2011 International Conference on Document Analysis and Recognition

Hybrid Approach to Adaptive OCR for Historical Books

Vladimir Kluzner, Asaf Tzadok, Dan Chevion, Eugene Walach
Document Processing and Management Group
IBM Research - Haifa
Haifa, Israel
{*kvladi, asaf, chevion, walach*}*@il.ibm.com*

***Abstract*—Optical character recognition (OCR) technology is widely used to convert scanned documents to text. However, historical books still remain a challenge for state-of-the-art OCR engines.**

This work proposes a new approach to the OCR of large bodies of text by creating an adaptive mechanism that adjusts itself to each text being processed. This approach provides significant improvements to the OCR results achieved.

Our approach uses a modified hierarchical optical flow with a second-order regularization term to compare each new character with the set of super-symbols (character templates) by using its distance maps. The classification process is based on a hybrid approach combining measures of geometrical differences (spatial domain) and distortion gradients (feature domain).

***Keywords*-hybrid classifier, character classification, adaptive OCR, hierarchical optical flow, second order regularization term, distance map.**

I. Introduction

Optical character recognition (OCR) technology has been used for decades to convert scanned images of documents to indexable text. Although the accuracy of commercially available OCR engines has improved to the point where many regard the OCR problem as having been solved, in practice, this statement is far from true. The mean word-level error rates for most OCR engines range roughly from 1% to 10% (see [1]). This level of OCR accuracy is inadequate for massive information retrieval applications. In addition, many commercial OCR packages have been optimized for short texts. As a result, these packages fail to utilize redundancies inherent to large bodies of text.

Thus, there is a growing need for improved methods for whole-book recognition. One of the popular approaches in this field is adaptive OCR, when the system uses an adaptive mechanism that attunes itself to the book text being processed. Various techniques using the adaptivity idea were developed in recent years. Khoubyari and Hull [2] introduced the word image matching method, which included the creation of improved word prototypes. Spitz [3] presented an algorithm that involves the transformation of text images into character shape codes. Xu and Nagy [4] proposed an automatic prototype extraction method, which is based on comparing the bitmaps of pairs of words that contain the same characters. Marinai et. al. [5] proposed an adaptive word-level indexing of modern printed documents, which are difficult to recognize using current OCR engines. Xiu and Baird [6] use a character-image classifier and word-occurrence probability model to describe an approach to the unsupervised high-accuracy recognition of the textual contents of an entire book. Kae and Learned-Miller [7] propose iterative document-specific (contextual) modeling, which first recognizes the least ambiguous characters and then iteratively refines the model to recognize more difficult characters. A document level word recognition technique is presented by Rasagna et al. [8], which makes use of the context of similar words to improve the word level recognition accuracy. Eventually, Kluzner et. al. [9] introduced the whole-book word-recognition-based adaptive OCR, assuming the existence of non-linear (elastic) distortions in the appearing words.

In this paper, we describe our extension of the work by Kluzner et. al. [9] which consists of a new adaptive OCR mechanism for large bodies of text, adopting a character-based clustering. Although the approach itself is general, we focus our testing on the particularly challenging problem of analyzing historical books containing a relatively large body of homogenous material printed using rare old fonts. In this context, the use of adaptation is especially effective.

However, because the basic recognition atom is smaller in size, the basic techniques presented in [9] are not robust enough. As a result, we strengthened our image processing and classification tools. These modifications included:

- Modified hierarchical optical flow with newly developed **second-order regularization term** (used for character comparisons)
- Use of **distance maps** of compared characters as an entry data for optical flow
- Use of a **hybrid classifier**, combining both spatial and feature domains

The structure of this paper is as follows. In Section II, we describe our system architecture. Section III is devoted to the training process, including the creation of super symbols. The recognition engine, which is the core of our system, is

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

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.183

presented in Section IV, which includes the description of modified hierarchical optical flow process, comparison of the distance maps, and the scoring methodology. Section V presents our recognition results for the chosen benchmark. Section VI is devoted to conclusions and future directions.

II. OCR System Architecture Overview

Although the book recognition process should start with image enhancement and layout analysis, these stages are beyond the scope of this paper. The Omni-font OCR approach (in our case ABBYY FineReader) is performed at the beginning of the process. Our system first segments the scanned (text) book pages into individual word images. This stage is straightforward since inter-word separation is relatively large in most texts. Our experiments use word segmentation provided by the Omni-font OCR engine. Each word is treated by our adaptive OCR engine, as described in Section IV. The system automatically extracts high confidence characters, clusters them, and performs auto-training. During this process, the so-called super-symbols are created. The correction of the OCR results, which may be performed either manually or automatically, based on existing dictionary or language model, is beyond the scope of this paper.

The recognition process is next repeated using the adaptive approach. Each character is recognized by finding the closest matching super-symbol template. Since this approach does not assume any a priori font knowledge and pre-defined feature set, it is particularly well-suited for historical fonts printed in rare typefaces, and a set of invariant features is not limited by size.

The flowchart of the process is presented in Figure 1.

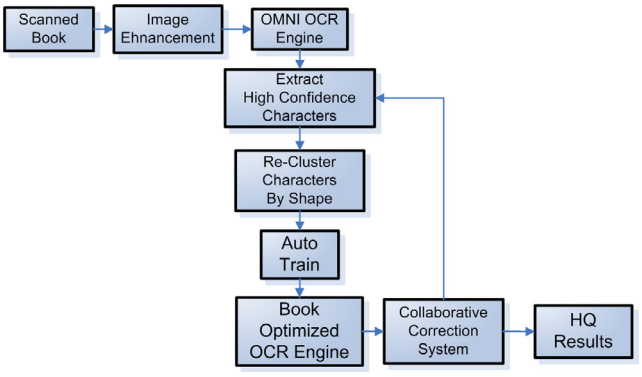

Figure 1. Adaptive OCR system architecture

III. Training Process

The goal of the training process is to create a font resource for the recognition engine. In our case, it incorporates the ideal representation of each character/symbol (super-symbol). The input data for the training process is high confidence characters automatically extracted by the system or received by manual correction of initial OCR results. The training process clusters the above characters into equivalence groups, using the cross-correlation technique as a comparison tool. The high confidence characters clustered in the same group are registered and averaged. The accepted mean character image is called a super-symbol (see Figure 2) for a given group of characters and is used by the recognition engine during the comparison process. The accepted super-symbols are close to ideal templates - particularly, they have a high signal-to-noise ratio (SNR) due to the averaging process (see, for example, three Old Gothic lowercase ”f” characters and their super-symbol on Figure 2).

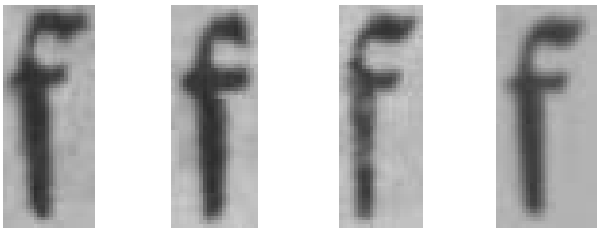

Figure 2. Three regular ”f” characters and their super-symbol (on the right)

IV. Recognition Engine

At the heart of our system lies the unique recognition engine, receiving as input two gray-level images: the current character being processed and the candidate super-symbol. The engine consists of two modules: character warping based on rigid and elastic image registration and character recognition based on newly developed hybrid classifier.

Character warping module finds the appropriate transformation between current character and candidate super-symbol. This transformation consists of image translation and elastic (non-rigid) registration, performed by modified hierarchical optical flow (see subsection IV-A) on the basis of images' distance maps (see subsection IV-B).

The transformed version of the current character is superimposed on the chosen super-symbol for classification purposes. Character classification module (see subsection IV-C) calculates the score, reflecting the similarity level between the current character and candidate super-symbol.

A. Modified Hierarchical Optical Flow-Based Second Order Regularization Term

Character warping module commences from the coarse registration, compensating for the translation difference between the two images. The compensation result is verified by means of cross-correlation metric. Stronger algorithms are needed in order to compensate for possible non-linear (elastic) differences. An attempt to solve the problem using a classic optical flow was made in [9]. Both images were treated as if they were obtained from a video sequence. Based on this notion, the distortion between the two images

was defined as an inter-frame motion. Accordingly, distortion can be estimated using a modified version of the optical flow technique.

The classical optical flow approach [10] is limited to distortions not exceeding four pixels. In order to overcome this limitation we adopted a hierarchical (pyramidal) optical flow technique (see [11]). The proposed scheme implements a three-level, coarse-to-fine warping strategy, and the standard down-sampling factor is 0.5 on each level.

Another crucial modification in our approach is an addition of a second order regularization term. Generally, the goal of regularization is to penalize transformations that are inconsistent with the known system properties. For example, for the fluid flow estimation, one may wish to penalize vorticity, divergence, or gradient of flow components (as in [9]). In order to achieve this goal, one may add a regularization term, consisting of the first order derivatives of the vector field. However, since fluid flows are not devoid of vortices, it may be desirable to limit their spatial variation. This can be achieved by the second-order regularization methods that are based on the second-order derivatives of the field function. Given the flow $\vec{f}$, the regularization term in this case can be expressed as $\nabla div(\vec{f})$ or $\nabla curl(\vec{f})$ (see [12]). To simplify the computational process, we use the Laplacian of the optical flow components.

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 will be as follows: given image $I(x(t), y(t), t)$, optimal values u, v are obtained by minimizing the following functional:

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

where Ω denotes the image domain. Assuming the desired optical flow (u, v) is a minimum of functional (1), we look for a solution of the following system of Euler-Lagrange equations

$$\begin{cases} I_x \left(\nabla I \cdot (u,v) + I_t\right) - \alpha \triangle u + \beta \triangle(\triangle u) = 0 \\ I_y \left(\nabla I \cdot (u,v) + I_t\right) - \alpha \triangle v + \beta \triangle(\triangle v) = 0 \end{cases},$$

with natural boundary conditions

$$\left.\frac{\partial u}{\partial \vec{n}}\right|_{\partial\Omega} = 0, \quad \left.\frac{\partial v}{\partial \vec{n}}\right|_{\partial\Omega} = 0$$

and

$$\left.\frac{\partial(\triangle u)}{\partial \vec{n}}\right|_{\partial\Omega} = 0, \quad \left.\frac{\partial(\triangle v)}{\partial \vec{n}}\right|_{\partial\Omega} = 0.$$

In our approach, we compute the optical flow for the distance maps of the two images being compared (see Fig. 3, upper row). The reasoning of this step is explained in subsection IV-B. We calculate the partial derivatives I_x, I_y on the mean of the images' distance maps, and derive the time derivative I_t from its usual definition - also using the distance maps as a basis.

Figure 3. Gothic "M" character distortion correction. Top row (from left to right): "M" super-symbol, original "M" character and "M" after correction; bottom: difference maps before (on the left) and after correction

Figure 3 illustrates the computed optical flow application to a character image (top row, in the middle), as compared to the super-symbol image (top row, on the left). The resulting modification of the character image (top row, on the right) more closely resembles the shape of super-symbol image than the shape of original character. However, the shape of the modified character is usually not identical to the shape of super-symbol. (See the difference map on the right in the bottom row.)

We introduced a significant modification to the traditional optical flow approach: in addition to modified regularization term (second order was added), we originally calculated the optical flow between two images basing on their distance maps.

B. Distance Maps Usage Motivation

Kluzner et. al. [9] showed that optical flow, applied to grey-level images, can be successfully used as an estimate of the difference between two words. However, we found that, for small templates (characters), this estimate is not sufficiently robust. The algorithm suffers from background abnormalities and irregularity of spatial gradients in the body of the character. In order to overcome the above problems for optical flow computation, we substituted the character grey-level images with their distance maps (see [13]). Given the binarized character image, its contour P was found, and then the distance map for every pixel q was calculated by

the formulae

$$T_P(q) = \inf_{p \in P} \|p - q\|_{\mathbb{L}^2}.$$

Substituting the image with its distance map changes the gradients map to the constant one (the gradient magnitude in the distance map is equal to 1 for all domain) and overcomes the background influence and object abnormalities. Another meaningful advantage of using the distance map is the correct influence of the background on the optical flow process, because it now has its own non-zero gradients (see Figure 4).

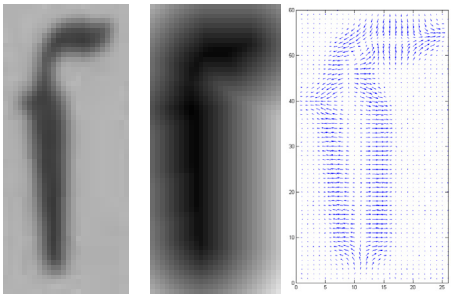

Figure 4. From left to right: the Old Gothic "s" character; its distance map; gradient field of original image; gradient field of distance map

C. Hybrid Classifier

Once the distortion (motion) vector is estimated, we need to translate it into a quantified difference measure. Unfortunately, for many fonts, different characters may be deceptively similar. As a result, the quantification of spatial differences alone may lead to mistaken conclusions. In order to mitigate this problem, we developed a unique hybrid classifier. It combines measurements in both spatial and feature domains: the similarity (computed for transformed character images) and warping, respectively. To estimate the difference between the current character and the super-symbol, we denote

$$diff = \left(\overline{B_1'} \bigcap B_2\right) \bigcup \left(\overline{B_2'} \bigcap B_1\right),$$

where B_1, B_2 are the binary images of the current character and the super-symbol, respectively, $(\cdot)'$ indicates the dilation operator with 3×3 element, and $\overline{(\cdot)}$ indicates a negative image. In this context, only large differences are considered. Then, given $\vec{f} = (u, v)$ as the calculated optical flow between the current character and the super-symbol, the score for a given super-symbol is

$$1 - \frac{diff + \gamma \sum_{B_2} div(\tilde{f})}{Area(B_2)}. \qquad (2)$$

The rationale behind this method is that, for each pair of character images, we wish to attain a high score when the similarity is high (see Figure 3, bottom row) and the divergence of optical flow is low (see Figure 5).

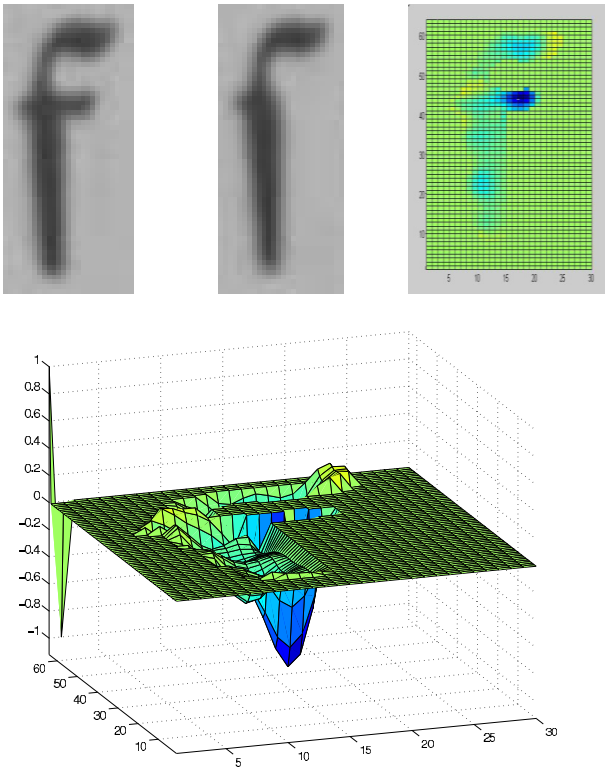

Figure 5. Divergence map between given "f" and "s" characters (top row on the left and in the middle, respectively) as intensity image (top row on the right), and its 3D presentation (bottom row)

The presence of divergence in the optical flow vector field indicates that there are differences in the qualitative features of the images, which goes beyond simple distortion. Figure 5 shows a comparison of the current Gothic font lowercase character "f" with the candidate super-symbol (Gothic font lowercase character "s"), including a qualitative difference - the horizontal line crossing the "f". A two-dimensional intensity graph of optical flow vector field divergence for this pair of images is shown on the right. A three-dimensional version of the above graph is presented at the bottom of Figure 5. The blue spot on the two-dimensional graph (or the local minimum of three-dimensional graph) indicates the divergence values that noticeably deviate from zero. This spot corresponds to the part of the horizontal line belonging to character "f", which differentiates it from the character "s". The sum of divergences over the entire graph (see (2)) indicates the non-trivial differences between these two characters. Thus, assigning a score in accordance with the above scoring formula to the pairing of current character image and candidate super-symbol image yields a relatively low score. A relatively low score indicates a poor match between the images.

Using the above method we can measure the degree of similarity between any pair of character images. This measure, in turn, divides all the characters into the corresponding

equivalence sets. Indeed, each character is associated with the highest score super-symbol (provided that the score exceeds a certain predetermined threshold). All the unassigned characters are compared to each other and used to create new classes. Each new class will yield a new super-symbol. The process is then repeated in such a manner until all the characters are assigned/recognized.

V. Results

To verify the validity of the above approach, we applied it to the benchmark of 101 scanned pages taken from an 18th century German book. We processed this benchmark twice: first, using a leading commercial OCR engine; and second, applying the adaptive OCR process described above. We then compared the OCR results. For simplicity, we performed all the measurements at the word level.

Naturally, any comparison of OCR engines must take into account the number of recognized words and the number of substitution errors. To facilitate the comparison process we need a single measure combining both of these key parameters. Accordingly, we used a Figure of Merit (FOM) defined in [9]:

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

where NOR is the number of rejects, NOF is the number of substitution errors, and NOW is the number of words. FOM serves as an indicator of the level of processing required to correct the data manually. Hence, a lower value of FOM indicates better performance of the recognition engine.

Our data set had 18,321 individual words. The recognition results are summarized in Table I. Note that adaptivity improved both the read rate and the error rate. Overall, the FOM was reduced by about 48% indicating that the optional manual correction effort is reduced by a factor of 2.

Table I
FOM results versus baseline

	Recogn. Rate	**Substitution Rate**	**FOM**
Commercial OCR	88.2%	2.1%	20.3%
Adaptive OCR	91.5%	0.5%	10.5%

VI. Conclusions and Future Work

We presented a new algorithm for book-oriented adaptive OCR that provides a significant enhancement with respect to the conventional (non-adaptive) OCR engines. Our character classification algorithm proved to be effective in recognizing characters that were highly distorted either because of poor quality of the printed material or because of scanning errors.

It should be noted that in our experiments we disregarded the issue of the system time performance. Processing a 101 page benchmark on a state-of-the-art server took approximately one hour. However, we believe that system software optimization would increase the system performance by an order of magnitude or more.

References

[1] A. Abdulkader and M. R. Casey, "Low cost correction of OCR errors using learning in a multi-engine environment," in *Proc. of 10th Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, August 2009, pp. 576–580.

[2] S. Khoubyari and J. J. Hull, "Keyword location in noisy document image," in *Second Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, 1993, pp. 217–231.

[3] A. L. Spitz, "An OCR based on character shape codes and lexical information," in *Proc. of 3d Int. Conf. on Document Analysis and Recognition*, Montreal, Canada, August 1995, pp. 723–728.

[4] Y. Xu and G. Nagy, "Prototype extraction and adaptive OCR," in *IEEE Trans. on Pattern Anal. Mach. Intell.*, ser. 12, vol. 21, 1999, pp. 1280–1296.

[5] S. Marinai, E. Marino, and G. Soda, "Font adaptive word indexing of modern printed documents," in *IEEE Trans. on Pattern Anal. Mach. Intell.*, ser. 8, vol. 28, 2006, pp. 1187–1199.

[6] P. Xiu and H. S. Baird, "Whole-book recognition using mutual-entropy-driven model adaptation," in *Document Recognition and Retrieval XV, Proc of SPIE*, vol. 6815, January 2008, pp. 06–10.

[7] A. Kae and E. Learned-Miller, "Learning on the fly: Font-free approaches to difficult OCR problems," in *Proc. of 10th Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, August 2009, pp. 571–575.

[8] V. Rasagna, A. Kumar, C. V. Jawahar, and R. Manmatha, "Robust recognition of documents by fusing results of word clusters," in *Proc. of 10th Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, August 2009, pp. 566–570.

[9] V. Kluzner, A. Tzadok, Y. Shimony, E. Walach, and A. Antonacopoulos, "Word-based adaptive OCR for historical books," in *Proc. of 10th Int. Conf. on Document Analysis and Recognition*, Barcelona, Spain, August 2009, pp. 501–505.

[10] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

[11] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *Proc. of 8th European Conf. on Computer Vision*, ser. LNCS 3024, T. Pajdla and J. Matas (Eds.), vol. 4, Prague, Czech Republic, May 2004, pp. 25–36.

[12] D. Suter, "Motion estimation and vector splines," in *Proc. of CVPR94*, Seattle, WA, June 1994, pp. 939–942.

[13] P. Danielson, "Euclidean distance mapping," *Computer Graphics and Image Processing*, vol. 14, pp. 227–248, 1980.