

Improving OCR Performance With Word Image Equivalence

Tao Hong[†] 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
taohong@cs.buffalo.edu

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

Abstract

This paper proposes a technique for improving the performance of an OCR system that uses information about equivalent word images inside a document. Words that are repeated inside a document are grouped into clusters by an image matching algorithm. The decisions of an OCR algorithm about the identities of those words are used to generate a common recognition result for each of the original word images. This technique thus combines information from the document image (word image clusters) with recognition results to correct errors made by OCR systems on different instances of the same word. Experimental results are presented that show about 50% of the words in a document are repeated two or more times. A clustering algorithm is able to reliably locate a large percentage of these words in the presence of noise. Experiments on images degraded with uniform noise show that the correct rate of a commercial OCR system can be improved from 79% to 92% on the

words in those clusters. An error analysis is given that shows with further development correct rates in the 98+% range are achievable.

1 Introduction

The objective of visual text recognition is to transform an arbitrary image of text into its symbolic equivalent correctly. Given a high quality text page, current commercial document recognition systems can recognize the words on the page at a high correct rate [2, 16, 17]. However, given a degraded text page, such as a multiple-generation photocopy or facsimile, their performance usually drops abruptly [1, 15].

Third was the loss of life at the South African-guarded Calueque Dam project close to the Angolan border, when Cuban Mig-23 jets attacked in June. The latter engagement established the Angolan and Cuban air forces' superiority over South Africa's for the first time.

South Africa's military setbacks resulted in a marked shift in the locus of its strategizing vis-à-vis Angola and Namibia—away from the military hardliners, such as Defense Minister Magnus Malan, armed forces chief Gen. Jannie J. Geldenhuys, and others in the State Security Council, and back to the South African cabinet and the more pragmatic approach of Foreign Minister Pik Botha and Neil Van Heerden, director-general of the Department of Foreign Affairs.

(a)

Third was the loss of life at the South African-guarded Calueque Dam project close to the Angolan border, when Cuban Mig 23 jets attacked in June. The latter engagement established the Angolan and Cuban air forces' superiority over South Africa's for the first time.
South Africa's military setbacks resulted in a marked shift in the locus of its strategizing vis-vis Angola and Namibia away from the military hardliners, such as Defense Minister Magnus Malan, armed forces chief Gen. Jannie J. Geldenhuys, and others in the State Security Council, and back to the South African cabinet and the more pragmatic approach of Foreign Minister Pik Botha and Neil Van Heerden, director-general of the Department of Foreign Affairs.

(b)

Third was the loss of life at the South African-guarded Calueque Dam project close to the Angolan border, when Cuban Mig 23 jets attacked in June. The latter engagement established the Angolan and Cuban air forces' superiority over South Africa's for the first time.
South Africa's military setbacks resulted in a marked shift in the locus of its strategizing vis-vis Angola and Namibia away from the military hardliners, such as defense Minister Magnus Malan, armed forces chief Gen. Jannie J. Geldenhuys, and others in the State Security Council, and back to the South African cabinet and the more pragmatic approach of Foreign Minister Pik Botha and Neil Van Heerden, director-general of the Department of Foreign Affairs.

(c)

Figure 1: A fragment of a degraded text page and its OCR results. (a) A text block; (b) OCR results with no dictionary; (c) OCR result when the dictionary is used.

Figure 1 (a) is a fragment of text extracted from a degraded journal page. Such an image can be read correctly by a human but the OCR system makes a significant number of mistakes. The recognition results generated by a commercial OCR system are also shown in the same figure. Figure 1 (b) shows the OCR result without using the built-in dictionary and Figure 1 (c) is the result with the dictionary.

Word	Image	OCR (no Dict.)	OCR (with Dict.)
Wimg.0007	the	le	de
Wimg.0012	project	proect	prolect
Wimg.0016	Angolan	Angolari	Angolarf
Wimg.0023	attacked	attacloed	attacked
Wimg.0039	Africa's	Aica's	Aica's
Wimg.0042	first	list	list
Wimg.0054	in	m	m
Wimg.0061	Angola	Angda	Angola
Wimg.0065	from	boin	Coin
Wimg.0069	such	such	such
Wimg.0071	Defense	11elense	defense
Wimg.0077	chief	chief	chief
Wimg.0090	and	ard	ard
Wimg.0102	ap-	aw	apt
Wimg.0103	proach	proach	proach

Figure 2: Examples of word recognition errors made by the OCR

OCR errors are supposed to be detected and corrected during contextual postprocessing. Most postprocessing methods use linguistic knowledge sources to detect and correct potential errors. Techniques such as character-n-gram-based relaxation, dictionary lookup and string editing have been developed [4, 5, 11, 12, 18].

Figure 2 lists some word images and their OCR results with and with-

out postprocessing. For some word images, such as "Wimg.0023", "Wimg.0061", "Wimg.0069" and "Wimg.0077", recognition results are not correct. By using the dictionary, some errors can be detected and corrected. However, for other word images in the figure, lexical knowledge is not helpful either to detect or to correct possible errors. For the word images "Wimg.0042" and "Wimg.0054", the errors were not detected, and therefore, were not corrected because the recognition results are valid dictionary words. For other images like "Wimg.0007", "Wimg.0012", "Wimg.0065", "Wimg.0071" and "Wimg.0102", recognition errors were detected because their postprocessing results are different from the original OCR results. But the decision words were chosen incorrectly among several word candidates suggested by dictionary lookup and string editing.

In this paper a method for postprocessing OCR results is proposed that combines visual inter-word constraints with OCR results. This approach is based on the following observations about the linguistic and typographic characteristics of document pages and the performance of OCR systems on degraded documents.

- In a text page, there are usually many occurrences of the same words. For example, in a normal English text, many function words and content words occur repeatedly. Because of the fact that the text on a given page is usually printed in a limited number of fonts, images of the same word are usually very similar.
- If two word images are equivalent, their recognition results should be the same. However, a commercial OCR often makes different decisions on different instances of the same word, especially when the document page is seriously degraded (see Figure 3). This

is understandable since variations in local noise, which may have no significant effect on the overall shape of a word, do make character segmentation and character classification difficult.

Based on these observations, an algorithm is proposed that first locates clusters of equivalent words in a document. A high accuracy is hypothesized for clustering since it uses global characteristics of word images that are not altered by local noise. The decisions of the OCR for the words in a cluster are then combined to generate a single consensus decision for these words. This should compensate for the case where an OCR makes different decisions on word images that are the same.

IMAGE	OCR	IMAGE	OCR
	arid		military
	and		military
	die		list
	the		first

Figure 3: Inconsistent OCR results on word images that are visually similar.

In the rest of this paper, we present the proposed approach. First, we discuss how to measure visual equivalence among word images through word image matching. Then, we describe how visual word equivalence constraints are used to postprocess OCR results so that many OCR errors can be detected and corrected. Then, some preliminary experimental results are reported. Finally, conclusions and future directions are presented.

2 Image Equivalence and Word Image Clustering

The visual similarity between two binary images of the same size can be measured quantitatively by how much the images match at the pixel level [10]. Let A and B be two $m \times n$ binary images. Inside an image, "1" and "0" denote "black" and "white" pixel respectively. We measure visual similarity between A and B as

$$r(A, B) = \frac{\sum_i^m \sum_j^n (A_{ij} \wedge B_{ij})}{\sum_i^m \sum_j^n (A_{ij} \vee B_{ij})}$$

where " \wedge " and " \vee " are *and* and *or* operators respectively. The higher the measurement r is, the better two images match. When two images, A and B , are slightly different in size, the similarity between them can be defined by the maximal matching obtained if A is shifted over B . By setting a proper threshold r_0 , we can define that two word images are visually equivalent if $r(A, B) > r_0$.

Given a sequence of word images from a text page, the visual equivalence among the word images can be computed by word image clustering [14]. After image clustering, images in the same cluster are visually equivalent. After clustering the word images extracted from Figure 1, the clusters with multiple images are listed in Figure 4.

3 Postprocessing OCR Output with Image Equivalence

The postprocessing method is applied to those word images that are equivalent to at least one other image in an input document (i.e, the images that occur in *large* clusters), This method is not applicable to the other word images that are included in clusters by themselves.

Figure 5 is the outline of the algorithm for postprocessing. For a cluster, if there

Cluster 0:	the the the the the the the the the the
Cluster 1:	and and and and and and
Cluster 2:	of of of of of
Cluster 3:	in in in in
Cluster 4:	South South South South
Cluster 5:	Cuban Cuban
Cluster 6:	Africa's Africa's
Cluster 7:	military military
Cluster 8:	Minister Minister
Cluster 9:	Foreign Foreign
Cluster 10:	to to
Cluster 11:	Angolan Angolan

Figure 4: Examples of word image clusters.

exists any disagreements among the recognition results for the word images in that cluster, some of the word images in the cluster have not been recognized correctly. After detecting a cluster with such an inconsistency, the error correction step is applied to choose a single decision for all the words in the cluster. There are several criteria for selecting the single decision. They are listed below:

- Majority voting: if there are several images in the cluster and most of them have the same decision as their OCR result, that decision will become the decision for the cluster (see Figure 6 (a));
- Dictionary lookup: if two candidates have the same number of votes, the one which is a valid dictionary entry is preferred(see Figure 6 (b));
- Rejection: if all candidates have an equal number of votes and none of them are valid words, the postprocessing method described here will not se-

lect a decision for the cluster(see Figure 6 (c)).

4 Preliminary Results

Sixteen document pages were selected as testing samples. Among them, twelve pages are from two multiple-page journal articles in CEDAR's journal page database. The remaining four pages are from UNLV's DOE image database that is available on the cdrom released by the University of Washington. More details about the testing samples are described in Table 1. We list the number of words for each page and the number of words which have at least another word instance in the page. On average, about sixty percent of the words in a page are repeated at least once in some other position.

Noise was added to the original images to simulate the effect of a multiple generation photocopy. We used the University of Washington document degradation model (DDM) [13] to add two different levels of

```

extract all word images from the text image;

word image clustering;

/* error detection through consistency analysis */
for each cluster  $C$ , where  $|C| > 1$ , do
    if there is disagreement among the images on recognition result
        then
            mark the cluster as inconsistent;
        end if
    end for

/* error correction */
for each cluster  $C$ , which was marked as inconsistent, do
    select a decision word for the cluster
        using majority voting, dictionary lookup and other criteria;
    all images use the decision of the cluster as their recognition result;
end for

```

Figure 5: Outline of the postprocessing algorithm

local noise, which are denoted as $dd1$ and $dd2$. The parameter settings for $dd1$ are (420, 0.0, 1.0, 2.5, 1.0, 1.0, 2); the parameter settings for $dd2$ are (820, 0.0, 1.0, 1.0, 1.0, 1.0, 3). The image generated using $dd2$ is more noisy than $dd1$. Figure 7 shows the effect of $dd1$ and $dd2$ on a fragment of an example page.

A Caere OCR system was used to compute OCR results and the coordinates of bounding boxes for the word images in each document. The ASCII truth data was generated from on the OCR output on the original images because the Caere OCR performance was very good (on average, word recognition accuracy is 98.3% and 96.6% according to whether or not the dictionary is used).

For degraded images generated by $dd1$

and $dd2$, the OCR performance dropped significantly in two respects: word segmentation accuracy and word recognition correct rate. For pages with noise at the $dd1$ level, if the OCR's dictionary is turned off, the word recognition correct rate is 71.4%. If the dictionary is used, the word recognition correct rate is 80.6%. For pages with noise at the $dd2$ level, the word recognition correct rate is 60.1% when the dictionary is disabled and 70.9% when it is used.

Word image clustering was performed using the coordinates of the word bounding boxes provided by the OCR package. After clustering, all *large* clusters (i.e., those clusters that contain two or more images) were located. The threshold r_0 was set uniformly as 0.60. About half the word images in a typical page are included in the clus-

ID	TYPE	NUM OF WORDS	NUM OF REPEATED WORDS	NOTES
A1	Text	822	455	A1-A5 are from a five-page journal article AA052532 in CEDAR's database
A2	Text	1119	683	
A3	Text	564	355	
A4	Text	1230	846	
A5	Text	735	444	
D1	Text	350	175	D1-D7 are from a seven-page journal article DK074491 in CEDAR's database
D2	Text	469	217	
D3	Text	693	406	
D4	Text	835	521	
D5	Text	536	260	
D6	Text	683	336	
D7	Text	1021	572	
N1	Reference	856	508	N022BIN in UW's CDROM
N2	Text & Reference	807	477	N036BIN in UW's CDROM
N3	Text	766	515	N03MBIN in UW's CDROM
N4	Text	1052	707	N04FBIN in UW's CDROM
Total		12538	7477	

Table 1: 16 document pages are used for testing

ters.

The three different postprocessing strategies were used to select a decision for each large cluster. They are: (1). majority vote; (2). majority voting plus dictionary lookup; and (3). majority voting plus dictionary lookup with the option of rejection. The dictionary used here is the word list collected from the Brown Corpus and the Penn Treebank. There are more than 70,000 words in the list. The best improvement was made by the third strategy.

Tables 2 and 3 show the performance on the words in large clusters. On this subset of word images, the improvement of the proposed postprocessing algorithm was significant. For the OCR results on pages with noise at the *dd1* level and without using the dictionary, the third postprocessing strategy improved the correct rate to 91.5% from the original 78.7%. For the OCR results on pages with noise at the *dd1* level

and with using the dictionary, the third postprocessing strategy improved the correct rate to 92.5% from original 86.4%. For the OCR results on pages with noise at the *dd2* level and when using the dictionary, the third postprocessing strategy improved the correct rate to 91.1% from original 69.8%. For the OCR results on pages with noise at *dd2* level and with using the dictionary, the third postprocessing strategy improved the correct rate to 92.3% from original 79.3%;

5 Error Analysis and Future Improvements

The potential improvement in performance possible with further refinements of the algorithm presented in this paper are illustrated by an analysis of the errors made when page A1 with *dd2* noise was processed. Out of the 380 words in large clus-

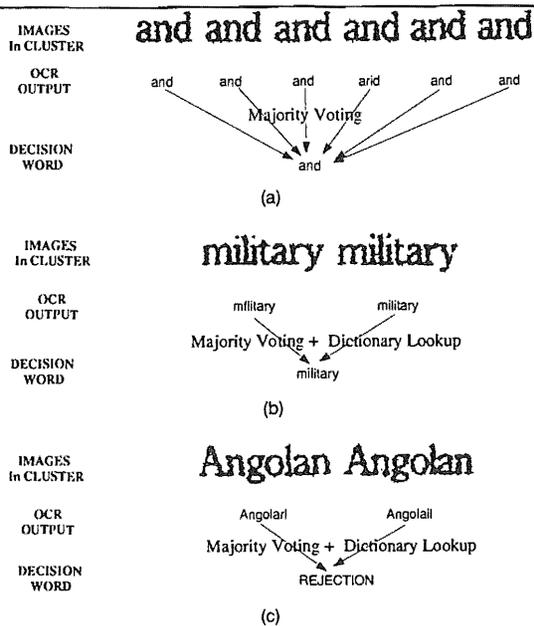


Figure 6: Determining a decision for a cluster with multiple entries. (a). Majority voting; (b). Majority voting + dictionary lookup; (c). Rejection

ters, 24 were recognized incorrectly and 15 were rejected when the M+D+R postprocessing method was used. Figure 8 shows the contents of the clusters calculated from document A1 that were incorrectly recognized. The decision made by the OCR package, the output of the postprocessing algorithm, and the truth value for each word are given. The decision made by the postprocessing algorithm for each cluster is shown after the identifier for the cluster.

Cluster 16 shows that the first character "A" was correctly recognized in three out of the six words in the cluster and the last four characters "ica's" were correctly recognized in four out of six cases. If the locations of the character segmentation points were available in those word images, the portion of the words between the "A" and the "ica's" could be further inspected and

Third was the loss of life at the South African-guarded Calueque Dam project close to the Angolan border, when Cuban Mig-23 jets attacked in June. The latter engagement established the Angolan and Cuban air forces' superiority over South Africa's for the first time.

(a). Original Image

Third was the loss of life at the South African-guarded Calueque Dam project close to the Angolan border, when Cuban Mig-23 jets attacked in June. The latter engagement established the Angolan and Cuban air forces' superiority over South Africa's for the first time.

(b). DDM = dd1

Third was the loss of life at the South African-guarded Calueque Dam project close to the Angolan border, when Cuban Mig-23 jets attacked in June. The latter engagement established the Angolan and Cuban air forces' superiority over South Africa's for the first time.

(c). DDM = dd2

Figure 7: Degraded document images were generated with the UW document degradation model (DDM).

compared to portions of words that were successfully recognized in other words to derive the decision that the missing characters are the ligature "fr". Such a divide-and-conquer strategy may also be applicable to clusters 45, 47, 51, and 80.

Another subset of the errors are attributable to proper nouns that are not in the dictionary. In clusters 41, 43, 59, and 66, the correct choice was among the decisions in the cluster. It was just not chosen by the postprocessing algorithm. This could be solved by including more proper

nouns in the dictionary or perhaps detecting proper nouns a-priori [3] and recognizing them with a specialized form of divide-and-conquer. Cluster 52 would be best solved by the second form of this strategy.

The remainder of the errors illustrate the need for improving the rejection strategy or that some errors are probably non-recoverable. Cluster 38 contains three different legal dictionary words. In this case all three words should be rejected. Clusters 68 and 79 contain alternative decisions that are legal dictionary words. The best way to solve these choices is probably to employ a different form of postprocessing such as word transition probabilities [8, 9] or parsing [7].

Thus, there is the potential with further development of the algorithm proposed in this paper to correct all but four of the errors. This would yield a 99% correct rate with no errors if the divide-and-conquer strategy always made the correct decision.

6 Conclusions and Future Directions

An OCR postprocessing method based on word image equivalence was proposed. The method combines information about which word images are equivalent with the recognition results calculated on those words by a commercial OCR package. The result is an algorithm that corrects errors made when different instances of the same word are assigned different identities by an OCR.

Experiments on 16 degraded document pages (with over 12,000 words in total) show that on a large portion of the word images (more than half the words in the pages), the method improved the word recognition accuracy from 70% to 92%. By analyzing the errors on one page, it was found that there is still room to further improve the performance of the proposed approach. With further refinements, the pro-

posed algorithm could have an accuracy as high as 99% correct if the errors made on the selected page are representative of the errors in general.

The method works only for a portion of the words in a document page. But the approach can be extended to the more general situation. Word image equivalence is only one of several visual inter-word relations that we have observed [6]. The other types of relations concern partial similarity. For example, one word image can be a subpattern of another word image, or two word images can match very well on their left parts, and so on. The coverage of relations based on visual partial similarity is much larger than that of relations based on visual word equivalence. Future work will be directed toward exploring the use of partial relations further.

7 Acknowledgment

The authors gratefully acknowledge the assistance provided by Carl Alsing and Morris Jones of the Caere Corporation.

References

- [1] H. S. Baird. Document image defect models and their uses. In *Proceedings of the Second International Conference on Document Analysis and Recognition ICDAR-93*, pages 62-67, Tsukuba, Japan, October 20-22 1993.
- [2] S. Chen, S. Subramaniam, R. M. Haralick, and I. T. Phillips. Performance evaluation of two ocr systems. In *Symposium on Document Analysis and Information Retrieval*, pages 299-317, Las Vegas, Nevada, April 26-28 1993.
- [3] G. DeSilva and J. J. Hull. Proper noun location in document images. *Pattern Recognition*, pages 311-320, February 1994.
- [4] J. Esakov, D. P. Lopresti, J. S. Sandberg, and J. Zhou. Issues in automatic ocr error classification. In *Symposium on Document*

- Analysis and Information Retrieval*, pages 401-412, Las Vegas, Nevada, April 26-28 1993.
- [5] A. Goshtasby and R. W. Ehrich. Contextual word recognition using probabilistic relaxation labeling. *Pattern Recognition*, 21(5):455-462, 1988.
- [6] T. Hong. Integration of visual inter-word constraints and linguistic knowledge in degraded text recognition. In *Proceedings of 32nd Annual Meeting of Association for Computational Linguistics, in student sessions*, pages 328-330, Las Cruces, New Mexico, June 27-30 1994.
- [7] T. Hong and J. J. Hull. Text recognition enhancement with a probabilistic lattice chart parser. In *Proceedings of the Second International Conference on Document Analysis (ICDAR-93)*, pages 222-225, Tsukuba, Japan, October 20-22 1993.
- [8] T. Hong and J. J. Hull. Degraded text recognition using word collocation. In *Proceedings of the Conference on Document Recognition, 1994 ISE/T/SPIE Symposium*, pages 334-342, San Jose, CA, February 6-10 1994.
- [9] T. Hong and J. J. Hull. Degraded text recognition using word collocation and visual inter-word constraints. In *Proceedings of ANLP94*, Stuttgart, Germany, October 13-15 1994.
- [10] J. J. Hull, S. Khoubyari, and T. K. Ho. Word image matching as a technique for degraded text recognition. In *Proceedings of 11th International Conference on Pattern Recognition*, pages 665-668, Hague, Netherlands, August 30 - September 3 1992.
- [11] M. A. Jones, G. A. Story, and B. W. Ballard. Integrating multiple knowledge sources in a bayesian ocr post-processor. In *Proceedings of ICDAR-91*, pages 925-933, St. Malo, France, 91.
- [12] S. Kahan, T. Pavlidis, and H. S. Baird. On the recognition of printed characters of any font and size. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(2):274-288, 1987.
- [13] T. Kanungo, R. M. Haralick, and I. Phillips. Global and local document degradation models. In *Proceedings of the Second International Conference on Document Analysis and Recognition ICDAR-93*, pages 730-734, Tsukuba, Japan, October 20-22 1993.
- [14] S. Khoubyari and J. J. Hull. Keyword location in noisy document images. In *Proceedings of the Second Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, Nevada, April 26-28 1993.
- [15] T. Pavlidis. Problems in the recognition of poorly printed text. In *Symposium on Document Analysis and Information Retrieval*, pages 162-173, Las Vegas, Nevada, 1992.
- [16] S. V. Rice, J. Kanai, and T. A. Nartker. An evaluation of ocr accuracy. In *1993 Annual Report of ISRI, University of Nevada, Las Vegas*, pages 9-34, 1993.
- [17] S. V. Rice, J. Kanai, and T. A. Nartker. The third annual test of ocr accuracy. In *1994 Annual Report of ISRI, University of Nevada, Las Vegas*, pages 11-40, 1994.
- [18] S. N. Srihari. *Computer Text Recognition and Error Correction*. IEEE Computer Society Press, 1984.

ID	OCR		POSTPROCESSING		
	OPT	ACCURACY	M	M+D	M+D+R
A1	ND	79.4%(339/427)	88.1%(376/427)	88.8%(379/427)	95.6%(373/390)
	WD	85.0%(363/427)	90.9%(388/427)	91.6%(391/427)	96.0%(387/403)
A2	ND	77.4%(501/647)	91.0%(589/647)	92.6%(599/647)	96.2%(587/610)
	WD	83.6%(541/647)	93.2%(603/647)	94.0%(608/647)	96.3%(595/618)
A3	ND	73.1%(256/350)	74.6%(261/350)	76.6%(268/350)	81.9%(266/325)
	WD	83.4%(292/350)	82.3%(288/350)	83.4%(292/350)	85.6%(290/339)
A4	ND	74.5%(644/864)	77.8%(672/864)	81.1%(701/864)	84.3%(701/832)
	WD	84.3%(728/864)	85.5%(739/864)	87.5%(756/864)	87.9%(756/860)
A5	ND	70.1%(283/404)	70.8%(286/404)	73.3%(296/404)	79.5%(295/371)
	WD	78.5%(317/404)	77.0%(311/404)	80.9%(327/404)	84.5%(326/386)
D1	ND	90.5%(143/158)	94.3%(149/158)	96.8%(153/158)	96.8%(153/158)
	WD	95.6%(151/158)	96.8%(153/158)	96.8%(153/158)	96.8%(153/158)
D2	ND	88.4%(175/198)	94.4%(187/198)	95.5%(189/198)	97.9%(185/189)
	WD	95.5%(189/198)	95.5%(189/198)	95.5%(189/198)	96.9%(187/193)
D3	ND	90.7%(341/376)	92.3%(347/376)	94.4%(355/376)	95.4%(353/370)
	WD	95.2%(358/376)	95.2%(358/376)	95.2%(358/376)	95.7%(356/372)
D4	ND	81.6%(399/489)	91.4%(447/489)	95.3%(466/489)	96.7%(463/479)
	WD	93.9%(459/489)	95.9%(469/489)	96.3%(471/489)	96.9%(468/483)
D5	ND	90.9%(209/230)	94.4%(217/230)	96.1%(221/230)	97.4%(220/226)
	WD	97.4%(224/230)	97.0%(223/230)	97.0%(223/230)	97.0%(223/230)
D6	ND	82.2%(221/269)	88.9%(239/269)	90.3%(243/269)	95.2%(240/252)
	WD	89.6%(241/269)	92.6%(249/269)	92.6%(249/269)	95.4%(248/260)
D7	ND	82.2%(425/517)	90.7%(469/517)	94.4%(488/517)	96.4%(485/503)
	WD	92.5%(478/517)	95.2%(492/517)	95.9%(496/517)	96.9%(493/509)
N1	ND	69.0%(342/496)	74.0%(367/496)	78.2%(388/496)	87.4%(376/430)
	WD	72.5%(361/498)	75.9%(378/498)	80.1%(399/498)	87.8%(387/441)
N2	ND	79.2%(357/451)	86.7%(391/451)	89.6%(404/451)	93.0%(396/426)
	WD	85.6%(386/451)	90.0%(406/451)	90.5%(408/451)	92.6%(401/433)
N3	ND	73.3%(291/397)	80.1%(318/397)	84.9%(337/397)	90.0%(331/368)
	WD	84.1%(334/397)	85.1%(338/397)	87.4%(347/397)	89.8%(343/382)
N4	ND	80.9%(499/617)	85.7%(529/617)	88.2%(544/617)	92.0%(530/576)
	WD	86.7%(535/617)	88.7%(547/617)	89.3%(551/617)	92.5%(540/584)
Total	ND	78.7%(5425/6890)	84.8%(5844/6890)	87.5%(6031/6890)	91.5%(5954/6505)
	WD	86.4%(5957/6892)	89.0%(6131/6892)	90.2%(6218/6892)	92.5%(6153/6651)

ND: without using dictionary

WN: using dictionary

M: Majority Voting

D: Dictionary Lookup

R: Rejection

Table 2: Postprocessing results on words from large clusters($DDM = dd1$)

ID	OPT	OCR	POSTPROCESSING		
		ACCURACY	M	M+D	M+D+R
A1	ND	66.8%(253/379)	84.7%(321/379)	88.7%(336/379)	94.8%(331/349)
	WD	76.6%(291/380)	89.0%(338/380)	89.5%(340/380)	93.4%(341/365)
A2	ND	64.4%(369/573)	89.0%(464/573)	86.7%(497/573)	97.0%(486/501)
	WD	75.1%(432/575)	88.5%(509/575)	91.5%(526/575)	95.6%(520/544)
A3	ND	70.7%(239/338)	78.1%(264/338)	82.5%(279/338)	87.4%(277/317)
	WD	80.1%(270/337)	84.0%(283/337)	86.1%(290/337)	90.3%(288/319)
A4	ND	69.0%(586/849)	73.4%(623/849)	76.4%(649/849)	83.7%(649/775)
	WD	79.7%(677/849)	84.7%(719/849)	87.4%(742/849)	90.2%(742/823)
A5	ND	63.2%(239/378)	68.5%(259/378)	70.6%(267/378)	76.2%(266/349)
	WD	73.4%(279/380)	77.6%(295/380)	80.3%(305/380)	83.5%(304/364)
D1	ND	87.3%(131/150)	93.3%(140/150)	96.0%(144/150)	98.6%(144/146)
	WD	96.7%(145/150)	97.3%(146/150)	97.3%(146/150)	98.7%(146/148)
D2	ND	79.6%(152/191)	88.5%(169/191)	94.2%(180/191)	96.7%(173/179)
	WD	89.1%(170/191)	93.7%(179/191)	94.8%(181/191)	95.1%(176/185)
D3	ND	84.7%(322/380)	92.1%(350/380)	95.8%(364/380)	96.5%(363/376)
	WD	96.1%(365/380)	96.8%(368/380)	96.8%(368/380)	97.1%(367/378)
D4	ND	72.7%(354/487)	86.9%(423/487)	90.8%(442/487)	94.6%(439/464)
	WD	87.1%(424/487)	93.0%(453/487)	95.1%(463/487)	95.4%(460/482)
D5	ND	80.8%(181/224)	89.7%(201/224)	93.3%(209/224)	96.8%(209/216)
	WD	93.3%(209/224)	95.5%(214/224)	97.3%(218/224)	98.2%(218/222)
D6	ND	62.4%(169/271)	77.9%(211/271)	82.3%(223/271)	89.6%(223/249)
	WD	62.4%(169/271)	77.9%(211/271)	82.3%(223/271)	89.6%(223/249)
D7	ND	73.7%(373/506)	86.2%(436/506)	89.9%(455/506)	93.2%(454/487)
	WD	87.6%(444/507)	90.9%(461/507)	93.3%(473/507)	95.0%(472/497)
N1	ND	57.1%(265/464)	66.2%(307/464)	73.9%(343/464)	88.6%(319/360)
	WD	57.3%(266/464)	66.2%(307/464)	73.9%(343/464)	88.6%(319/360)
N2	ND	69.1%(304/440)	78.4%(345/440)	84.1%(370/440)	93.7%(359/383)
	WD	76.1%(335/440)	83.9%(369/440)	87.5%(385/440)	92.6%(374/404)
N3	ND	67.0%(240/358)	78.5%(281/358)	83.8%(300/358)	91.0%(283/311)
	WD	77.7%(278/358)	83.5%(299/358)	85.5%(306/358)	88.3%(295/334)
N4	ND	70.2%(415/591)	78.7%(465/591)	83.6%(494/591)	91.6%(488/533)
	WD	78.8%(465/590)	86.1%(508/590)	87.0%(513/590)	90.5%(507/560)
Total	ND	69.8%(4592/6579)	79.9%(5259/6579)	84.4%(5552/6579)	91.1%(5463/5995)
	WD	79.3%(5219/6583)	86.0%(5659/6583)	88.4%(5822/6583)	92.3%(5746/6227)

ND: without using dictionary

WD: using dictionary

M: Majority Voting

D: Dictionary Lookup

R: Rejection

Table 3: Postprocessing results on words from large clusters($DDM = dd2$)

CLUSTER 16	---	DECISION: cats		
WImg.0608		OCR: Afinica's	CORRECTION: cats	TRUTH: Africa's
WImg.0044		Ahica's	cats	Africa's
WImg.0172		f*;ca's	cats	Africa's
WImg.0723		kica's	cats	Africa's
WImg.0133		A*ica's	cats	Africa's
WImg.0038		cats	cats	Africa's
CLUSTER 36	---	DECISION: tune		
WImg.0042		tune	tune	time.
WImg.0549		tine	tune	time
WImg.0738		tirrxe	tune	time
CLUSTER 38	---	DECISION: hats		
WImg.0796		has	hats	has
WImg.0801		bus	hats	has
WImg.0626		hats	hats	has
CLUSTER 41	---	DECISION: Rejected		
WImg.0714		Brazzaville	Rejected	Brazzaville
WImg.0325		13razzaville	Rejected	Brazzaville
WImg.0519		Brazzavdle	Rejected	Brazzaville
CLUSTER 43	---	DECISION: Pi		
WImg.0107		Pik	Pi	Pik
WImg.0188		Pi	Pi	Pik
WImg.0393		-	Pi	Pik
CLUSTER 45	---	DECISION: Rejected		
WImg.0219		.Mica.	Rejected	Africa.
WImg.0410		Airica,	Rejected	Africa,
WImg.0691		Afros,	Rejected	Africa,
CLUSTER 47	---	DECISION: notary		
WImg.0598		Sitars	notary	military
WImg.0045		notary	notary	military
WImg.0065		rniEtary	notary	military
CLUSTER 51	---	DECISION: Rejected		
WImg.0513		Mmst.er	Rejected	Minister
WImg.0106		Mmster	Rejected	Minister
WImg.0070		Moister	Rejected	Minister
CLUSTER 52	---	DECISION: Rejected		
WImg.0514		Boeta	Rejected	Botha
WImg.0242		Boffia	Rejected	Botha
CLUSTER 59	---	DECISION: Man		
WImg.0128		Man	Man	Van
WImg.0111		Van	Man	Van
CLUSTER 66	---	DECISION: Rejected		
WImg.0129		Heerden,	Rejected	Heerden,
WImg.0112		HeerdenF	Rejected	Heerden,
CLUSTER 68	---	DECISION: agam		
WImg.0718		agam	agam	again
WImg.0683		agam	agam	again
CLUSTER 79	---	DECISION: economy		
WImg.0788		economy	economy	economic
WImg.0825		egonomic	economy	economic
CLUSTER 80	---	DECISION: Rejected		
WImg.0526		Abican	Rejected	African
WImg.0656		A*ican	Rejected	African

Figure 8: Examples of postprocessing errors from page A1 with DD2 and WD

Hull

Fourth Annual Symposium on Document Analysis and Information Retrieval

April 24 - 26, 1995

Desert Inn Hotel
Las Vegas, Nevada

Sponsored by:

Information Science Research Institute

and

Howard R. Hughes College of Engineering



University of Nevada, Las Vegas

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