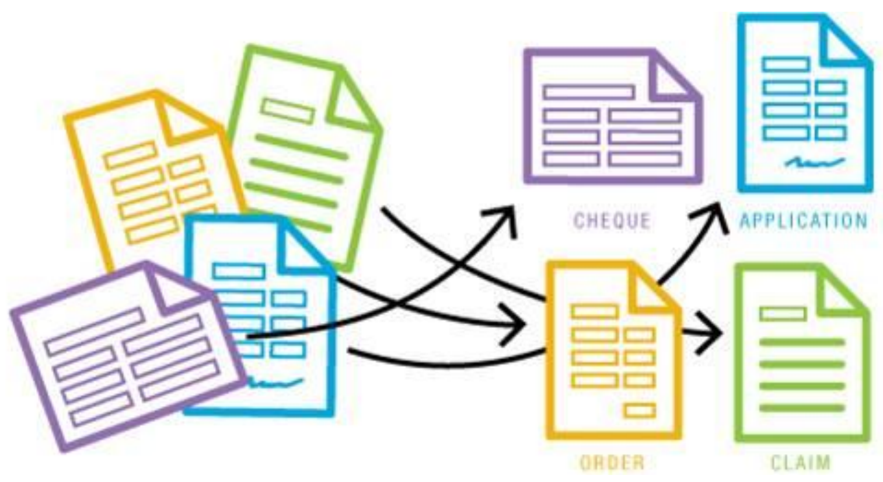
OCR-Based Document Classification

Gillian Pan
Mark Greene, Dr. Yash Shrivastava
School of Electrical and Information Engineering
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES
THE UNIVERSITY OF SYDNEY

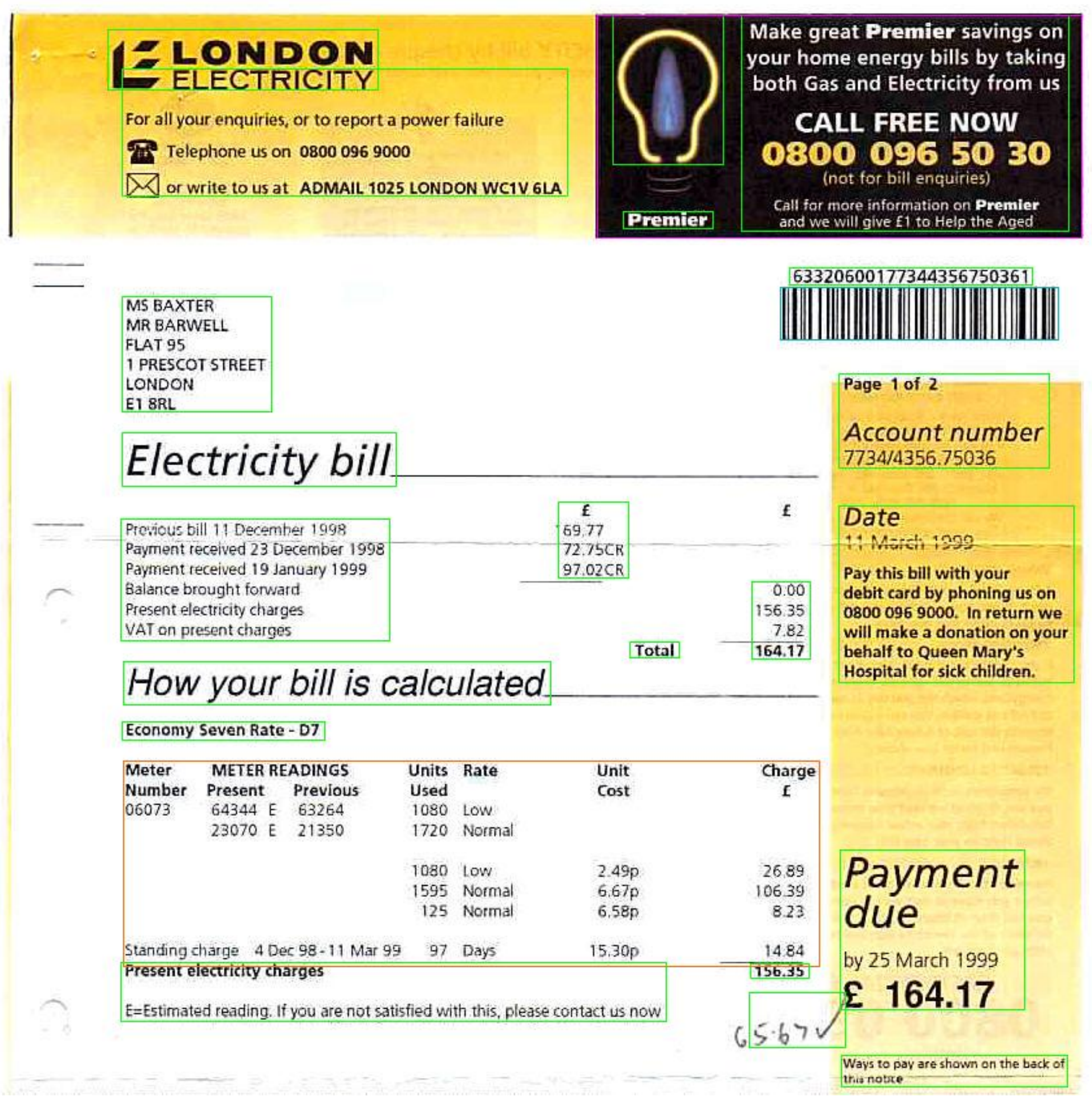
DOCUMENT CLASSIFICATION

- Document classification is the sorting of documents into predefined categories.
- **T**he trend of the paperless office in the recent years prompted a large growth in more efficient document processing and retrieval, driven largely by the incentive of reducing labor costs.
- Document classification has been studied previously to support the classification of wartime archival documents, medical journals, identity documents, forms and financial documents
- Toshiba hopes to integrate document classification into their e-BRIDGE Re-Rite product to enable workflow automation.



OPTICAL CHARACTER RECOGNITION (OCR)

- Extracts text , blocks from scanned images
- Segments the Text in region blocks



LONDON ELECTRICITY
For all your enquiries, or to report a power failure
☎ Telephone us on 0800 096 9000
✉ or write to us at ADMAIL 1025 LONDON WC1V 6LA

Make great **Premier** savings on your home energy bills by taking both Gas and Electricity from us
CALL FREE NOW
0800 096 50 30
(not for bill enquiries)
Call for more information on **Premier** and we will give £1 to Help the Aged
Premier

63320600177344356750361

MS BAXTER
MR BARWELL
FLAT 95
1 PRESCOT STREET
LONDON
E1 8RL

Page 1 of 2
Account number
7734/4356.75036

Electricity bill

	£	£
Previous bill 11 December 1998	69.77	
Payment received 23 December 1998	72.75CR	
Payment received 19 January 1999	97.02CR	
Balance brought forward		0.00
Present electricity charges		156.35
VAT on present charges		7.82
Total		164.17

Date
11 March 1999
Pay this bill with your debit card by phoning us on 0800 096 9000. In return we will make a donation on your behalf to Queen Mary's Hospital for sick children.

How your bill is calculated

Economy Seven Rate - D7

Meter Number	METER READINGS Present		Previous	Units Used	Rate	Unit Cost	Charge £
06073	64344	E	63264	1080	Low		
	23070	E	21350	1720	Normal		
				1080	Low	2.49p	26.89
				1595	Normal	6.67p	106.39
				125	Normal	6.58p	8.23
Standing charge	4 Dec 98 - 11 Mar 99			97	Days	15.30p	14.84
Present electricity charges							156.35

E=Estimated reading. If you are not satisfied with this, please contact us now

65·67✓

Payment due
by 25 March 1999
£ 164.17

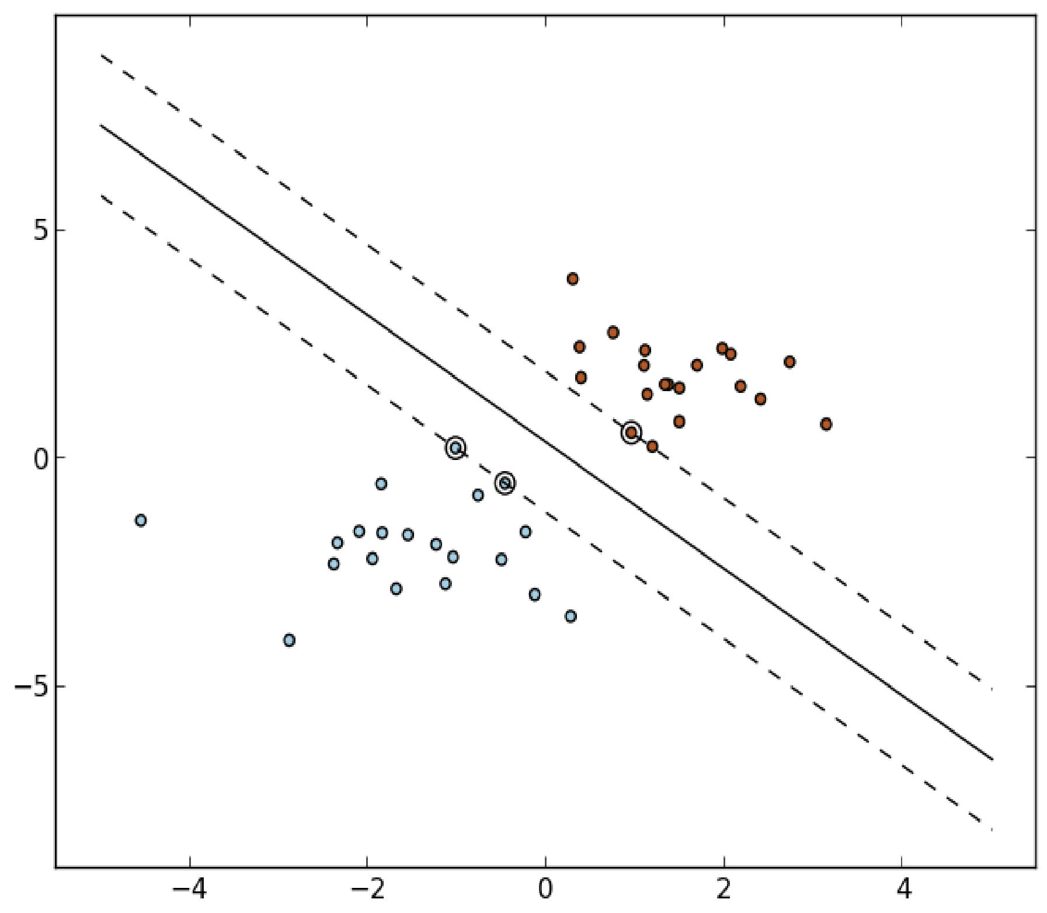
Ways to pay are shown on the back of this notice

MACHINE LEARNING

- A branch of Artificial Intelligence, Machine Learning is good at making predictions with large volumes of data.
- Also used for speech recognition, handwriting recognition and spam filtering

Support Vector Machine (SVM)

- A support vector machine constructs a hyper-plane or set of hyper-planes in a high or infinite dimensional space and finds a separating hyperplane with the maximum margin i.e. the largest distance to the nearest training data points of any class



CLASSIFIER FEATURES

Preprocessing

- Levenshtein Distance, Thesaurus

Semantic Features

- Keyword Exist, Number Words, Number Numeric Words

Visual Features

- Number of blocks of each type, Total area taken by each block type, Block Widths

Combination Semantic and Visual Features

- Word Size, Word Position

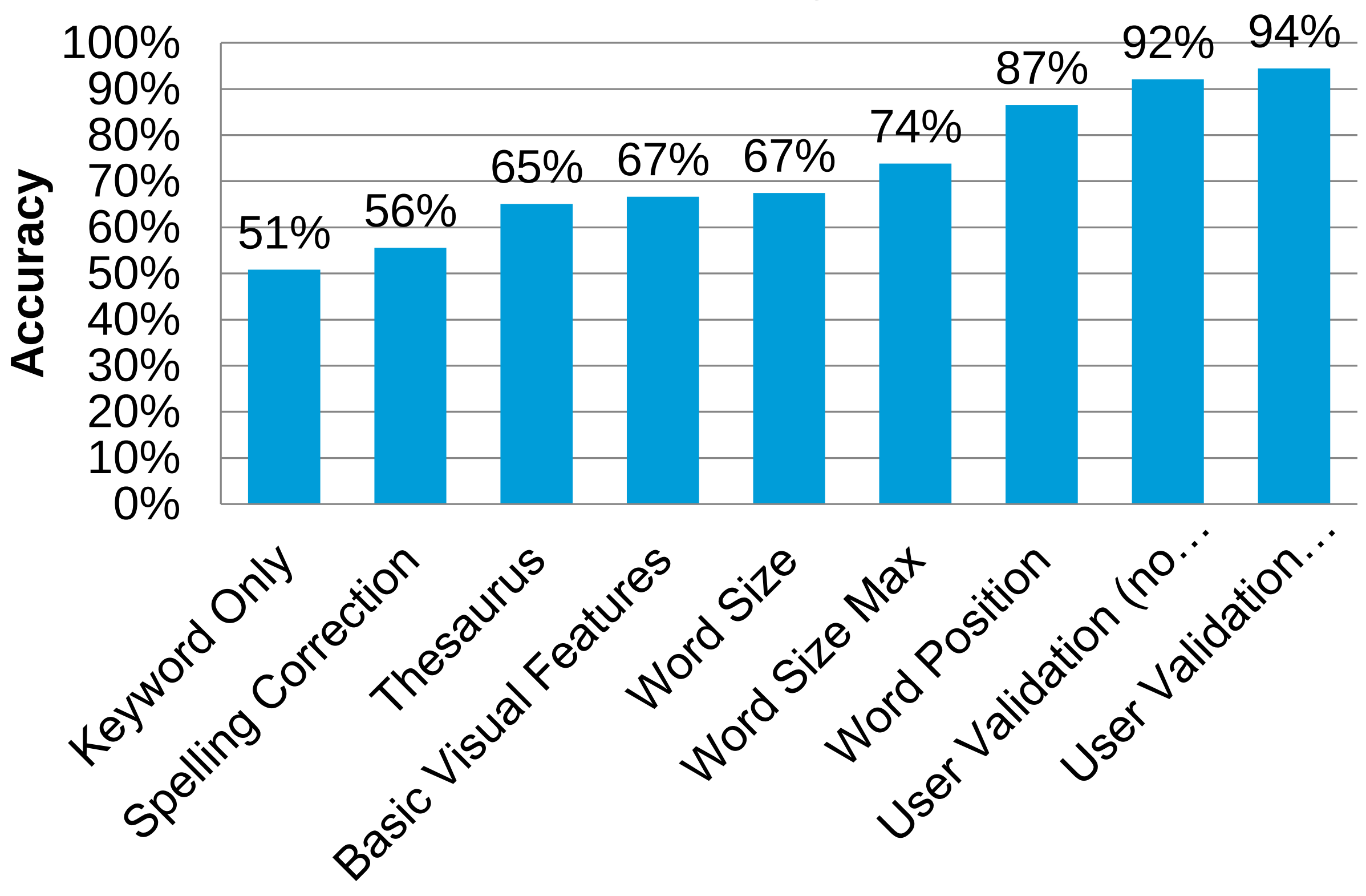
User Validation

- Keyword Search , Prediction Probabilities

ACCURACY

- Below shows the graph showing the amount of improvement gained at each revision of the classifier. Each revision trained the model containing an additional feature

Classifier Accuracy During Each Revision

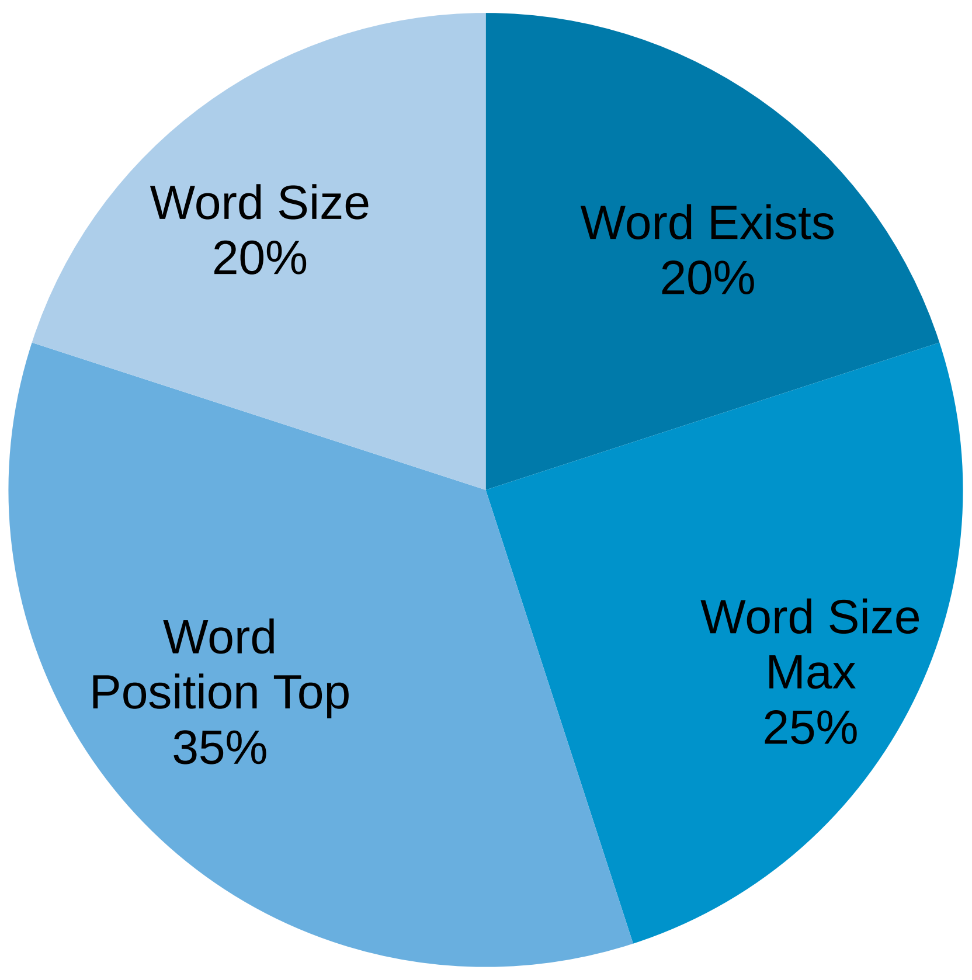


Accuracy Improvement

Feature Name	% Improvement
Word Size	0.79%
Basic Visual Features	1.59%
User Validation (probabilities)	2.38%
Spelling Correction	4.76%
User Validation (no keyword)	5.56%
Word Size Max	6.35%
Thesaurus	9.52%
Word Position	12.70%

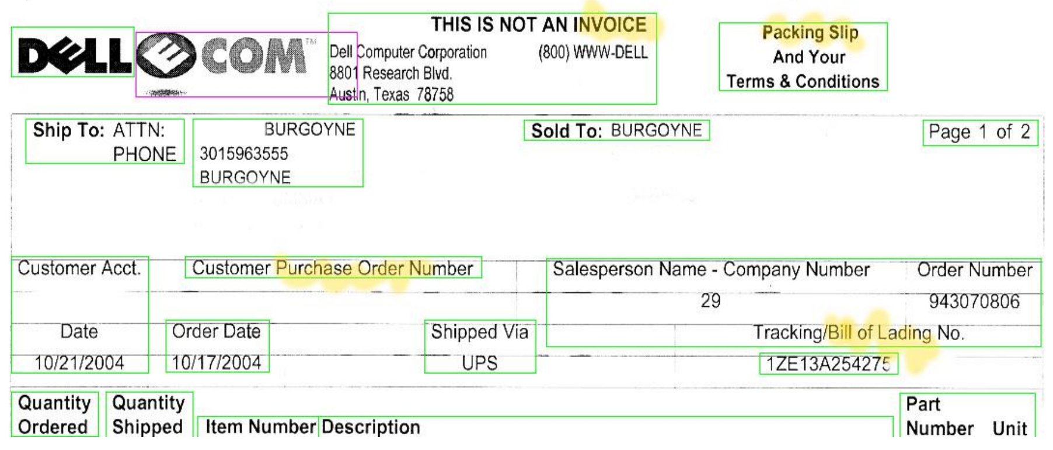
FEATURE IMPORTANCE

20 Most important Features



THINGS TO LOOK OUT FOR

Multiple words in document.



DELL.COM
THIS IS NOT AN INVOICE
Dell Computer Corporation (800) WWW-DELL
8801 Research Blvd.
Austin, Texas 78758

Packing Slip
And Your
Terms & Conditions

Ship To: ATTN: BURGOYNE
PHONE 3015963555
BURGOYNE

Sold To: BURGOYNE

Page 1 of 2

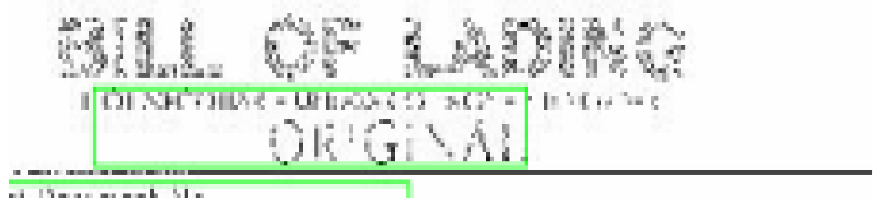
Customer Acct.	Customer Purchase Order Number	Salesperson Name - Company Number	Order Number
		29	943070806

Date	Order Date	Shipped Via	Tracking/Bill of Lading No.
10/21/2004	10/17/2004	UPS	1ZE13A254275

Quantity Ordered	Quantity Shipped	Item Number	Description	Part Number	Unit

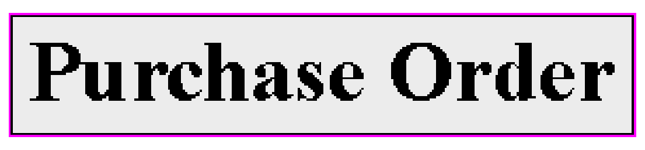
OCR Errors

- Blocks are not always identified by the OCR



BILL OF LADING
ORIGINAL

- Below words were identified as an image.



Purchase Order

More OCR Errors (Spelling)

	Words (Levenshtein Distance < 3)
Incorrect spelled instances of the word 'invoice'	Nvoiceno, iioic, invuiife, voce, iioic, rvoices, nvoi, Invoc, invoceb, invocf, invoco, jlnvoice, jlnvoice, nwoice, tnvotce, Invoicc, lnvoice, mnvoice, unvoice, nvoice
Total	20

Form Errors

- Forms that do not follow the normal conventions of the form type. E.g. Packing slips does not always contain the words 'packing slip' in it.

GENERALIZATION

- Tries to find out how well the model might perform in practice and attempts to determine if there is any bias caused by any unlucky splits of the available data.

K-Folds Cross Validation

Test ID	Predictions Correct	Errors	Accuracy
1	121	4	96.80%
2	122	3	97.60%
3	121	4	97.58%
4	119	6	95.20%
5	116	9	92.80%
Average	119.8	5.2	96.00%

CONCLUSIONS

- Accuracy of the classification and generalization results were consistently high.
- The features that the classifier considered important were reflective of how humans classify documents.
- A large accuracy improvement was seen from fixing OCR errors.
- The amount of user interaction needed has to be considered to integrate the technologies described here into Toshiba's e-BRIDGE Re-Rite.

THIS RESEARCH IS SPONSORED BY
TOSHIBA