

Shape-Free Statistical Information in Optical Character Recognition

Scott Leishman Sam Roweis

Machine Learning Group Talk
University of Toronto

April 2nd, 2007

Optical Character Recognition

Definition

Optical Character Recognition (OCR) is the process by which digital images of textual symbols are translated into a machine-readable representation.

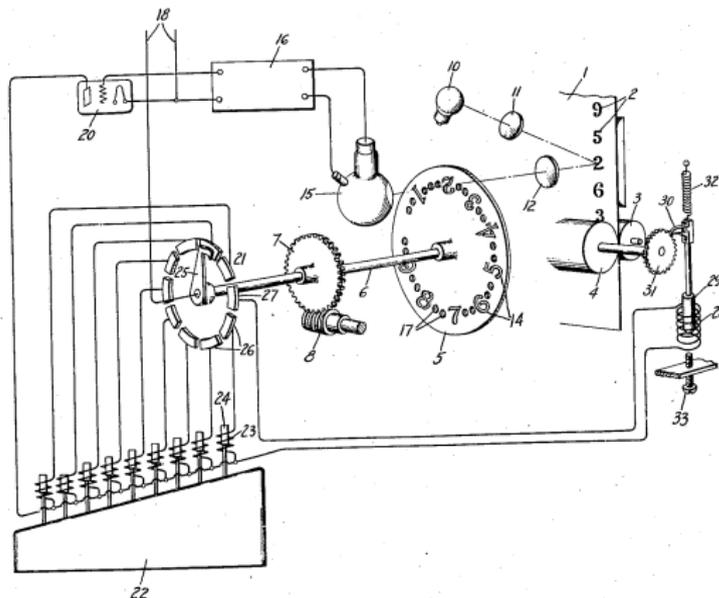


e
0x65



Winning the Washington State Lottery's Scratch game is one of the nicer things that can happen in life, as opposed to some of the little things that can go wrong. That's the ...

The Early Years - "Optical" Recognition Systems



System Proposed in Handel's 1933 Patent Filing

- Classic pattern recognition problem
- Idea dates back at least to patent filings in the early 1930's
- Initial uses included telegraph processing and as an aid to the blind

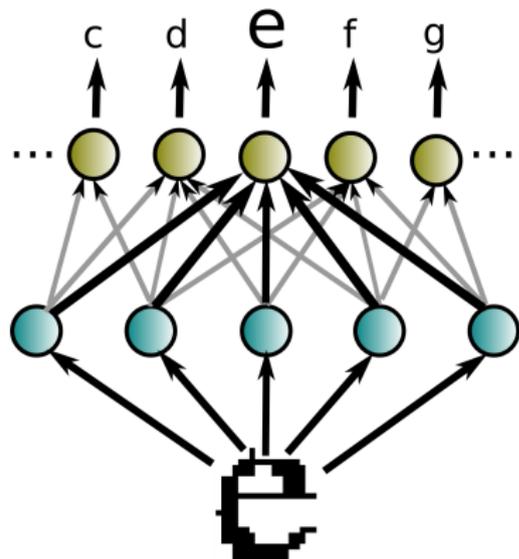
The "Middle" Ages - Constrained Template Matching

A	B	C	D	E	F	G	H	I	J	K	L	M	N
O	P	Q	R	S	T	U	V	W	X	Y	Z	a	b
c	d	e	f	g	h	i	j	k	l	m	n	o	p
q	r	s	t	u	v	w	x	y	z	0	1	2	3
4	5	6	7	8	9	!	@	#	\$	%	^	&	*

The OCR-A Font

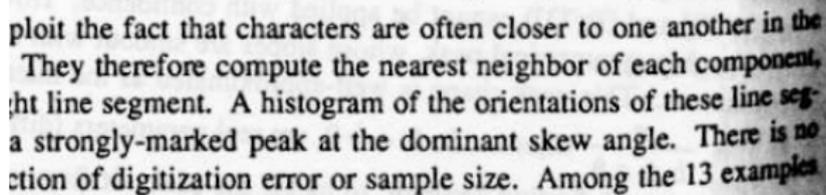
- First commercial OCR systems appear in the mid 1950's
- System designs heavily influenced by computer logic circuitry and electronics
- Input documents extremely constrained
- Most systems used template matching (limited to a few fonts and sizes)

Current Systems - Visual Classification



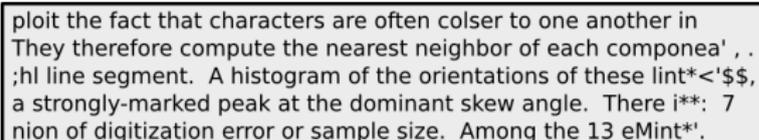
- Neural network or other shape-based classifiers trained on a large variety of fonts.
- Incorporation of some contextual information after initial recognition phase
- Many commercial offerings, some claiming accuracy rates $> 99\%$
- Problem solved??

Some Problem Cases - Noisy Curled Characters



exploit the fact that characters are often closer to one another in the
They therefore compute the nearest neighbor of each component,
right line segment. A histogram of the orientations of these line seg-
a strongly-marked peak at the dominant skew angle. There is no
tion of digitization error or sample size. Among the 13 examples

Acrobat



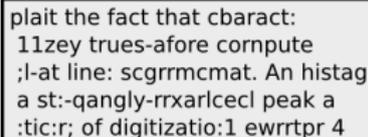
exploit the fact that characters are often colser to one another in
They therefore compute the nearest neighbor of each componea', .
;hl line segment. A histogram of the orientations of these lint* < '\$\$,
a strongly-marked peak at the dominant skew angle. There i** : 7
nion of digitization error or sample size. Among the 13 eMint*1.

Any2DjVu



a stangly-madced peak at the dominant skew angle.

Tesseract



plait the fact that cbaract:
11zey trues-afore compute
;l-at line: scgrrmcmat. An histag
a st:-qangly-rrxarlcecl peak a
:tic;r; of digitizatio:1 ewrrtpr 4

Some Problem Cases - Textured Background



Acrobat



Any2DjVu



Tesseract



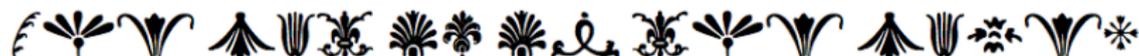
abcdef 012345 ! \$ % " ,

Acrobat
Any2DjVu
Tesseract

abeder 0L2345
g;;;sa;;;g;;m:zu...> \& `~li{\\$lll;z;;;1ij;;;:g;;=..h 4===au=;}}}

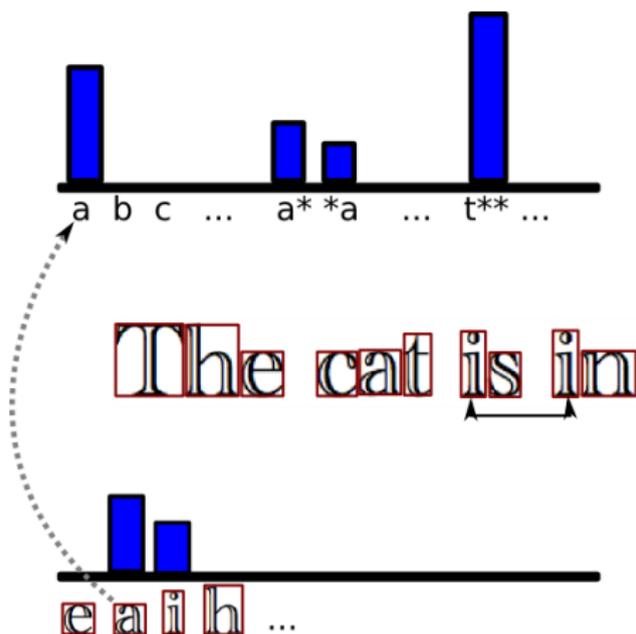
- Clearly these systems have not been trained on a font in this style/size!

The cat is in the cage.



- Ideally want a system that **adapts** to each document to be recognized
- Exploit language consistency within a document (true unless dealing with dictionaries, machine translation papers, etc.)
- Exploit glyph shape and font consistency within a document (relatively true for most cases except things like ransom notes and font manuals)

Our Approach



- Estimate frequency distributions of character co-occurrences from a large text corpus
- Locate and cluster together similar looking glyph images from a document to be recognized
- Estimate frequency distributions over the clusters
- Determine the mapping between these two

Preprocessing - Denoising

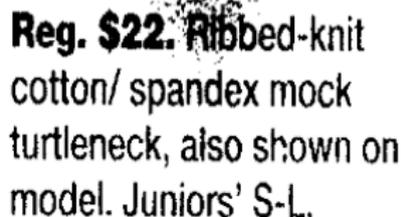
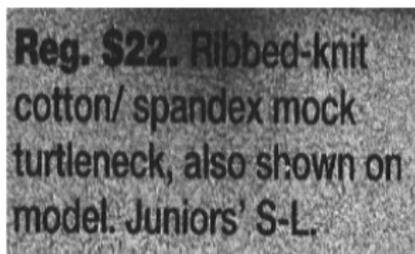
- Input document images can contain noise from a number of sources
 - Fax line-noise
 - Scanning Sensor noise
 - Other unwanted marks (ex. staples, large book gutters)
- Find components and use aspect ratio to throw out very large or very small objects thus removing additive noise (must be careful not to throw out small punctuation symbols)
- Convolve image to smooth over dropout pixels (must be careful not to join non-touching symbols, fill small holes)

Bayesian

Bayesian

Preprocessing - Binarization

- Input document images often given in full colour or grayscale
- Reducing them to two intensities improves processing speed and certain algorithms require binary images (e.g. Hausdorff distance)
- Methods fall under one of two categories: global or local methods
- No magic bullet!

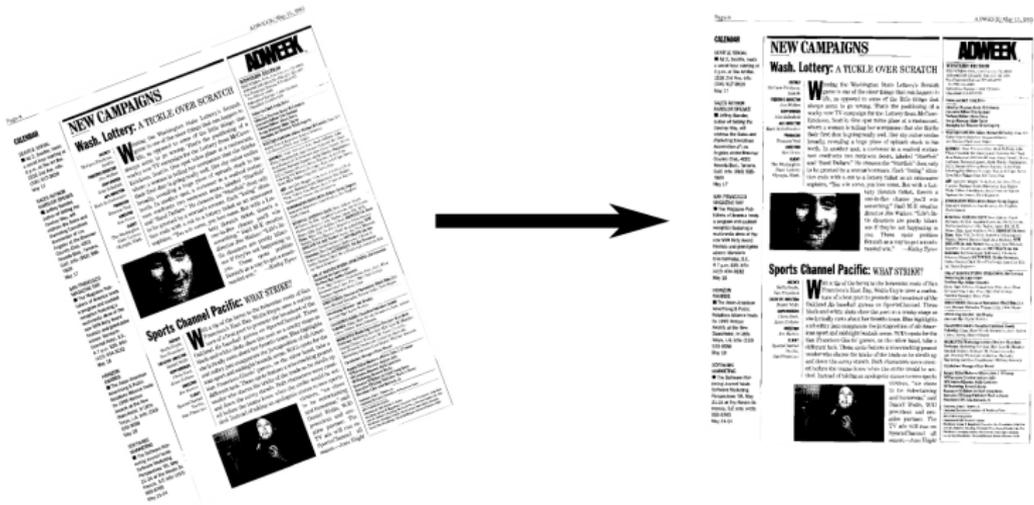


Reg. \$22. Ribbed-knit
cotton/ spandex mock
turtleneck, also shown on
model. Juniors' S-L.

The image shows the same grayscale photograph as the previous one, but it has been converted to a binary (black and white) image. The text is now much sharper and more legible, with the background being mostly white and the text being black. The text reads: "Reg. \$22. Ribbed-knit cotton/ spandex mock turtleneck, also shown on model. Juniors' S-L."

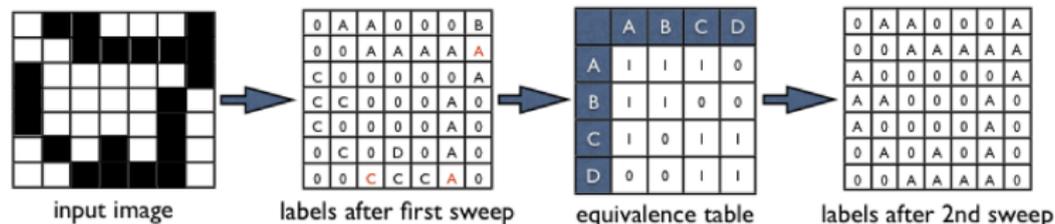
Preprocessing - Page Deskewing

- Manually placed scanned pages typically vary up to $\pm 15^\circ$
- Some algorithms for line finding and other processing tasks will not work correctly on skewed documents
- Many early approaches based on the generalized Hough transform[Duda1972]



Isolating Individual Symbols - Connected Components

- Simple 2-pass sweep used to group pixels into connected components[Rosenfeld1966]
- Bounding box co-ordinates calculated and stored



we can implicitly integrate out the infinitely many

Isolating Individual Symbols - Neighbour / Line Finding

- Nearest neighbouring component (and distance) calculated for each component in 4 principal directions (requires 2 sweeps over component label image)
- Lines found by following chains of neighbouring components, then baseline and x-height offsets estimated from profile sums
- Attempts made to merge diacritical and other small vertical components belonging to the same symbol like i , \acute{e} , $:$, $=$, $?$

~~To accordance with Article III of the By-Laws of the Mesa Village Homeowners Association, and Paragraph 6 of the Declaration of Covenants, Conditions, and Restrictions for the property, notice is hereby given that the Annual Meeting of the Mesa Village Home-~~

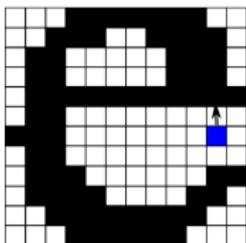
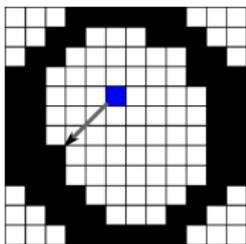
Clustering Connected Components

- Bottom-up, agglomerative clustering used
- Component pixel intensities compared to determine clustering
- Distance metrics considered include Euclidean, Hausdorff
- For nearly noiseless documents, initial sweep using small threshold performed to rapidly reduce number of clusters

~~In accordance with Article III of the By-Laws of the Mesa Village Homeowners Association, and Paragraph 6 of the Declaration of Covenants, Conditions, and Restrictions for the property, notice is hereby given that the Annual Meeting of the Mesa Village Home-~~



Hausdorff Distance Metric



- Would prefer to weight mis-matched pixel intensities so that “on” pixels lying further away from the nearest “on” pixel in the comparison image are charged a larger cost
- Hausdorff distance[Huttenlocher1992] does this, measuring the distance between two images A, B as follows:

$$D_H = \max(h(A, B), h(B, A)) \quad (1)$$

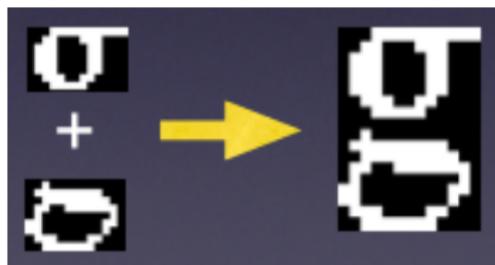
where

$$h(X, Y) = \max_{i \in X'} (\min_{j \in Y'} (d(X(i), Y(j)))) \quad (2)$$

and $d(x, y)$ is replaced by a distance metric like Euclidean distance, X' denotes the set of foreground pixels of X

Merging Fragmented Symbols

- After one round of match based processing, cluster information used to piece together fragmented glyphs
- Clusters containing components who share nearest neighbours belonging to the same cluster are marked for merger
- Provided they lie only a small distance apart, belong to the same line, and the neighbours in turn list components in the original cluster as their neighbour, then a merger is performed between these components
- Symbol segmentation often carried out as part of the recognition process



Splitting Apart Touching Symbols

- Clustering information also used to try and split components containing multiple symbols
- Candidate horizontal split points found based on component width and vertical projection sum
- Halves are recursively searched for matches against other cluster centroids (using Hausdorff distance for example)



Refining the Clusters

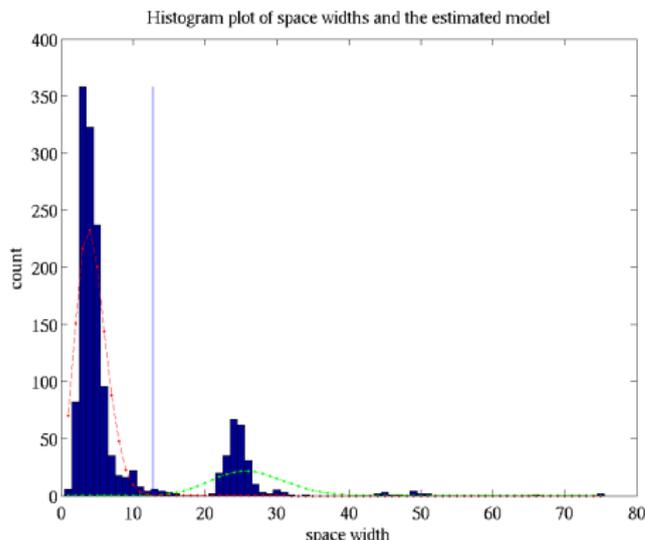
	e	t	a	i	o	n	s	r	h
2001	227	193	161	249	344	131	122	93	77
d	c	l	C	f	m	p	g	u	I
69	67	57	47	47	42	48	37	36	32
·	n	·	r	b	M	-	()	w
25	25	22	23	18	17	16	34	34	34
S	Y	·	z	5	v	E	o	A	T
34	13	12	12	12	11	11	11	11	10
P	V	mm	R	N	4	WM	a	4	c
10	18	18	9	9	9	8	8	7	7
2	D	P	k	3	7	x	6	j	1
6	6	6	5	4	4	3	3	3	3
·	9	0	3	u	mm	na	og	o	ru
3	2	2	2	2	1	1	1	1	1
rm	RA	w	b	L	H	-	un	M	8
1	1	1	1	1	1	1	1	1	1
Z	n	l	l	w	7	Q	U	R	r
1	1	1	1	1	1	1	1	1	1
^	Y	M	·						
1	1	1	1						

- Matching, merging, and splitting process is repeated over affected clusters, until no further changes are seen
- Conservative thresholds used to prevent cluster impurities (at the expense of leaving multiple clusters per character)
- No real attempts made at rescaling cluster averages

```

eeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeeee
oooooooooooooooooooooooooooooooooooooooooooooooooooooooooooo
ttttttttttttttttttttttttttttttttttttttttttttttttttttttttt
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
rrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrrr
aaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaaa
sssssssssssssssssssssssssssssssssssssssssssssssssssssssss
iiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiii
hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
lllllllllllllllllllllllllllllllllllllllllllllllllllllllllll
  
```

Determining Word Boundaries



- Determining word boundaries is critical for accurate contextual estimation since our approach makes use of within word positional frequency
- For most textual documents, the distribution over neighbouring horizontal component distances is bimodal with the smaller representing intercharacter spacing, and the second interword
- We attempt to determine cut-off point by fitting a mixture model to these frequency counts

Symbol Decoding Using Positional Frequency

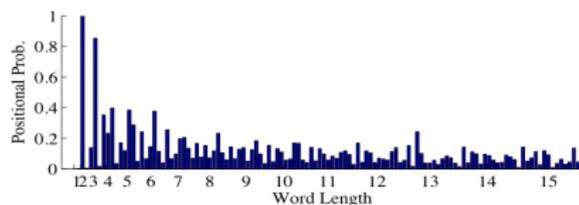
~~In accordance with Article III of the By-Laws of the Mesa Village Homeowners Association, and Paragraph 6 of the Declaration of Covenants, Conditions, and Restrictions for the property, notice is hereby given that the Annual Meeting of the Mesa Village Home-~~

23 7 2 15 15 10 12 29 2 7 15 1 25 4 3 11 55 12 3 4 15 13 1 23 23 23 10 19 3 11 1 44 32 67 56 2 25 6 10 19 3 11 1 33 1 6 2 62 4 13 13 2 24 1
58 10 30 1 10 25 7 1 12 6 55 6 6 10 15 4 2 3 4 10 7 66 2 7 29 54 2 12 2 24 12 2 20 11 49 10 19 3 11 1 40 1 15 13 2 12 2 3 4 10 7 10 19
...

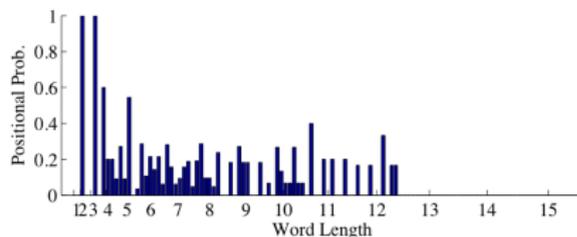
- Neighbours in reading order give cluster id sequence
- Common words like *the*, *of*, *and*, *it* dominate documents that are written in grammatically correct English prose (Zipf's "law")
- Certain character n -grams are much more common in English words than others (contrast *ing* with *zzz*)
- Given word breaks, certain characters often occur in particular word positions; uppercase letters tend to be seen in the first position of a word, letters like *s* and punctuation symbols like *.* are common choices for the last position of a word.

Positional Frequency (Cont'd)

- We exploit these regularities by estimating the **word positional frequencies** for each symbol
- Estimated first for each symbol in a large labelled text corpus, then on each cluster using component and estimated word boundary information
- Counts normalized to create distributions over each word length (unless no counts seen of that symbol at a particular word length)



e



e

Matching Observed Cluster Vectors to Reference Vectors

- Counts define a point in an $\frac{x(x+1)}{2}$ dimensional “positional” feature space (where x is max word length considered)
- Mapping from observed counts to reference counts can be carried out via likelihood or cross entropy minimization
- Our current approach
 - Normalize counts within each word length
 - Feature vectors re-weighted based on word-length to minimize affect of wild positional differences on seldom seen long words
 - Euclidean distance measured between each cluster feature vector and symbol vectors to find closest map

Improving the Mappings via Dictionary Lookup

- Positional mapping alone often provides good match for frequently seen symbols like lower case vowels and some consonants, however other symbols regularly exhibit high variance between cluster and corresponding reference vector
- Causes for this include the short length of most input documents coupled with document type and subject matter
- Use dictionary lookup to improve matchings

the α uick

$\alpha = ?$

Dictionary Lookup Procedure

- Greedily assign clusters to symbols based on occurrence frequency
- Try candidates in order based on positional feature distance

~~In accordance with Article III of the By-Laws of the Mesa Village Homeowners Association, and Paragraph 6 of the Declaration of Covenants, Conditions, and Restrictions for the property, notice is hereby given that the Annual Meeting of the Mesa Village Home-~~

237 2 15 15 10 12 29 2 7 15 1 25 4 3 11 55 12 3 4 15 13 1 23 23 23 10 19 3 11 1 44 32 67 56 2 25 6 10 19 3 11 1 33 1 6 2 62 4 13 13 2 24 1
58 10 30 1 10 25 7 1 12 6 55 6 6 10 15 4 2 3 4 10 7 66 2 7 29 54 2 12 2 24 12 2 20 11 49 10 19 3 11 1 40 1 15 13 2 12 2 3 4 10 7 10 19

...

237 a 15 15 10 12 29 a 7 15 e 25 it 11 55 12 ti 15 13 e 23 23 23 10 19 t 11 e 44 32 67 56 a 25 6 10 19 t 11 e 33 e 6 a 62 i 13 13 a 24 e
58 10 30 e 10 25 7 e 12 6 55 6 6 10 15 i a t i 10 7 66 a 7 29 54 a 12 a 24 12 a 20 11 49 10 19 t 11 e 40 e 15 13 a 12 a t i 10 7 10 19

...

Dictionary Lookup Procedure

- Lookup each partially assigned word cluster sequence in which this cluster appears, and check for it in the dictionary (use wildcards for unmapped clusters)
- If at least some threshold of these words match, permanently map this symbol to each component in this cluster
- If multiple symbols produce the same word lookup score, ties are first broken based on line offset information (each symbol belongs to one of 4 classes: ascenders, descenders, normal, and short x-height)
- If ties still remain, or no symbol achieves the threshold score, closest positional match taken as final symbol

Experimental Setup

ISRI

ID:702-895-1185

DEC 13 '95

9:58 No.017 P.01

BY LAWS OF

_____ OWNERS ASSOCIATION

ARTICLE I

PURPOSE AND POWERS

Section 1.01. Name. The name of the Corporation is _____.

Section 1.02. Purpose. The Corporation shall have such purposes as are now or may hereafter be set forth in its Articles of Incorporation.

Section 1.03. Powers. The Corporation shall have such powers as are now or may hereafter be granted by the Nonprofit Corporation Act of the State of Nevada. The laws applicable to other Nevada rights, privileges and duties thereunder shall apply to private corporation organized under chapter 79 of NRS and all professional corporations, except where such laws are in conflict with or inconsistent with the provisions of the chapter under which the corporation is organized. In case of conflict, the provisions of the chapter under which the corporation is organized shall apply.

ARTICLE II

MEMBERS

Section 2.01. Membership. The membership of the corporation shall consist of all Condominium Owners in the Project, as defined in that Declaration of Covenants, Conditions and Restrictions for _____ recorded on _____.

is _____, as Instrument No. _____, Book No. _____, Official Records, Office of the County Recorder, Clark County, Nevada, and all amendments thereto (hereinafter the "Covenant").

Section 2.02. Certificate And Transfer Of Membership. The corporation shall issue a certificate of membership to each member, but the certificate thereof shall not be transferred, pledged, assigned or alienated in any way; provided, however, that upon the sale of any Condominium the corporation shall cancel the certificate of the seller and shall issue a new certificate to the buyer thereof. Any prohibited transfer shall be void and shall not be reflected upon the books and records of the corporation. In the event any owner of a Condominium shall fail or refuse to surrender his certificate of membership upon sale of his Condominium, the corporation shall have the right to record the transfer of the

- Symbol alphabet contains 92 different symbols including upper and lower case letters, digits, punctuation, brackets, simple arithmetic operators
- Reference positional symbol counts and the word lookup dictionary were constructed from a 17,601 word chunk of the Reuters-21578 News corpus[Lewis2004] (744,522 symbols). All symbols left intact, but trailing "Reuter" byline removed
- Initial input tests performed against 10 15 page Legal documents from the UNLV ISRI OCR dataset[Nartker2005] (fine fax mode dpi)

	% Correct
word	92.49
character	90.72
low letters	97.79
upper letters	4.02
digits	7.51
other sym.	31.44

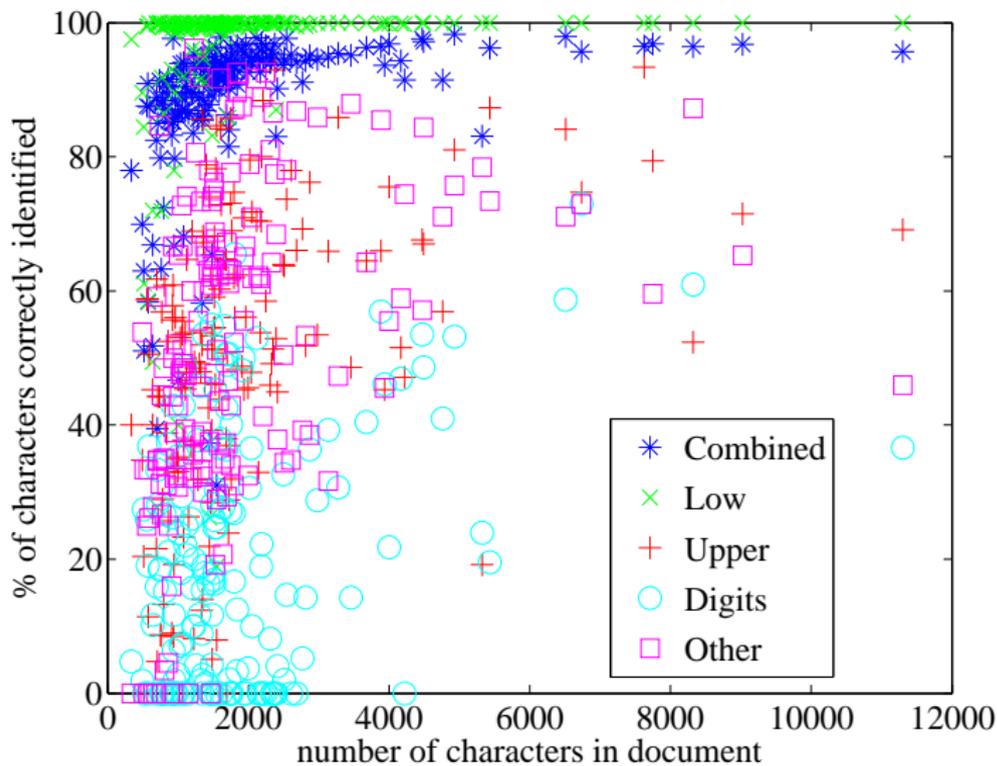
- Text aligned to ground truth as best as possible, then string edit operations used to determine accuracy by class
- Results for lowercase letters found to be roughly on par with shape based approaches, but other symbols and overall performance worse
- Lowercase letters dominate this dataset (84.2% of all symbols)

Impact of Segmentation and Clustering on Performance

	Regular	Perfect
word	92.49	95.30
character	90.72	96.07
low letters	97.79	100
upper letters	4.02	65.66
digits	7.51	23.64
other sym.	31.44	61.72

- To determine impact of segmentation and clustering on the performance, the ASCII codes of each symbol were used to group symbols together
- This represents a perfect clustering (no split or merged symbols, only a single cluster for each distinct symbol)

Impact of Document Length on Performance



Short Document Results

- 159 short 1-2 page business letters sampled from the UNLV ISRI OCR dataset[Nartker2005] (fine fax mode dpi)
- 2010 symbols per document on average

	Bus.	Bus (Perfect)	Legal	Legal (Perfect)
word	50.50	79.91	92.49	95.30
character	67.84	88.24	90.72	96.07
low letters	73.17	95.71	97.79	100
upper letters	9.21	50.65	4.02	65.66
digits	6.84	20.52	7.51	23.64
other sym.	24.55	52.36	31.44	61.72

Conclusions and Future Work

- Can't build a complete system using context alone
- Works well when characters appear enough times that they start to somewhat approximate their reference counts
- Positional counts and word lookup scores provide a fairly useful source of information to the recognition process, something that isn't being exploited much by current approaches
- Biggest improvements can be had by improving segmentation performance (rather than improving contextual collection techniques). Majority of errors due to merged symbols not being separated during clustering phase
- Future work involves exploring more detailed models for statistical information comparison

Advantages Of Our Approach

- Big gain is that our approach is font and resolution independent (though we have tested using baseline and x-height offset to improve tie-breaking)
- Can also be re-targetted to other phonetic languages by plugging in appropriate lookup dictionary
- Works (faster) on symbolically compressed documents (see next slide)

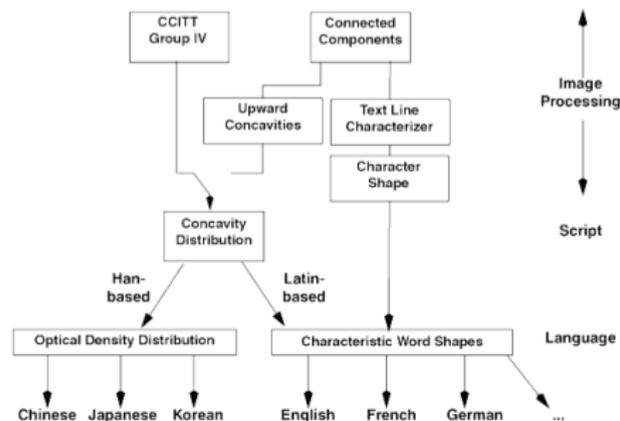
Symbolically Compressed Documents

- JBIG2 image compression standard for binary images, specially suited for images that are composed of repeated subimages (like textual document images)
- Lossless compression scheme stores a single template image, as well as co-ordinate offsets on each page, and image differences at those offsets
- Lossy versions also store templates and offsets, but don't store the residual image differences
- Used in PDF (1.4 and higher), DjVu, xpdf, and others
- Our OCR approach can work directly with these compressed documents without having to perform clustering (though split and merge refinements may be required)

- Cryptogram decoding[Nagy1984]
- Ho and Nagy Symbol Class identification
- Huang Entropy based approach[Huang2006]

Automatic Script and Language Determination

- Recognizing glyphs via statistical language features is only possible if we know the underlying language of the input region
- Possible to distinguish Han scripts from Latin-based scripts by measuring frequency and height of upward concaving runs of pixels [Spitz1997]
- Able to distinguish amongst 23 Latin-based languages using character codes from frequently occurring word images



	English		French		German	
WST	Rank	Word	Rank	Word	Rank	Word
Aax	1	the				
xA	2	of	7			
Ax	3	to	1	la le	6	
ix	4	is			10	
xxA	5	and			3	auf
xx	6		2	en	2	an
Axx	9		3	les	1	der das
xxx	8		4	aux	5	wer
gxx			5	pas		
Aix					4	die