Eds.): SETN 2006, LNAI 3955, pp. 490 – 493, 2006.

characters [4]. A holistic approach is followed in [2][5][6][8][11] where line and word segmentation is used for creating an index based on word matching.

In the case of historical documents, Manmatha and Croft [7] presented a holistic method for word spotting wherein matching was based on the comparison of entire words rather than individual characters. In this method, an off-line grouping of words in a historical document and the manual characterization of each group by the ASCII equivalence of the corresponding words are required. The volume of the processed material was limited to a few pages. This process can become very tedious for large collections of documents.

Typing all unique words as well as constructing an index is an almost impossible task for large document collections. To eliminate this tedious process, we propose a novel holistic method for keyword-guided word spotting which is based on: (i) creation of synthetic image words; (ii) word segmentation using dynamic parameters; (iii) efficient feature extraction for each image word and (iv) a retrieval procedure that is improved by user's feedback. The synthetic keyword image is used as the query image for the retrieval of all relevant words, initializing in this way, the word spotting procedure. The retrieval accuracy is further improved by the user's feedback. Combination of synthetic data creation and user's feedback leads to satisfactory results in terms of precision and recall.

2 Synthetic Data Creation

Synthetic data creation concerns the synthesis of the keyword images from their ASCII equivalences. Prior to the synthesis of the keyword image, the user selects one example image template for each character. This selection is performed “once-for-all” and can be used for entire books or collections. During manual character marking, adjustment of the baseline for each character image template is applied in order to minimize alignment problems.

3 Word Segmentation

The process involves the segmentation of the document images into words. This is accomplished with the use of the Run Length Smoothing Algorithm (RLSA) [10] by using dynamic parameters which depend on the average character height. In the proposed method, the horizontal length threshold is experimentally defined as 50% of the average character height while the vertical length threshold is experimentally defined as 10% of the average character height. The application of RLSA results in a binary image where characters of the same word become connected to a single connected component. In the sequel, a connected component analysis is applied using constraints which express the minimum expected word length. This will enable us to reject stopwords and therefore eliminating undesired word segmentation. More specifically, the minimum expected word length has been experimentally defined to be twice the average character height.

4 Feature Extraction

The feature extraction phase consists of two distinct steps; (i) normalization and (ii) hybrid feature extraction. For the normalization of the segmented words we use a bounding box with user-defined dimensions. For the word matching, feature extraction from the word images is required. Several features and methods have been proposed based on strokes, contour analysis, zones, projections etc. [2][3][9]. In our approach, we employ two types of features in a hybrid fashion. The first one, which is based on [3], divides the word image into a set of zones and calculates the density of the character pixels in each zone. The second type of features is based on the work in [9], where we calculate the area that is formed from the projections of the upper and lower profile of the word.

5 Word Image Retrieval

The process of word matching involves the comparison/matching between the query word (a synthetic keyword image) and all the indexed segmented words. Ranking of the comparison results is based on L_1 distance metric. Since the initial results are based on the comparison of the synthetic keyword with all the detected words, these results might not present high accuracy because a synthetic keyword cannot a priori perform a perfect match with a real word image. Motivated by this, we propose a user intervention where the user selects as query the correct results from the list produced after the initial word matching process. Then, a new matching process is initiated. The critical impact of the user's feedback in the word spotting process lies upon the transition from synthetic to real data. Furthermore, in our approach user interaction is supported by a simplified and user friendly graphical interface that makes the word selection procedure an easy task.

6 Experimental Results

For the evaluation of the performance of the proposed method for keyword guided word spotting in historical typewritten documents, we used the following methodology. We created a ground truth set by manually marking certain keywords on a subset of the available document collection. The performance evaluation method used is based on counting the number of matches between the words detected by the algorithm and the marked words in the ground truth. For the experiments we used a sample of 100 image document pages. The total number of words detected is 27,702. The overall system performance given in Fig. 1 shows the average recall vs. average precision curves in the case of single features as well as in the case of the proposed hybrid scheme. In Fig. 1 we demonstrate the improvement achieved due to user's feedback mechanism. It is also clearly illustrated that the hybrid scheme outperforms the single feature approaches.

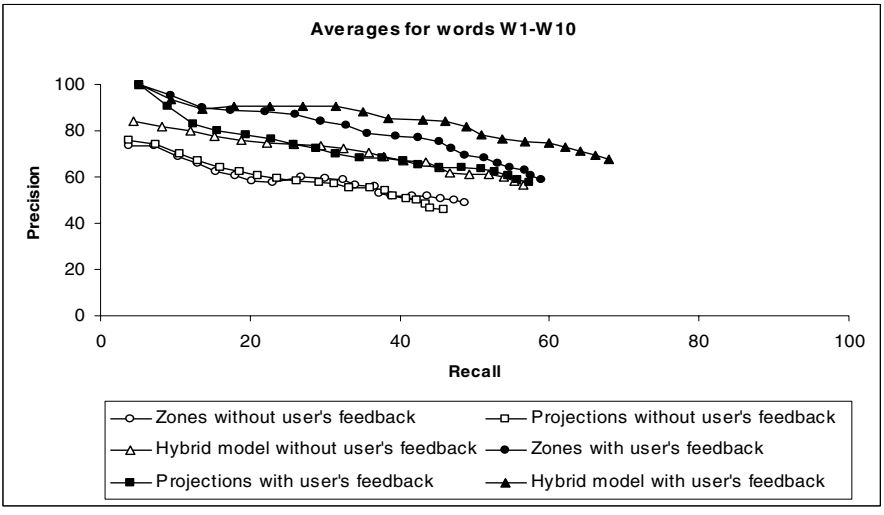

Fig. 1. Average Precision/Recall rates for all words

References

1. Baird H. S.: The state of the art of document image degradation modeling. IARP 2000 Workshop on Document Analysis Systems (2000) 10-13
2. Bhat D.: An evolutionary measure for image matching. In: Proceedings of the Fourteenth International Conference on Pattern Recognition, ICPR'98, volume I (1998) 850-852
3. Bokser M.: Omnidocument technologies. In: Proceedings of the IEEE, 80(7) (1992) 1066-1078
4. Gatos B., Papamarkos N. and Chamzas C.: A binary tree based OCR technique for machine printed characters. Engineering Applications of Artificial Intelligence, 10(4) (1997) 403-412
5. Lu Y., Tan C., Weihua H., Fan L.: An approach to word image matching based on weighted Hausdorff distance. In: Sixth International Conference on Document Analysis and Recognition (ICDAR'01) (2001) 10-13
6. Madhvanath S., Govindaraju V.: Local reference lines for handwritten word recognition. Pattern Recognition, 32 (1999) 2021-2028
7. Manmatha R.: A scale space approach for automatically segmenting words from historical handwritten documents. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, No. 8 (2005) 1212-1225
8. Marcolino A., Ramos V., Ármalo M., Pinto J. C.: Line and Word matching in old documents. In: Proceedings of the Fifth IberoAmerican Sympsium on Pattern Recognition (SIARP'00) (2000) 123-125
9. Rath T. M., Manmatha R.: Features for word spotting in historical documents. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03) (2003) 218-222
10. Waked B., Suen C. Y., Bergler S.: Segmenting document images using diagonal white runs and vertical edges. In: Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR'01) (2001) 194-199
11. Weihua H., Tan C. L., Sung S. Y., Xu Y.: Word shape recognition for image-based document retrieval. International Conference on Image Processing, ICIP'2001 (2001) 8-11