Greek Alphabet Recognition Technique for Biomedical Documents

Daniel X. Le, Scott R. Straughan and George R. Thoma
National Library of Medicine
8600 Rockville Pike, Bethesda, MD 20894

ABSTRACT

Most current commercial optical character recognition (OCR) systems can accurately recognize the text in documents written in a single language. However, when dealing with Greek characters embedded in predominantly English text, these systems do not perform well, and most OCR systems do not recognize the characters as belonging to the Greek alphabet. As a result, the degree of manual review required to validate and correct OCR errors is high. To handle this problem, we propose a new technique based on features calculated from the output of multiple OCR systems, and combined with string pattern matching and document content analysis to improve the recognition of both Greek characters and regular text. Our proposed technique uses two passes of a document page image through OCR systems that use different recognition languages.

Experiments carried out on a sample of medical journals show the feasibility of using the proposed technique for Greek character recognition. Preliminary evaluation conducted on a sample of medical journal page images shows that our approach improves the recognition of Greek characters embedded within predominantly English language text.

Keywords: Greek character recognition, Optical character recognition, Automated document data entry, MEDLINE® database, National Library of Medicine.

1. INTRODUCTION AND BACKGROUND

The conversion of paper-based document information to electronic format is important for automated document delivery, journal distribution, document preservation, and other document related applications. In biomedical documents, especially documents in molecular biology, biochemistry, pharmacology, drug development, chemical analysis, etc., we encounter Greek characters embedded in predominantly English text. When bibliographic databases such as MEDLINE® are created, Greek characters that occur in journal articles need to be converted to English words to support searches by keyword. For example, Greek characters “a” and “ß” in the articles are replaced by “alpha” and “beta” in the database respectively. Therefore, to support automated document searching and electronic publishing (converting papers from one format to another, or modifying manuals and references, etc.), techniques are required to recognize these Greek characters.

The Lister Hill National Center for Biomedical Communications, a research and development division of the National Library of Medicine (NLM), is developing a system, the Medical Article Records System (MARS), to automatically extract bibliographic information from printed biomedical journals for inclusion in the MEDLINE® database used by biomedical professionals worldwide. This paper describes research toward one aspect of the recognition problem in MARS: the detection of Greek characters from scanned document images. This recognition technique is based on features calculated from the output of several OCR systems, string pattern matching, and a set of rules that is derived from an analysis of document content information.

Earlier approaches to detect embedded Greek characters within English language text [1, 2, 3] have been reported. As an example, Liang et al. [1] use a technique based on image processing and probability map in a binary image to refine the character segmentation and detect special symbols including Greek characters. While this technique was reported to perform well on the character segmentation, the misdetection and the false alarm rates are quite high. Other techniques [2, 3] only handle the segmentation and extraction of Greek characters from printed mathematical formulas, and

not from sources consisting of mixed English and Greek text.

Current commercial OCR systems can accurately recognize the text in documents written in a single language. However, when dealing with Greek characters embedded in predominantly English text, these systems do not perform well, and most OCR systems do not recognize the characters as belonging to the Greek alphabet. As a result, a considerable degree of manual review is required to validate and correct the OCR errors.

Our proposed technique uses two passes of a document page image through OCR systems that use different recognition languages. Furthermore, document contextual information is analyzed to improve the recognition results. The low-confidence Greek character and its associated low-confidence word (word that contains the low-confidence character) recognized from previous sentences and/or paragraphs are analyzed, and their features (contents, attributes, and frequencies of occurrence) are recorded for use in recognizing future Greek characters. Preliminary evaluation results show that the system is capable of improving the recognition of Greek characters embedded within predominantly English language text.

The rest of this paper is organized as follows. Section 2 provides a system overview. Section 3 presents system features including low-confidence characters and words. Section 4 describes the Greek alphabet recognition process. Section 5 gives the experimental results. Section 6 contains a summary.

2. SYSTEM OVERVIEW

The recognition technique described in this paper is one component of our second-generation MARS system developed at NLM [4]. The recognition process takes a scanned binary image as input and performs the first OCR pass using English as the recognition language for each text zone. From this OCR output, all recognized low-confidence characters and their associated words are identified and used as inputs for the second OCR pass using English and/or Greek as a recognition language. Finally, string pattern matching and document content analysis are applied between these low-confidence words and characters and similar words and characters obtained from previous steps to improve the recognition of both Greek characters and regular text

In the first pass, we use a commercial 5-engine OCR system developed by Prime Recognition [5], to segment scanned binary document images into rectangular text zones. Each zone is then processed to deliver an OCR output (including zone coordinates, text line information, characters and their bounding boxes, confidence levels, font sizes, and certain style attributes). From this output, all low-confidence characters and their associated words are identified and extracted for the Greek alphabet recognition operation.

Based on preliminary experiments on a small set of journal page images during the second OCR pass, we discovered two interesting results that may affect the outcome of the OCR process.

1. The OCR output may differ depending on how a character is submitted to the OCR engine: either as a stand-alone character or as an embedded character within a word. For example, the Greek character "α" in the word "α-subunit" can be recognized as "a" if it is a stand-alone character, or as "α" if the entire word is submitted.
2. The OCR output for a character or a word using English and Greek as a combined recognition language may be different from that using only Greek as a recognition language. For example, the word "factor-α" can be recognized as "factor-α" for the first language setting, or "fαcτor-α" for the second language setting.

In order to collect all possible OCR outputs of low-confidence characters and their associated words, we will incorporate these two results into the second OCR pass as follows:

1. Separately submitting both the low-confidence characters and the low-confidence words of which they are part, and
2. Recognizing each word or character using both English and Greek as the combined recognition language, and Greek only as a recognition language.

The final OCR result for a low-confidence character will be obtained from the multiple OCR outputs, using string pattern matching and a set of rules derived from document-specific information to analyze and compare among these outputs.

Two OCR engines that can recognize Greek characters were chosen for the second OCR operation: Recognita™ from ScanSoft® [6] and FineReader™ from ABBYY® [7]. The outputs of these two engines are analyzed to get the final result.

3. LOW-CONFIDENCE CHARACTERS AND WORDS

Features of low-confidence characters and their associated low-confidence words calculated for this recognition technique are based on the character confidence analysis of the output of the commercial 5-engine OCR system.

For each character, the output of the Prime OCR system [5] includes the following: the 8-bit code for the recognized character, confidence level (1= lowest, 9 = highest), bounding box, font size and font attributes. In this paper, a "low-confidence" character is one whose confidence is less than the highest confidence level of 9. So any character having a confidence level of 8 or less is considered a low-confidence character.

The coordinates of a low-confidence word are derived from its character coordinates as follows:

Left coordinate:	the left coordinate of the first character in a word
Top coordinate:	the smallest top coordinate of the characters in a word
Right coordinate:	the right coordinate of the last character in a word
Bottom coordinate:	the largest bottom coordinate of the characters in a word

A list of 6 features for each low-confidence character and its associated low-confidence word at "character level" and "word level" used in the Greek alphabet recognition process is:

Character Level

Recognized 8-bit character
Confidence level
Character coordinates (Left, Top, Right, Bottom)

Word Level

Total number of characters
Recognized 8-bit characters
Word coordinates (Left, Top, Right, Bottom)

Usually, OCR systems do not perform well on small image areas and many of these systems require a certain minimum font size for better results. For example, the OmniPage Pro™ OCR software from ScanSoft® [6] can reliably recognize characters with font sizes exceeding 5 points. Since the character bounding box represents the left-most/top-most/right-most/bottom-most pixels of a character, characters without ascenders and/or descenders such as **a, o, u** sometimes may be smaller than the minimum size required for processing. In order to deal with this problem, the top and bottom of low-confidence characters bounding boxes are modified to satisfy the font size requirement. The modification of a character bounding box is done in such a way that the modified bounding box does not overlap any neighboring bounding boxes. For example, consider a low-confidence character in a word; its top/bottom coordinates are expanded up/down to equal the top/bottom coordinates of the word.

4. GREEK ALPHABET RECOGNITION PROCESS

The Greek alphabet recognition technique described here consists of six steps: (1) scan journal pages, (2) perform the first OCR pass using English as the recognition language, (3) identify all low-confidence characters and their associated low-confidence words, and calculate their features, (4) perform the second OCR pass on these low-confidence characters and words using both English and Greek as the combined recognition language, and Greek only as a recognition language, (5) apply string pattern matching between these low-confidence words and similar words obtained from previous steps, and (6) finally, apply a set of rules derived from document content analysis to recognize these low-confidence characters and words. In the following subsection, each step is discussed in detail.

4.1 Scan journal images

In this step, the first page of each article of a journal issue is scanned and saved as a binary document image. Image processing operations such as page orientation and skew detection are then applied to improve the quality of a scanned image. The images of pages in landscape mode are automatically rotated

to be in portrait mode, and skewed page images are rotated to correct the skew angle.

4.2 Perform the first OCR pass

Using a commercial 5-engine OCR system, each image is segmented into text and graphics zones. With English as the recognition language, each text zone is processed to deliver an OCR output (including characters, their bounding boxes, and their confidence levels).

4.3 Identify low-confidence words and calculate word features

In this step, using the character confidence levels obtained from the OCR output, all low-confidence characters and their associated low-confidence words are identified, and for each character, the six features as defined in Section 3 are calculated. The character and word coordinates features will be used in the second OCR pass, while the other four features will be employed in the string pattern matching.

4.4 Perform the second OCR pass

For each low-confidence character, using the character and word coordinates features calculated in the above step, the second OCR pass is performed using two Greek-enabled OCR engines: Recognita and FineReader.

1. For the Recognita OCR engine
   a. Character Level
      - Set English and Greek as the combined recognition language
      - Obtain the first Recognita OCR character-based output
      - Set Greek only as a recognition language
      - Obtain the second Recognita OCR character-based output
   b. Word Level
      - Set English and Greek as the combined recognition language
      - Obtain the first Recognita OCR word-based output for the associated low-confidence word
      - Set Greek only as a recognition language
      - Obtain the second Recognita OCR word-based output for the associated low-confidence word
2. For the FineReader OCR engine
   a. Character Level
      - Set English and Greek as the combined recognition language
      - Obtain the first FineReader OCR character-based output
      - Set Greek only as a recognition language
      - Obtain the second FineReader OCR character-based output
   b. Word Level
      - Set English and Greek as the combined recognition language
      - Obtain the first FineReader OCR word-based output for the associated low-confidence word
      - Set Greek only as a recognition language
      - Obtain the second FineReader OCR word-based output for the associated low-confidence word

4.5 Apply string pattern matching

For each associated low-confidence word, the word output from the first OCR pass is matched with that from the second OCR pass using low-confidence characters as alignment guidelines. Based on the matching result, the third and fourth Recognita/FineReader OCR character-based outputs are extracted from the first and second Recognita/FineReader OCR word-based outputs.

4.6 Apply rule-based analysis

In this final step, the rule-based analysis algorithm compares four Recognita OCR character-based outputs with the four FineReader OCR character-based outputs to obtain a final decision on the low-confidence character. Assume that there are Greek characters in these 8 OCR character-based outputs given a predefined weight threshold, if both OCR engines agree then the detected Greek character is assigned to the low-confidence character. Otherwise, a list of possible Greek characters derived from these 8 OCR character-based outputs is used for the low-confidence character. The list shall be ordered according to the character weight. In case there is no Greek character in any of these 8 OCR character-based outputs, the low-confidence character is not considered as a candidate for Greek alphabet detection.

The following algorithm summarizes the rule-based algorithm to detect a Greek alphabet for a low-confidence character using a predefined weight threshold of 50% for each OCR engine.

1. For each of four Recognita OCR character-based outputs
    If it is a Greek character then
        Assign a weight of 25%
    Else
        Discard the output
    End If
   End For
2. Get the Recognita Greek character whose weight is maximum
3. For each of four FineReader OCR character-based outputs
    If it is a Greek character then
        Assign a weight of 25%
    Else
        Discard the output
    End If
   End For
4. Get the FineReader Greek character whose weight is maximum
5. If both the maximum Recognita and FineReader Greek character weights equal to 0 then
    The low-confidence character is not considered as a Greek character
   Else
    If the Recognita Greek character equal to the FineReader Greek character then
        Assign the detected Greek character to the low-confidence character
    Else
        Assign a list of possible Greek characters derived from these 8 OCR outputs to the low-confidence character, where the list is ordered according to character weight.
    End If
   End If

5. EXPERIMENTAL RESULTS

The Greek alphabet recognition technique has been implemented and experiments have been conducted with binary document images selected from several different biomedical journals. All documents used in these experiments are 8.5 x 11 inches in size and were scanned at 300 dpi resolution.

A test sample consisting of 301 page images from 28 different journal types was used in the experiment. Only abstract text zones are extracted for the experiment, and there are 3,800 low-confidence characters detected. Among these low-confidence characters, there is a total of 157 Greek characters.

Using a predefined weight threshold of 50% for each OCR engine, the following is the summarized experimental results:

Above the weight threshold

86	detected by both Recognita and FineReader OCR engines
20	detected by Recognita OCR engine only
16	detected by FineReader OCR engine only
1	detected error

Below the weight threshold

6	detected by both Recognita and FineReader OCR engines:
5	detected by Recognita OCR engine only
7	detected by FineReader OCR engine only

There is only one error for Greek character detection and 86 Greek characters (55%) correctly detected by both Recognita and FineReader OCR engines. However, out of 70 Greek characters that are not detected by both OCR engines with the given weight threshold of 50%, 54 Greek characters (77%) appear in the list from which the correct Greek character may be selected by the user. Therefore, the actual number of Greek characters detected by the proposed Greek alphabet recognition technique is 140, giving a detection rate of about 89%.

6. SUMMARY

A Greek alphabet recognition technique employing two passes of a document page image through OCR systems that use different recognition languages has been presented. The technique yielded encouraging performance on 28 different journal types, and showed the possibility of extension to other journals. Even though 89% of Greek characters are detected by either the Recognita™ or the FineReader™ OCR engines, the total Greek characters detected by both engines amounts to about 55%. Future research will incorporate a third Greek-enabled OCR engine to improve the system performance.

7. REFERENCES

[1] J. Liang et al., A Methodology for Special Symbol Recognitions, Proceedings of the International Conference on Pattern Recognition (ICPR'2000), Barcelona, Spain, 2000.

[2] R. J. Fateman et al., Progress in recognizing typeset mathematics, Proceedings of SPIE, volume 2660, pp. 37-50, San Jose, CA 1996.

[3] A. Kacem et al., Automatic extraction of printed mathematical formulas using fuzzy logic and propagation of context, Proceedings of ICDAR'99, India, pp. 116-122, 1999.

[4] G. R. Thoma, Automating the production of bibliographic records for MEDLINE. Internal R&D report, CEB, LHNCBC, NLM, 2001.

[5] Prime Recognition Inc., Prime OCR Access Kit Guide, version 2.70, San Carlos, CA, 1997.

[6] ScanSoft Corporation, Developer's Kit 2000, Version 10.

[7] ABBYY Corporation, FineReader Developer's Kit, Version 4.0c.