The OCRopus open source OCR system

Conference Paper *in* Proceedings of SPIE - The International Society for Optical Engineering · January 2008

DOI: 10.1117/12.783598 · Source: DBLP

1 author:

Thomas Breuel
Google Inc.
260 PUBLICATIONS **6,446** CITATIONS

SEE PROFILE

The OCRopus Open Source OCR System

Thomas M. Breuel
DFKI and U. Kaiserslautern
Kaiserslautern, Germany
tmb@iupr.dfki.de

ABSTRACT

OCRopus is a new, open source OCR system emphasizing modularity, easy extensibility, and reuse, aimed at both the research community and large scale commercial document conversions. This paper describes the current status of the system, its general architecture, as well as the major algorithms currently being used for layout analysis and text line recognition.

1. INTRODUCTION

There has been a resurgence of interest in optical character recognition (OCR) in recent years, driven by a number of factors. Search engines have raised the expectation of universal access to information on-line, and cheap networking and storage have made it technically and economically feasible to scan and store the books, newspapers, journals, and other printed materials of the world.

Commercial OCR engines have traditionally been optimized for desktop use–scanning of letters, memos, and other end-user documents; some other engines have been optimized for special applications like bill scanning. However, OCR engines for large-scale digital library applications differ in their requirements from such traditional OCR systems. In addition, OCR systems traditionally have usually been developed for specific scripts and languages. These issues limit the usefulness of such existing OCR systems for large scale digital library applications.

The goal of the OCRopus OCR system is to overcome these limitations.

- OCRopus is an open source OCR system allowing easy evaluation and reuse of the OCR components by both researchers and companies.
- The particular open source license used by OCRopus, the Apache 2 license, simplifies collaboration between commercial and academic researchers, since contributions can be used commercially with few restrictions.
- The system is designed from the ground up with multi-lingual and multi-script recognition; for example, it uses Unicode throughout and relies on the HTML[1] and CSS[2] standards for the representation of typographic phenomena in a wide variety of languages and scripts.
- The system relies on only a small number of intermediate representations and interfaces, most of them image based,[3] making it easy to integrate both existing and new algorithms.
- The system is extensible and programmable in a built-in scripting language.

The rest of this paper will provide an overview of the methods used in OCRopus, as well as some other information of interest to potential users or contributors to OCRopus. Please note that this paper is not a review of OCR; for example, there are many worthwhile and competitive algorithms for layout analysis and character recognition, but this paper will focus on algorithms that are actually used within the OCRopus system.

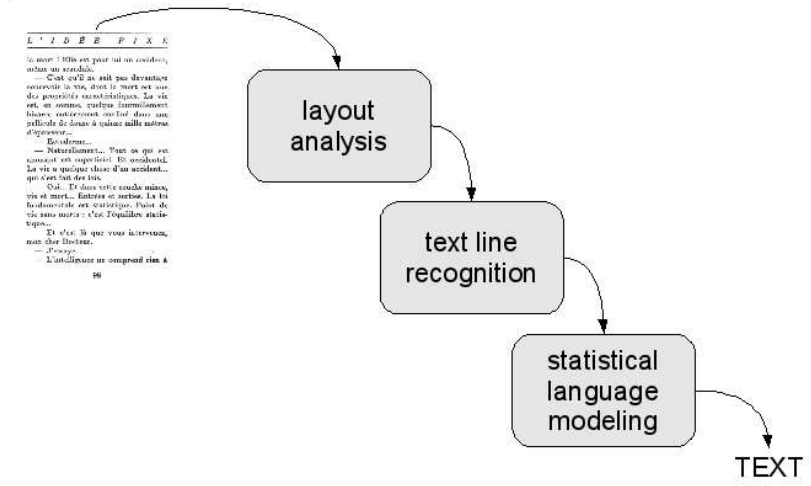

Figure 1. A rough flow diagram of the OCRopus system. The system is strictly feed-forward and consists of three major components: layout analysis, text line recognition, and statistical language modeling. (In addition, it also contains tools for preprocessing and classifier combination.)

2. ARCHITECTURE

The overall architecture of the OCRopus OCR system is a strictly feed-forward architecture (no backtracking) with three major components: (physical) layout analysis, text line recognition, and statistical language modeling; it is similar to a previous handwriting recognition system.[4] The individual steps are (Figure 1):

- Physical layout analysis is responsible for identifying text columns, text blocks, text lines, and reading order.
- Text line recognition is responsible for recognizing the text contained within each line (note that lines can be vertical or right-to-left) and representing possible recognition alternatives as a hypothesis graph.
- Statistical language modeling integrates alternative recognition hypotheses with prior knowledge about language, vocabulary, grammar, and the domain of the document.

Text line recognition itself either relies on black-box text line recognizers, including pre-existing ones, or by recognition using over-segmentation and construction of a hypothesis graph (below). An important aspect of the OCRopus system is that we attempt to approximate well-defined statistical measures at each processing step. For example, the default layout analysis component is based on maximum likelihood geometric matching under a robust Gaussian error model. And the MLP-based character recognition, followed by statistical language modeling, approximates posterior probabilities and segmentation probabilities based on training data and then attempts to find the Bayes-optimal interpretation of the input image as a string, given prior knowledge encoded in statistical language models.

3. PROCESSING STEPS

Above, we saw generally how the processing steps of the OCRopus system fit together. Let us now look at each of the processing steps in more detail.

$$
\begin{aligned}
& P(W, S|x) \\
&= \frac{P(x|W,S)\,P(W,S)}{P(x)} \\
&\approx P(W) \prod_i \frac{P(x_i|w_i,s_i)\,P(s_i|w_i)}{P(x_i)} \\
&= P(W) \prod_i \frac{P(w_i,s_i|x_i)}{P(w_i)}
\end{aligned}
$$

Figure 2. The Bayesian foundations of the OCRopus OCR system. The original approach of combining a discriminative classifier that estimates posterior probabilities with a segmenter was developed for speech recognition.[5] It has been adapted to handwriting recognition[4] and OCR.

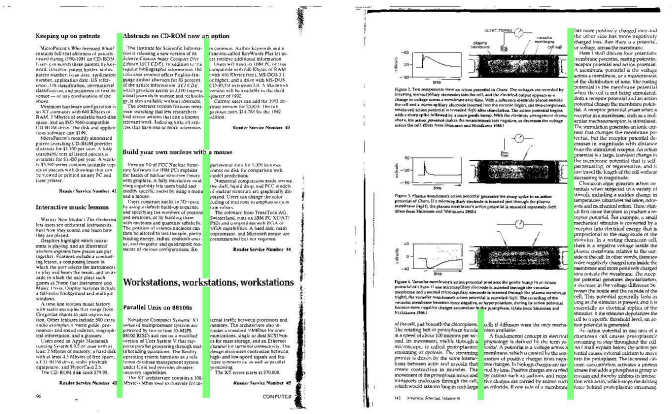

Figure 3. Example of column finding. The whitespace between the columns is identified as maximum area whitespace rectangles with a high aspect ratio and large numbers of adjacent, character-sized connected components.

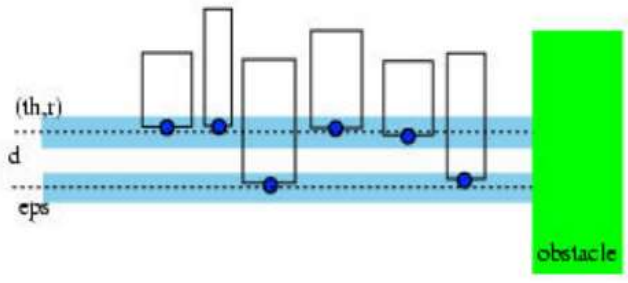

Figure 4. The currently best performing text line extractor in OCRopus searches for globally optimal matches to a precise text line model under a parameterized error model.

Algorithm	Error rates [%]		
	Train	Test	
	Mean	Mean	Stdev
X-Y cut	14.7	17.1	24.4
Smearing	13.4	14.2	23.0
Whitespace	9.1	9.8	18.3
Text-line	5.6	7.0	13.3
Docstrum	4.3	6.0	15.2
Voronoi	4.7	5.5	12.3
Whitespace-cuts	**1.7**	**4.4**	**11.1**

Figure 5. Error rates of the current layout analysis system when applied to the documents in the UW3 database. Error rates are compared to the performance of other, standard methods. (see[11] for more information about experimental conditions.)

3.1 Preprocessing and Cleanup

Preprocessing (binarization, image cleanup, skew correction, etc.) is an important part of OCR systems; good preprocessing can greatly improve overall OCR error rates. However, preprocessing is often quite dependent on the application domain and the image capture methods used. In the long term, we would like to automate image preprocessing and cleanup as much as possible (e.g., using a generate-and-test method similar to the one described in[4]). However, in the short term, we provide a scriptable toolbox that permits users to rapidly construct preprocessing and image cleanup pipelines for their specific needs.

OCRopus provides a standard toolbox of binary and grayscale image processing and mathematical morphology routines, all easily invocable from the scripting interface. In addition, OCRopus provides four tools that are less commonly found in other systems:

- A run-length based binary morphology toolbox permitting fast operations using large masks and using non-rectangular masks[6]
- An efficient local adaptive thresholding algorithm based on integral images[7]
- High-accuracy RAST-based skew detection and correction[8]
- Page frame detection,[9] permitting noise in non-content areas to be cropped away and removed

3.2 Layout Analysis

The overall goal of layout analysis is to take the raw input image and divide it into non-text regions and "text lines"–subimages of the original page image that each contain a linear arrangement of symbols in the target language. Text lines need not be horizontal, left-to-right; the system imposes no constraints on the shape or direction of the text lines generated by the layout analysis. However, layout analysis modules must indicate the correct reading order for the collection of text lines, and the text line recognizer needs to be able to cope with the direction and nature of the text lines returned by page layout analysis.

3.2.1 Text-Image Segmentation

OCRopus contains a simple text-image segmentation system.[10] It operates by first dividing the input image into candidate regions. Then, features are extracted for each candidate region. Finally, each region is classified using logistic regression into text, grayscale image, line drawing, ruling, and other kinds of regions.

3.2.2 RAST-Based Layout Analysis

The primary layout analysis method used by OCRopus is currently based on two related algorithms, one for whitespace identification, and the other for constrained text line finding.[12] Both methods operate on bounding boxes computed for the connected components of the scanned input page image.

Figure 6. A dynamic programming algorithm generates character segmentation hypotheses[19] (here illustrated for the case of handprinted letters); the algorithm does not require characters to be segmentable using a straight line and hence works in the presence of kerning and italics.

Column Finding In the first step, the column finder uses a maximal whitespace rectangle algorithm[12] to find vertical whitespace rectangles with a high aspect ratio. Among those, the system selects those rectangles that are adjacent to character-sized components on the left and the right side. These whitespace rectangles represent column boundaries with very high probability. In a post-processing step, we eliminate geometrically implausible column boundaries. Columns separators found in this way are shown in Figure 3.

Text Line Modeling Text-line finding matches a geometrically precise text line model (Figure 4) against the bounding boxes of the source document; each text line is constrained not to cross any column boundary.[8] Search is currently carried out by a memory-intensive best-first branch-and-bound method; to conserve memory, it will be altered to perform a depth-first search. Each text line is found independent of every other text line, so the orientations of individual text lines can vary across the page (it is, however, possible to constrain text lines to all share the same orientation). The method is specific to Latin script, but the approach generalizes to other scripts.

Reading Order Determination The output of column finding and constrained text line matching is a collection of text line segments; it remains putting them in reading order. Reading order is determined by considering pairs of text lines. For certain pairs of text lines, reading order can be determined unambiguously. These pairwise reading order constraints are then extended to a total order by topological sorting.

Performance The RAST-based layout analysis summarized above computes exact maximum likelihood solutions to the whitespace and constrained text line matching problems under a robust Gaussian error model; of course, although this model is plausible, it represents a simplification of actual page statistics. When applied to standard ground truthed databases of document images with Manhattan layouts, it yields high performance compared to other standard methods[11–18] (Figure 5).

3.2.3 Other Approaches to Layout Analysis

Although the RAST-based approach to layout analysis described above has turned out to be a reliable workhorse, it has a number of limitations. Most importantly, when it fails, it can only be adapted by modifying a small number of parameters. It would be desirable to have a system that is trainable or adaptable to specific layouts. It is also only applicable to Manhattan layouts. Finally, it only represents one of many desirable cost/performance tradeoffs for layout analysis.

To address these issues, we will likely incorporate the implementations of other layout analysis methods (Voronoi-based, XY-cuts, etc.) into the system. In addition, we are currently developing trainable layout analysis methods that permit explicit representation of a wide variety of layout structures and their geometric variation.

3.3 Text Line Recognition

Layout analysis transforms the problem of OCR from a 2D problem into a 1D problem by reducing the page to a collection of "text lines", spatially linear arrangements of characters. Each of these text lines is passed to a text line recognizer for the input document's language. There are already two text line recognizers incorporated into OCRopus, and more will be added in the future.

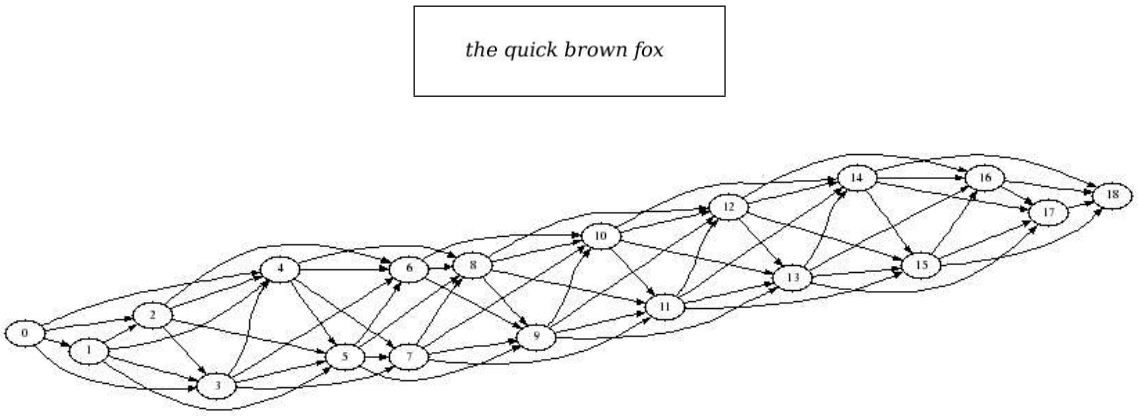

Figure 7. Example of oversegmentation of an input string and representation of the oversegmentation as a finite state transducer (equivalent to a hypothesis graph).

Figure 8. Features used by the MLP-based recognizer include gradients, singular points of the skeleton, the presence of holes, and unary-coded geometric information, such as location relative to the baseline and original aspect ratio and skew prior to skew correction.

3.3.1 Tesseract

The Tesseract[20,21] text line recognizer is based on a mature OCR system developed at Hewlett and Packard (HP) and open sourced recently. Internally, it uses a two-stage shape comparison between prototype character shapes and characters in the input image. Tesseract integrates segmentation of the input image into individual characters with their recognition, using backtracking in case a subsequent state determines that the segmentation is geometrically or linguistically implausible.

Tesseract's character recognition error rates still do not quite achieve the same performance as current commercial systems, but are getting closer, as Tesseract is still under active development. Tesseract is now capable of recognizing many variants of the Latin script, and will likely soon be able to handle other alphabetic languages.

Tesseract does not attempt to estimate character likelihoods, posterior probabilities, or probabilistic segmentation costs, but does return "match scores". Tesseract also does not compute a complete segmentation graph for the input. Both of these factors limit the ability to use Tesseract with the statistical natural language models used by OCRopus.

3.3.2 MLP-Based Recognition

A second text line recognizer integrated into OCRopus uses multi-layer perceptrons (MLPs) for character recognition. It proceeds in several steps, first attempting oversegmentation of the input string, then recognizing each potential character hypothesis, and finally expressing both the recognition results and the geometric relationships between character hypotheses as a graph structure.

Oversegmentation is achieved using a dynamic programming algorithm[?,?] (Figure 6), which also permits kerned and italic character pairs to be segmented. The dynamic programming algorithm identifies potential cut

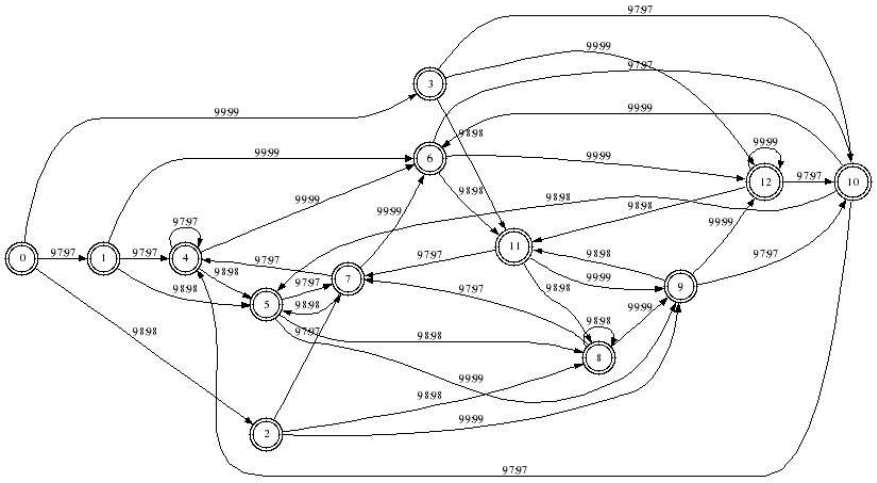

Figure 9. An two letter bigram language model represented as a probabilistic finite state transducer.

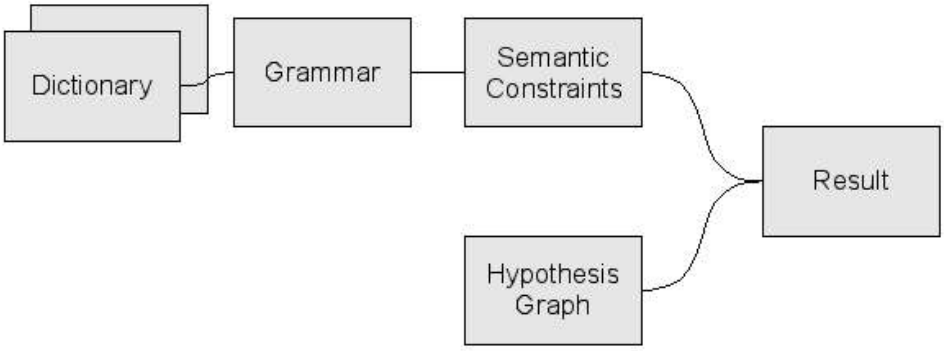

Figure 10. Language models based on finite state transducers can be composed modularly from dictionaries, n-grams, grammatical patterns, and semantic patterns. This allows OCRopus to be retargeted and adapted quickly to new document types and languages.

paths between characters. Pairs of nearby cut paths are then considered to be the left and right boundaries of character hypotheses; some of these character hypotheses will correspond to actual characters, while others represent mis-segmented characters. The adjacency relationships between character hypotheses are encoded in a hypothesis graph (Figure 7). For each character subimage, the corresponding image features are computed, determining both the probability that it represents a valid character, and, assuming that it is a valid character, the posterior probability for each class. Features used by the system currently includes gradients, singular points of the skeleton, the presence of holes, and unary-coded geometric information, such as location relative to the baseline and original aspect ratio and skew prior to skew correction.

This combination of oversegmentation, hypothesis graph construction, and later best path identification using dynamic programming has become fairly common in handwriting and OCR systems (although some systems prefer to use backtracking methods or other control structures instead). In particular for text line recognition, the system performs either *maximum a-posterior* (MAP) recognition, or recognition based on full Bayesian posterior probability estimates. The statistical foundations of the system were formulated for speech recognition in[5] and applied to handwriting recognition in[4] (Figure 2).

The MLP currently does not perform as well as Tesseract, either in terms of error rates or speed; however, the network has not been trained extensively, and there is considerable potential for performance improvements.

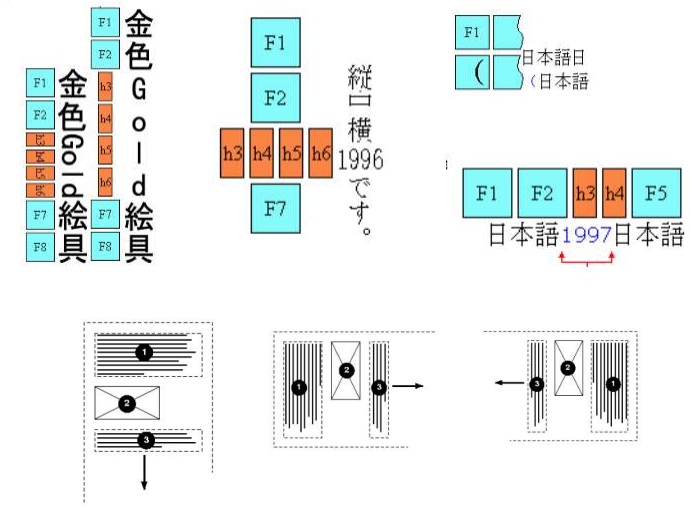

Figure 11. Examples of typographic phenomena that have standard representations in HTML and XHTML–and hence hOCR–but that are not well handled by many other OCR output formats. (Examples taken from W3C.[2])

3.3.3 Other Text Line Recognizers

We have also developed a number of other recognizers that we will incorporate into future versions of OCRopus. These include an HMM-based recognizer, suitable for recognizing small font sizes, and an alternative shape-based recognizer, based on prior work on shape-based character recognition.[22]

3.4 Language Modeling

The third major component of OCRopus is statistical language modeling. Examples of statistical language models are dictionaries, character-level and word-level n-grams, and stochastic grammars. Statistical language models associate probabilities with strings; their function in an OCR system is to resolve ambiguous or missing characters to their most likely interpretation.

Since its alpha release, OCRopus includes the open source OpenFST library[23,24] as the basis of its statistical language modeling tools. OpenFST represents statistical language models as `weighted finite state transducers`. Weighted finite state transducers are a generalization of both hidden Markov models and finite state transducers; they can also be thought of as a form of "weighted regular expression" with the ability to perform limited substitutions. By representing language models as finite state transducers, OCRopus separates language modeling from recognition; that is, complex language models can be prepared off-line using a rich set of language modeling tools, compiled into a weighted finite state transducer representation, and then used as a language model within OCRopus, without OCRopus having to contain any explicit code for complex language modeling tasks.

Weighted finite state transducers can also be used for a variety of other important tasks in OCR:

- expression of character set transformations
- representation of ligature rules and probabilities
- robust information extraction on noisy OCR output
- robust information retrieval on noisy OCR output
- direct statistical machine translation on noisy OCR output