

A Valiant Attempt to Reconstruct *Don Quixote*

A. Lawrence Spitz

DocRec Ltd, 34 Strathaven Place, Atawhai, Nelson, New Zealand

Email: spitz@docrec.com

Abstract

Despite the current practice of re-keying most documents placed in digital libraries, we continue to try to improve accuracy of automated recognition techniques for obtaining document image content. This task is made more difficult when the document in question has been rendered in letterpress, subjected to hundreds of years of the ageing process and been microfilmed before scanning.

In supporting the accurate capture of textual information, we endeavoured to leave intact a previously described document reconstruction technique, where we combine information from a language model and character image pattern matching to iteratively reduce ambiguity in document images. Combining word shape information and lists of similar bitmap patterns in a document at least partially resolves the character content without optical character recognition.

We enhance the document image to bring the perceived production values up to a more modern standards in order to process a novel of historic importance: *Don Quixote* by Miguel de Cervantes Saavedra. Pre-processing of the page images before application of the reconstruction techniques was performed to accommodate early 17th century typography and low-quality scanned microfilm images.

Though our technology easily outstripped the capabilities of commercial OCRs, it too was found lacking, at this stage of development, for automated processing of historical documents for digital libraries. We had hoped to develop a useful transcription of the text and a lexicon of Spanish contemporary with the composition of this novel. However the actual accomplishment was limited to making improvements in the recognizability of the page images involved and providing a basis for further research.

1 Introduction

This paper is presented in two parts: a review of the previously developed document reconstruction process and the specific application of this basic technology to the problem of reconstructing a 400 year old document of historic significance.

Reconstruction [10] is based on an integration of information from multiple sources: word shape, character bitmaps and lexical information. Implicit non-lexical language model information is dynamically inferred based on the character shape distributions.

We have obtained 8 digitizations, from microfilm, of the novel *Don Quixote* performed by the Biblioteca Nacional, Spanish Royal Academy, Oxford, Harvard and Yale Universities, the British Library, and two copies from the Hispanic Society of America.

The process of dealing with these problematic images, difficult production values and unusual linguistic content are described in Section 3.

1.1 Document Reconstruction

Reynar, Spitz and Sibun [4] previously described a process of reconstructing a document from its image without resorting to Optical Character Recognition (OCR), leaving ambiguities to be resolved by a downstream process or by a human reader. This process relies on a low computational cost transformation of character images into Character Shape Codes (CSCs).

The document reconstruction technique (RECON) described here is an integration of information from multiple sources: word shape, character bitmaps and lexical information. Implicit non-lexical language model information is dynamically inferred based on the character shape distributions.

Because of its inability to deal with non-lexical tokens such as number strings and punctuation, document reconstruction cannot ever be the sole basis for conversion of documents to for use in digital libraries. However for largely text documents such as novels, document reconstruction has been shown to provide a significant contribution to the faithful rendering of the text of the document.

In Section 2.1 we will describe the character shape coding process and the word shape encoding resulting from the agglomeration of these CSCs. In Section 2.3 we will describe the collection of character bitmaps and the construction of lists of similar bitmaps. Section 2.4 shows the data structure that represents the pattern and word-based references to the lists of patterns

Section 2.5 will describe the initial labelling of these lists with character codes. Section 2.6 describes the progressive reduction of ambiguity on a word-by-word basis.

We present preliminary accuracy results in Section 2.7 and present some conclusions in Section 4.

1.2 Application of RECON to *Don Quixote*

In this paper we apply RECON to the 6 pages that comprise Chapter 5.

In Section 3.1 we will describe in some detail the challenge in terms of production values and image quality. Section 3.2 will show the results of processing these difficult images using commercial OCRs. Starting in Section 3.3 we will describe the special purpose image processing that was required in order to be able to reconstruct this document.

Section 3.8 describes the structure and contents of two lexica used in the reconstruction process.

We present preliminary accuracy results in Section 3.10 and present some conclusions in Section 4.

2 RECON

2.1 Character shape coding

The CSCs encode whether or not the character in question fits between the baseline and the x-line or if not, whether it has an ascender or descender and the number and spatial distribution of the connected components. For a more complete discussion of this process see [8]

Table 1: Definitions of Character Shape Codes for alphabetical characters

Characters	CSC	Definition
A-Zbdhkl	A	ascender
acemnor suvxwz	x	between baseline and x-line
gpqy	g	descender
i	i	x-height plus mark above
j	j	descender plus mark above

There are a few CSCs, not shown here, that represent punctuation.

An implicit part of the shape coding function is the development of a data structure describing the hierarchy of text line, word box and character cell coordinates.

CSCs are used internally within the document reconstruction process and their word-level aggregations into Word Shape Tokens (WSTs) are used as indices in lexica.

The results of CSC processing are shown in figure 1.

2.2 Lexical information

A RECON lexicon is simply a series of word lists, each indexed by a common WST. For a more complete discussion of lexical structure see [8]. Though in operation RECON is not totally reliant on the correctness of the WST, its performance is highly dependent on this transformation. In many instances

chapter I 2 LOOMINGS Call me Ishmael. Some years ago--never mind how long

chapter I 2 LOOMINGS Call me Ishmael. Some years ago--never mind how long

xXgAxx A A AAAAAAAAA AxAA xx AxAxxxA. Axxx gxxxx xgx--xxxxx xixA Axx Axxg

Figure 1: Character codes, their printed and scanned representation and the result of the character shape coding process.

there may be a single surface form word that maps to the WST thereby defining all of the character positions simultaneously. However even if there are multiple words in the lexical entry, there are often some character positions where the character identity is unambiguous.

We encode this ambiguity, or lack of it, in two data structures. The Degenerate Regular Expression (DRE) shows the possible alternative characters in each character position, while the Simplified Regular Expression (SRE) merely shows the fact that ambiguity does or does not exist in each character position, viz.:

Table 2: Ambiguity resulting from character shape coding

WST	gAxxxxx
Lexical entry	glances glasses phrases pleases
DRE	[gp][lh][are][nsa][cs]es
SRE	?????es

whereas where there is a single word in the lexicon with that shape:

Table 3: Some words have no ambiguity

WST	gixAxxixA
Lexical entry	pictorial
DRE	pictorial
SRE	pictorial

2.3 Character bitmaps

We process the line, word, character cell structure by comparing the bitmaps found in each character cell with all extant lists of characters. Using a very strict criterion for declaring a match between a candidate bitmap with the bitmap characterizing the list, we either accept the candidate as a member of that list, or if no match is found, start a new list.

We assessed the method of Xu and Nagy [11] for constructing exemplar templates. Though it worked well, its improvement over that described in detail in [8] was marginal and the time penalty was intolerably large, so the method described in [8] was retained.

When we have finished generating the pattern lists, we sort them so that the longest list, that with the greatest

number of matching bitmaps, is first and the populations of the lists monotonically decrease. We show in figure 2 the first 10 instances of each list. Note that at this stage we do not yet know the character identity for the lists. Also note that there might be more than one list containing slightly different bitmaps of a particular character due to the tight threshold for inclusion in an extant list.t.

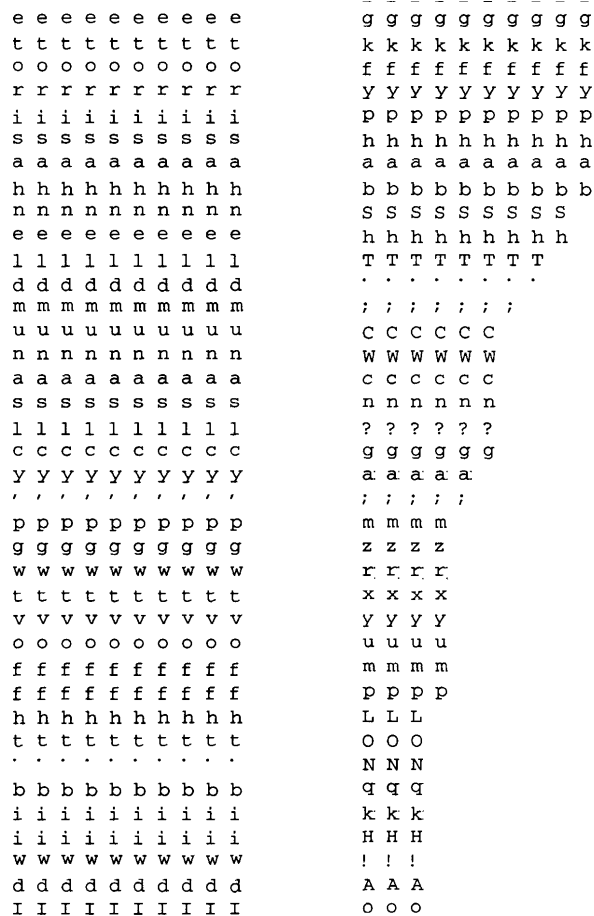


Figure 2: Character cell bitmap lists in order of decreasing frequency

From this point forward it is presumed that there are no lists containing bitmaps arising from different characters.

2.4 Data structure

The data structure used for storing information is shown graphically in figure 3. Each character image is linked to the ccell element of the word structure. Each member of the pattern list is likewise linked to its ccell. Thus it is possible to make associations both from the shape coded words to the patterns and from the patterns to the words.

2.5 Labeling bitmap lists

Looking at the unambiguous character positions in the SREs we label the corresponding individual bitmap in the lists with the character code. If no mistakes were

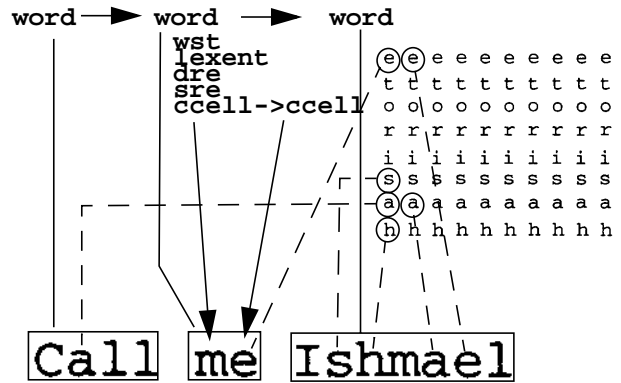


Figure 3: Representation of the data structures linking the character images to both the words and the pattern lists.

made in character shape coding and if the lexicon is adequate, all of the bitmap list entries will be labelled correctly (if at all).

However neither CSC generation nor lexical contents are perfect, sometimes resulting in character positions being labelled when they should not have been.

The lists themselves are now labelled if at least 2 individual bitmaps are consistently labelled and there are no inconsistent labels. If there are inconsistencies in the labelling, the list is assigned the label of the preponderance of bitmap labels and those bitmaps with the inconsistent labels are marked for further processing. Preponderance is defined as the number of consistent labels being greater than 4 times the number of inconsistent labels

Note again that there may be more than one list with a particular character label. And it is possible that some lists, particularly short ones, will remain unlabelled. Lists labelled with the same character are concatenated if any pair of characters, one from each list, have a small enough difference. This allows for different degradations of the character form to be handled, even if the characters on the heads of the two lists differ by an amount that exceeds the threshold.

Unlabelled lists are likewise concatenated with labelled ones if a (nearly) matching pair of bitmaps is found.

Once a list has been labelled all of the individual bitmaps on that list are likewise labelled. Doing this materially reduces ambiguity.

Resolution of infrequently occurring characters such as capitals is difficult because they occur so infrequently.

2.6 Word recognition

It may be that the labelling of the pattern lists will result in all of the characters in a word also being labelled. If not, and only some of the characters are labelled, the lexical entries for the WST are examined

and those that are inconsistent with the defined characters are removed. Often this will result in characters becoming defined that have not yet been directly examined. From this new information it is possible to label some previously unlabelled pattern lists.

Each reduction of the lexical entry results in a recalculation of the SRE and DRE.

It may also happen that the application of the character identities will result in the removal of all of the elements of the lexical entry indicating either a CSC error or the presence of a word not in the lexicon. In either case the word is marked for character by character resolution without further recourse to the lexicon.

Residual ambiguity is reflected in the DRE. In each character position, the bitmaps on the appropriately labelled lists for the legitimate alternative characters can be checked against the character cell bitmap. Each time character position ambiguity is resolved, the lexical entries are pruned.

At the end of the process the defined character bitmaps are shown in figure 4

a	b	c	d	e	f	g	h	i	l	m	n	o	p	r	s	t	u	v	w	y	z
	C					G	I		L	M	N	O		S				W			

Figure 4: Character bitmaps from first page of “Moby Dick”

An example of the entire pattern lists has already been shown in figure 2. Unresolved, that is unlabelled, pattern lists are shown in figure 5

T	T	T	T	T	T	T	T	k	k
?	?	?	?	?	?			2	
r	r	r	r					R	
x	x	x	x					P	
y	y	y	y					N	
q	q	q						j	
k	k	k						a	
!	!	!						Y	
:	:	:						x	
A	A	A						Y	
t	t							B	
r	r							H	
H	H							r	
B	B								

Figure 5: Unlabeled pattern lists

2.7 Accuracy

Alphabetic character accuracy exceeds 99.4%, Word accuracy is approximately 96%. Most of the errors are due to missing labels from capital letters.

Since RECON relies on a lexicon, its performance will be somewhat affected by the appropriateness of that lexicon to the document being processed. See Spitz [8] for a discussion of lexical intersection and specificity.

3 Tackling *Don Quixote*

Encouraged by the excellent results on modern printing of *Moby Dick*, we went on to try to apply the same principles to *Don Quixote*.

3.1 Typography

There are several challenges in the production values used in the printing of this book.

Chapters start with an illuminated drop capital character followed by a capitalized second character. figure 6 shows the rendering of the initial word of Chapter 5 “Viendo”.



Figure 6: Chapters start with an illuminated drop cap (in this case: V) followed by a capitalized second

There appear to be no intentional ligatures in the text such as **fl**, but the image degradation was severe enough that many character pairs touched, and in the italic chapter and page headings long runs of touching characters were not uncommon.

The **t** in the font used has a very small ascender. It is not unusual for **t** ascenders to be shorter than those for **b**, **d**, **h**, **k** etc., but in this font it is very difficult to consistently classify **ts** as either ascender or x-height characters.

Hyphenation is inconsistent. The text was set fully justified but in some instances a word broken across lines was hyphenated and sometimes not. See for example, figure 7. It was hard to devine the rules, if

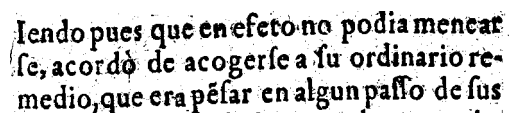


Figure 7: Note that the last word on the first line “menearse” is not hyphenated while the last word on the next line is hyphenated.

there were any, for when hyphenation was used or not.

There were multiple examples of words that were broken across pages. In this instance they were hyphenated and the terminal part of the word appeared right justified on the last line of the page as well as at the top of the following page.

There are no paragraph indents or vertical spacing, and no quotation marks.

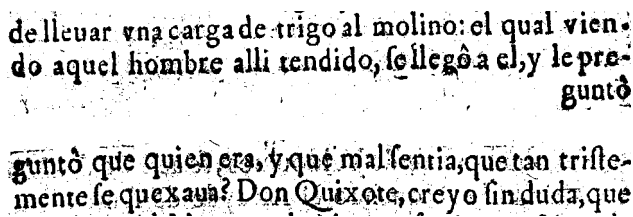


Figure 8: The word “pregunto” broken across pages. The top panel shows the bottom of Page 15 Verso, and the bottom panel shows the top of Page 16 Recto. Note that “gunto” appears redundantly. Also note the deep kerning of the initial letter of “Quixote” in the second line of 16R.

3.2 Conventional OCR

We submitted partially processed page images to two commercial OCR packages. These images had been segmented to isolate the text from the page edges and had the italics and illuminated drop caps removed. These same images were used as input to the OCRs and the RECON process.

The OCRs used were Omnipage Professional 9.0 and Finereader Pro 6.0. The character and word accuracy rates are shown in table 4. Clearly Omnipage inserted many bogus characters while Finereader deleted a significant number of legitimate character images. In any case the word accuracy rates are sufficiently low to render these data useless for a digital library (or any other) application.

Table 4: Character and word accuracy for two OCRs.

	Number of Characters	Character Accuracy (%)	Number of Words	Word Accuracy (%)
Truth	8800		1598	
Omnipage	15212	40.6	3228	0.12
Finereader	6330	28.5	774	1.2

3.3 Pre-recognition Processing

The images supplied us were scanned from microfilm with two facing pages per frame. It was necessary to develop document-specific techniques for segmenting the individual pages and for reducing the noise level in the document images. These techniques turned out to be pretty trivial. First a connected component analysis was performed and then all components too tall to be of text were removed. This had the effect of removing the black areas outside of the page frames and illuminated drop caps as well, but since we had no technology to deal with drop caps anyway, this was considered to be appropriate for the application. The page was deskewed based on a single angle for the whole page and projection profiles of the remaining connected components determined the boundaries of the text frame.

The image processing specifically developed for this project took the form of a pre-processing filter on the input image. This was done in order to obviate the need to adapt the reconstruction processor to the aberrant typography.

The spatial resolution varied such that there was a range of approximately 14 to 45 pixels in the x-height. This results in a font height of 30 to 100 pixels. The

Table 5: Resolution characteristics of the 8 scans of the Don Quixote text

Source	x-height (pixels)	font height (pixels)	dpi at 12 pt
bd	19	48	288
bm	27	60	360
bn	22	48	288
hs1	45	100	600
hs2	26	58	348
hu	14	30	180
ra	21	43	258
yl	14	32	192

image quality by modern standard was poor, due to the degradation inherent in microfilming and subsequent scanning, the paper and print quality and the fact that the book had been subjected to nearly 400 years of aging. We noted that the best perceived image quality was not, in this instance, associated with the highest spatial resolution.

3.4 Noise Removal

There are multiple sources of noise in the images we analysed, the two principal ones being the speckle created by the printing process and those inherent to microfilming. We dealt with speckle in much the same way as Cannon, *et al.* described [1].

There is considerable print-through on some pages.

3.5 Line Straightening

All of the images are of bound volumes resulting in significant warp of some text lines as they near the gutter. In addition, apparently the pages were often distorted by application of a clear platen, resulting in a distortion that can be characterized as continuously variable skew as a function of position on the page. The variable skew was removed first using techniques described in Spitz [7]. The line warp was removed in the process of character shape coding described in [5].

3.6 Word Spacing

Word spacing is highly variable and in many instances is so tight as to make it hard to distinguish word spacing from letter spacing. The use of a lexicon-based recognizer requires that we be able to accurately delimit word images. A partial solution to this problem was effected by applying a non-linear expansion of the

inter-character spacing to make small increments somewhat larger.

However the text was set without a word space following a comma as is current practice. Here it was necessary to explicitly insert a word space after the shape coding process had detected a comma.

3.7 Noise Cancellation

An attempt was made to take advantage of the multiple images available of the same pages. It was hoped that by scaling and superimposing images that noise would drop out. We first cropped the page frames to the text and scaled them to the same dimensions. We then multiplied the images together.

The result obtained from superimposing the two best images of page 16 recto is shown in figure 9. Note that though the images are very closely superimposed at both the top and bottom of the page, the vertical centre region of the page is not in registration. The reason for this result is unknown but is likely to be due to differential and non-linear paper shrinkage though it might also be due to non-affine optical distortion in the microfilming process.

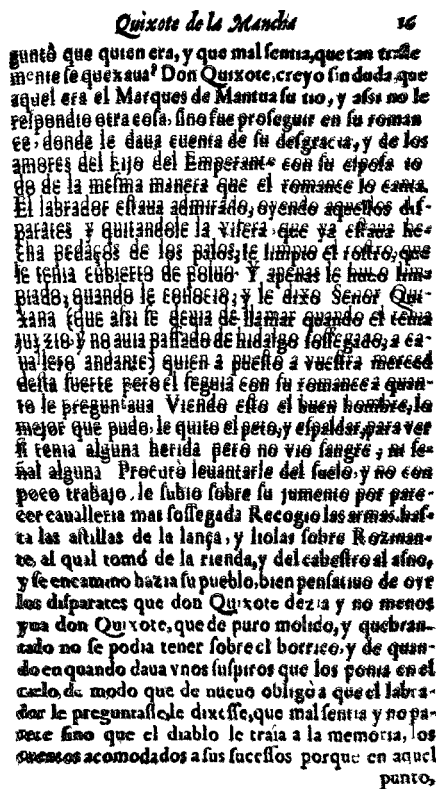


Figure 9: Result of attempt to register two images of the same page showing good registration at the top and bottom, poor registration in the middle due to

Kerning, especially on characters like Q leads to deep overlaps in character bounding boxes. Baselines wander. An attempt was made to correct all of these

image characteristics by enforcing constant intra-word spacing (thereby dekerneling character pairs), straightening baselines, inserting word spaces after a comma and inserting leading to keep descenders and ascenders from tangling. The results of this processing are shown in figure 10

que quien era, y que mal sentia, que tan triste
se quezaua? Don Quixote creyo sin duda que

que quien era, y que mal sentia, que tan triste
se quezaua? Don Quixote creyo sin duda que

Figure 10: Separating image elements for more accurate word detection. Note that the i and ? dots have been removed in the noise reduction process.

3.8 Lexicon

Don Quixote was written around the turn of the 17th century using Spanish spellings some of which are not acceptable by today's standards. For this reason, use of a modern lexicon is inappropriate. Were it available, we would use a lexicon consistent with the language and spelling prevalent at the time the book was written. However we had no access to such a lexicon.

As Spitz has shown [6], the performance of our reconstruction technique is improves with the use of the optimum lexicon: one with a large intersection of the words in the document and with a minimal number of words not found in the document.

For this application we developed the lexicon iteratively, starting with a bootstrap list of frequently-occurring words: articles, pronouns and some proper nouns, shown in figure [5].

Dulzinea Mancha Mantua Marques Quixote Rozinante Valdouinos a al como con cura de del don donde dos el ella ello ellos en es estas este esto estos la las le lo los mas me mi muchas muy ni no o por qual quando quatro que se señor señora si su tio todo y yo

Figure 11: 54 common words in *Don Quixote* comprising the bootstrap lexicon

This tiny lexicon contains 48.5% of the words found in Chapter 5. The shape coded lexicon is shown in table 6. Note that all of the uncapitalized words shown in figure 11 are present in the lexicon in both uncapitalized and capitalized form. Thus the size of the lexicon expands from 54 words to 101 surface forms.

These 101 surface forms are indexed by 44 WSTs, 20 of which are singleton representations. That is, 20 WSTs unambiguously define the lexical form that they encode

Table 6: WST indexed bootstrap lexicon (abbreviated for space)

WST	Words
A	A O Y
AAAx	Ella Ello
Aix	Tio tio
Ax	De En Es La Le Lo Me No Se Su Yo de la le lo
AxAxixxxx	Valdounos
AxAx	Este Esto Todo todo
AxAxixxxx	Dulzinea
AxAxx	Estas Estos
Axixx	Señor
AxixxAx	Quixote
Axixxxx	Señora
Axx	Con Don Dos Las Los Mas Por Que don dos las los
AxxAx	Donde donde
AxxAxx	Mantua Muchas Quatro
AxxixxAx	Rozinante
AxxxAx	Mancha Quando
g	Y
gxx	por que
x	a o
xxAx	este esto
xxAxx	estas estos
xxg	muy
xxxAx	muchas

3.9 Character Bitmaps

We process the line, word, character cell structure by comparing the bitmaps found in each character cell with all extant lists of characters. Because it is very important that a list contain only the bitmaps of a single identity character (see exception below) we employ a very strict criterion for declaring a match between a candidate bitmap with the bitmap characterizing the list, we either accept the candidate as a member of that list, or if no match is found, start a new list.

When we have finished generating the pattern lists, we sort them so that the longest list, the one with the greatest number of matching bitmaps, is first and the populations of the lists monotonically decrease. We show in figure 12 the first 10 instances of each list. Note that at this stage we do not yet know the character identity for the lists. Also note that there might be more than one list containing slightly different bitmaps of a particular character due to the tight threshold for inclusion in an extant list.

The highly variable character morphology extant in these images, combined with the strict matching criterion, resulted in many lists being generated for each intended glyph. Since the **long s**, which looks much like an **f** without a full crossbar, was indistinguishable from the **f** using the RECON bitmap comparison technique, we lumped the **long s** and **f** together in the bitmap domain and resolved the

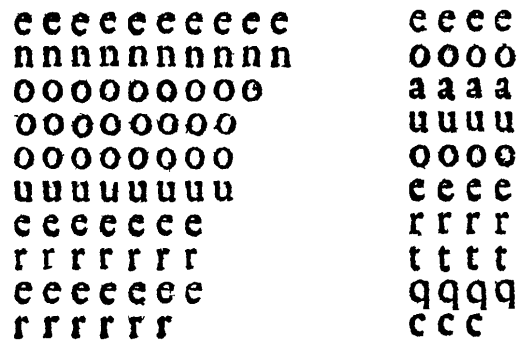


Figure 12: Character cell bitmap lists in order of decreasing frequency. Note duplication.

ambiguity at the lexical level. Examples of the two character forms are shown in figure 13.

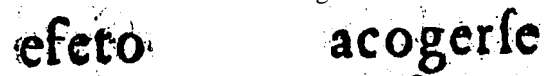


Figure 13: Two words from the text. The word on the left contains an f. The word on the right contains a long

3.10 Accuracy

This is an instance where it is very difficult to know the truth. The author was supplied with eight manually produced transcriptions for Chapter 5. These transcriptions differ in small but significant ways. For tests of character accuracy we used the transcription supplied with the Hispanic Society digitizations.

Since RECON relies on a lexicon, its performance will be affected by the appropriateness of that lexicon to the document being processed. See Spitz [8] for a discussion of lexical intersection and specificity.

Word accuracy using the document reconstruction technique with the bootstrap lexicon was 11.4%. This abysmal result was only notable with respect to the even greater failures exhibited by the commercial OCRs.

4 Conclusions

Integration of linguistically-based and image-based information can be quite tricky but is extremely powerful. Rigorous congruity of character bitmap pattern lists allows extension of the reconstruction even to those words not found in the lexicon.

RECON works well on documents of significant length such as those of one page or more but because it relies on having a statistically significant number of character bitmaps, is not suitable for very short documents.

This paper does not present scientific results of great significance. Indeed while it is still possible that the paradigm of pre-processing page images to suit the RECON process, and to post-process the character code information returned from RECON, will result in

a useful transcription of the source text, this paper merely sets out some of the problems encountered when *traditional* document recognition techniques were applied to the difficult images of *Don Quixote*.

We posit that application of these techniques to longer passages of text than the single chapter available to us at this time, would show that the progressive resolution paradigm used here would increase performance, both in speed and accuracy, the longer the text.

Having multiple digitizations of the same pages has allowed us to do explore techniques to take advantage of the redundancy of information in those images: something smarter than processing them all and taking the results from those that produce the “best” results. Since there are several variables between digitizations it might be possible to overlay these images in an attempt to cancel out the noise in the images.

5 Acknowledgments

The document images are the property of the various sources credited in the introduction, and were obtained from Prof. Richard Furuta of Texas A & M University.

Thanks also go to Luke Hutchison of Brigham Young University for his application of the Fourier-Mellin transform to multiple images in an attempt to register them.

6 References

- [1] Michael Cannon, Judith Hochberg, Patrick Kelly, “Quality assessment and restoration of typewritten document images”, *International Journal on Document Analysis and Recognition*, 2, 2/3, pp 80-89, 1999.
- [2] David J. Ittner, David D. Lewis, David D. Ahn, “Text Categorization of Low Quality Images”, *Symposium on Document Analysis and Information Retrieval*, Las Vegas, pp 301-315, 1995.
- [3] G. Nagy, S. Seth & K. Einspahr, “Decoding substitution ciphers by means of word matching with application to OCR”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9, 5, pp 710-715, 1987.
- [4] Jeffrey C. Reynar, A. Lawrence Spitz, Penelope Sibun, “Document Reconstruction: A Thousand Words from One Picture”, *Symposium on Document Analysis and Information Retrieval*, pp 367-385, 1995.
- [5] A. Lawrence Spitz, “Generalized Line, Word and Character Finding”, in *Progress in Image Analysis and Processing III*, S. Impedovo ed., World Scientific, Singapore, pp 377-383, 1994.
- [6] A. Lawrence Spitz, *Moby Dick* meets GEOCR: Lexical considerations in word recognition, *International Conference on Document Analysis and Recognition*, Ulm, Germany, 1997.
- [7] A. Lawrence Spitz, “Analysis of compressed document images for dominant skew, multiple skew and logotype detection”, *Computer Vision and Image Understanding*, 70, 3, pp 321-334, 1998.
- [8] A. Lawrence Spitz, “Shape-based Word Recognition”, *International Journal of Document Analysis and Recognition*, pp 178-190, 1999.
- [9] A. Lawrence Spitz, J. Paul Marks, “Measuring the robustness of character shape coding”, in *Lecture Notes in Computer Science*, Nakano and Lee ed., Springer-Verlag, 1655, pp 1-12, 1999.
- [10] A. Lawrence Spitz, “Progress in document reconstruction”, *International Conference on Pattern Recognition*, Quebec City, pp 464-467, 2002.
- [11] Yihong Xu, George Nagy, “Prototype extraction and adaptive OCR”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99-01-07, pp 1280-1296, 1999.