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 hull@cs.buffalo.edu

Abstract

OCR is an error-prone process when input images are degraded. Most current OCR techniques use linguistic information such as character n-grams or dictionaries to postprocess character recognition results. These methods essentially discard the input image after the character recognition is complete. This paper proposes a new 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 reliable 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.

Fourth Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, April 24-26, 1995, pp. 177-190.

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]).

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.

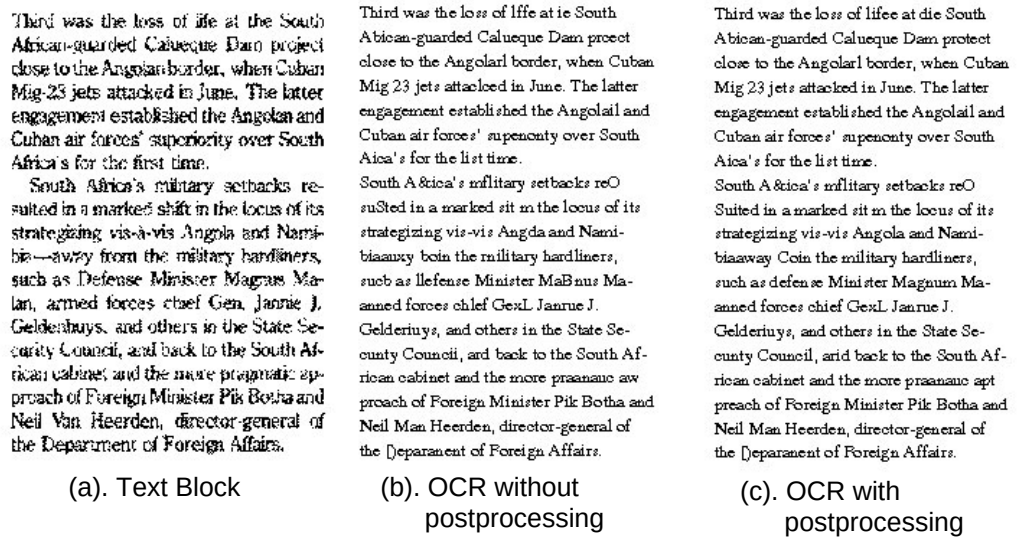


Figure 1: A fragment of text and its OCR results (with and without using a dictionary)

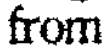
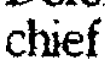
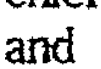
Word	Image	OCR (no Dict.)	OCR (with Dict.)
WImg.0007		ie	die
WImg.0012		prcect	protect
WImg.0016		Angolarl	Angolarl
WImg.0023		attaclced	attacked
WImg.0039		Aica's	Aica's
WImg.0042		list	list
WImg.0054		m	m
WImg.0061		Angda	Angola
WImg.0065		boin	Coin
WImg.0069		sucb	such
WImg.0071		11efense	defense
WImg.0077		chlef	chief
WImg.0090		ard	arid
WImg.0102		aw	apt
WImg.0103		proach	preach

Figure 2: OCR performance example (with or without using dictionary)

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 without 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 affected 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		mflitary
	and		rnilitary
	die		list
	the		first

Figure 3: Inconsistency of OCR results

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.

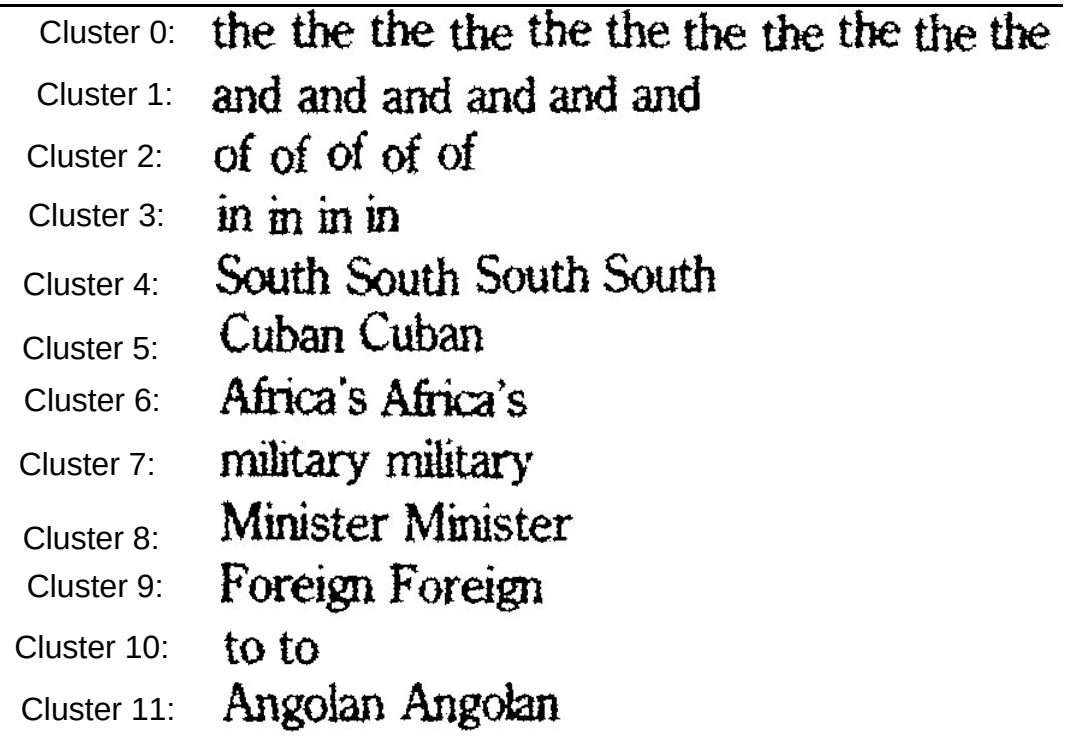


Figure 4: Word Image Clustering

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 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

```
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

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

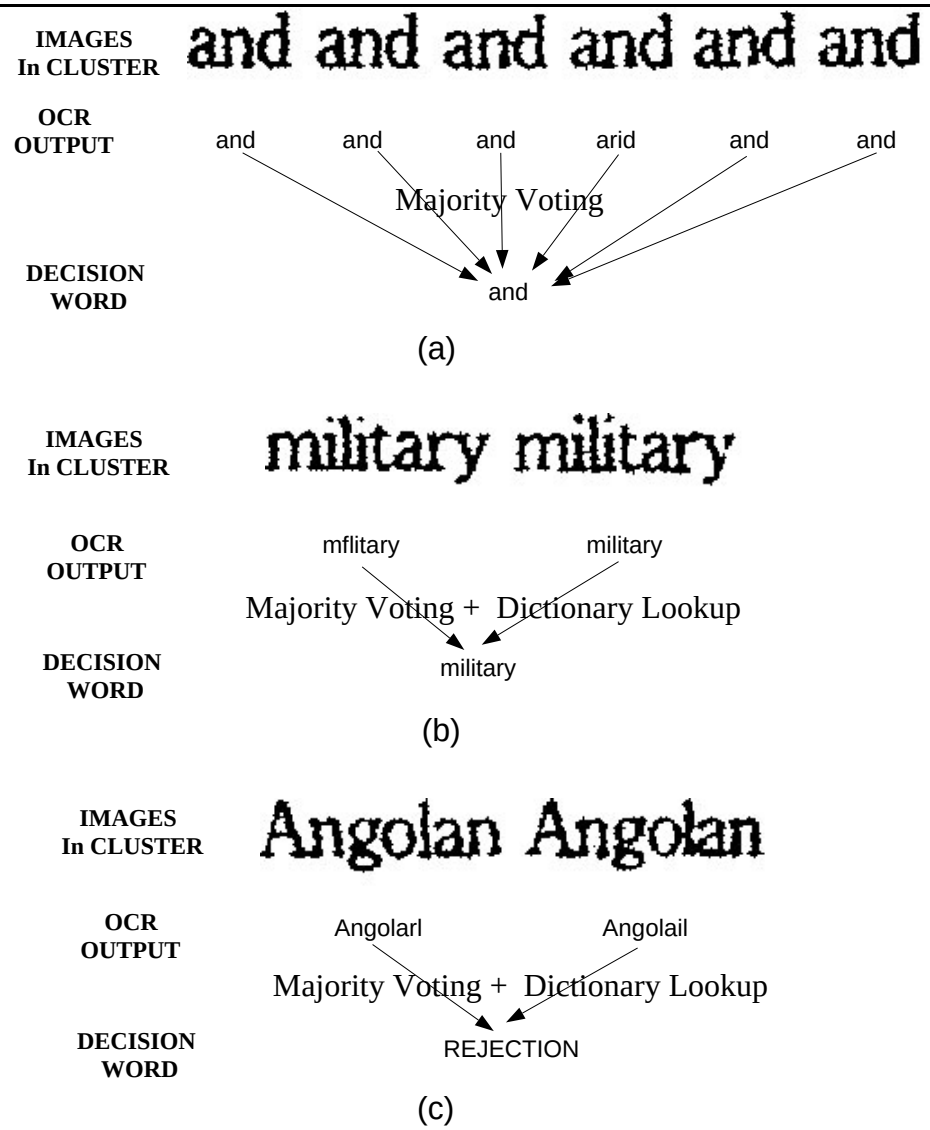


Figure 6: Selecting a Decision for a Cluster

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 which is available on the cdrom released by 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 words in a page are repeated at least once in some other position.

ID	TYPE	NUM OF WORDS	NUM OF REPEATED WORDS	NOTES
A1	Text	822	455	A1-A5 are from
A2	Text	1119	683	a five-page journal
A3	Text	564	355	article AA052532
A4	Text	1230	846	in CEDAR's database
A5	Text	735	444	
D1	Text	350	175	D1-D7 are from
D2	Text	469	217	a seven-page journal
D3	Text	693	406	articles DK074491
D4	Text	835	521	in CEDAR's database
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: Information about Sixteen Testing Document Pages

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 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.

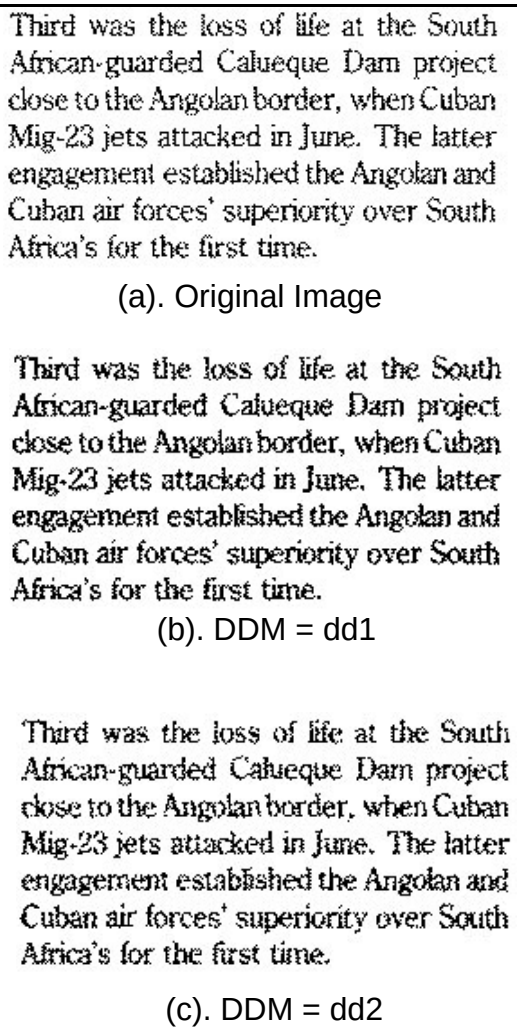


Figure 7: Image Degradation

Caere's AllFont OCR package was used to compute OCR results and the coordinates of bounding-boxes for the word images in each document. The coordinates of word bounding-boxes were used in our program for word image clustering. The truth data of the testing samples were generated based on the OCR output on the original images because the Caere OCR performance

on those high quality pages 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 71.38%. If the dictionary is used, the word recognition correct rate is 80.61%. For pages with noise at the $dd2$ level, the word recognition correct rate is 60.13% when the dictionary is disabled and 70.89% 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, which 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 clusters.

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 91.52% from the original 78.73%. For the OCR results on pages with noise at $dd1$ level and with using the dictionary, the third postprocessing strategy improved the correct rate to 92.51% from original 86.43%. For the OCR results on pages with noise at $dd2$ level and without using dictionary, the third postprocessing strategy improved the correct rate to 91.12% from original 69.80%. 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 92.28% from original 79.28%;

	OCR		POSTPROCESSING		
ID	OPT	ACCURACY	M	M+D	M+D+R
A1	ND	79.39%(339/427)	88.06%(376/427)	88.76%(379/427)	95.64%(373/390)
	WD	85.01%(363/427)	90.87%(388/427)	91.57%(391/427)	96.03%(387/403)
A2	ND	77.43%(501/647)	91.04%(589/647)	92.58%(599/647)	96.23%(587/610)
	WD	83.62%(541/647)	93.20%(603/647)	93.97%(608/647)	96.28%(595/618)
A3	ND	73.14%(256/350)	74.57%(261/350)	76.57%(268/350)	81.85%(266/325)
	WD	83.43%(292/350)	82.29%(288/350)	83.43%(292/350)	85.55%(290/339)
A4	ND	74.54%(644/864)	77.78%(672/864)	81.13%(701/864)	84.25%(701/832)
	WD	84.26%(728/864)	85.53%(739/864)	87.50%(756/864)	87.91%(756/860)
A5	ND	70.05%(283/404)	70.79%(286/404)	73.27%(296/404)	79.51%(295/371)
	WD	78.47%(317/404)	76.98%(311/404)	80.94%(327/404)	84.46%(326/386)
D1	ND	90.51%(143/158)	94.30%(149/158)	96.84%(153/158)	96.84%(153/158)
	WD	95.57%(151/158)	96.84%(153/158)	96.84%(153/158)	96.84%(153/158)
D2	ND	88.38%(175/198)	94.44%(187/198)	95.45%(189/198)	97.88%(185/189)
	WD	95.45%(189/198)	95.45%(189/198)	95.45%(189/198)	96.89%(187/193)
D3	ND	90.69%(341/376)	92.29%(347/376)	94.41%(355/376)	95.41%(353/370)
	WD	95.21%(358/376)	95.21%(358/376)	95.21%(358/376)	95.70%(356/372)
D4	ND	81.60%(399/489)	91.41%(447/489)	95.30%(466/489)	96.66%(463/479)
	WD	93.87%(459/489)	95.91%(469/489)	96.32%(471/489)	96.89%(468/483)
D5	ND	90.87%(209/230)	94.35%(217/230)	96.09%(221/230)	97.35%(220/226)
	WD	97.39%(224/230)	96.96%(223/230)	96.96%(223/230)	96.96%(223/230)
D6	ND	82.16%(221/269)	88.85%(239/269)	90.33%(243/269)	95.24%(240/252)
	WD	89.59%(241/269)	92.57%(249/269)	92.57%(249/269)	95.38%(248/260)
D7	ND	82.21%(425/517)	90.72%(469/517)	94.39%(488/517)	96.42%(485/503)
	WD	92.46%(478/517)	95.16%(492/517)	95.94%(496/517)	96.86%(493/509)
N1	ND	68.95%(342/496)	73.99%(367/496)	78.23%(388/496)	87.44%(376/430)
	WD	72.49%(361/498)	75.90%(378/498)	80.12%(399/498)	87.76%(387/441)
N2	ND	79.16%(357/451)	86.70%(391/451)	89.58%(404/451)	92.96%(396/426)
	WD	85.59%(386/451)	90.02%(406/451)	90.47%(408/451)	92.61%(401/433)
N3	ND	73.30%(291/397)	80.10%(318/397)	84.89%(337/397)	89.95%(331/368)
	WD	84.13%(334/397)	85.14%(338/397)	87.41%(347/397)	89.79%(343/382)
N4	ND	80.88%(499/617)	85.74%(529/617)	88.17%(544/617)	92.01%(530/576)
	WD	86.71%(535/617)	88.65%(547/617)	89.30%(551/617)	92.47%(540/584)
Total	ND	78.73%(5425/6890)	84.82%(5844/6890)	87.53%(6031/6890)	91.52%(5954/6505)
	WD	86.43%(5957/6892)	88.96%(6131/6892)	90.22%(6218/6892)	92.51%(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($DDM = dd1$)

	OCR		POSTPROCESSING		
ID	OPT	ACCURACY	M	M+D	M+D+R
A1	ND	66.75%(253/379)	84.70%(321/379)	88.65%(336/379)	94.84%(331/349)
	WD	76.58%(291/380)	88.95%(338/380)	89.47%(340/380)	93.42%(341/365)
A2	ND	64.40%(369/573)	80.98%(464/573)	86.74%(497/573)	97.01%(486/501)
	WD	75.13%(432/575)	88.52%(509/575)	91.48%(526/575)	95.59%(520/544)
A3	ND	70.71%(239/338)	78.11%(264/338)	82.54%(279/338)	87.38%(277/317)
	WD	80.12%(270/337)	83.98%(283/337)	86.05%(290/337)	90.28%(288/319)
A4	ND	69.02%(586/849)	73.38%(623/849)	76.44%(649/849)	83.74%(649/775)
	WD	79.74%(677/849)	84.69%(719/849)	87.40%(742/849)	90.16%(742/823)
A5	ND	63.23%(239/378)	68.52%(259/378)	70.63%(267/378)	76.22%(266/349)
	WD	73.42%(279/380	77.63%(295/380)	80.26%(305/380)	83.52%(304/364)
D1	ND	87.33%(131/150)	93.33%(140/150)	96.00%(144/150)	98.63%(144/146)
	WD	96.67%(145/150)	97.33%(146/150)	97.33%(146/150)	98.65%(146/148)
D2	ND	79.58%(152/191)	88.48%(169/191)	94.24%(180/191)	96.65%(173/179)
	WD	89.01%(170/191)	93.72%(179/191)	94.76%(181/191)	95.14%(176/185)
D3	ND	84.74%(322/380)	92.11%(350/380)	95.79%(364/380)	96.54%(363/376)
	WD	96.05%(365/380)	96.84%(368/380)	96.84%(368/380)	97.09%(367/378)
D4	ND	72.69%(354/487)	86.86%(423/487)	90.76%(442/487)	94.61%(439/464)
	WD	87.06%(424/487)	93.02%(453/487)	95.07%(463/487)	95.44%(460/482)
D5	ND	80.80%(181/224)	89.73%(201/224)	93.30%(209/224)	96.76%(209/216)
	WD	93.30%(209/224)	95.54%(214/224)	97.32%(218/224)	98.20%(218/222)
D6	ND	62.36%(169/271)	77.86%(211/271)	82.29%(223/271)	89.56%(223/249)
	WD	62.36%(169/271)	77.86%(211/271)	82.29%(223/271)	89.56%(223/249)
D7	ND	73.72%(373/506)	86.17%(436/506)	89.92%(455/506)	93.22%(454/487)
	WD	87.57%(444/507)	90.93%(461/507)	93.29%(473/507)	94.97%(472/497)
N1	ND	57.11%(265/464)	66.16%(307/464)	73.92%(343/464)	88.61%(319/360)
	WD	57.33%(266/464)	66.16%(307/464)	73.92%(343/464)	88.61%(319/360)
N2	ND	69.09%(304/440)	78.41%(345/440)	84.09%(370/440)	93.73%(359/383)
	WD	76.14%(335/440)	83.86%(369/440)	87.50%(385/440)	92.57%(374/404)
N3	ND	67.04%(240/358)	78.49%(281/358)	83.80%(300/358)	91.00%(283/311)
	WD	77.65%(278/358)	83.52%(299/358)	85.47%(306/358)	88.32%(295/334)
N4	ND	70.22%(415/591)	78.68%(465/591)	83.59%(494/591)	91.56%(488/533)
	WD	78.81%(465/590)	86.10%(508/590)	86.95%(513/590)	90.54%(507/560)
Total	ND	69.80%(4592/6579)	79.93%(5259/6579)	84.39%(5552/6579)	91.12%(5463/5995)
	WD	79.28%(5219/6583)	85.96%(5659/6583)	88.44%(5822/6583)	92.27%(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($DDM = dd2$)

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 clusters, 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 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.

```
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

6 Conclusions and Future Directions

A new 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 proposed 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.

References

[1] Henry 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, 1993.

[2] Su Chen, Suresh Subramaniam, Robert M. Haralick, and Ihsin T. Phillips. Performance evaluation of two ocr systems. In *Symposium on Document Analysis and Information Retrievals*, pages 299–317, 1993.

[3] G. DeSilva and J. J. Hull. Proper noun location in document images. *Pattern Recognition*, pages 311–320, February 1994.

[4] Jeffrey Esakov, Daniel P. Lopresti, Jonathan S. Sandberg, and Jiangying Zhou. Issues in automatic ocr error classification. In *Symposium on Document Analysis and Information Retrievals*, pages 401–412, 1993.

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

[6] Tao 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, 27-30 June 1994.

[7] Tao Hong and Jonathan J. Hull. Text recognition enhancement with a probabilistic lattice chart parser. In *Proceedings of the Second International Conference on Document Analysis (ICDAR-93)*, 1993.

[8] Tao Hong and Jonathan J. Hull. Degraded text recognition using word collocation. In *Proceedings of the Conference on Document Recognition of 1994 IS&T/SPIE Symposium*, pages 334–342, February 6-10 1994.

[9] Tao Hong and Jonathan J. Hull. Degraded text recognition using word collocation and visual inter-word constraints. In *proceedings of ANLP94*, October 1994.

[10] Jonathan Hull, Siamak Khoubyari, and Tin Kam Ho. Word image matching as a technique for degraded text recognition. In *Proceedings of 11th International Conference on Pattern Recognition*, 1992.

[11] Mark A. Jones, Guy A. Story, and Bruce W. Ballard. Integrating multiple knowledge sources in a bayesian ocr post-processor. In *Proceedings of ICDAR-91*, pages 925–933, 91. It shows the knowledge of the recognition device (OCR here) and the knowledge of the source material (English text) can be used to correct the output word from the pattern recognition system. It also discusses about how to integrate those different type informations. Very interesting article.

[12] Simon Kahan, Theo Pavlidis, and Henry S. Baird. On the recognition of printed characters of any font and size. *IEEE Transactions on pattern analysis and machine intelligence*, 9(2), 1987.

[13] Tapas Kanungo, Robert M. Haralick, and Ihsin Phillips. Global and local document degradation models. In *Proceedings of the Second International Conference on Document Analysis and Recognition ICDAR-93*, pages 730–734, 1993.

[14] Siamak Khoubyari and Jonathan J. Hull. Keyword location in noisy document images. In *Proceedings of the Second Annual Symposium on Document Anaysis and Information Retrieval*, Las Vegas, Nevada, April 26-28 1993.

[15] Theo Pavlidis. Problems in the recognition of poorly printed text. In *Symposium on Document Analysis and Information Retrievals*, pages 162–173, 1992.

[16] Stephen V. Rice, Junichi Kanai, and Thomas A. Nartker. An evaluation of ocr accuracy. In *1993 Annual Report of ISRI, University of Nevada, Las Vegas*, pages 9–34, 1993.

[17] Stephen V. Rice, Junichi Kanai, and Thomas A. Nartker. The third annual test of ocr accuracy. In *1994 Annual Report of ISRI, University of Nevada, Las Vegas*, 1994.

[18] S. N. Srihari. *Computer Text Recognition and Error Correction*. IEEE Computer Society Press, 1984.