technische universität
dortmund

Bag of Features for Text Detection in Natural Scene Images

Rashik Thalappully

Master Student,
Informatik XII, Technische Universität Dortmund
18. Mai 2015

Overview

- Motivation
- Evaluated Methods
- Scale Detection Using Maximally Stable Extremal Regions
- Experiments and Results
- Conclusion
- Future Improvements

Text Detection in Images

- Recognizing the text regions in an arbitrary image.

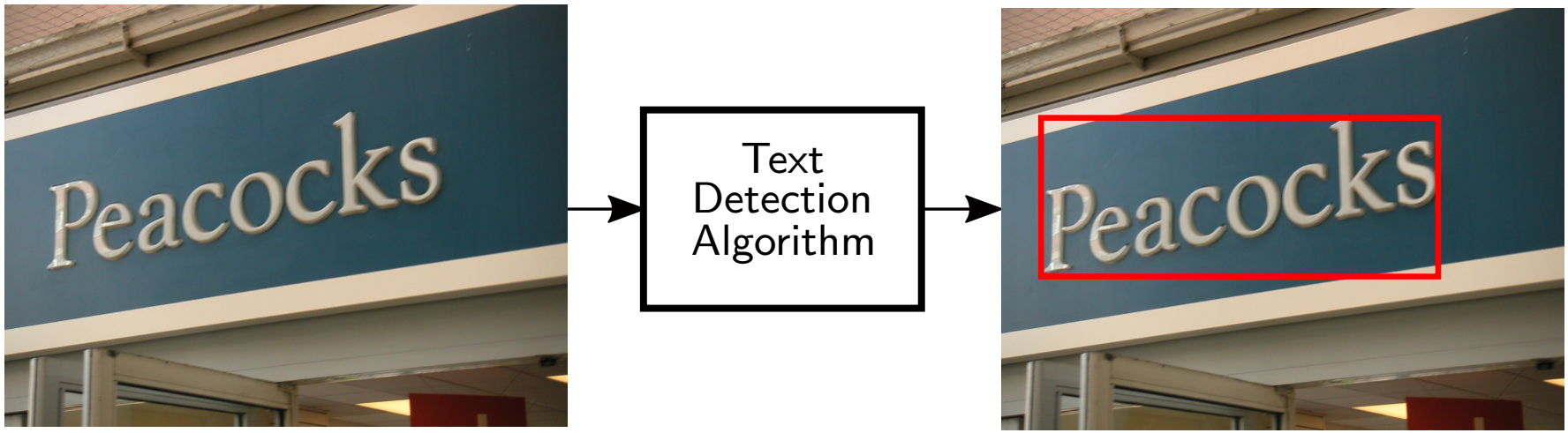


Image Credits: *ICDAR 2003*

Text Detection in Images

- Why recognizing the text regions in an arbitrary image is difficult for computers?

What We See

0	0	8	41	77	119	178	221	234	248
0	8	28	41	77	128	192	234	248	255
0	8	28	41	119	178	221	248	255	248
0	8	41	77	128	192	248	255	248	234
8	28	77	119	178	234	255	248	221	192
28	77	119	178	221	248	255	234	221	192
51	95	128	192	234	255	248	234	178	150
77	119	150	221	248	248	234	192	150	128

What Computers See

Image Credits: *ICDAR 2003*

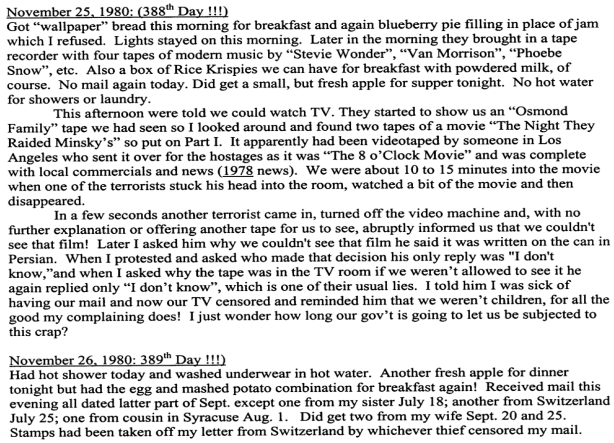
Text Detection in Images

- Problem is quite similar to Optical Character Recognition (OCR)
- Numerous solutions available for OCR for images of digital documents
- Use same solutions for natural scene images?

Image Credits: *mybigguide.com*

Natural Scene Image OCR != Traditional OCR

Image Credits: *http://www.northernsound.ie/*



November 25, 1980: (388th Day !!!)
Got "wallpaper" bread this morning for breakfast and again blueberry pie filling in place of jam which I refused. Lights stayed on this morning. Later in the morning they brought in a tape recorder with four tapes of modern music by "Stevie Wonder", "Van Morrison", "Phoebe Snow", etc. Also a box of Rice Krispies we can have for breakfast with powdered milk, of course. No mail again today. Did get a small, but fresh apple for supper tonight. No hot water for showers or laundry.

This afternoon were told we could watch TV. They started to show us an "Osmond Family" tape we had seen so I looked around and found two tapes of a movie "The Night They Raided Minsky's" so put on Part I. It apparently had been videotaped by someone in Los Angeles who sent it over for the hostages as it was "The 8 o'Clock Movie" and was complete with local commercials and news (1978 news). We were about 10 to 15 minutes into the movie when one of the terrorists stuck his head into the room, watched a bit of the movie and then disappeared.

In a few seconds another terrorist came in, turned off the video machine and, with no further explanation or offering another tape for us to see, abruptly informed us that we couldn't see that film! Later I asked him why we couldn't see that film he said it was written on the can in Persian. When I protested and asked who made that decision his only reply was "I don't know,"and when I asked why the tape was in the TV room if we weren't allowed to see it he again replied only "I don't know", which is one of their usual lies. I told him I was sick of having our mail and now our TV censored and reminded him that we weren't children, for all the good my complaining does! I just wonder how long our gov't is going to let us be subjected to this crap?

November 26, 1980: 389th Day !!!)
Had hot shower today and washed underwear in hot water. Another fresh apple for dinner tonight but had the egg and mashed potato combination for breakfast again! Received mail this evening all dated latter part of Sept. except one from my sister July 18; another from Switzerland July 25; one from cousin in Syracuse Aug. 1. Did get two from my wife Sept. 20 and 25. Stamps had been taken off my letter from Switzerland by whichever thief censored my mail.

Image Credits: *http://www.docstoc.com/*

Natural Scene images

- Artistic fonts
- Extreme lighting variation
- Large variation in color and texture
- Wide range of viewing angles

Images of digital documents

- Typical fonts
- Structured text
- Captured under controlled settings

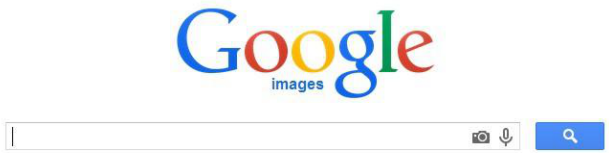
Natural Scene Image OCR != Traditional OCR

- Additional challenges
- Solutions of traditional OCR can't be applied to natural scene image OCR

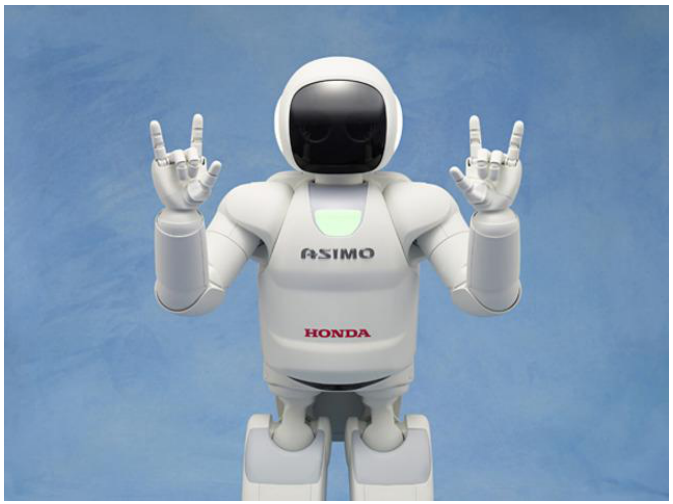
Input Image	Traditional OCR	Input Image	Natural Image OCR
	0 i-n C D 0 iz z 0 0 *CD* H) Oc 0 m *(I)* *CD* U 0 *(1 ')* CDCD >< ITW I • I		University of Essex Day Nursery The Houses Keynes R gh Tawney and William Morris Towers Wolfson Court Buses and Cycles Only CONFERENCE CAR PARK EntTance 4

Image Credits: *web.iit.edu*

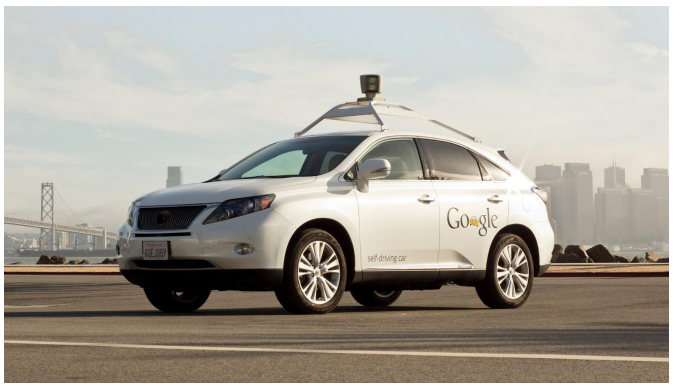
Applications - Text Detection in Natural Scene Images



Content Based Image Retrieval



Robotics



Self Driving Automobiles

Image Credits: *Google Images*

Text Detection Using Learned Feature Descriptor

- Creating visual vocabulary

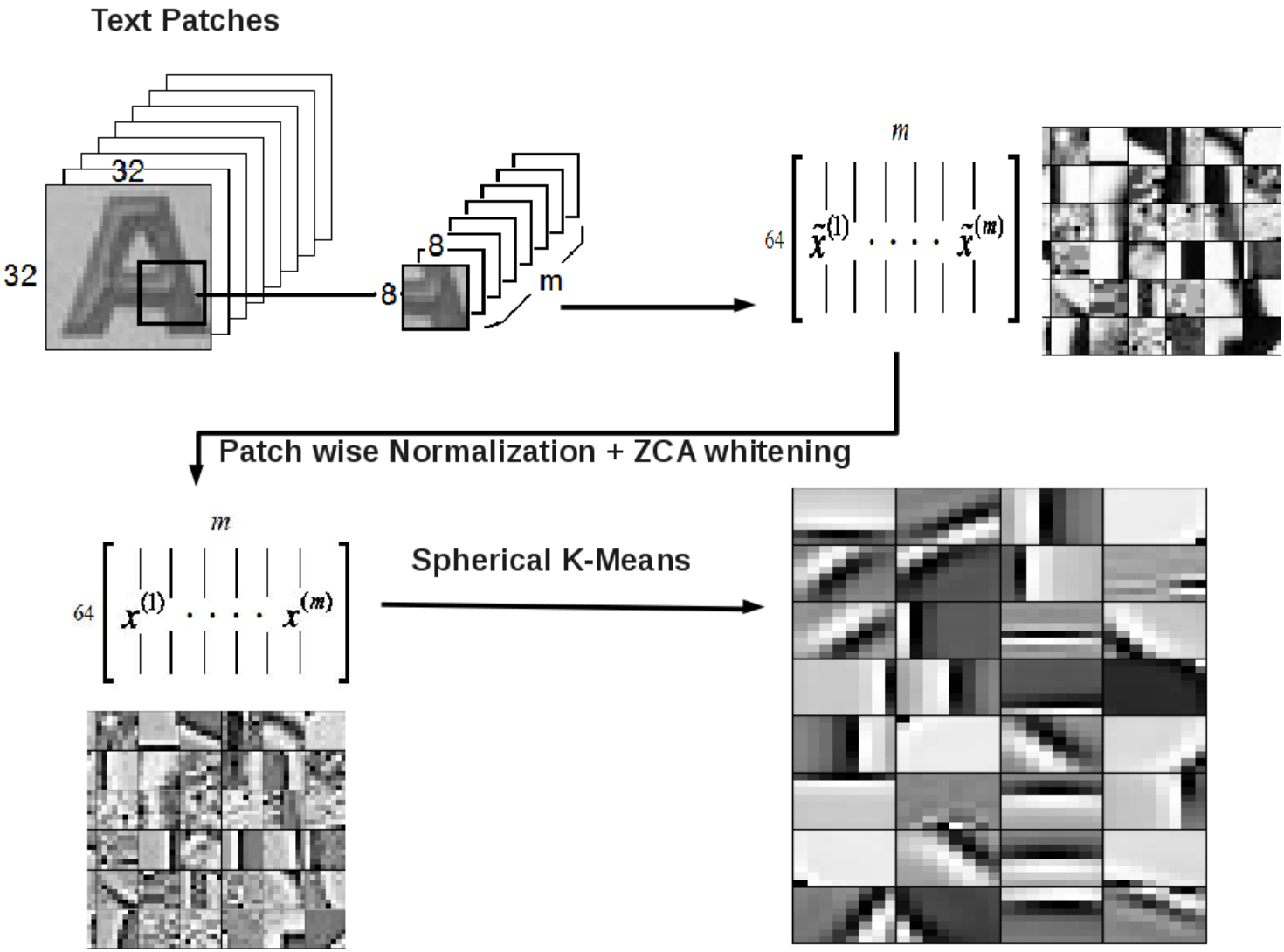


Image Credits: *ICDAR 2003*

Text Detection Using Learned Feature Descriptor

- Feature extraction

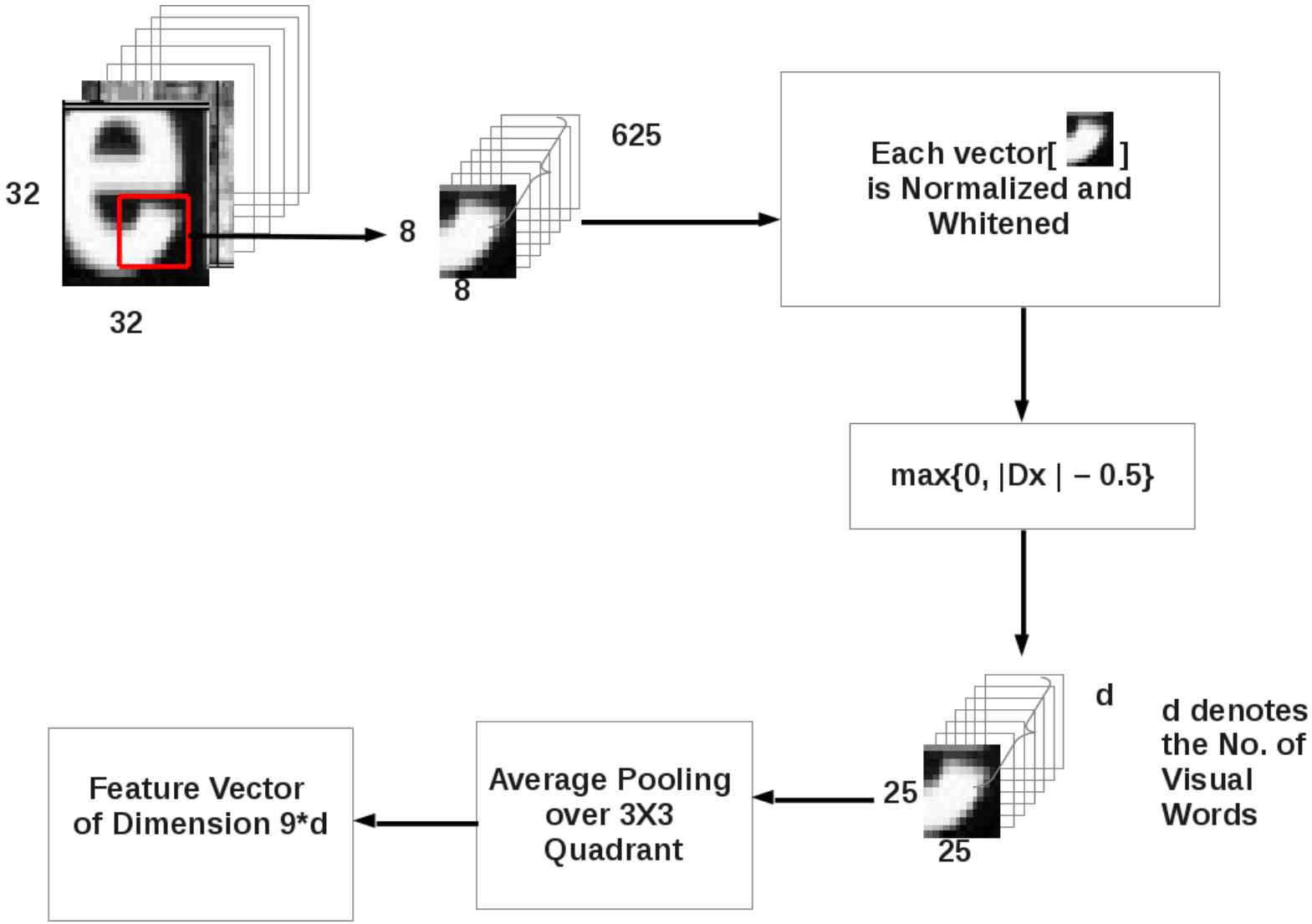


Image Credits: *ICDAR 2003*

Training Data Generation [1]

A 32-by-32 gray image patch is labeled as text region, only if it satisfies the following conditions:

1. 80% of the image patch region is within the text region defined by the ground truth.
2. Character bounding box width or height is within 30% of the width or height of the image patch region respectively

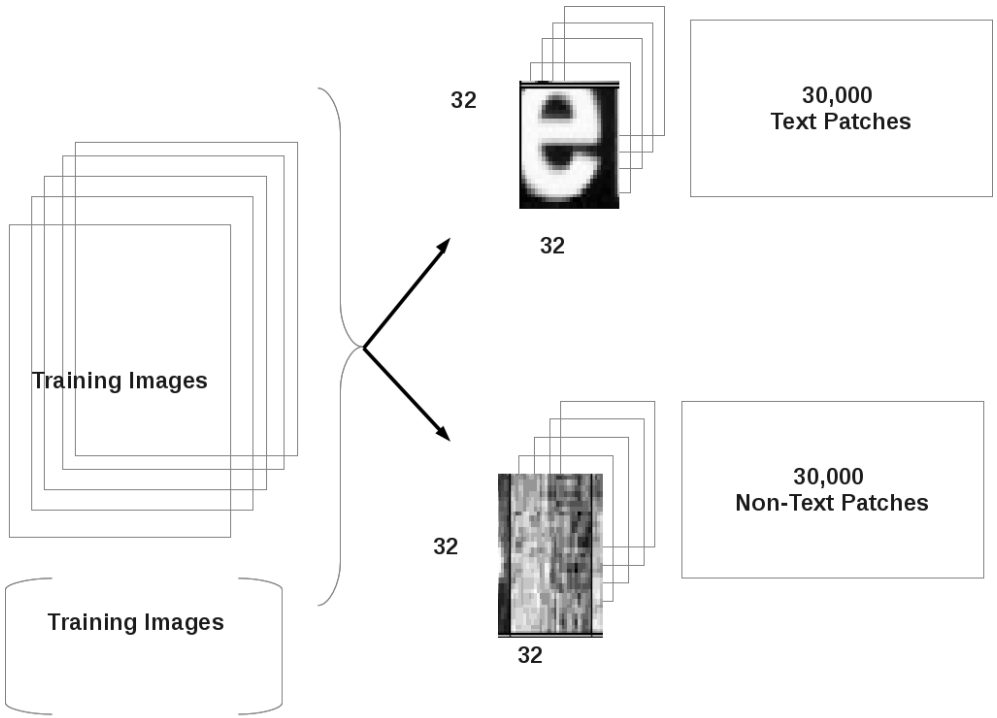
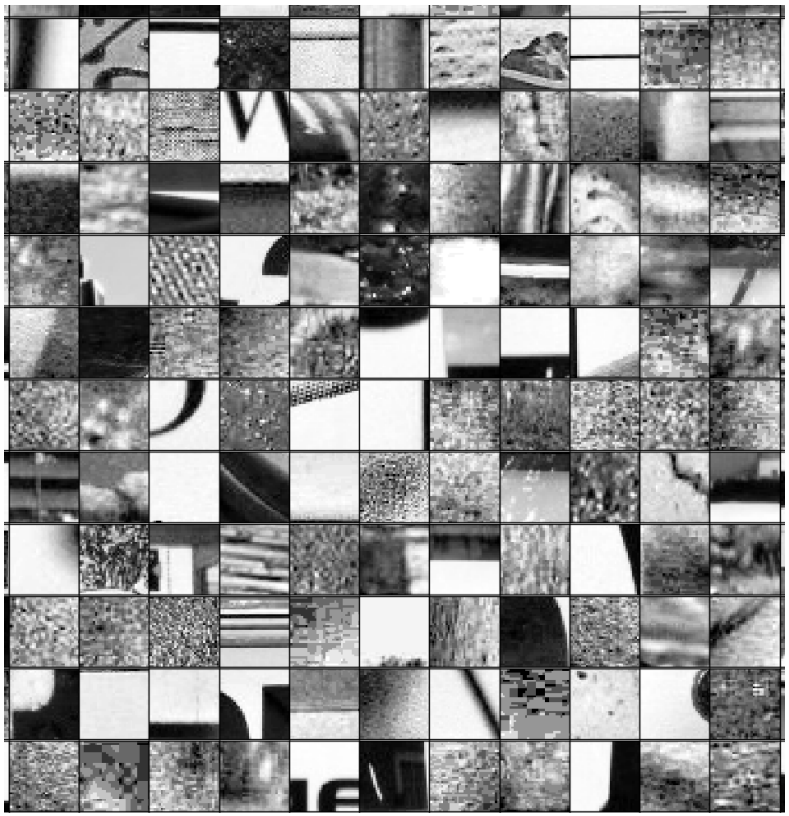


Image Credits: *ICDAR 2003*

[1] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu und A. Y. Ng: *Text detection and character recognition in scene images with unsupervised feature learning*. In: *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, S. 440–445. IEEE, 2011

Text Patches and Non-Text Patches



Non-Text Patches

Text Patches

Image Credits: *ICDAR 2003*

SVM Classifier Training Using Learned Feature Descriptor [2]

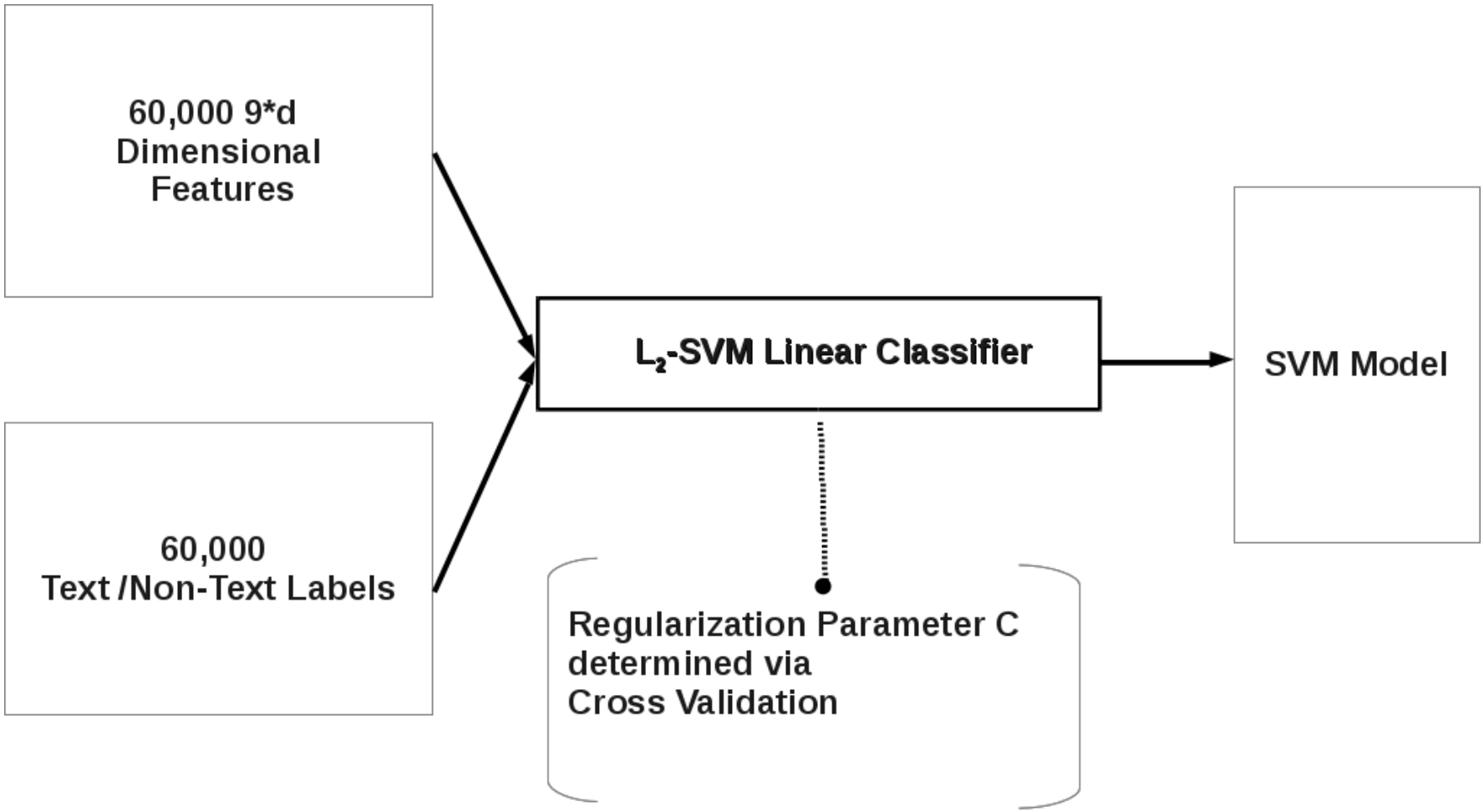


Image Credits: *ICDAR 2003*

[2] A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, D. J. Wu und A. Y. Ng: *Text detection and character recognition in scene images with unsupervised feature learning*. In: *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, S. 440–445. IEEE, 2011

Text Detection Using Hand Crafted Feature Descriptor

- SIFT Descriptor [3]
- 16x16 region around the key point is considered
- Each 4x4 sub region, histogram of oriented gradients is calculated with 8 bins each
- All the values from these histograms (16*8=128) form the 128 element descriptor

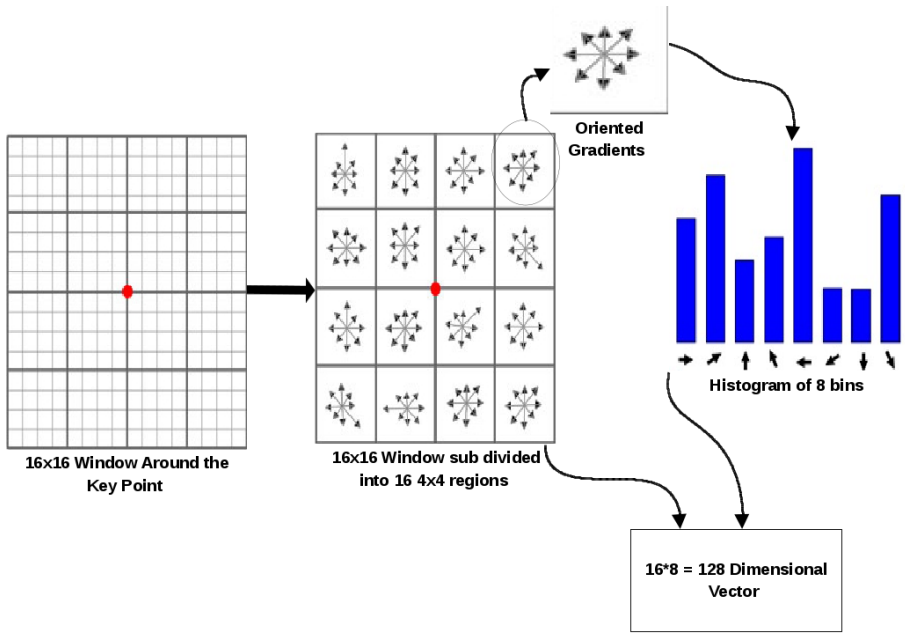


Image Credits: *[4]*

[3] D. G. Lowe: *Distinctive image features from scale-invariant keypoints*. International journal of computer vision, 60(2):91–110, 2004

[4] D. G. Lowe: *Distinctive image features from scale-invariant keypoints*. International journal of computer vision, 60(2):91–110, 2004

SVM Classifier Training Using SIFT Descriptor

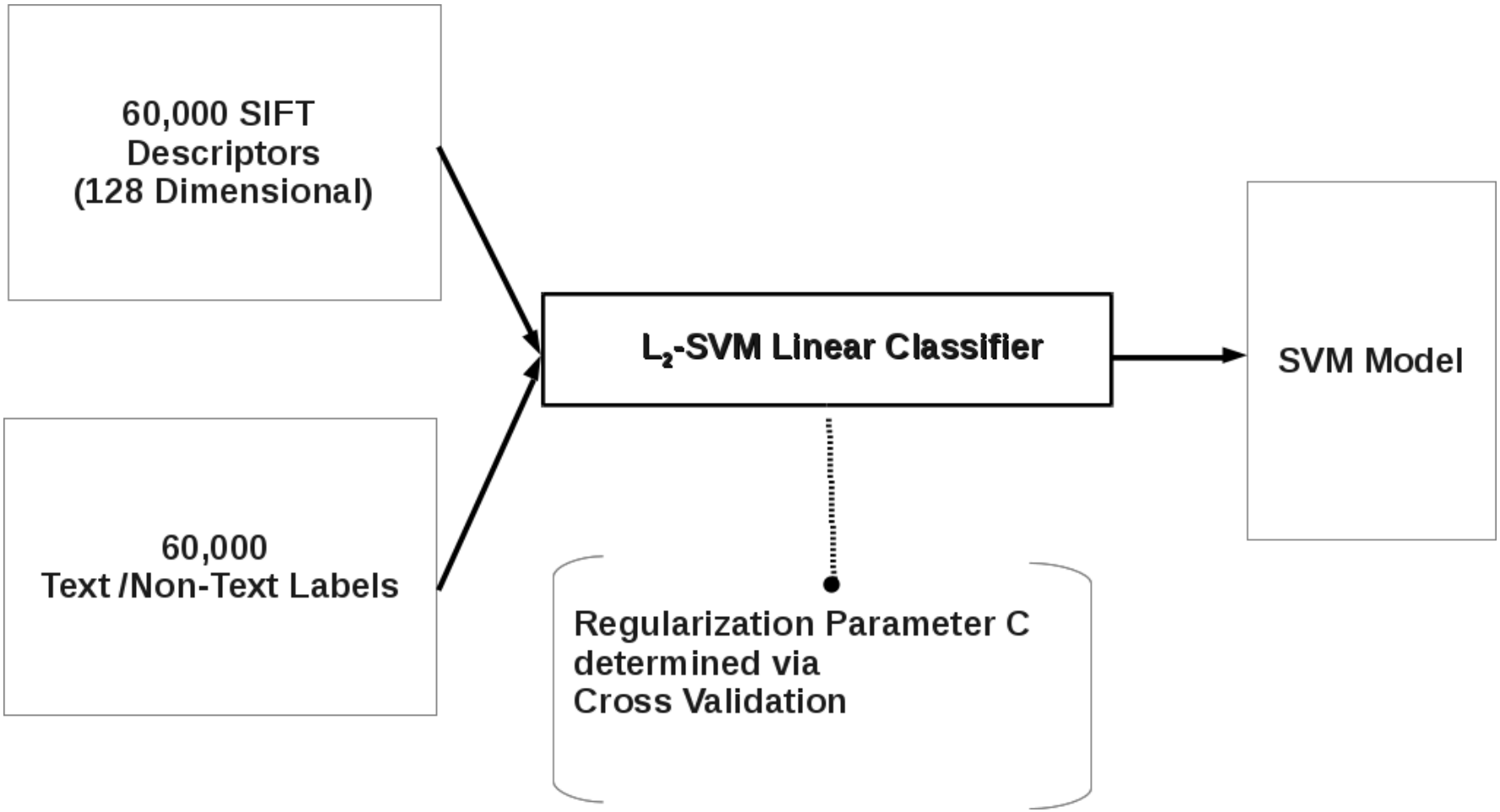


Image Credits: *ICDAR 2003*

Evaluation of Generated Text Detector Models

- Both the Text Detectors are evaluated in similar fashion

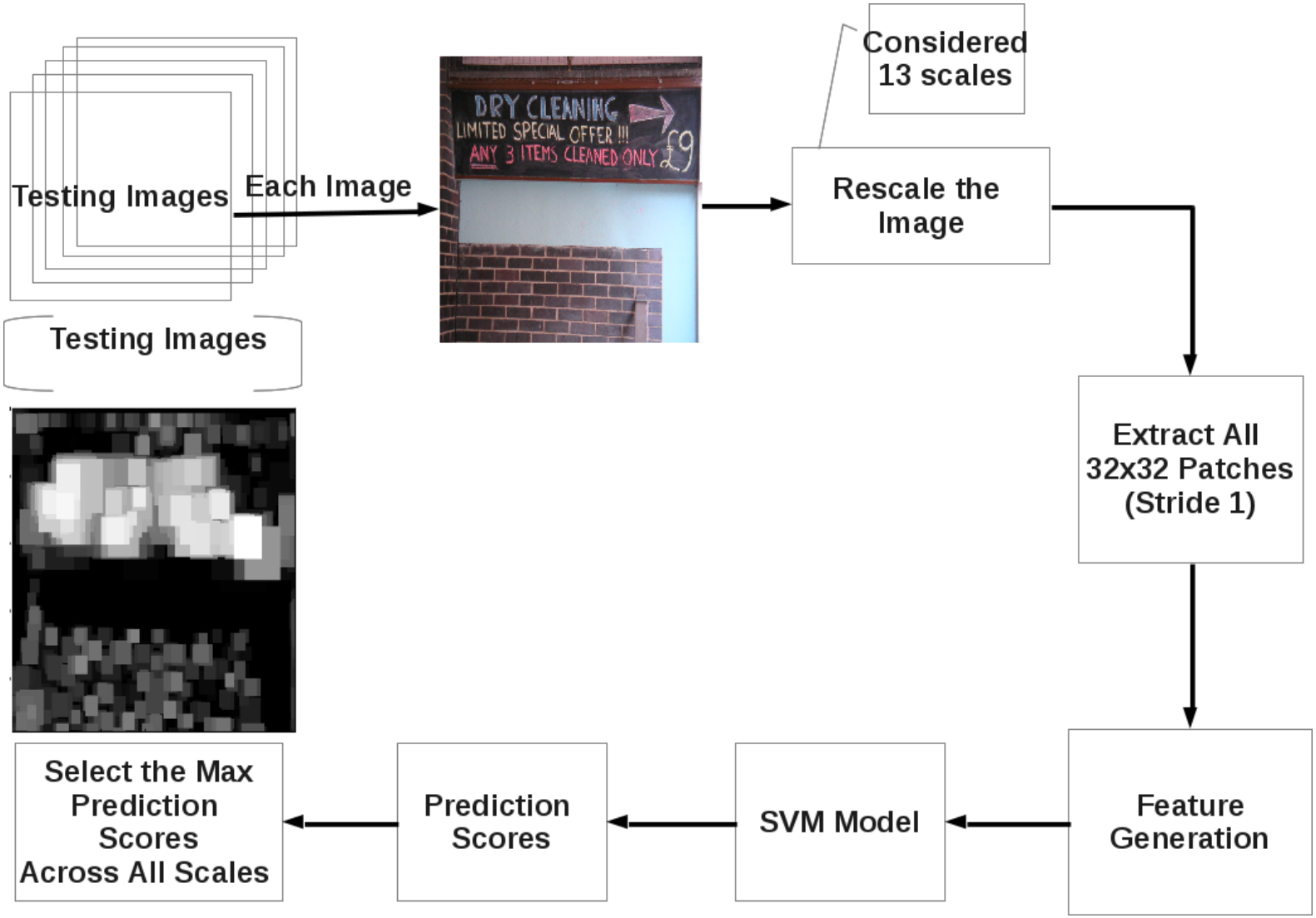


Image Credits: *ICDAR 2003*

Demo of Sliding Window Evaluation

TEXT

Drawbacks

- Experimentally determined scales
- Computationally expensive

Maximally Stable Extremal Regions (MSER) [5]

Characteristics of MSER

- Connected components with similar intensity values bounded by contrasting backgrounds

Image Credits: *Matas et al.*

[5] J. Matas, O. Chum, M. Urban und T. Pajdla: *Robust wide-baseline stereo from maximally stable extremal regions*. Image and vision computing, 22(10):761–767, 2004

Maximally Stable Extremal Regions (MSER) [6]

- Connected components virtually unchanged over a range of thresholds

[6] J. Matas, O. Chum, M. Urban und T. Pajdla: *Robust wide-baseline stereo from maximally stable extremal regions*. Image and vision computing, 22(10):761–767, 2004

Scale detection using Maximally Stable Extremal Regions (MSER)

- Open CV MSER detector [7]

Image Credits: *ICDAR2003*

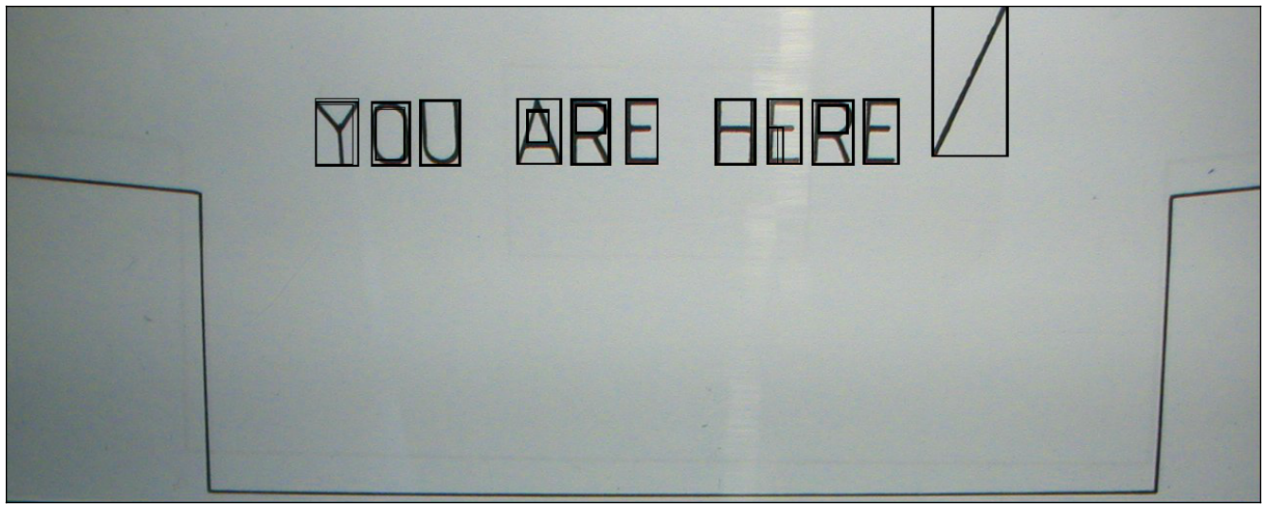
[7] G. Bradski: *The opencv library*. Doctor Dobbs Journal, 25(11):120–126, 2000

Scale detection using Maximally Stable Extremal Regions (MSER)

- Cluster (width of bounding box, height of bounding box) using Lloyds algorithm
- Run the clustering for number of centeroids $= 1$ till number of centeroids $= 13$
- Select the best cluster using GAP statistics [8]
- Determined centroids represents the size of character candidates
- Scale factors are determined from the size of character candidates

[8] R. Tibshirani, G. Walther und T. Hastie: *Estimating the number of clusters in a data set via the gap statistic*. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2):411–423, 2001

Process of Scale detection



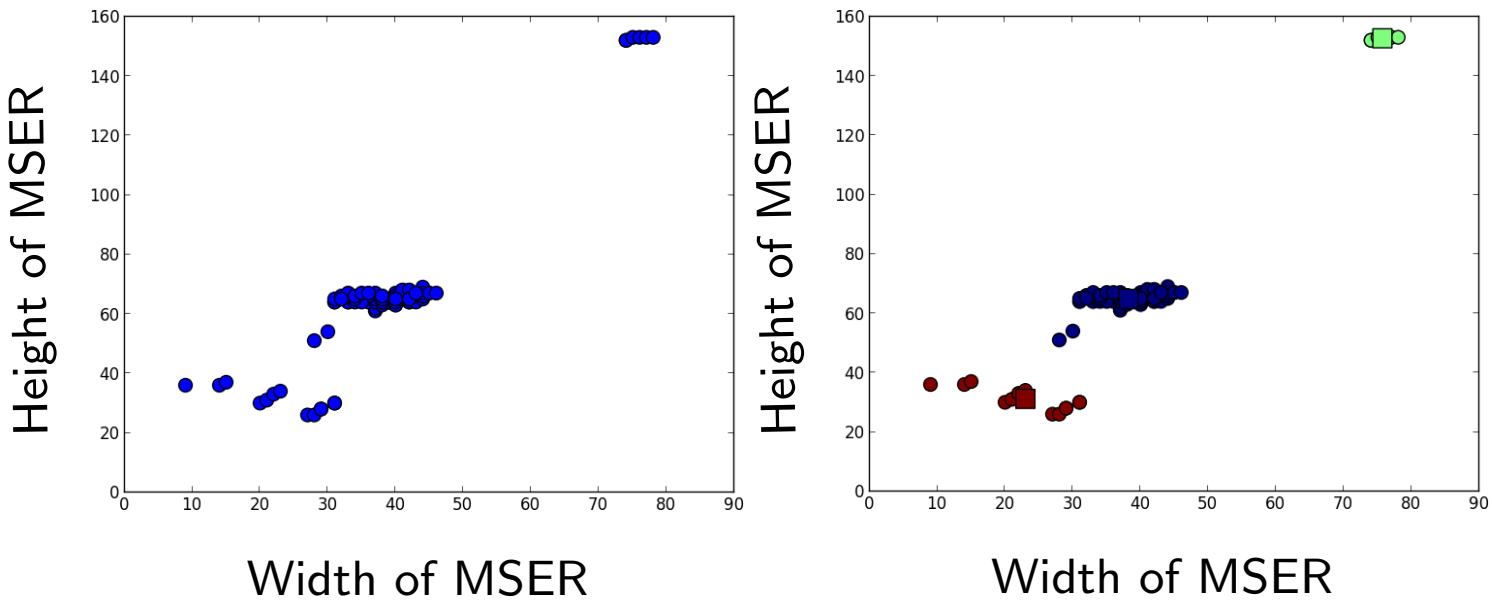


Image Credits: *ICDAR2003*

Experiments and Evaluation Datasets

Experiments

- Hand-Crafted Feature Representation Vs Learned Feature Representation
- Effect of Increasing Number of Visual Words
- MSER Integrated Sliding Window Vs Standard Sliding Window

Evaluation Datasets

- ICDAR 2003 [9]
- ICDAR 2013 [10]

ICDAR 2003

ICDAR 2013

[9] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong und R. Young: *ICDAR 2003 robust reading competitions*. In: *2013 12th International Conference on Document Analysis and Recognition*, Bd. 2, S. 682–682. IEEE Computer Society, 2003

[10] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, L. P. de las Heras et al.: *ICDAR 2013 Robust Reading Competition*. In: *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, S. 1484–1493. IEEE, 2013

Evaluation Metric

- Fixed thresholds considered across all test images
- For each test image, at each threshold level, True Positives, False Positives and False Negatives values were calculated
- For whole dataset, at each threshold level accumulate True Positives, False Positives and False Negatives values and then calculate Precision and Recall

	Predicted Text	Predicted Non-text
Actual Text	True Positives	False Negatives
Actual Non-text	False Positives	True Negatives

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negative}}$$

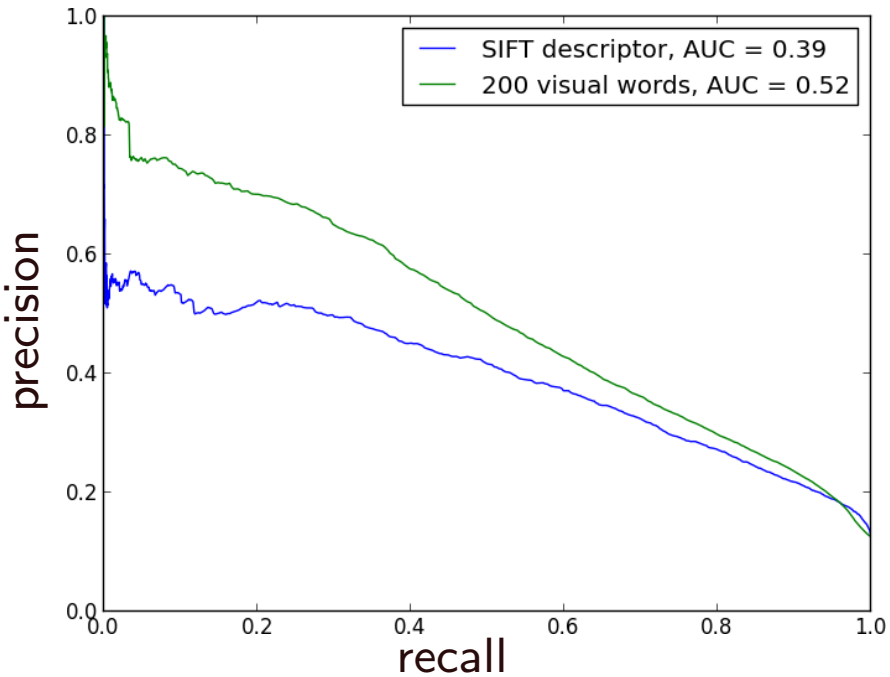
Hand-Crafted Feature Representation Vs Learned Feature Representation

- Results for the performance of hand-crafted and learned feature representation methods on ICDAR 2003 and ICDAR 2013

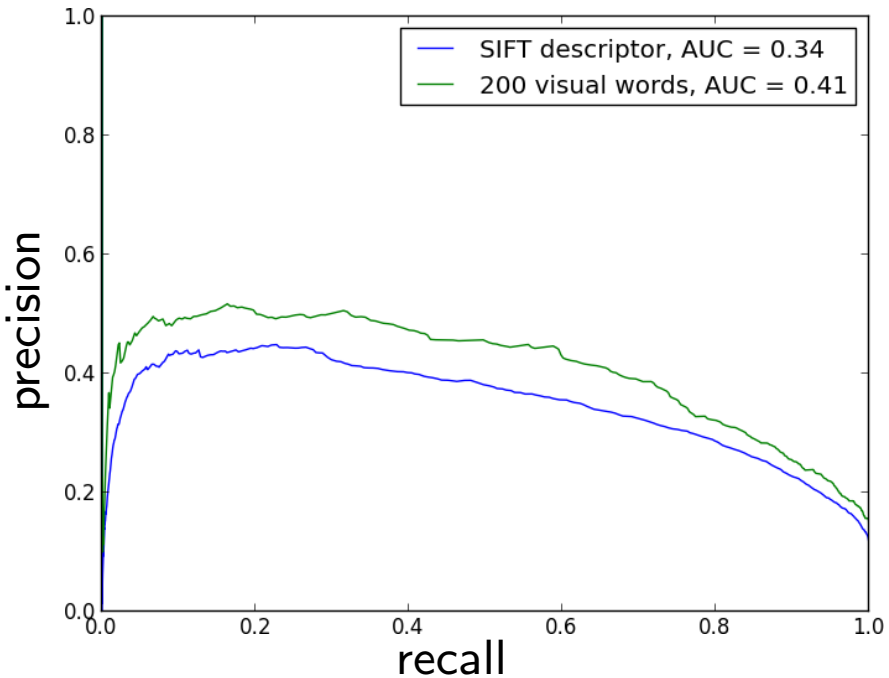
Feature representation	ICDAR 2003		ICDAR 2013	
	Classification accuracy	AUC	Classification accuracy	AUC
Hand-crafted (SIFT descriptor)	73.4%	0.39	70.5%	0.34
Learned (200 visual words)	87.2%	0.52	82.4%	0.41

Hand-Crafted Feature Representation Vs Learned Feature Representation

- Precision-Recall curves for Hand-Crafted Feature Representation (SIFT) and Learned Feature Representation



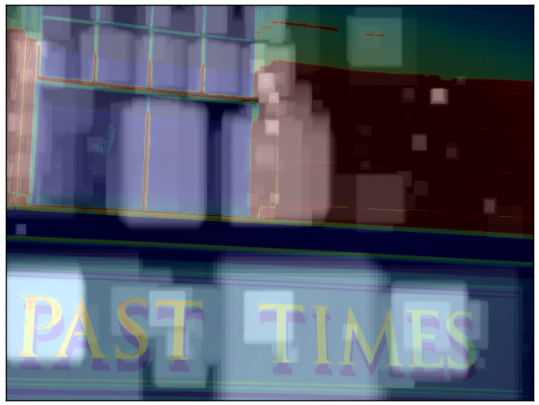
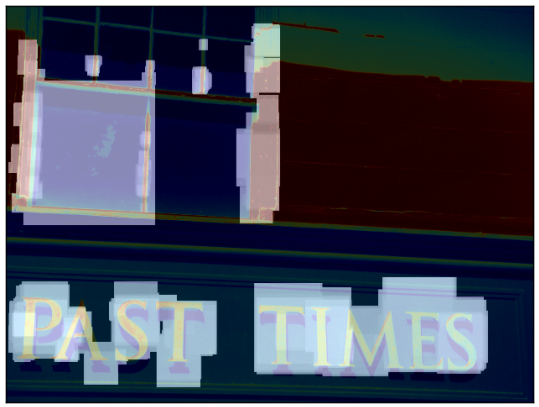
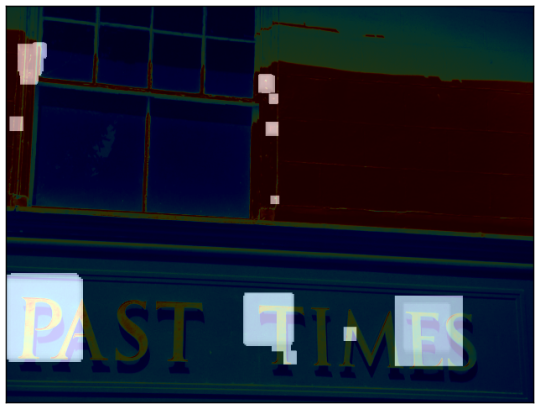
ICDAR 2003



ICDAR 2013

Hand-Crafted Feature Representation Vs Learned Feature Representation

- The per-pixel predictions overlaid over the original image

	Learned	Hand-Crafted
Threshold = 0		
Threshold = 1		

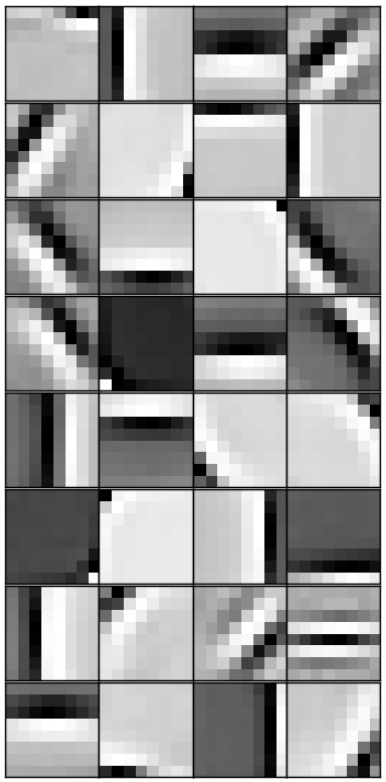
Effect of Increasing Number of Visual Words

- Results for the performance of learned features with varying number of visual words on ICDAR 2003 and ICDAR 2013

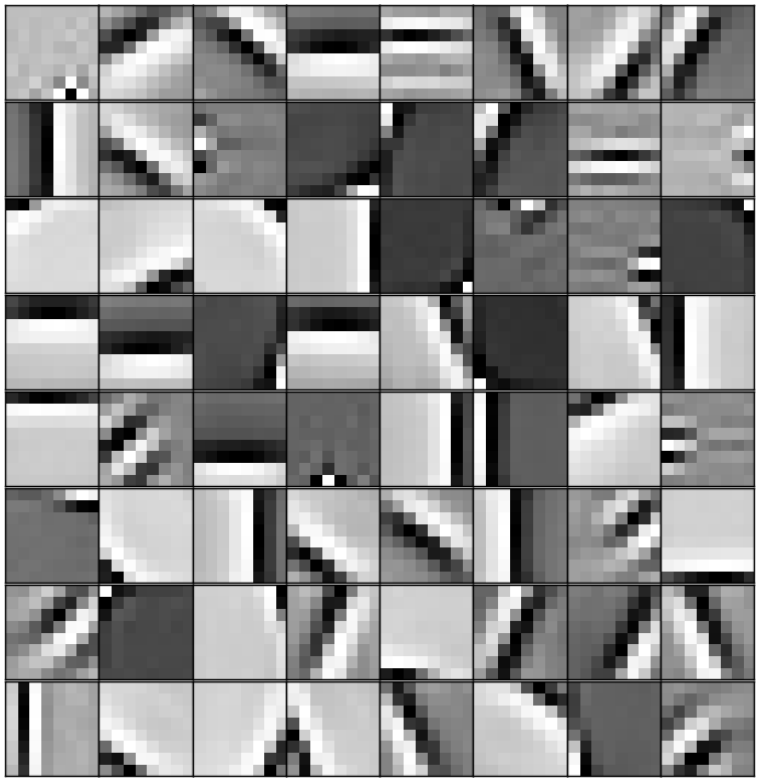
No. of visual words	ICDAR 2003		ICDAR 2013	
	Classification accuracy	AUC	Classification accuracy	AUC
32	83.3%	0.45	77.2%	0.36
64	84.3%	0.47	78.1%	0.38
200	87.2%	0.52	82.4%	0.41

Effect of Increasing Number of Visual Words

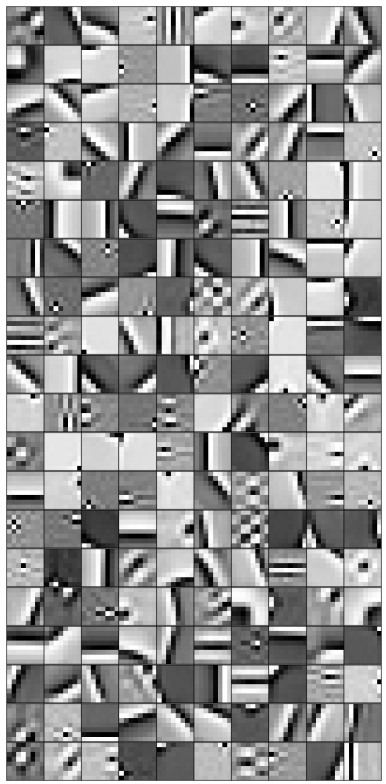
- Visual vocabulary comparison



32 Visual Words



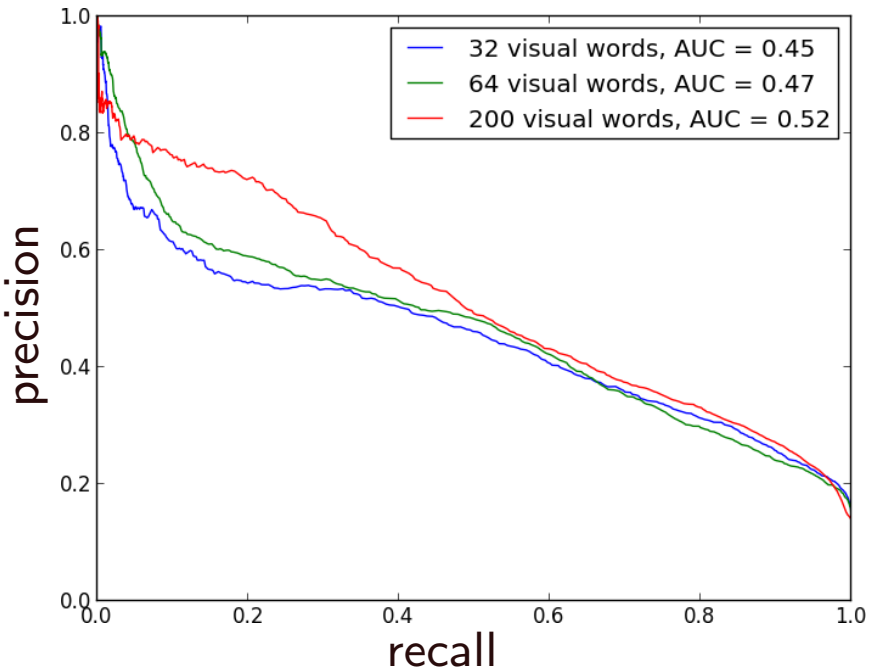
64 Visual Words



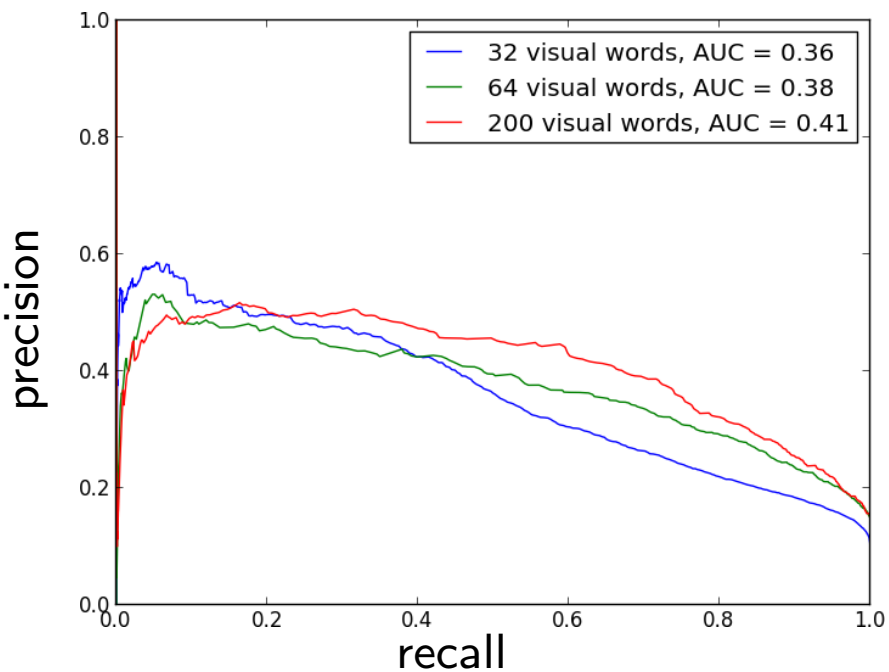
200 Visual Words

Effect of Increasing Number of Visual Words

- Precision-Recall curve for varying number of visual words on ICDAR 2003 and ICDAR 2013 dataset



ICDAR 2003



ICDAR 2013

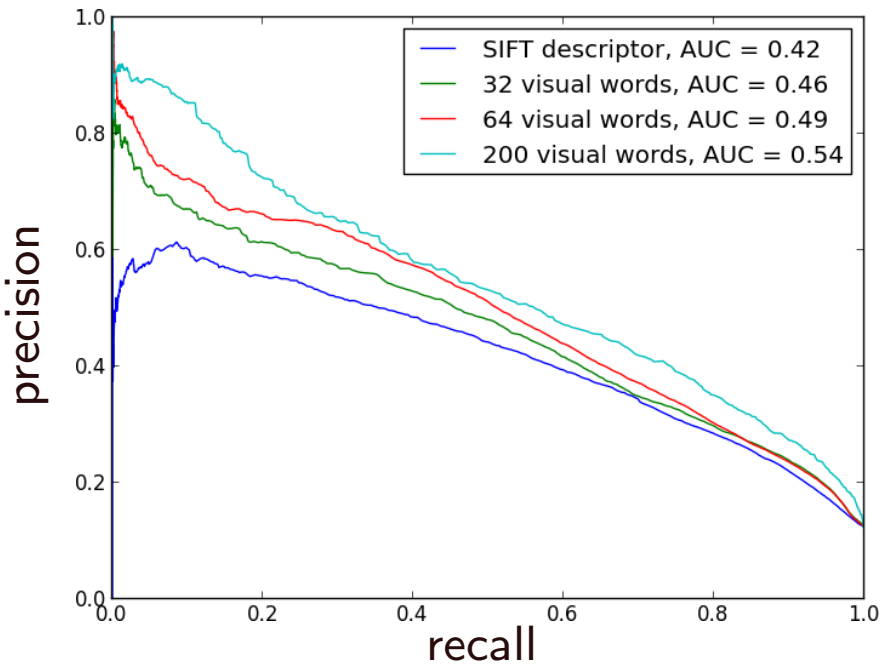
MSER integrated Sliding Window Vs Standard Sliding Window

- Results of the performance of standard sliding window and MSER integrated sliding window on ICDAR 2003 and ICDAR 2013

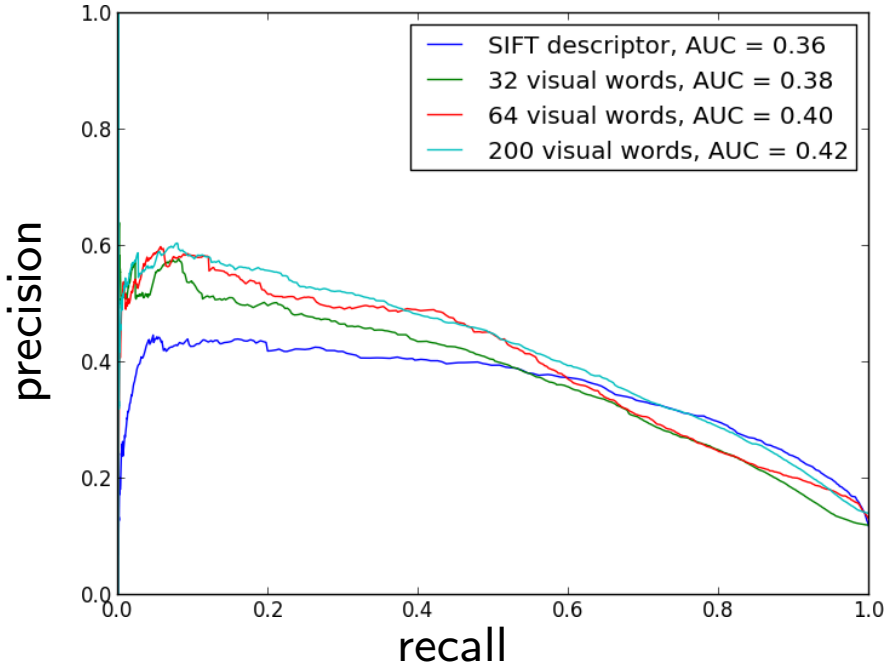
	AUC					
	ICDAR 2003			ICDAR 2013		
	13 scales	MSER scales	Increase in %	13 scales	MSER scales	Increase in %
SIFT descriptor	0.39	0.42	7.7	0.34	0.36	5.9
32 visual words	0.45	0.46	2.2	0.36	0.38	5.6
64 visual words	0.47	0.50	6.4	0.38	0.40	5.3
200 visual words	0.52	0.55	5.8	0.41	0.42	2.4

MSER Integrated Sliding Window Vs Standard Sliding Window

- Precision-Recall curves obtained for standard sliding window and MSER integrated sliding window on ICDAR 2003 and ICDAR 2013 dataset



ICDAR 2003



ICDAR 2013

MSER Integrated Sliding Window Vs Standard Sliding Window

- The per-pixel predictions overlaid over the original image

Standard Sliding Window

MSER integrated Sliding Window

MSER Integrated Sliding Window Vs Standard Sliding Window

- Evaluation time of standard sliding window and MSER integrated sliding window on ICDAR 2003 and ICDAR 2013

	Time taken for evaluation					
	ICDAR 2003			ICDAR 2013		
	13 scales (s)	MSER scales (s)	Reduction in %	13 scales (s)	MSER scales (s)	Reduction in %
SIFT descriptor	155.8	65.4	58.0	157.3	67.1	57.3
32 visual words	254.7	148.3	41.7	312.2	194.3	37.8
64 visual words	571.3	322.4	43.5	639.3	367.2	43.6
200 visual words	1021.8	547.8	46.3	1147.4	586.6	48.9

Conclusion

- Learned feature representation outperformed the hand-crafted feature representation
- Increase in performance was observed with the increase in the number of visual words in the visual vocabulary
- MSER integrated sliding window not only reported higher performance but also reduced the execution time

Future Improvements

- Improvement of MSER detector especially in images with repeating patterns or vegetations
- Exploring the feasibility of proposing a guideline for parameter configuration involved in the training data generation

Questions?

Thank you