2009 10th International Conference on Document Analysis and Recognition

Word-Based Adaptive OCR for Historical Books *

Vladimir Kluzner, Asaf Tzadok,
Yuval Shimony, Eugene Walach
IBM Corporation, Haifa Research Lab
{*kvladi, asaf, yshimony, walach*}*@il.ibm.com*

Apostolos Antonacopoulos
School of Computing, Science and Engineering
University of Salford
A.Antonacopoulos@primaresearch.org

Abstract

The aim of this work is to propose a new approach to the recognition of historical texts by providing an adaptive mechanism that automatically tunes itself to a specific book. The system is based on clustering together all the similar words in a book/text and simultaneously handling entire class. The paper describes the architecture of such a system and new algorithms that have been developed for robust word image comparison (including registration, optical flow based distortion compensation, and adaptive binarization). Results for a large dataset are presented as well. Over 23% *recognition improvement is demonstrated.*

1. Introduction

There is a tremendous amount of historical information contained in world's libraries, museums, and archives. These documents contain valuable information forming the collective memory of our societies over the years. However, access to this invaluable information is limited. Indeed, since the overwhelming majority of historical documents is not yet digitized, access is usually limited to a few experts who are able to view the physical documents on site.

Accordingly, over the last several years, leading world libraries have begun an extensive digitization process, including historical books and newspapers. This mass digitization can only be achieved by full-text digitization: transforming digital images of scanned books into electronic text. However, these efforts are hampered by the high cost of extracting text from document images. Automated text recognition, carried out by Optical Character Recognition (OCR), does not produce satisfactory results for historical documents (see [3]).

The preservation of cultural and historical heritage residing in document archives is a widespread focus of interest among the Document Analysis and Recognition (DAR) community. Baird et al. [2] stated the DAR challenges in historical digital libraries collections. Professional literature includes a number of outstanding contributions in these domains [1, 6, 7, 9, 10, 15]. Nevertheless, as mentioned above, in many cases the results are still far from being satisfactory.

In some cases, the aforementioned techniques are augmented by the addition of word spotting techniques [12, 13, 16]. Here, the system is trained to search for specific words of interest. Hence, rather than performing character-based recognition (used in the conventional OCR engines) word images are searched for. Typically, word spotting is based on a set of features invariant to font shape and size and to various geometrical distortions. Another example of an existing OCR augmentation technique is [14], which transforms text images into character shape codes and uses special lexica containing information on the shape of words.

The purpose of this paper is to describe a new approach to the extraction of full text from the historic documents. State-of-the-art commercial OCR engines focus on recognizing diverse materials printed using modern fonts. In contrast, we intend to focus on historical books containing a relatively large body of homogenous material printed using rare old fonts. In this context it makes sense to create an adaptive OCR that would automatically tune itself to each book or document.

The structure of this paper is as follows. In Section 2 we describe our system architecture. In the next two sections we describe our word comparison method. In Section 3 we discuss registration between two word images and in Section 4 we compare the registered words. Section 5 describes the adaptive OCR engine based on the word comparison algorithm. Then Section 6 presents our recognition results for the chosen benchmark. Finally, Section 7 is devoted to conclusions and future directions.

*The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-201 under grant agreement 215064.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.133

2. Adaptive OCR System Architecture

2.1. Basic Concept

The system works on large bodies of homogenous material. If the given text contains several languages and/or several fonts, the system applies the adaptation in parallel, to each text type. The system works with manual correction of the OCR results (although a similar approach can be applied to purely automatic applications). OCR receives results of the manual correction and uses them to improve recognition results (so the error rate for the last page would be much lower than for the first page). However, the adaptation itself is transparent to the final user so that application is as simple and straightforward as that for conventional OCR systems.

In principle, adaptivity can be applied to conventional (character-based) OCR engines. However, we chose to apply a word-based recognition approach. This way we avoid the error-prone character segmentation process. In this, our approach is similar to that of conventional word spotting. However, in our case, the keyword library includes all the unique words in the given book. Hence, the system is not limited to a small set of invariant features. Instead, it can compare directly between the various word images, thereby yielding superior recognition results. Moreover, since this approach does not assume any a priori font knowledge, it is particularly well suited for historical fonts printed in rare typefaces.

It is useful to apply adaptive word-based OCR in connection with a conventional OCR engine. Thus, an adaptive engine is used to improve conventional results.

2.2. System Architecture

The book recognition process should start from the image enhancement and layout analysis. However, since these stages go beyond the scope of this paper, we do not describe them in detail.

Next, the scanned (text) book pages are segmented into the individual word images. Since in typical texts, inter-word separation is relatively large, this stage fairly straightforward. Our experiments use word segmentation provided by the FineReader engine.

Once individual word images are created, the system proceeds to determine equivalence classes such that each class contains images presumed to show different instances of the same word. Hence, recognition is performed simultaneously for the entire class. For instance, one can combine conventional recognition results for all the words in the given class. If no automatic recognition results exist, manual data entry is used. However, only a single class member requires manual handling (as the remaining class members are presumed to be the same).

Naturally, the efficiency of our approach hinges on having a large number of word repetitions in the given text. Figure 1 depicts the ratio between the number of distinct words and the total number of words in the dataset used. As expected, this ratio decreases exponentially, reaching (for texts above 40000 words) its limit of 1/7. This ratio illustrates the potential for savings due to the word-based adaptivity. Note, that adaptation process is expected to converge after about 30000 words.

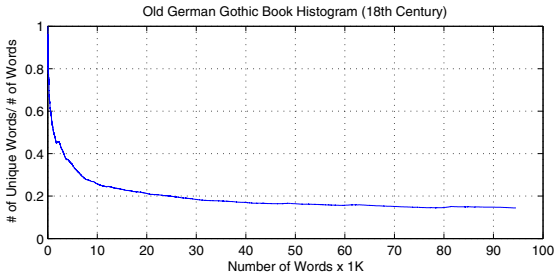

Figure 1. Ratio of unique words as a function of the text size

Clearly, the core element of our system is an algorithm for word image comparison, which determines which words are the same. To achieve good quality word image comparison, our algorithm is split into two stages: *Image Distortion Compensation* and *Difference Detection*.

Image Distortion Compensation commences from the coarse registration (based on the cross-correlation technique). The fine registration is performed using a modified optical flow method. Once the pixel displacement vector is computed, compensation is performed.

The Image Distortion Compensation module yields two gray (or color) images that can be superimposed on one other. However, for comparison purposes it is useful to create a binary version of both word images. In principle, one can apply separate binarization to each of the images. However, it is advantageous to perform simultaneous binarization of both images optimized for Difference Detection. Finally, the difference between the two binary images is computed using a non-linear difference measure to ensure that even minute single character differences are detected.

3. Word Image Registration

3.1. Coarse Registration

This stage is aimed at finding the optimal translation between the two word images. There are two known main approaches to this problem. The first is based on the sum of the squared distance measure. The second is based on normalized cross-correlation coefficients. Lewis in [11] shows

that there is a close connection between these two methods. However, the distance measure technique suffers from inherent normalization problems. Accordingly, we chose to adopt the cross-correlation approach, where the correlation coefficient is normalized already.

3.2. Modified Optical Flow

Fast normalized cross-correlation presented above cannot compensate for geometric distortions (other than translation). An example of frequently occurring word corruptions, taken from an 18th century book, is displayed in Fig. 2, top line.

Despite the coarse registration, there is still a significant degree of geometrical distortion. To overcome this problem, certain non-rigid (elastic) registration is needed. A variety of image registration techniques is described in [5, 17]. Taking into account that, in our case, relatively slight compensation is sufficient, we have chosen an alternative approach. We treat both images as if they are obtained from a video sequence. Hence, we can apply a modified version of the optical flow technique, developed for motion estimation between the two consecutive frames ([4]).

Given two similar images, the optical flow process calculates a velocity vector (u, v) for each pixel (x, y) that represents the speed and direction of the estimated pixel movement. The variational formulation of this problem is widely considered in [8]; given image $I(x(t), y(t), t)$, optimal values u, v are obtained by minimizing the following functional:

$$\begin{aligned} F(u,v) &= \int_{\Omega} \left(\nabla I \cdot (u,v) + I_t\right)^2 dxdy \qquad (1) \\ &+ \alpha \int_{\Omega} \left(|\nabla u|^2 + |\nabla v|^2\right) dxdy, \end{aligned}$$

where Ω denotes image domain. In our case, the optical flow is computed for the two images being compared. The partial derivatives $I_x,\ I_y$ are calculated on the mean of these two images, and the time derivative I_t is derived from its usual definition.

We introduced a significant modification to the traditional optical flow approach. Indeed, in our case, images originated from a bi-level source. Hence, it is possible to minimize the influence of the image background by performing preprocessing including: (i) binarization (ii) low pass filtering by the Gaussian 3×3 filter (to get smooth spatial derivatives).

The results of word distortion correction are shown in Fig. 2.

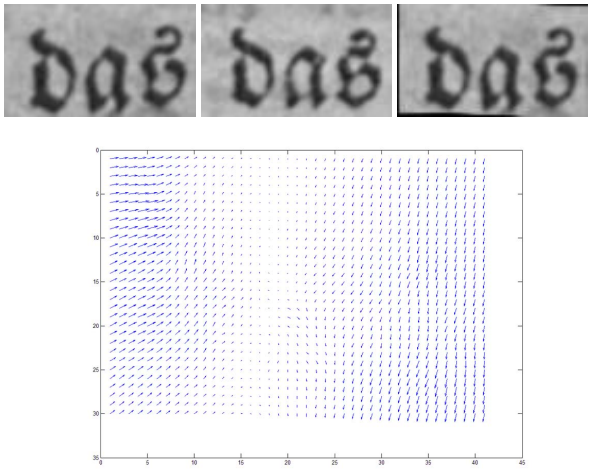

Figure 2. Word distortion correction. Top row (left and center): two original word images; bottom: distortion matrix; top right: compensation result

4. Image Difference Estimation

4.1. Adaptive Binarization

At this stage we have two registered word images that can be superimposed on top of one another. Since documents of that age are presume to be bitonal (dark text on light background), it is useful to perform comparisons on the binarized images.

There is a wide variety of existing binarization methods, all of them dealing with every image separately. We propose the joint (adaptive) binarization technique optimized for image comparison purposes.

Let's denote

$$B_1 = Thresh(I_1, t_1), \quad B_2 = Thresh(I_2, t_2),$$

where $I_1,\ I_2$ - given grey-level images, $t_1,\ t_2$ - threshold values, $B_1,\ B_2$ - resulting binary images after thresholding grey-level images $I_1,\ I_2$ with threshold values $t_1,\ t_2$ respectively. Then the joint binarization goal is to solve the following minimization problem:

$$(t_1^*, t_2^*) = arg \min_{(t_1,\ t_2)} \left[\left(\overline{B_1'} \bigcap B_2 \right) \bigcup \left(\overline{B_2'} \bigcap B_1 \right) \right],$$

where $(\cdot)'$ means the dilation operator with 3×3 element, and $\overline{(\cdot)}$ - negative image. The rationale behind this method is that we are looking for a pair of thresholds that should provide the best possible result (for a minimal difference between the two images, the naturally trivial solution $t_1^* > 255,\ t_2^* > 255$ is excluded). In case of multiple minima, we choose the one close to the result of conventional binarization method. The advantage of the adaptive

binarization is shown in Fig. 3. We show the joint part in black. Small differences are shown in blue and red. Large differences are magenta and rose. Clearly, large difference areas (magenta and rose) are much smaller when applying the adaptive joint binarization algorithm.

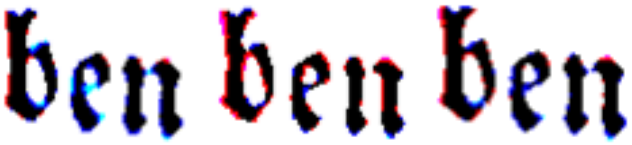

Figure 3. Adaptive joint binarization. On the left: threshold pair $(t_1,\ t_2^*),\quad t_1 < t_1^*$; in the middle: threshold pair $(t_1,\ t_2^*),\quad t_1 > t_1^*$; on the right: threshold pair $(t_1^*,\ t_2^*)$

4.2. Word Comparison

Now we are ready for the image difference estimation. In this context only large differences are considered (rose and magenta areas). For these areas, connected components are computed. The images are deemed to be the same only if the largest connected component is not greater than four pixels. This tight measure is chosen so that even minute differences in a single character are detected.

Some illustrative examples are shown in Fig. 4.

Figure 4. Equivalence classes created by the word classification algorithm

5. Word-based Adaptive OCR

In the previous sections we described our core algorithm for the comparison of two word images. The final result of this stage is a binary decision: two words are deemed to be the same or not. Based on this approach, the system goes over all the words in the given book and identifies equivalence classes i.e., groups of words that are deemed to be the same.

To reduce computational complexity, an early rejection mechanism is applied. In other words, no detailed comparison is performed when two word images are clearly different.

Then each class is recognized simultaneously. To facilitate the adaptive recognition process we utilize conventional recognition results for the state-of-the-art OCR engine. This engine was applied for all the words in each class. There are three possibilities: (i) No word in the class has a high confidence OCR result, meaning that manual data entry must be applied. However, since all the class members are deemed to be the same, it suffices to perform manual data entry for a single group member. (ii) One or more words have high confidence results and all the results are the same. Here, all class members are assigned the high confidence value provided by the OCR engine (including members that are not recognized by OCR to begin with) (iii) Inside the group there is more than one different recognition result so it is split into three subclasses (A - words having the first recognition result, B - words having the second recognition result, C - words having no high confidence results). Based on the number of words in each subclass, the system decides whether manual correction is invoked.

6. Results

To verify the validity of the above approach, we applied it to the chosen dataset by taking 101 scanned pages from a book printed in 18th century Old German Gothic font. We processed this benchmark twice: once using a leading commercial OCR engine and secondly applying the adaptive OCR process described above. Then, we compared the OCR results. For simplicity, we performed all the measurements on the word level.

Naturally, comparison of OCR engines must take into account both the number of recognized words and the number of substitution errors. To facilitate the comparison process we needed a single measure combining both these key parameters. Accordingly, we introduced a Figure of Merit (FOM):

$$FOM = (NOR + 5 * NOF)/(NOW), \qquad (2)$$

where NOR is number of rejects, NOF is number of substitution errors and NOW is number of words. The reasoning behind (2) is to 'penalize' substitution errors due to the extra manual correction required. FOM serves as an indicator of the level of processing required to correct the data manually. Hence, a lower value of FOM indicates a better recognition engine performance.

Our dataset counted 18,934 individual words, which the adaptive word recognition engine grouped into 7250 equivalence classes. Recognition results are summarized in Table 1. The first line shows the recognition results for the conventional (non-adaptive) OCR engine: a correct recognition result of 82.5% and 1.85% substitution errors yielding $FOM = 24.9\%$. The adaptive word recognition engine

improved these parameters to a 86.6% recognition rate with 1.7% substitution errors, yielding a FOM of 20.2%. Note that adaptivity improved both the read rate and error rate. Overall, FOM was reduced by about 23% indicating that the manual correction effort is reduced by this ratio.

	Reco. Rate	**Subst. Rate**	**FOM**
Commercial OCR	82.5%	1.85%	24.9%
Commercial OCR after addition of Adaptive OCR	86.6%	1.7%	20.2%

Table 1. FOM results versus baseline

7. Conclusions and Future Work

We presented a new algorithm for book-wide adaptive OCR that provides a significant enhancement to the conventional (non-adaptive) OCR engines. We believe that the results prove the validity of the chosen approach.

Our word classification algorithm proved to be quite effective in avoiding false joins (i.e., putting two different words in the same equivalence group). However there are a significant number of cases where the same words are put into different classes. So, the challenge is to reduce the number of false splits without a concomitant increase in the number of false joins. We believe that this goal can be achieved mainly by improving our method for geometrical distortion compensation. We would like to address the cases of distortions for words located near the middle of the book where binding may cause severe distortions.

Another challenge is to handle words split in two and complex words that are created by a combination of shorter words. An important example of complex words is a word created by the addition of a prefix or suffix.

Finally, we must address the issue of rare words with only one or two words present in the given book. In these cases, the adaptivity principle must be applied at a lower level: adaptation can be performed either on the character or on the ligature level.

It should be noted that in our experiments we disregarded the issue of system time performance. Indeed, an existing non-optimized system runs for several hours on a state-of-the-art server for the 101 page benchmark. However, we expect that after optimization, we can achieve a significant improvement.

Present work focuses on recognition of historic texts. However, similar approach has been tested successfully for modern text as well.

Acknowledgements: The authors thank Tal Drory and Ami Ben-Horesh of the IBM Haifa and Gūnter Mūhlberger of Innsbruck University for many valuable discussions.

References

[1] A. Antonacopoulos and D. Karatzas. A complete approach to the conversion of typewritten historical documents for digital archives. In *Document Analysis Systems VI*, volume 3163 of *LNCS*, pages 90–101. Springer-Verlag, 2004.

[2] H. Baird, V. Govindaraju, and D. Lopresti. Document analysis systems architectures for digital libraries. In *Document Analysis Systems VI*, volume 3163 of *LNCS*, pages 1–16. Springer-Verlag, 2004.

[3] H. Balk. Responding to the challenges in mass digitisation of historical printed text. In *The IMPACT workshop at the EVA MINERVA Conf.*, Jerusalem, Israel, November 2008.

[4] S. S. Beuchemin and J. L. Barron. The computation of optical flow. *ACM Computing Surveys*, 27:433–467, 1995.

[5] L. Brown. A survey of image registration techniques. *ACM Computing Surveys*, 24:326–376, 1992.

[6] B. Couasnon, J. Camillerapp, and I. Leplumey. Making handwritten archives documents accessible to public with a generic system of document image analysis. In *Proc. of First Int. Workshop on DIAL*, pages 270–277, Palo Alto, California, January 2004.

[7] J. He and A. Downton. Evaluation of a user assisted archive construction system for online natural history archives. In *Proc. of 8th Int. Conf. on Document Analysis and Recognition*, page 4246, Seoul, Korea, August 2005.

[8] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[9] N. Journet, V. Eglin, J. Ramel, and R. Mullot. Text/graphic labelling of ancient printed documents. In *Proc. of 8th Int. Conf. on Document Analysis and Recognition*, pages 1010–1014, Seoul, Korea, August 2005.

[10] F. Le-Bourgeois and H. Kaileh. Automatic metadata retrieval from ancient manuscripts. In *S. Marinai and A. Dengel, editors, Document Analysis Systems VI*, volume 3163 of *LNCS*, pages 75–89. Springer-Verlag, 2004.

[11] J. P. Lewis. Fast template matching. *Vision Interface*, pages 120–123, 1995.

[12] J. Lladós and G. Sánchez. Indexing historical documents by word shape signatures. In *Proc. of 9th Int. Conf. on Document Analysis and Recognition*, pages 362–366, Curtuba, Brazil, September 2007.

[13] T. Rath and R. Manmatha. Word image matching using dynamic time warping. In *Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 521–527, Madison, WI, June 2003.

[14] L. Spitz. Shape-based word recognition. *International Journal on Document Analysis and Recognition*, 1(4):178–190, 1999.

[15] U. Surapong, M. Hammound, C. Garrido, P. Franco, and J. Ogier. Ancient graphic documents characterization. In *Proc. of Sixth IAPR Workshop on Graphics Recognition*, pages 97–105, Hong Kong, China, August 2005.

[16] C. Tomai, B. Zhang, and V. Govindaraju. Transcript mapping for historic handwritten document images. In *Proc. of 8th Int. Workshop on Frontiers in Handwriting Recognition*, pages 413–418, Ontario, Canada, August 2002.

[17] B. Zitova and J. Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003.