

Evaluation of a User-Assisted Archive Construction System for Online Natural History Archives

J. He and A. C. Downton

Department of Electronic Systems Engineering, University of Essex, UK. Email: {jhe, acd}@essex.ac.uk

Abstract

The creation of structured digital libraries from paper-based archives is an area of growing demand in many scientific and cultural fields, and is not satisfied either by off-the-shelf OCR or commercial form-processing systems. This paper describes and evaluates a configurable archive construction system, which integrates document image pre-processing and analysis with text post-processing tools and a standard OCR package. The prototype system is currently being used in conjunction with the UK Natural History Museum to help convert more than 500,000 cards of Lepidoptera and Coleoptera to a searchable digital archive. Evaluation results are summarised for two datasets comprising over 5,000 cards selected from different parts of this database, and indicate that overall end-to-end word recognition rates of 70-90% are readily achievable for key data fields, subject to availability of suitable electronic dictionaries.

1 Introduction

Digital archive construction from historic paper archives is a major image analysis application of interest both for cultural and scientific purposes. Archive documents are often stored in well-structured taxonomies (e.g. libraries, scientific specimen indexes and censuses) where the structure extends across the index as well as within the layout of each document. Documents are recorded using text which challenges off-the-shelf OCR, not only because of poor quality and/or decayed typescript or handwriting, but also because standard commercial OCR systems cannot infer the data structure inherent within the records without human guidance. Similarly, although some commercial form-processing OCR systems exist, these normally process fixed-format pre-defined forms designed for OCR using background drop-out colours and/or tabular guidelines to maximise performance, rather than arbitrary pre-existing document archives.

To address archive applications, we have designed a user-configurable archive document processing system which integrates image analysis and text post-processing tools with a configurable commercial OCR package, to generate text content that can be fed direct into a target online database. The pattern recognition aspects of the system (ranging from colour segmentation, to document structure classification, to stamp identification and removal) are uniformly implemented using a fuzzy classification scheme which is parameterised within the user interface.

Our system has been developed in conjunction with the UK Natural History Museum (NHM), and hence has largely been evaluated on their archive card indexes, which contain bibliographical data and other information for one scientific name on each card, laid out in a standardised format (Fig. 1) for each archive [1]. This paper describes each part of the overall archive document processing system and then evaluates its end-to-end system performance.

2. System Overview

The overall system (Fig. 2) consists of four main components, pre-processing, document analysis, OCR (using a commercial OCR engine) and post-processing. Pre-processing reads the original JPEG document images, and converts them from either colour or grey level into binary for semantic labelling purposes. Document analysis then segments and semantically labels important text fields. The output normally consists of labeled text field colour or grey-level sub-images in the system's internal image format (PNM). However, pre-processing can also be recalled after Document Analysis so that the labelled text field sub-images can be converted into binary under system control rather than relying on the OCR's internal binarization algorithm. The OCR system recognises labelled images text-field by text-field and converts them into raw text, which is further processed by regular expression post-processing to meet final database input

requirements. The components are integrated into a complete archive batch processing system with the user interface shown in Fig. 3.

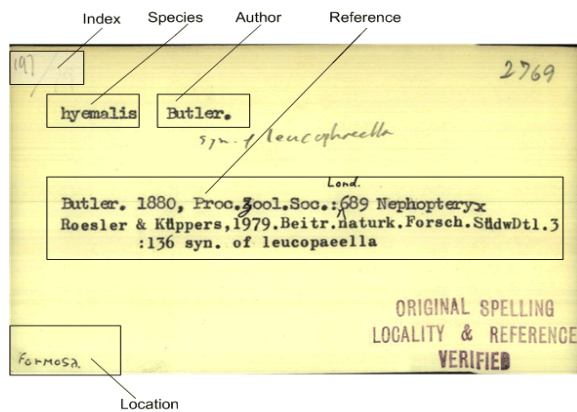


Figure 1. An index card with multiple hand print and handwriting annotations showing components to be extracted.

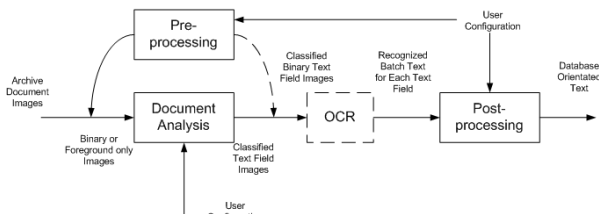


Figure 2. Overall System Diagram

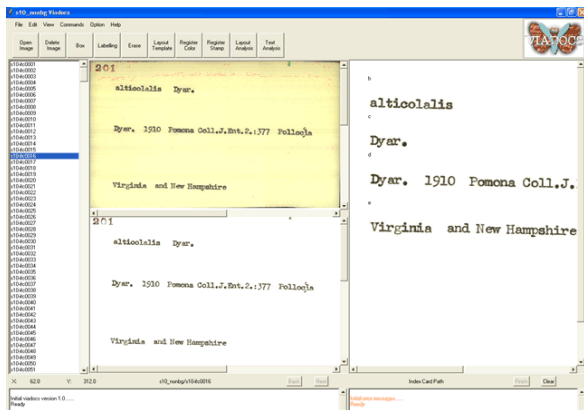


Figure 3. Archive Document Analysis System

3. Document Image Pre-Processing

Pre-processing provides a number of tools, including several alternative binarization and color segmentation algorithms, any of which can be applied to input images. Five alternative algorithms are available for binarization from grey-level images: global thresholding, Niblack's algorithm [2], adaptive Niblack, Sauvola's algorithm [3]

and adaptive Sauvola. A comparative performance evaluation of all these algorithms, and also the internal binarization algorithm used by the OCR system, established that the best OCR performance for the NHM archive dataset was achieved by our adaptive Niblack algorithm [4]. All system evaluation results reported in this paper are therefore based on using this segmentation method.

4. Document Image Analysis and OCR

The format of archive index cards consists of several independent blocks of text, and each block contains one or more logically related text fields. Blocks retain a fairly consistent mutual layout over a complete archive, but the layout of text fields within each block is not strictly fixed. Nor are there any tabular guidelines defining fixed block boundaries. The X-Y cuts algorithm [5] is therefore an appropriate segmentation algorithm for this class of document image structure. Pixel smearing [6], with a threshold sufficient to join adjacent text characters but not adjacent horizontal words or vertical lines, is first applied as a pre-processing non-linear low-pass filter to each archive card image. The X-Y cuts algorithm then extracts and stores the contents of each index card into a hierarchical tree structure (the so-called X-Y tree), consisting of text blocks, lines and words.

In addition to segmentation, DIA labels each segmented region of each card (as shown in Fig. 1), in this case based upon the best-match template layout pre-registered during system configuration for the batch of cards being processed. Labelled image fields allow the OCR system to be configured with field-specific dictionaries, and raw text output from the OCR to be fed to the correct database field.

The OCR used in the proposed system is a commercial product, Abbyy FineReader 6.0. Since it is designed for stand-alone use, it includes its own internal image processing (e.g. binarization) integrated with OCR, but can also accept pre-processed images in a variety of different image formats including binary, allowing us to substitute alternative binarization algorithms. We used it as middleware working in combination with the other components of our system. For example, Fig. 5 shows texts on the card that have been extracted and labelled into 5 classes of images: Index, Species, Author, Reference and Location. To recognise the class Author, a specific name dictionary (provided by NHM) is added to the default OCR English dictionary. A different field-specific dictionary is used for each semantic field. The OCR output of this processing is the raw text which is saved into a separate text file for each text image.

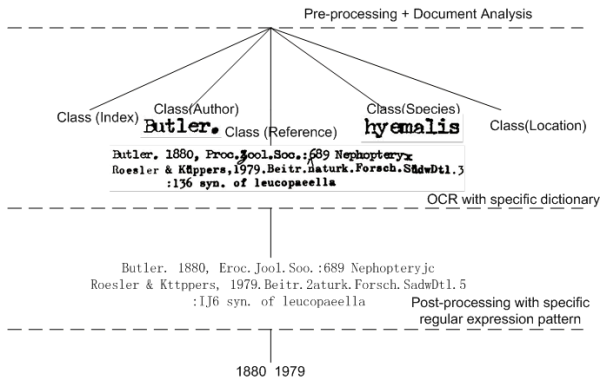
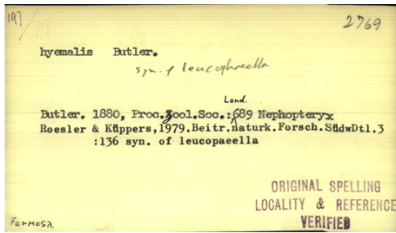


Figure 5. Example of Card Processing

5. Post-processing

The purpose of text post-processing is to generate database-oriented text strings for input to the online database. The texts obtained from OCR are regarded as “raw” in comparison with the requirements. For example, the NHM online database only needs the published year in the recognized reference to be stored as a search key; other contents can be ignored. Another example is the author name, where the database requires complete author names, but abbreviated author names (terminated with a full stop) are frequently found in the original images, e.g. the author “Warren” may be abbreviated to “Warr.”. The corresponding complete author names need to be retrieved and substituted for each abbreviation in the online database.

Tcl regular expressions specified within another part of the system interface (Fig. 6) are the main tool to manipulate the raw OCR text output. For example the regular expression to parse the published year from a reference is expressed by:

regexp {[1][7-9][0-9][0-9]} \$reference year

where *regexp* is the regular expression command, *{[1][7-9][0-9][0-9]}* is the parsing pattern (which searches for a year between 1700 and 1999), *\$reference* represents the reference raw text generated by OCR, and *year* contains the 4 matched digits output to the ‘year’ database field. Similarly, for author abbreviations, a regular expression is applied to the author name raw OCR output field to find any pattern terminated by a full stop:

regexp {[^\.] +} \$author pattern

Based on the detected *pattern*, another regular expression is used to search the specific Author dictionary to find the matching full Author name.

Finally, the post-processing combines the batch of separate text files generated by the OCR process, augmented with the results of regular expression matching, into a single file suitable for populating the online database.

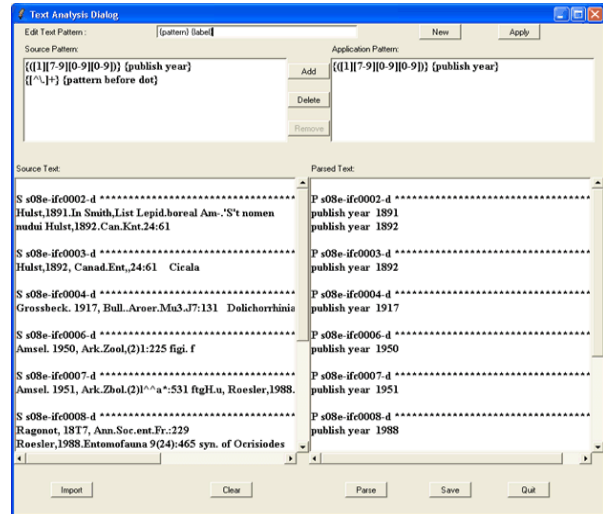


Figure 6 User Interface for Post-processing

6. System Evaluation

6.1 Evaluation Datasets

The system was evaluated on two sets of sample cards. One set of 4435 cards was randomly chosen from the Pyraloidea dataset of 27,578 archive cards, for which full truth data was independently available. A second set of 994 cards was processed from the Curculionidae subset of the Coleoptera archive, as part of the overall system trials. The Curculionidae testset used different card layouts and dictionaries from the Pyraloidea testset and therefore provided an independent dataset for validating system performance, and also for estimating whether sufficient user (re)configurability had been allowed in the system design.

6.2 Overall Evaluation Method

For both datasets, the text fields extracted from each archive card for evaluation were: genus/species name, author name and the date sub-field within the reference, since these fields are currently indexed in the museum’s online archive [1]. Electronic dictionaries (not always complete) are also available for genus names, species

names and author names. Evaluation was carried out using pre-processing and document analysis to generate three sets of binary text sub-images (genus/species name, author name and reference), which were binarized using our adaptive Niblack algorithm. The text sub-images were then fed into the Abbyy OCR for recognition class-by-class using respective class dictionaries, and the results were saved into three sets of text files.

The three sets of text files were parsed and merged into three single text files by post-processing with respective regular expressions for database input. The results produced by the system were then compared with the word-level truth data for the corresponding database fields. In the word-level evaluation, if any unmatched character was found, the whole text field (which could contain one or more words) was considered incorrect as shown in Fig. 7, where the german ü was unmatched, because it is not included in the OCR character set.

The same evaluation methodology was used for both evaluation datasets, except that only a single archive card template was required for the first dataset (Pyraloidea cards), whereas three different templates were required for the second dataset (Curculionidae cards), because three different card layouts were detected within this dataset (Fig. 9).

6.3 Results for First Evaluation Dataset

Analysis of errors in this dataset (Table 1) shows that 15% of overall errors occurred when document image analysis wrongly extracted or labelled text fields, and 13% resulted from truthing errors (e.g. see Fig. 8) including abbreviations. The remaining 72% of errors were generated by the OCR system, often caused by touching typewritten characters. 16.4% of errors were subsequently corrected by text post-processing.

Since author recognition was carried out with an incomplete Author dictionary, the word recognition rate for this field is lower than for Species/Genus, where a full dictionary was available. The poorer result for Year was mainly caused by the OCR, which was less accurate in recognising digits than characters (nearly 89% of total errors for Year were caused by OCR errors compared with the average of 72%). Another cause of poor performance is that quite a few years are handwritten.

6.4 Results for Second Evaluation Dataset

The other set of 994 cards were randomly chosen from the Curculionidae dataset of 10,000 archive cards. Three different layout formats were encountered as shown in Fig. 9. In this evaluation, three templates corresponding to these formats were registered for Document Analysis.



Figure 7. Example of Text Field Recognition

Table 1. Error Analysis for Pyraloidea Dataset

Text Field	Species/ Genus	Author	Year
Text fields	4435	4435	4435
Doc. Analysis	149/4435 -3.4%	166/4435 -3.7%	140/4435 -3.1%
OCR	460/4435 -10.4%	711/4435 -16.0%	1080/4435 -24.4%
Truthing Data	50/4435 -1.1%	365/4435 -8.2%	0
Post- processing	165/4435 +3.7%	346/4435 +7.8%	0
Correct Text.fields	3941/4435 88.9%	3539/4435 79.8%	3215/4435 72.5%

Original Image Recognized Text Truth Data

leucopilalis leucopilalis leucospilalis

Figure 8. Difference Between Original Image and Truth Data

Table 3 shows that 99% of cards' formats were identified correctly, and hence subsequently analysed with the correct template. Type (a) was by far the most common template, occurring in more than 88% of the sampled cards. Therefore, we carried out the same evaluation as in Section 6.3 just on the 881 cards of type (a).

The text fields extracted from each archive card for evaluation in the second evaluation dataset were genus name, species name and author name. On the image (Fig. 9(a)), the top block is Genus and the second (reference) block contains both Species and Author data. Most of the time, Species is the first word in the block. Author, in most cases, is located in the middle of the block, and terminated with a comma. Its initial letter is always capitalized. Suitable regular expressions are used to search for these fields embedded within the 'raw' OCR output for the reference sub-image.

Table 4 summarises the evaluation results. Analysis of errors in the evaluation dataset shows that 8.1% of overall errors occurred when document image analysis wrongly extracted or labelled text fields, and the remaining 91.9% of errors were generated by the OCR system, often caused by touching typewritten characters and complex

