# A Word-level Deciphering Algorithm for Degraded Document Recognition

Chi Fang[†] and Jonathan J. Hull[‡]
[†] Center of Excellence for Document Analysis and Recognition

Department of Computer Science
State University of New York at Buffalo
Buffalo, New York 14260
chifang@cs.buffalo.edu

[‡] RICOH California Research Center
2882 Sand Hill Road, Suite 115
Menlo Park, CA 94025
hull@crc.ricoh.com

## Abstract

*A text recognition algorithm is proposed that uses word-level language constraints in a deciphering framework to directly decode the identity of each word pattern in an input text. This is a font-independent approach that solves problems of touching characters and character fragmentation. The major difficulty of using a deciphering approach on the word level is that the existence of relatively stable and reliable language constraints on the character level, such as character n-grams and a vocabulary of common words, usually do not scale up to the word level. A word-level deciphering approach is presented in this paper that solves a selected portion of an input text using a word-level relaxation deciphering algorithm. Font information is then learned from the solved portion of the text and used to re-recognize the rest of the text. Tests of the proposed approach on both artificially generated and scanned documents show satisfactory performance in the presence of image degradation.*

## 1 Introduction

Current OCR systems are usually dependent on knowledge of many fonts. The need for large amounts of font training data is an impediment to the development of high performance text recognition algorithms([1, 12]). Also, touching characters and character fragmentation as the result of document image degradation pose severe problems to OCR algorithms. Despite much progress reported for constrained domains, character segmentation in the presence of image degradation is still a difficult problem ([11, 4]).

Character level deciphering algorithms have been proposed for OCR by solving a substitution cipher([2, 3, 9, 10]). A clustering step is usually assumed as a preprocessing stage that converts each character on an input text page into a computer readable code in such a manner that ev-



Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, April 24-26, 1995, 191-202.

ery shape corresponds to a distinct code. The basic idea with using a deciphering algorithm for OCR lies in utilizing language statistics and assigning alphabetic labels to the cipher codes in a way that the character repetition pattern in the input text passage best matches the letter repetition pattern provided by a language model. Because character patterns in the input text passage are used only for forming distinct pattern groups rather than for direct character recognition, this approach to OCR is font-independent.

However, character-level deciphering algorithms are also sensitive to touching characters and character fragmentation. Improvements have led to a character-level deciphering algorithm that is able to handle touching characters and is tolerant to clustering mistakes by combining visual constraints with language models ([5]). But sensitivity to document degradation is still a difficult problem.

A potential solution to touching characters and character fragmentation is to use word-level language constraints in a deciphering framework to directly decode the identity of words. Besides being a font-independent recognition approach, a word-level deciphering algorithm overcomes sensitivity to touching characters and character fragmentation by identifying words directly instead of building up results from character recognition.

However, little progress has been reported on using a deciphering algorithm on the word level for OCR. The major difficulty is that the stable and reliable language constraints that exist on the character level usually do not scale up to the word level. Word n-gram statistics are not as reliable as character n-gram statistics. Also, there does not exist a set of "legitimate English sentences" to serve as constraints for word-level deciphering algorithms just as a dictionary does in a character-level deciphering algorithm.

The solution to this problem lies in the following observation. While it is true that there usually does not exist a content-independent stable word n-gram repetition pattern for even extensive text passages, usually a portion of the words in the input text that are either subject-related key words or function words do repeat themselves in a way that they constitute relatively stable and reliable word bigrams that match relatively well with those provided by a matching language model.

Based on this observation, we proposed a novel word-level deciphering algorithm that solves a selected portion of the input text by using a word-level relaxation deciphering algorithm. Font information is learned from the decrypted text and then used for precise re-recognition of the rest of the text that was not available for cipher solution. This results in a word-level deciphering approach to OCR that overcomes the need for font training through the use of substitution cipher decryption. Sensitivity to touching characters and character fragmentation is accounted for by identifying words directly using word-level language constraints. This proposed approach provides a recognition technique that has the ability to automatically adjust to a given input document and achieves high levels of performance in the presence of image degradation.

The rest of this paper contains three sections. Section 2 discusses each component of the proposed approach. Section 3 presents and discusses the experimental results of applying the word-level deciphering algorithm to degraded documents. Finally, conclusions and some future directions are pointed out in section 4.

## 2 Proposed Approach

The proposed word-level deciphering approach and its major components are shown



Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, April 24-26, 1995, 191-202.

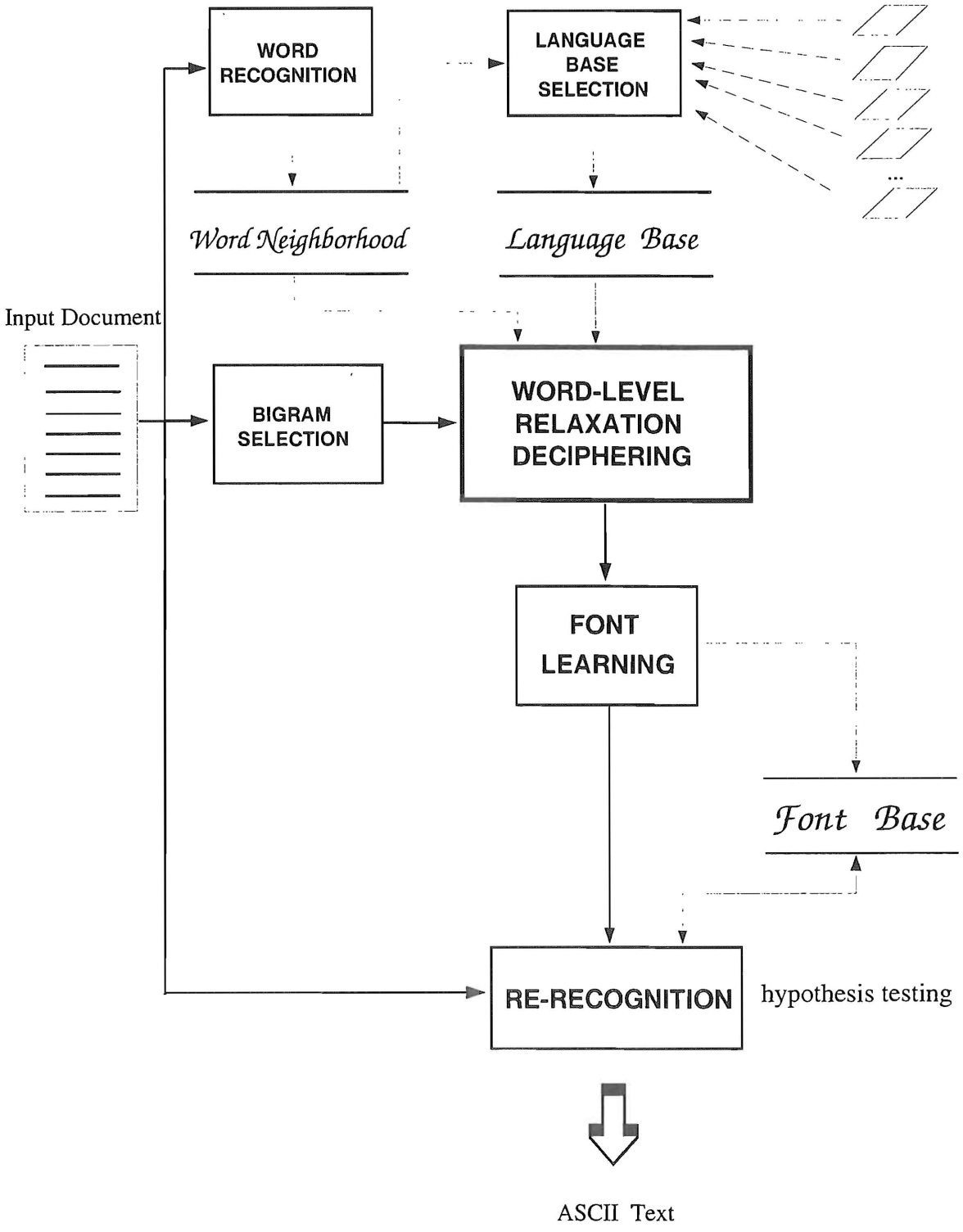


**Figure 1: Word-level deciphering approach**

in Figure 1.

A word clustering algorithm is first applied to an input document to generate word image clusters. Cipher word bigrams with relatively high frequency in the text are selected for deciphering by the word-level relaxation algorithm. Two sources of information are used by the word-level relaxation algorithm. One of them are the neighborhoods generated by a word recognizer. These neighborhoods may not be 100% accurate but they still provide reliable constraints for the relaxation algorithm.

Another source of information used by the relaxation algorithm are the word bigram statistics compiled from a language database. The matching language database is acquired by a vocabulary matching algorithm from a corpus based on the neighborhoods generated for each word in the input text ([8]).

The word-level deciphering stage solves the selected portion of text using the two constraints mentioned above with a probabilistic relaxation algorithm. Font information is then learned from the recognized portion of text and used to re-recognize the rest of the text.

The following sub-sections contain detailed descriptions for each of the major components in the proposed approach.

## 2.1 Selecting a Portion of the Text with Reliable Word Bigram Statistics

A word clustering algorithm groups visually similar word patterns into distinct clusters. Word patterns within a cluster thus share a common cluster identity. The text passage is considered a substitution cipher with each word being substituted by its cluster identity. The number of occurrences of each two-cipherword sequence (cipherword bigram) occurring in the text is then calculated. Those cipherword bigrams that have relatively high bigram frequencies are selected from the text and solved by the relaxation process. The cipherwords selected in this way will usually consist of two categories of words: function words like "the", "of", "and" etc. and subject-related key words.

## 2.2 Probabilistic Relaxation Deciphering Algorithm

As part of the pre-processing to the proposed word-level relaxation labeling algorithm, each word pattern is transformed into a cipher code and assigned probabilities that the cipher code represents each word in the neighborhood. A word-level probabilistic relaxation labeling algorithm is used to update the probabilities for each word using the word bigram constraints from the matching language database. Iterating the updating scheme results in improved estimates that finally lead to the decryption of the selected portion of text. The core idea of the relaxation process here is to use information about the labels of highly confident words and combine them with the bigram constraints from a matching language database to correct or improve the labels of less confident words.

The updating scheme is based on Goshtasby's procedure on the character level ([6]). We changed the updating scheme so that it works on the word level by using word transition probabilities within each word bigram. Figure 2 shows the outline of the word-level relaxation labeling algorithm.

## 2.3 Font Learning

A portion of the cipherwords are recognized as a result of the word-level relaxation deciphering stage. Information regarding the text font is learned by segmenting the word images to construct character prototypes using word image profile analysis guided by known word identities. One

```
Algorithm Word_Level_Relaxation(Selected_Bigram_Set)
BEGIN
        get word transition probabilities from language base
        get initial candidate neighborhood for each word
        FOR (each word bigram in Selected_Bigram_Set)
            REPEAT
                FOR (each candidate in word_0)
                    compute the contribution factor
                    update the probability of candidate
                FOR (each candidate in word_1)
                    compute the contribution factor
                    update the probability of candidate
            UNTIL(converge)
END
```

**Figure 2: Outline of word level relaxation labeling algorithm**

thing is worth noticing here: because at this stage the word identities are already known from the deciphering stage, the segmentation step here to learn the font is different and much easier than the character segmentation problem in most of the segmentation-based OCR methods where the clue for word identities is not available to verify and constrain the many segmentation hypotheses. The visual image and other information regarding each character in a recognized word are acquired to update the character prototypes in the font base.

During font learning, in case an incoming character image pattern turns out to be radically different by a certain similarity measure from the character prototype in the font base, the font learning algorithm matches this character with all the other learned character prototypes in the font base. If no acceptable match is found, this indicates a deformed character image resulted from a mistake made in character segmentation. In this case, the deformed incoming character is simply discarded and excluded from the font learning process. This achieves selective font learning and will guarantee that the fonts learned are of good quality and consistency.

If, on the other hand, the incoming character matches well with another character prototype in the font base, this indicates a possible mistake in the word recognition or word clustering stage. In this case, the suspicious incoming character pattern is re-assigned a character identity of the best match if the new assignment conforms with a legal word identity, which also causes the identity of the word to be changed. This implementation of font learning achieves "self-correction" of mistakes made at the clustering stage and the relaxation stage that are difficult to detect and correct by other means.

Selective learning and self-correction in font learning help guarantee the quality of the learned font prototypes and contribute to the overall tolerance of the approach to

image degradation.

## 2.4 Re-recognition Using Learned Font Information

A hypothesis testing approach using an $A^*$ search algorithm is proposed to re-recognize the rest of the words that were not deciphered. The information that is available for the hypothesis testing algorithm are the neighborhoods for each word and the learned partial font base. For each word to be recognized, the algorithm starts by examing the character patterns at both ends of the partial word image and proceeds inward, matching the character patterns with font prototypes constrained by the neighborhood. Each legal combination of matches at both ends results in a partial word decision. Each partial word decision is associated with a cost that measures the confidence of this partial word decision. It is also associated with a reduced candidate neighborhood that conforms with the accumulated partial decisions of this word made so far.

The cost associated with each partial word decision $f^*$ is the combination of the accumulated cost $g^*$ of the series of character matches as the algorithm proceeds inward, and the estimate of the potential cost $h^*$ to reach a complete word decision. Each partial word decision is saved into an open queue with its associated cost $f^*$ and the reduced word candidate neighborhood. The algorithm chooses the node from the open queue with the optimal cost $f^*$ to expand further, until a complete word decision with a globally optimal cost is reached or the open queue becomes empty.

For a partial word decision, the estimated cost $h^*$ is assigned a value of 0 for most cases where there is no obvious indication of mismatch between the remaining portion of the word image and the remaining characters of its associated candidate set. However, it will be assigned a large value if it becomes certain that the remaining portion of the word image doesn't match any of its associated candidates, preventing this partial word decision from being considered further.

The decision about whether the remaining portion of a word image matches its associated set of candidates is based on visual characteristics such as the overall size of the pattern, the existence of descenders and ascenders, and the existence of isolated spots at the top. These visual constraints are relatively stable under various image degradations.

This method compensates for local character degradations by relying on an $A^*$ search algorithm that seeks a globally optimal word decision. Multiple visual constraints that are relatively stable across image degradations are used to estimate the potential cost $h^*$, which enables the algorithm to use as much as possible the visual heuristic power to speed up the search, while maintaining the admissibility of the $A^*$ algorithm.

It is likely that font information of some of the characters that appear in the text will not be learned because the relaxation deciphering stage only recognizes a portion of the text. This will cause difficulty when words that include characters not learned in the font learning stage are to be re-recognized. This problem is accounted for by using general geometric heuristics of characters to estimate the match, and also accounted for by a second phase of font learning during re-recognition. Prototypes of characters not available in the original font base can be learned when some words that include these characters are reliably recognized by the hypothesis testing algorithm.

Figure 3: Test document for word level relaxation labeling

# 3 Experimental Results and Discussion

The proposed word-level deciphering algorithm was tested on both artificially created documents and on scanned documents with varying image degradations.

## 3.1 Artificially Created Document

The document text used to test the proposed word-level deciphering algorithm consists of two pieces of recent news about the Korean Peninsula nuclear crisis. The text contains 921 words and 380 unique words. The image of this text was generated using a text formatting package. Noise that simulates image degradation was added to the clean image. As a result of the image degradation, the word recognizer we used to generate the top ten neighborhood ([7]) achieved a 79% correct rate for the top choice. The correct rate for the top ten choices was 99.6%. The degraded document image is shown in Figure 3.

The matching language database that provides word bigram constraints consists of 31 recent news reports on the Korean Peninsula nuclear crisis which is generated by a vocabulary matching system ([8]) from a comprehensive language corpus acquired from Clarinet. The language database contains 12595 words and 2464 unique words. The word bigram set constructed from it contains 23907 different word bigrams, from which the bigrams with frequencies greater than a threshold are selected as the word bigram database. The frequencies of this selected set of word bigrams are then normalized and used in the relaxation deciphering algorithm.

The word-level relaxation deciphering stage correctly deciphers 301 out of the 921 words. These include 24 unique words. Ten of them are function words and 14 of them are key words. This limited number of function words and key words constitute about one third of the whole input text, and the proposed word-level deciphering algorithm is able to recognize them with high reliability.

From the recognized portion of text, the font learning algorithm developed prototypes for 25 of the 32 characters that appeared in the input text. The results are shown in Figure 4.

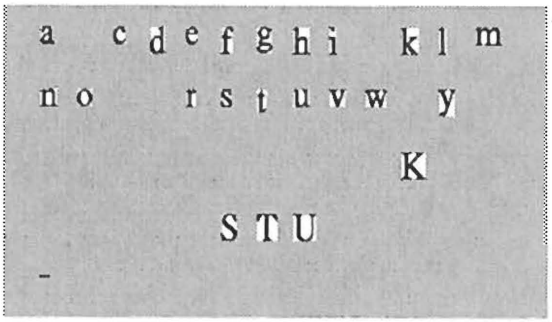


**Figure 4: Font learned from the deciphered portion of words**

For the re-recognition stage using the learned font information, only eight words were not recognized correctly. The eight cases where the algorithm failed are: "be" recognized as "he", "an" recognized as "au", "21" was incorrectly recognized because digits had not been recognized previously and "Hans" was recognized as "Kaus". Four other words were not correctly recognized because the correct candidates were not included in the initial neighborhood.

Some interesting observations can be drawn from these errors. In case the word to be recognized is short, such as the words "be" and "an", the hypothesis testing algorithm loses its ability to compensate for local mismatches and deformations by relying on global characteristics. This makes the word decisions heavily dependent on local character recognitions that are unreliable in the presence of image degradations.


Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, April 24-26, 1995, 191-202.

Short words combined with unknown character prototypes aggravate this problem, as shown in the case of the word "Hans".

The second phase of font learning in the re-recognition stage was able to learn four more character prototypes from re-recognition results that were reliable. The updated font base is shown in Figure 5.

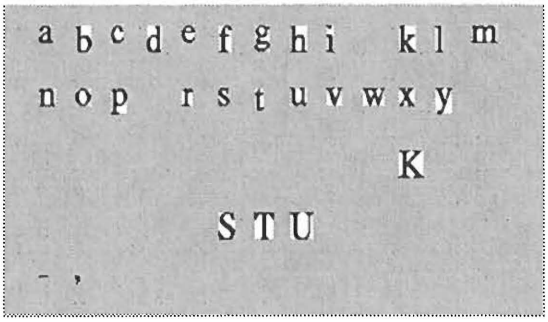

**Figure 5: Updated font base after font learning in the re-recognition stage**

The eight unique word mistakes resulted in 13 words that were not recognized correctly. The proposed word-level relaxation deciphering approach as a whole was able to achieve a correct word recognition rate of 98.6% for the artificially created and degraded document.

### 3.2 Scanned Document

The text of the scanned document was acquired the same way as presented in last section. The document image was first generated with the `ditroff` text formatting package in an 11 pt. Times Roman font. The clean image was printed on plain paper using a laser printer and scanned at 300 ppi resolution. The final binary image was the result of thresholding the scanned gray-scale image by choosing a threshold such that the character distortion and touching characters that were produced test the ability of the algorithm to handle these degradations. An enlarged portion the final binary document image used for the test is shown in Figure 6.

The matching language base and the word bigrams are generated in the same way as presented in the last section.

The image degradation in the scanned document resulted in reduced word clustering performance, which resulted in only 258 words being correctly recognized by the relaxation deciphering stage. These 258 words consist of 21 unique words. Among them, 12 are key words and 9 are function words. From the deciphered words, the font learning stage was able to learn 26 out of the 32 character font prototypes.

The re-recognition stage using the learned font information failed to recognize only four words. The four cases where the re-recognition failed were: "an" recognized as "au", "21" incorrectly recognized because digits had not been recognized previously, "Hans" recognized as "Kaus", and "sufficient" recognized as "sediment". The errors were caused by similar reasons as in the case of the artificially created document discussed in the last section. Also, the alignment of the font sometimes makes it difficult to deal with the problem where an 'i' following an 'f' is mistaken as part of the 'f' pattern and completely cut off when the 'f' is recognized, as in the case of the word "sufficient". This calls for more sophisticated use of character typeface and alignment information in the re-recognition algorithm.

Reliable recognition of some of the words also enabled the font re-learning phase to learn three more character and punctuation prototypes.

The four unique word mistakes in the re-recognition stage resulted in 8 words that were not recognized correctly. The proposed word-level relaxation deciphering approach as a whole was able to achieve a correct word recognition rate of 99.1% for the scanned document.



Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, April 24-26, 1995, 191-202.

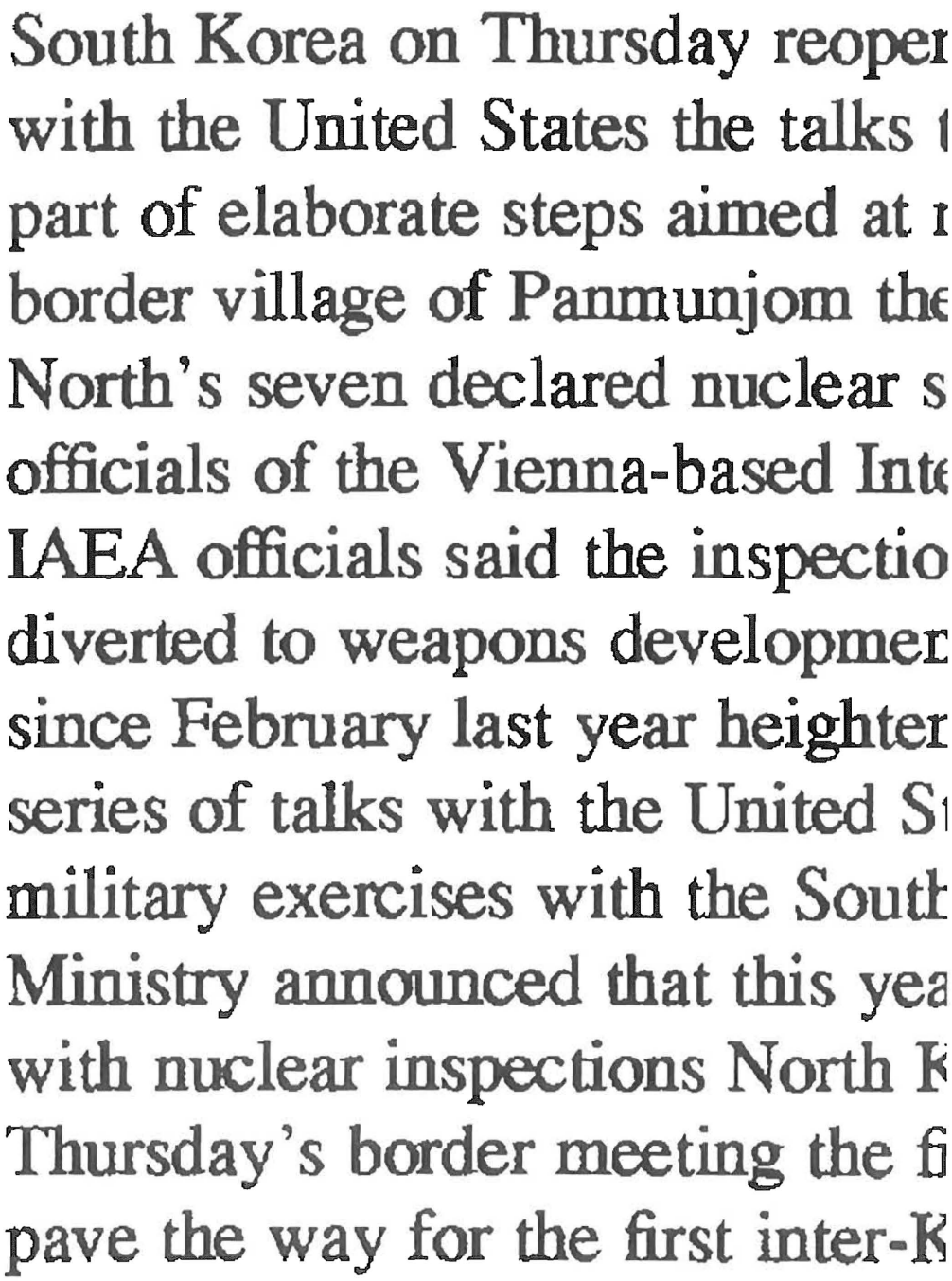


**Figure 6: An enlarged portion of the scanned test document.**



Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, April 24-26, 1995, 191-202.

# 4 Conclusions and Future Directions

A novel word-level deciphering algorithm for OCR was proposed in this paper that solves a selected portion of text with high reliability by using word-level language constraints within a relaxation deciphering algorithm. Font information is learned from the decrypted portion of the text and used to re-recognize the rest of the text that was not solved in the deciphering stage. This word-level deciphering approach overcomes the need for font training by solving a substitution cipher and by learning font information dynamically from the input document. Sensitivity to touching characters and character fragmentation is reduced to the minimum by decoding words directly without relying on character segmentation. Tests of the proposed approach on both an artificially created document and a scanned document showed reliable performance in the presence of image noise.

More extensive tests of the proposed approach on more varieties of real scanned documents are underway. Improvements in performance of the word clustering algorithm and the font learning algorithm are needed. Further improvement of the re-recognition stage to take into consideration the font typeface and alignment information is underway. Expansion of the system so that it can process the full range of punctuation and digits is also being planned.

# References


[1] Henry S. Baird, "Document image defect models and their uses," *Proceedings of the Second International Conference on Document Analysis and Recognition ICDAR-93*, 1993.

[2] R. Casey and G. Nagy, "Autonomous reading machine," *IEEE Trans. Comput.*,vol. C-7, May 1968

[3] R. Casey, "Text OCR by solving a cryptogram," *Proc. ICPR-8*, Paris, 1986, pp. 349-351

[4] R. Casey and K. Wong, "Document-analysis system and techniques," In R. Kasturi and M. Trivedi, editors, *Image Analysis and Applications*, pages 1–35. New York, 1990.

[5] Chi Fang and Jonathan J. Hull, "A modified character level deciphering algorithm for OCR in degraded documents," to appear in *Proceedings of the Conference on Document Recognition of 1995 IS&T/SPIE Symposium*, 1995.

[6] A. Goshtasby and R. W. Ehrich, "Contextual word recognition using probabilistic relaxation labeling," *Pattern Recognition*, 21(5):455–462, 1988.

[7] Tin Kam Ho, Jonathan J. Hull, and Sargur N. Srihari, "A word shape analysis approach to lexicon based word recognition," *Pattern Recognition letters*, 13:821–826, 1992.

[8] Jonathan J. Hull and Y. Li, "Interpreting Word Recognition Decisions with a Document Database Graph," *International Conference on Document Analysis and Recognition*, Tsukuba, Japan, October 20-22, 1993.

[9] G. Nagy, "Efficient algorithms to decode substitution cipher with application to OCR," *Proc. ICPR-8*, Paris, 1986, pp. 352-355

[10] G. Nagy, S. Seth and K. Einspahr, "Decoding substitution cipher by means of word matching with application to OCR," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 5, Sept. 1987

[11] G. Nagy. At the frontiers of OCR. *Proceedings of the IEEE*, 80(7), 1992.

[12] R. A. Wilkenson et al, "The first census optical character recognition system conference," NIST internal report, U.S. National Institute of Standards and Technology, Gaithersburg, Maryland, 1992.


Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, April 24-26, 1995, 191-202.

# Fourth Annual Symposium on Document Analysis and Information Retrieval

April 24 - 26, 1995

Desert Inn Hote
Las Vegas, Nevada

Sponsored by:

**Information Science Research Institute**

and

**Howard R. Hughes College of Engineering**

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

University of Nevada, Las Vegas

Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, April 24-26, 1995, 191-202.

# CONFERENCE CHAIRMAN'S MESSAGE

The Fourth Annual Symposium on Document Analysis and Information Retrieval is the latest in a series of very successful symposiums started in March of 1992. The goal of these meetings is to encourage communication between two very different disciplines — the ability to transform an image of a paper document into an electronically "understandable" file, and the ability to locate specific information in that file. Cross-communication is essential if these two areas plan to do more than the serial process of feeding paper into retrieval machines. For example, retrieval systems need to take advantage of the structural information that can be provided by good document analysis, and this can happen only if both groups have some detailed knowledge of each other's techniques and problems.

The success of these meetings comes from three factors. The first is an ever increasing interest in merging results from the disciplines of document analysis and information retrieval into a new discipline of document understanding. The quality of the papers submitted to the program committee has increased each year, with more papers trying to "bridge the gap" between the two areas. This year there were 46 papers submitted, with 25 accepted for oral presentation and 9 for poster presentation. Both oral presentations and poster papers are included in these proceedings.

The second factor is the continued heavy involvement of the entire UNLV/ ISRI staff in research and testing in this area. This enthusiasm permeates the conference, and the annual reports from this group on the third day of the meeting have always been a highlight for me.

The third factor is the hard work put in by many people to make these symposiums happen. My special thanks go to the two program chairs, Larry Spitz and David Lewis, and to their committees for their diligence in creating a worthy program for the symposium. These meetings would not happen without the work of the staff at ISRI who do the many organizational tasks needed and produce the proceedings. Special recognition should go to Tom Nartker, Junichi Kanai, Debbie Wallace, Andy Bagdanov, and Mary Guirsch for the immense amount of effort that has gone into making this fourth Symposium a success.

Donna Harman
Symposium Chair



Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, April 24-26, 1995, 191-202.

# Conference Committee

## Symposium Chair

Donna Harman
National Institute of Standards and Technology

## Information Retrieval Chairman

David D. Lewis
AT&T Bell Laboratories

## Document Analysis Chairman

Larry Spitz
Fuji Xerox Palo Alto Laboratory

**Document Analysis Committee:**
Henry Baird, AT&T Bell Laboratories
Andreas Dengel, German Research Center for Artificial Intelligence
Hiromichi Fujisawa, Hitachi, Central Research Laboratory
Jonathan Hull, RICOH California Research Center
Junichi Kanai, University of Nevada, Las Vegas
Juergen Schuermann, Daimler Benz Research Center
Suzanne Taylor, Unisys Corporation
Karl Tombre, INRIA, Lorraine, France

**Information Retrieval Committee:**
Christopher Buckley, Cornell University
Kenneth Church, AT&T Bell Laboratories
Robert Korfhage, University of Pittsburgh
Fausto Rabitti, CNR-IEI
Kazem Taghva, University of Nevada, Las Vegas
Takenobu Tokunaga, Tokyo Institute of Technology
Howard Turtle, West Publishing
Ross Wilkinson, RMIT
Peter Willett, University of Sheffield



Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, April 24-26, 1995, 191-202.