Representation in Scene Text Detection and Recognition



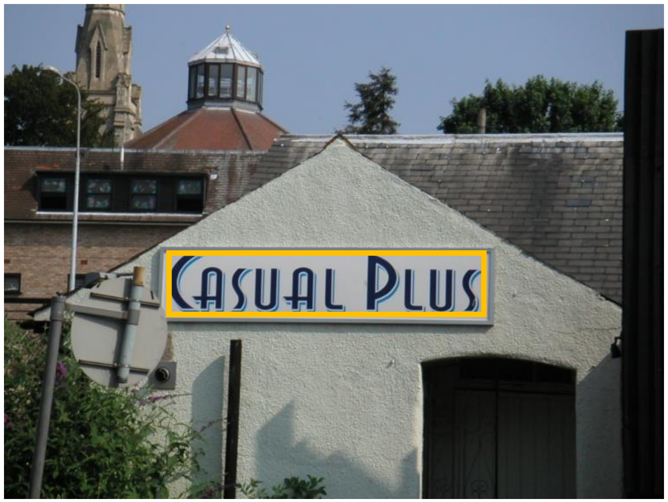

Prof. Xiang Bai

Huazhong University of Science and Technology

Contents

Contents

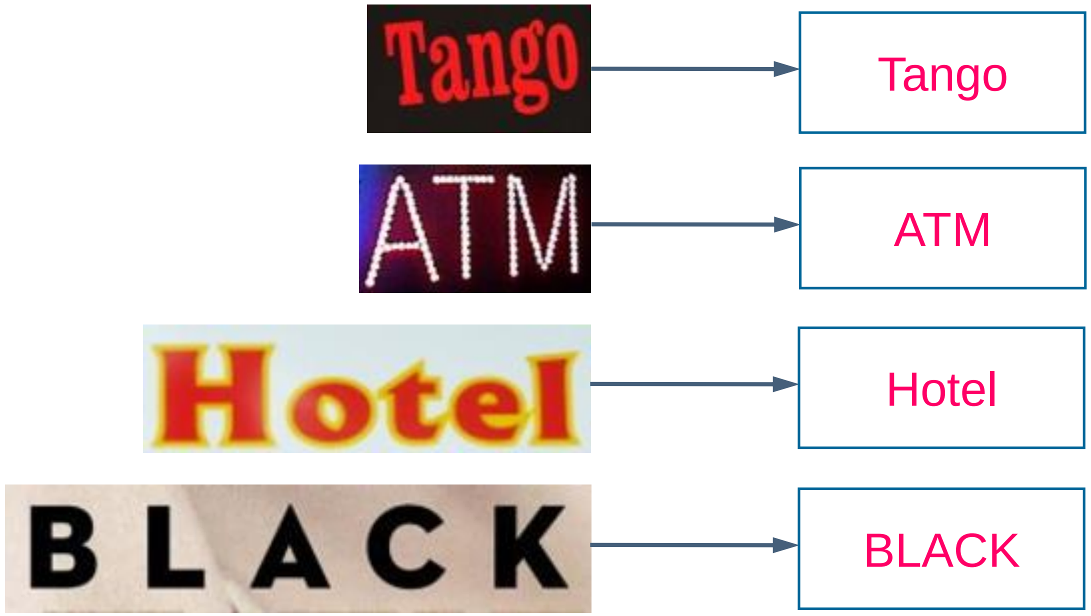
Problem definition



Scene text detection:

the process of predicting the presence of text and localizing each instance (if any), usually at word or line level, in natural scenes

Problem definition



Scene text recognition:

the process of converting text regions into computer readable and editable symbols

Contents

- Problem definition
- Significance and challenges
- Previous works
- Our algorithms
- Conclusion

Significance

- text in natural scenes carries rich and precise high level semantics
- text information can be useful to a variety of applications:
  scene understanding, product search, HCI, virtual reality…

challenges

Diversity of scene text:
different colors, scales, orientations, fonts, languages…

challenges

Complexity of background:

elements like signs, fences, bricks, and grasses are virtually undistinguishable from true text

challenges

Various interference factors:

noise, blur, non-uniform illumination, low resolution, partial occlusion...

challenges

These challenges make
scene text detection and recognition
extremely difficult problems

Contents

- Problem definition
- Significance and challenges
- Previous works
- Our algorithms
- Conclusion

Previous works

Three categories:

1. text detection

only localize text regions, no need to recognize the content

2. text recognition

only recognize the content, assume text regions are given

3. end-to-end text recognition

perform both text detection and recognition

Previous works

In the following slides, we will review
a number of previous algorithms, mainly from
the perspective of representation

Text Detection

MSER

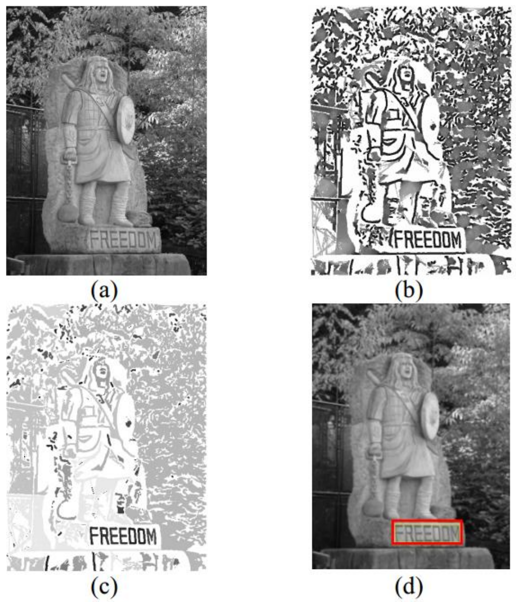
(a) (b) (c) (d)

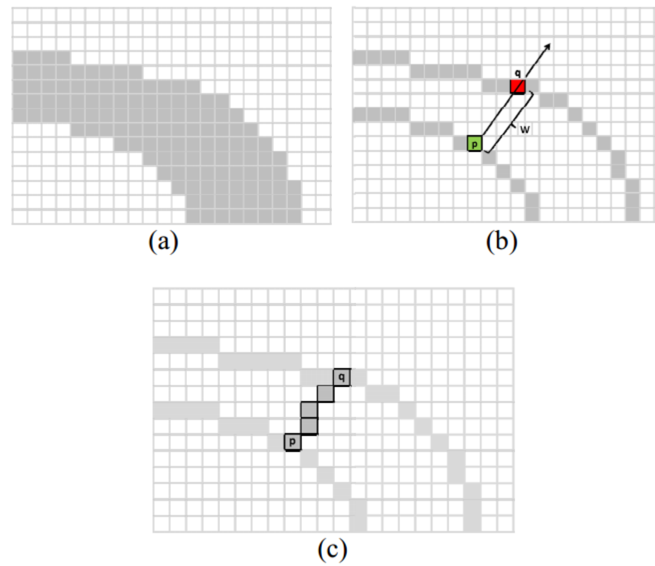
[Neumann and Matas, ACCV 2010]

- extract character candidates using Maximally Stable Extremal Regions, assuming similar color within each character
- robust, fast to compute, independent of scale and orientation

Text Detection

SWT





[Epshtein et al., CVPR 2010]

- extract character candidates with Stroke Width Transform, assuming consistent stroke width within each character
- robust, fast to compute, independent of scale and orientation

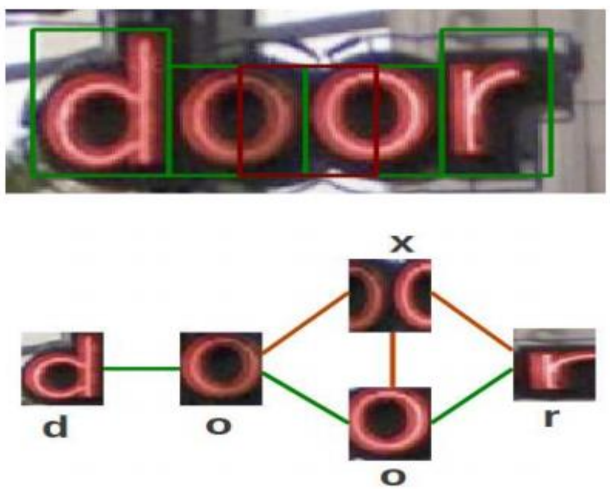
Text Detection

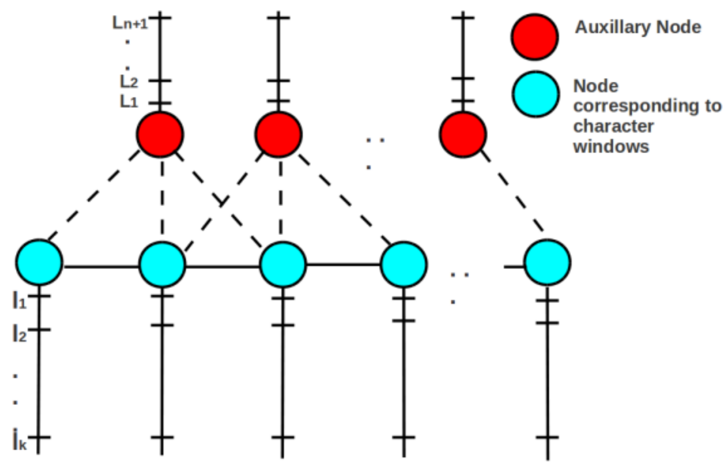
MSER and SWT are representative methods in scene text detection, which constitute the basis of a lot of subsequent works

[Chen et al., ICIP 2011],
[Yao et al., CVPR 2012],
[Neumann and Matas, CVPR 2012],
[Novikova et al., ECCV 2012],
[Huang et al., ICCV 2013],
[Yinet al., SIGIR 2013],
[Koo et al., TIP 2013],
[Yin et al., TPAMI 2014],
[Yao et al., TIP 2014],
[Huang et al., ECCV 2014],
.....

Text Recognition

Top-Down and Bottom-up Cues

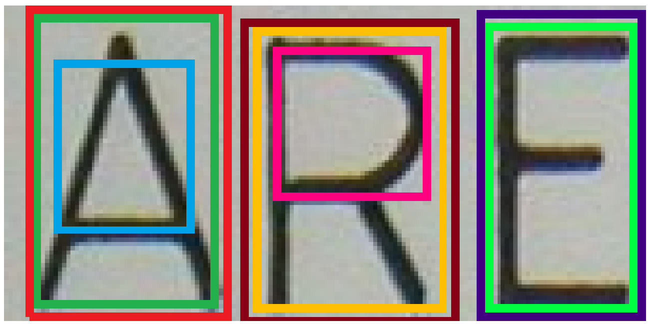


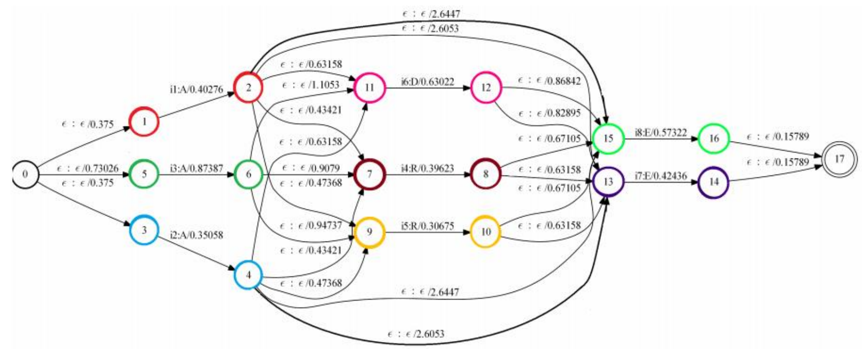


[Mishra et al., CVPR 2012]

- seek character candidates using sliding window, instead of binarization
- construct a CRF model to impose both bottom-up (i.e. character detections) and top-down (i.e. language statistics) cues

Large-Lexicon Attribute-Consistent



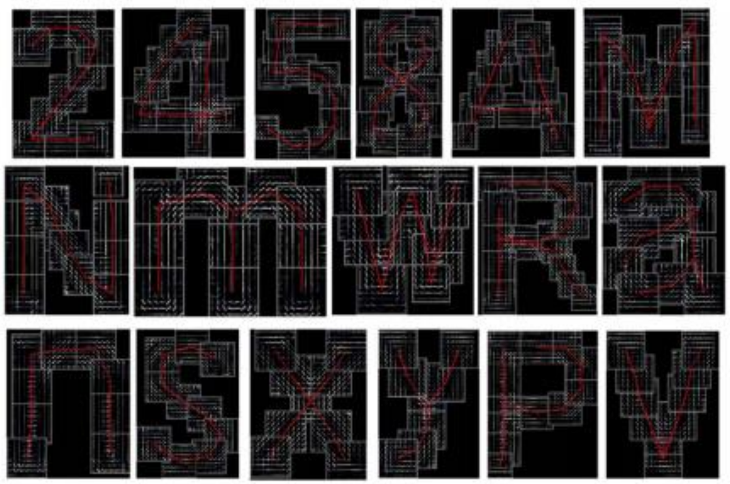


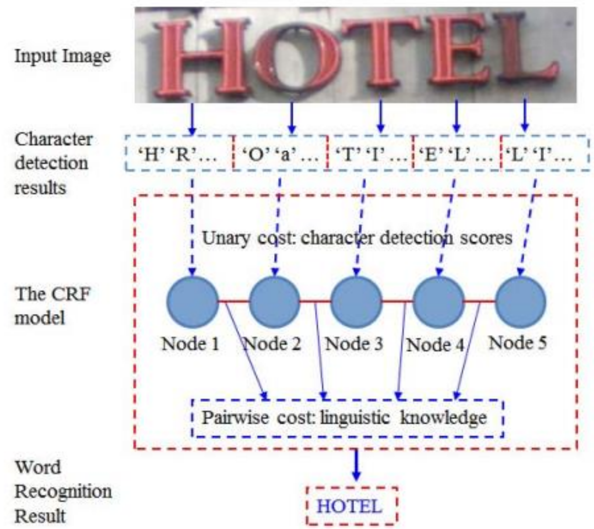
[Novikova et al., ECCV 2012]

- seek character candidates via MSER extraction
- utilize Weighted Finite-State Transducers, to simultaneously introduce language prior and enforce attribute consistency between hypotheses.

Text Recognition

Tree-Structured Model





[Shi et al., CVPR 2013]

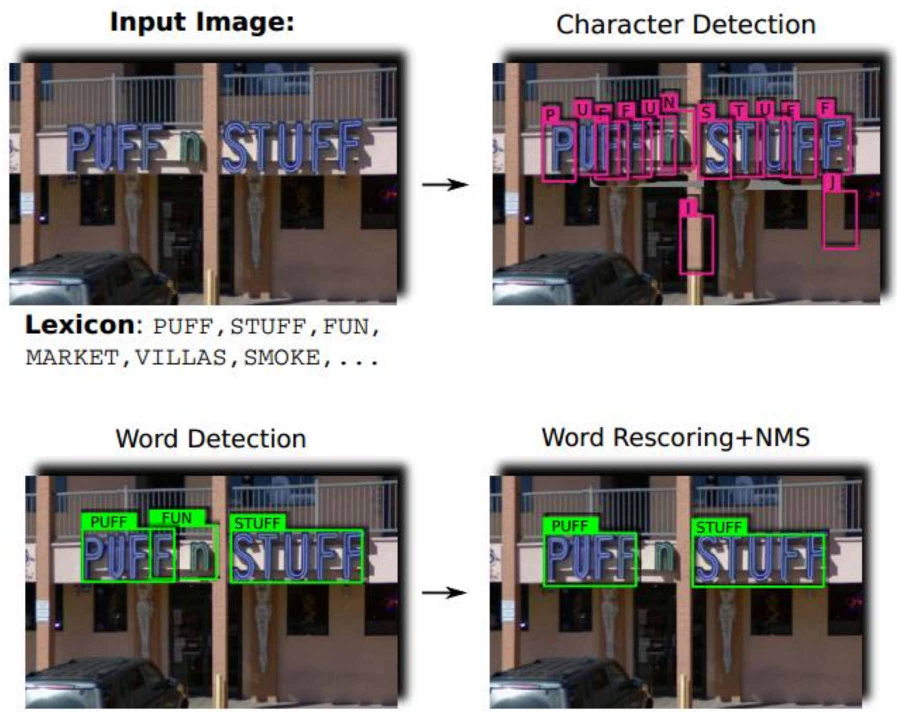
- DPM for character detection, human-designed character structure models and labeled parts
- build a CRF model to incorporate the detection scores, spatial constraints and linguistic knowledge into one framework

Text Recognition

Best practice in scene text recognition:
redundant character candidate extraction
+
high level model for error correction

End-to-End Text Recognition

Lexicon Driven



[Wang et al., ICCV 2011]

- detect characters using Random Ferns + HOG
- find an optimal configuration of a particular word via Pictorial Structure with a Lexicon

End-to-End Text Recognition

Real-Time

[Neumann and Matas, CVPR 2012]

- pose character detection a as sequential selection from the set of Extremal Regions (ERs)
- achieve real-time performance with incrementally computable descriptors

End-to-End Text Recognition

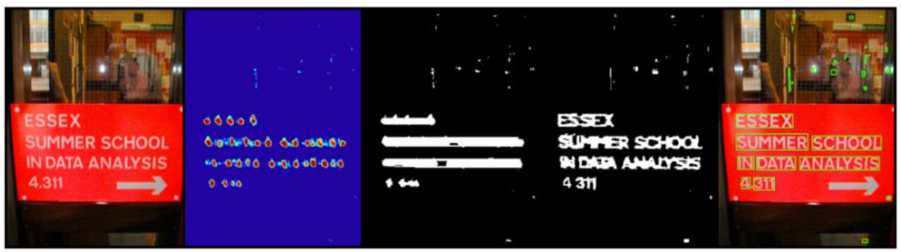
PhotoOCR

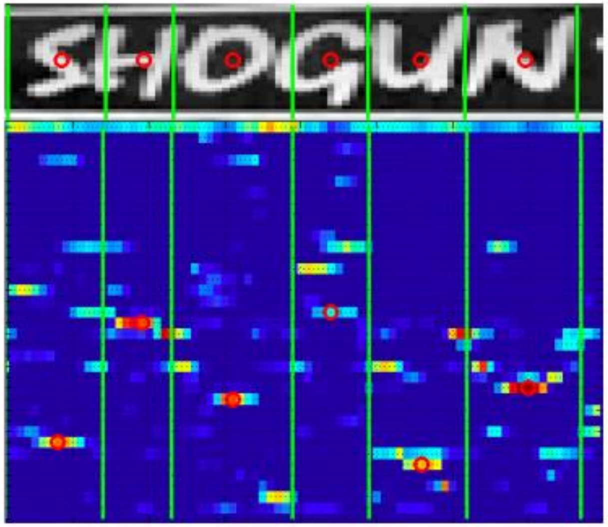
[Bissacco et al., ICCV 2013]

- localize text regions by integrating multiple existing detection methods
- recognize characters with a DNN running on HOG features, instead of raw pixels
- use 2.2 million manually labelled examples for training

End-to-End Text Recognition

Deep Features





[Jaderberg et al., ECCV 2014]

- propose a novel CNN architecture, enabling efficient feature sharing for text detection and character classification
- generate word and character level annotations via automatic data mining of Flickr

End-to-End Text Recognition

Deep learning + Big data
seem to dominate this field

For more details:
[1] Y. Zhu, C. Yao, and X. Bai, Scene Text Detection and Recognition: Recent Advances and Future Trends, Frontier of Computer Science, to appear.

Contents

- Problem definition
- Significance and challenges
- Previous works
- Our algorithms
- Conclusion

Our algorithms

We will introduce three of our works
that propose novel representations
for better text detection and recognition

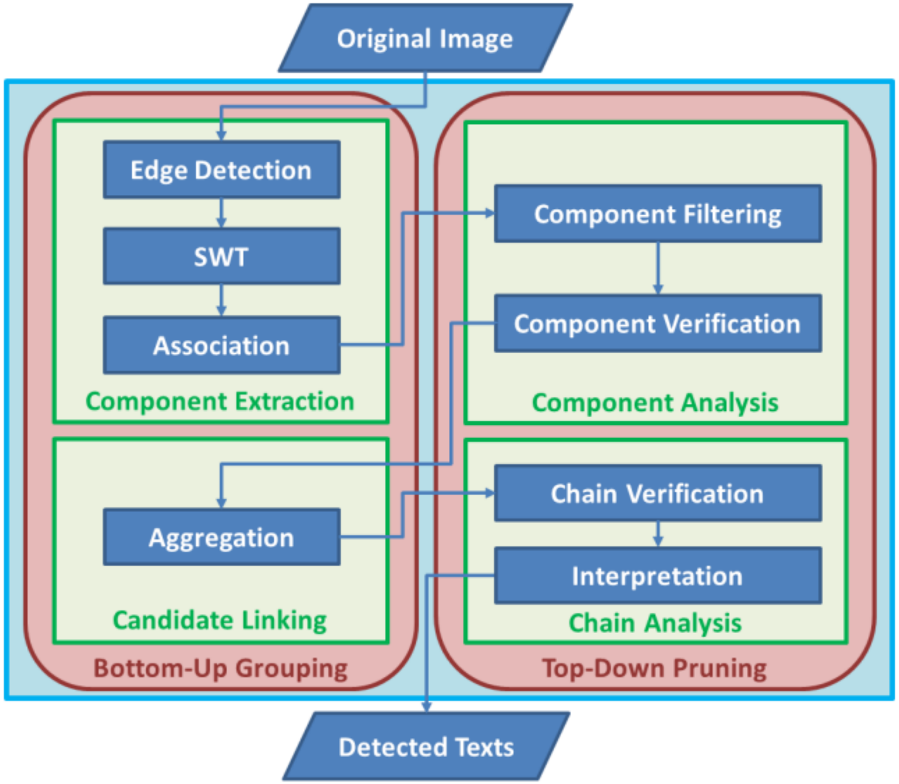
Multi-Oriented Text Detection

detect texts of different orientations, not limited horizontal ones, from natural scenes

[1] Cong Yao, Xiang Bai, Wenyu Liu, Yi Ma, and Zhuowen Tu. Detecting texts of arbitrary orientations in natural images. CVPR, 2012.

[2] Cong Yao, Xiang Bai, and Wenyu Liu. A Unified Framework for Multi-Oriented Text Detection and Recognition. TIP, 2014.

Multi-Oriented Text Detection



algorithmic pipeline

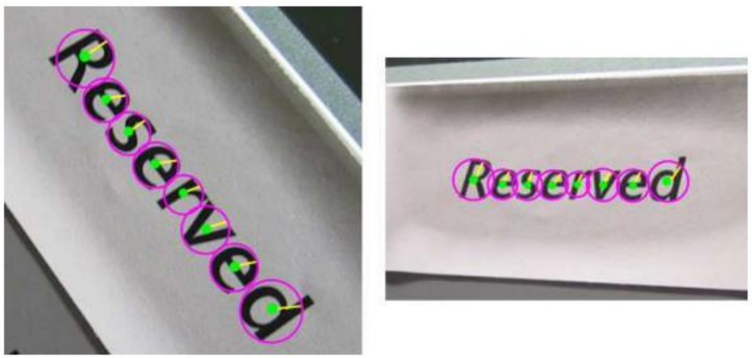
Multi-Oriented Text Detection

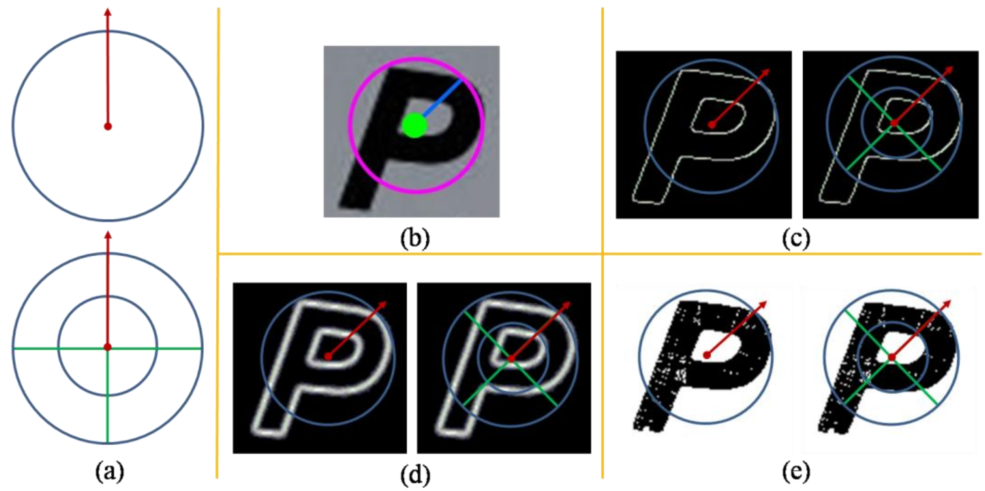
(a) Original image (b) Edge detection (c) SWT (d) Association (e) Component filtering
(f) Component verification (g) Aggregation (h) Chain verification (i) Interpretation (j) Detected texts

full process of text detection

Multi-Oriented Text Detection

Main Contribution





two sets of rotation-invariant features that facilitate multi-oriented text detection:

•component level: estimate center, scale, and direction before feature computation…

•chain level: size variation, color self-similarity, structure self-similarity…

Multi-Oriented Text Detection

Main Contribution

(a) (b) (c)

an evaluation protocol that is suitable for assessing algorithms designed for texts of arbitrary orientations

Multi-Oriented Text Detection

Qualitative Results

detection examples on the ICDAR 2003 dataset
(mainly horizontal texts)

Multi-Oriented Text Detection

Qualitative Results

detection examples on the MSRA TD-500 dataset
(including texts of different orientations)

Multi-Oriented Text Detection

Qualitative Results

detected texts in various languages

Multi-Oriented Text Detection

Quantitative Results

Algorithm	Precision	Recall	F-measure
TD-Mixture	0.69	0.66	**0.67**
TD-ICDAR	0.68	0.66	0.66
Epshtein et al. [7]	**0.73**	0.60	0.66
Yi et al. [29]	0.71	0.62	0.62
Becker et al. [20]	0.62	**0.67**	0.62
Chen et al. [6]	0.60	0.60	0.58
Zhu et al. [20]	0.33	0.40	0.33
Kim et al. [20]	0.22	0.28	0.22
Ezaki et al. [20]	0.18	0.36	0.22

compare favorably with the state-of-the-art algorithms when handling horizontal texts

Multi-Oriented Text Detection

Quantitative Results

Algorithm	Precision	Recall	F-measure
TD-Mixture	**0.63**	**0.63**	**0.60**
TD-ICDAR	0.53	0.52	0.50
Epshtein et al. [7]	0.25	0.25	0.25
Chen et al. [6]	0.05	0.05	0.05

achieve much better performance on texts of arbitrary orientations

Mid-Level Elements for Text Recognition

a learned multi-scale mid-level representation for scene text recognition

[1] Cong Yao, Xiang Bai, Baoguang Shi, and Wenyu Liu. Strokelets: A Learned Multi-Scale Representation for Scene Text Recognition. CVPR, 2014.

Mid-Level Elements for Text Recognition

the discriminative clustering algorithm proposed in [Singh et al, ECCV 2012] is adopted to learn a set of mid-level primitives, called strokelets, which capture the substructures of characters at different granularities

Mid-Level Elements for Text Recognition

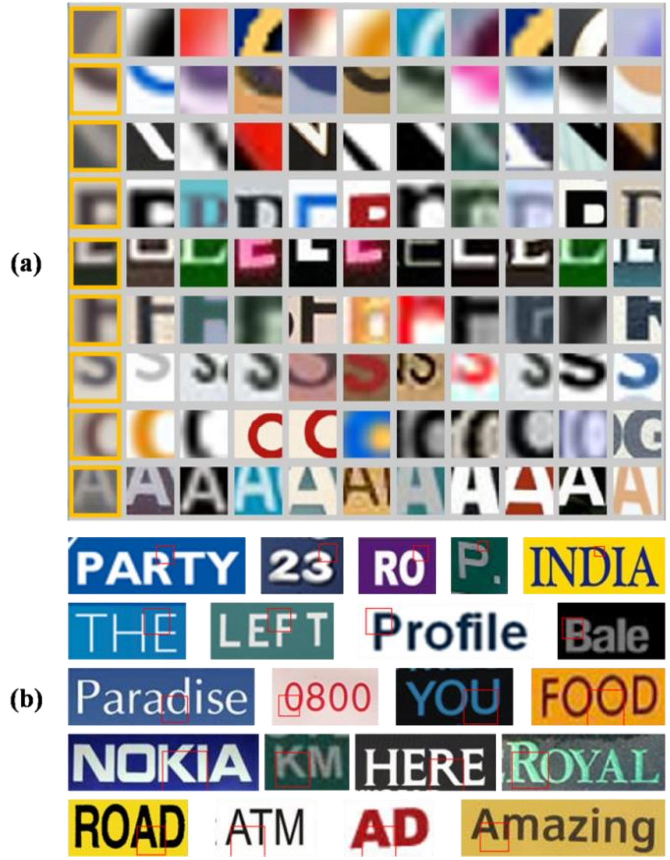
Algorithm 1 Algorithm for strokelet generation

Require: Training set S, interval $[a, b]$, strokelet count Γ

- $\{\mathcal{D}, \mathcal{N}\} \Leftarrow construct(S)$ ▷ Construct Discovery set $\mathcal{D}$ and Natural World set $\mathcal{N}$ from S
- $\mathcal{D} \Rightarrow \{D_1, D_2\}$; $\mathcal{N} \Rightarrow \{N_1, N_2\}$ ▷ Split $\mathcal{D}$ and $\mathcal{N}$ into equal sized disjoint subsets
- $R \Leftarrow random\_sample(D_1, [a, b])$ ▷ Sample patches with scale ratio randomly drawn from $[a, b]$
- $K \Leftarrow cluster(R, \lambda\Gamma)$ ▷ Cluster sampled patches, the initial cluster number is set to $\lambda\Gamma$ ($\lambda > 1$)
- **repeat** ▷ Iterate until convergence
  - **for all** i such that $size(K[i]) \geq \tau$ **do** ▷ Maintain clusters with enough members, τ is a predefined threshold
    - $C_{new}[i] \Leftarrow train(K[i], N_1)$ ▷ Train classifier for each cluster
    - $K_{new}[i] \Leftarrow detect\_top(C[i], D_2, q)$ ▷ Find top q new members in the other discovery subset
  - **end for**
  - $K \Leftarrow K_{new}$; $C \Leftarrow C_{new}$ ▷ Update clusters and classifiers
  - $swap(D_1, D_2)$; $swap(N_1, N_2)$ ▷ Swap the two subsets
- **until** converged
- $A[i] \Leftarrow score(K[i])\ \forall i$ ▷ Compute score for each cluster, see [28] for details
- $\Omega \Leftarrow select\_top(K, C, A, \Gamma)$ ▷ Sort according to scores and select top Γ clusters and classifiers
- **return** Ω

algorithmic pipeline for learning strokelets

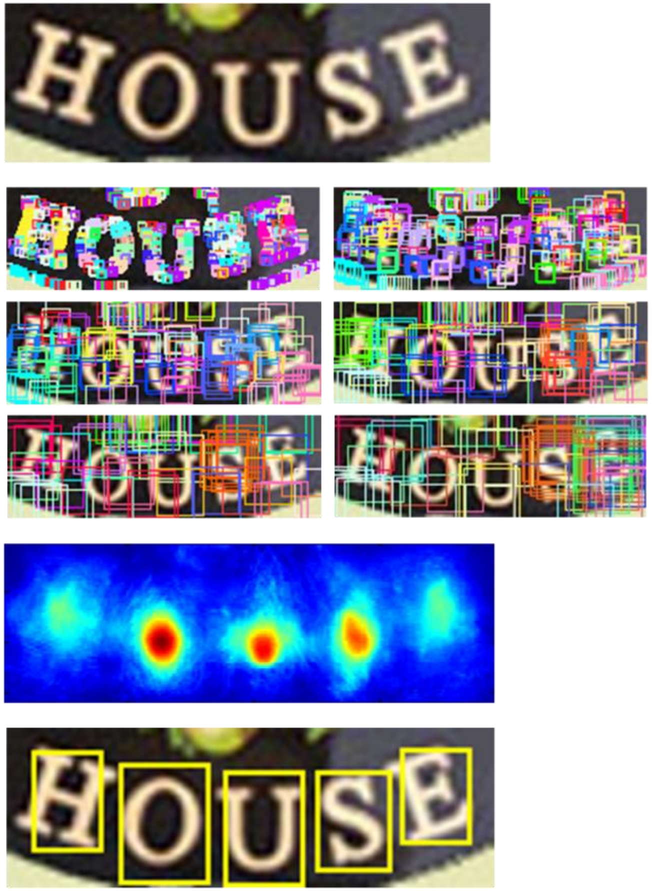
Mid-Level Elements for Text Recognition



learned strokelets and the instances shown in the original images

Mid-Level Elements for Text Recognition

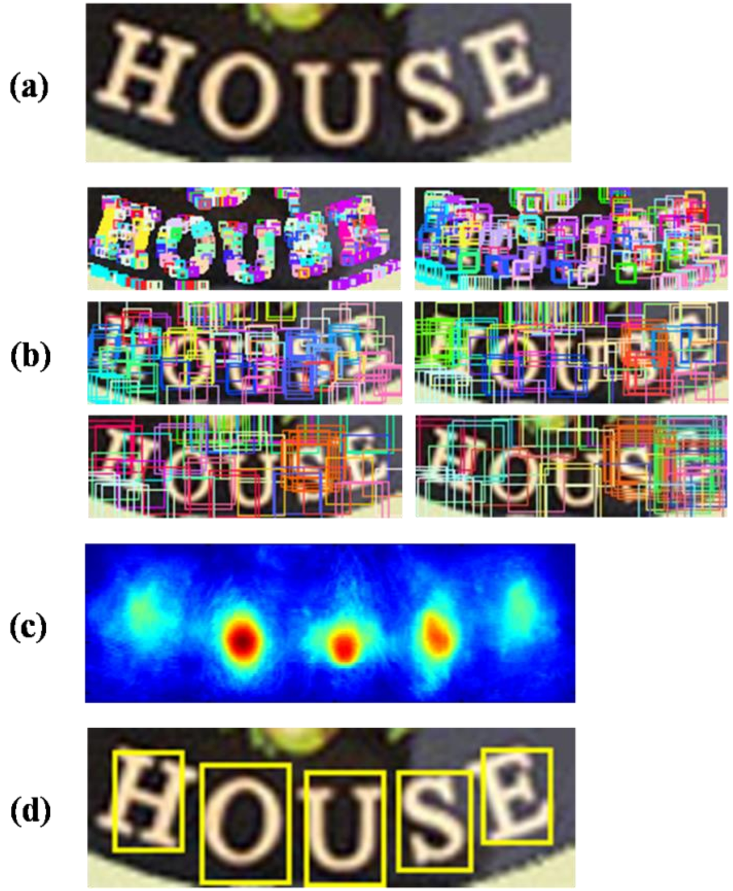
(a)
(b)
(c)
(d)



Character detection:
Identify candidates via multi-scale strokelet detection and voting

character detection and description with strokelets

Mid-Level Elements for Text Recognition



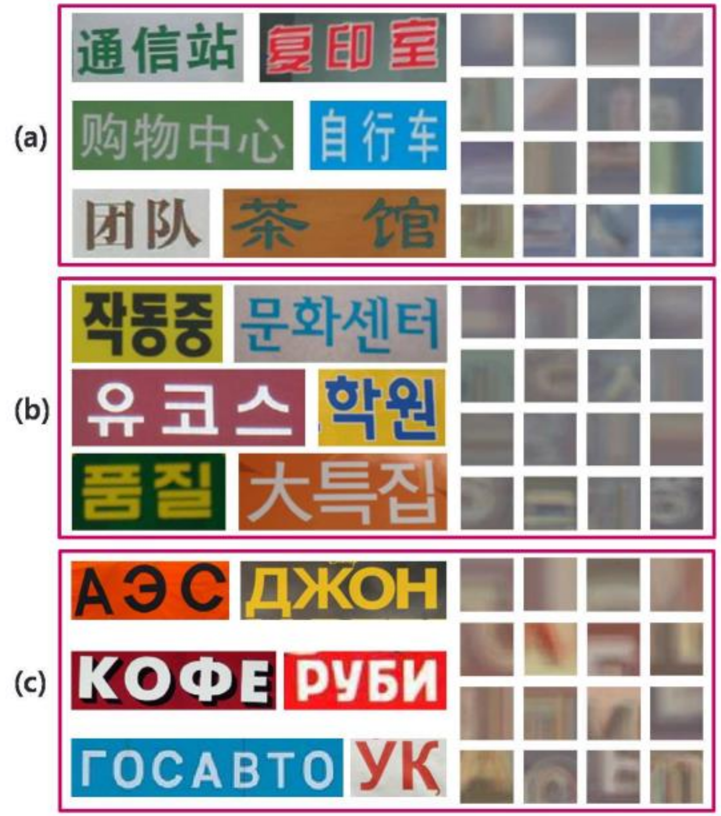
Character description:

•Bag of Strokelets: A histogram feature is formed by binning all the strokelets

• HOG: A template is constructed for each character candidate

character detection and description with strokelets

Qualitative Results



learned strokelets on different languages: Chinese, Korean, Russian

Mid-Level Elements for Text Recognition

Qualitative Results

robust to interference factors like noise, blur, non-uniform illumination, partial occlusion, font variation, scale change

Quantitative Results

Lexicon	Small	Medium	Large
Proposed	**80.2**	**69.3**	38.3
Higher Order (with edit distance)	68.25	55.50	28
Higher Order (without edit distance)	64.10	53.16	**44.30**
Pairwise CRF (with edit distance)	66	57.5	24.25
Pairwise CRF (without edit distance)	55.50	51.25	20.25
ABBYY9.0	24.33	-	-

achieve state-of-the-art performance on IIIT 5K-Word, a large, challenging dataset in this field

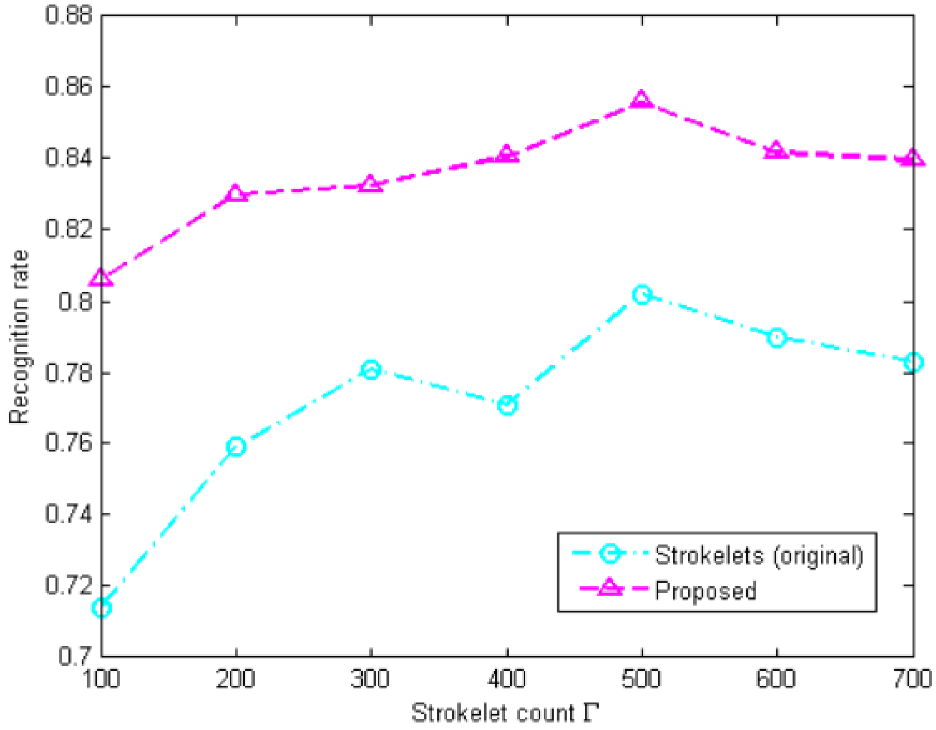
Quantitative Results

Dataset	ICDAR 2003(FULL)	ICDAR 2003(50)	SVT
Proposed	80.33	88.48	75.89
CNN	**84**	**90**	70
Whole	-	89.69	**77.28**
TSM+CRF	79.30	87.44	73.51
TSM+PLEX	70.47	80.70	69.51
Multi-Class Hough Forests	-	85.70	-
Large-Lexicon Attribute-Consistent	82.8	-	72.9
Higher Order (with edit distance)	-	80.28	73.57
Higher Order (without edit distance)	-	72.01	68.00
Pairwise CRF (with edit distance)	-	81.78	73.26
Pairwise CRF (without edit distance)	-	69.90	62.28
SYNTH+PLEX	62	76	57
ICDAR+PLEX	57	72	56
ABBYY9.0	55	56	35

achieve highly competitive performance on ICDAR 2003 and SVT

Mid-Level Elements for Text Recognition

Recent Advance



achieve significantly enhanced performance (5% improvement on average) after modification

Quantitative Results

Γ	100	200	300	400	500	600	700
Accuracy(%)	71.4	75.9	78.1	77.1	**80.2**	79.0	78.3

impact of strokelet set size

Quantitative Results

Scale(a=b)	0.2	0.3	0.4	0.5	0.6	0.7	multi-scale
Accuracy(%)	59.9	71.9	74.1	74.4	74.8	74.3	**80.2**

advantage of multi-scale representation

Symmetry-Based Text Line Detection in Natural Scenes

- Text lines always bear distinctive symmetry and self-similarity properties. By considering these properties, we could find text region without seeking for individual characters.

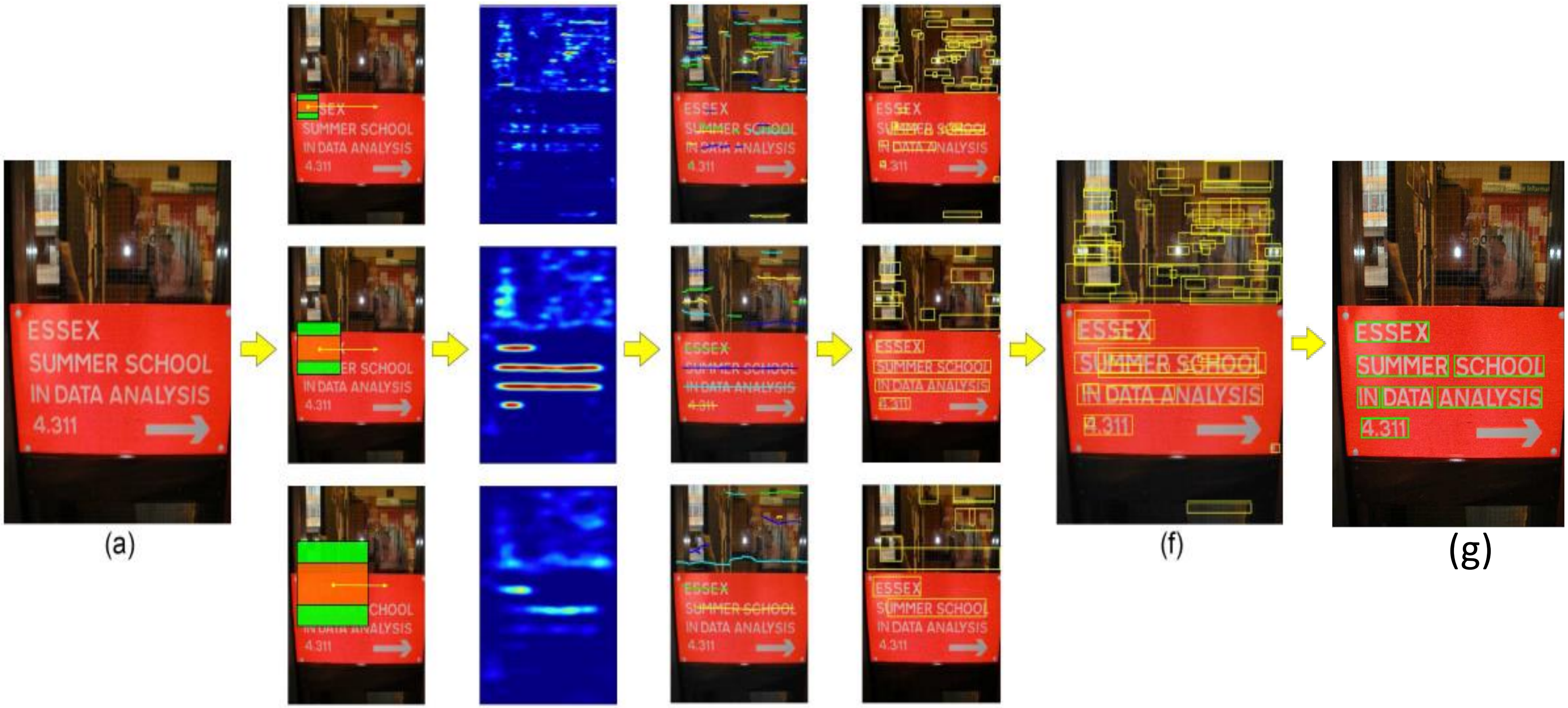
[1] Zheng Zhang, Wei Shen, Cong Yao, Xiang Bai. Symmetry-based Text Line Detection in Natural Scenes, submitted to IEEE CVPR, 2015. (2,2,3)

Overview of the proposed methodology

1. Feature extraction at multiple scales.
2. Symmetry probability estimation.
3. Axes sought in the symmetry probability maps.
4. Bounding box estimation and proposals generation.
5. False positive removal and word partition

Symmetry-Based Text Line Detection in Natural Scenes

Overview of the proposed methodology



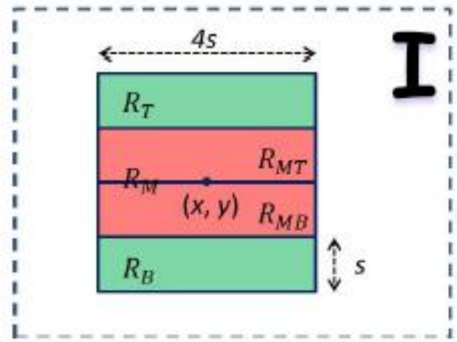
(a) (b) (c) (d) (e) (f) (g)

Feature Extraction and Symmetry probability estimation

1.Symmetry feature
2.Appearance Feature (LBP)
3.Probability estimation by Random Forest at Multiple scales

Symmetry-Based Text Line Detection in Natural Scenes

Symmetry feature



- Self-Similarity

$$S^c_{x,y} = \chi^2(h^c_{x,y}(R_{MT}), h^c_{x,y}(R_{MB}))$$

- Disimilarity

$$Ct^c_{x,y} = \chi^2(h^c_{x,y}(R_T), h^c_{x,y}(R_{MT}))$$

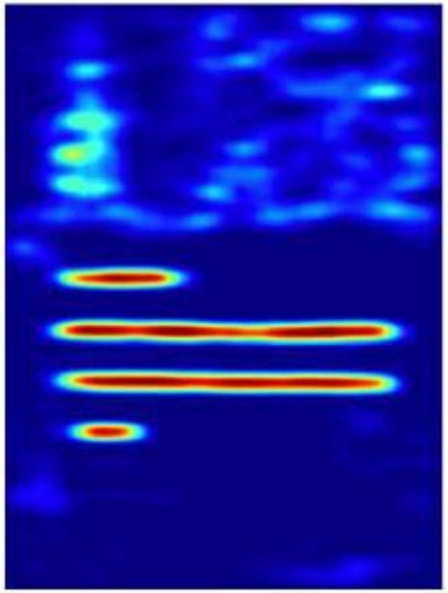
$$Cb^c_{x,y} = \chi^2(h^c_{x,y}(R_B), h^c_{x,y}(R_{MB}))$$

- Calculation at LAB, Gradient and Textons channels

Symmetry-Based Text Line Detection in Natural Scenes

Axes sought in the symmetry probability maps

1. Non-Maximum Suppression
2. Axes linking
*Angular Difference Constraint
*Distance Constraint
3. Above two steps are applied at each scale respectively



Symmetry-Based Text Line Detection in Natural Scenes

Bounding box estimation and proposals generation

False positive removal and word partition

1.Character level CNN classifier(Text Spotting, ECCV2014, Zisserman)

   * Word partition

   * Preliminary false positive removal

2.Textline level CNN classifier for further filter

Experimental result

ICDAR 2011

Algorithm	Precision	Recall	F-measure
Proposed	0.84	**0.76**	**0.80**
Huang *et al.* [9]	**0.88**	0.71	0.78
Yin *et al.* [40]	0.863	0.683	0.762
Koo *et al.* [13]	0.814	0.687	0.745
Yao *et al.* [35]	0.822	0.657	0.730
Huang *et al.* [8]	0.82	0.75	0.73
Neumann *et al.* [24]	0.793	0.664	0.723
Shi *et al.* [29]	0.833	0.631	0.718
Kim *et al.* [28]	0.830	0.625	0.713
Neumann *et al.* [23]	0.731	0.647	0.687
Yi *et al.* [38]	0.672	0.581	0.623
Yang *et al.* [28]	0.670	0.577	0.620
Neumann *et al.* [28]	0.689	0.525	0.596
Shao *et al.* [28]	0.635	0.535	0.581

ICDAR 2013

Algorithm	Precision	Recall	F-measure
Proposed	**0.88**	**0.74**	**0.80**
iwrr2014 [41]	0.86	0.70	0.77
USTB TexStar [40]	**0.88**	0.66	0.76
Text Spotter [23]	**0.88**	0.65	0.74
CASIA_NLPR [1]	0.79	0.68	0.73
Text_Detector_CASIA [29]	0.85	0.63	0.72
I2R_NUS_FAR [1]	0.75	0.69	0.72
I2R_NUS [1]	0.73	0.66	0.69
TH-TextLoc [1]	70	0.65	0.67

Symmetry-Based Text Line Detection in Natural Scenes

Contributions of different types of feature

Feature	Precision	Recall	F-measure
symmetry	0.80	0.65	0.72
appearance	0.79	0.57	0.66
symmetry+appearance	0.84	0.76	0.80

Character detection rates of different methods on the ICDAR 2013 dataset

Algorithm	Detection Rate	Proposal Number
Proposed	**0.977**	**1310**
MSER (Gray+LUV)	0.964	8415

Examples


CAMOUFLAGE
EINWURF
EURO 0,05-2,00
1) GELD EINWERFEN
2) NUMMER DES FACHES WÄHLEN
3) GELDRÜCKGABE : TASTE (R) DRÜCKEN
iiyama
POCKET FLASH READER
中華料理
パール
2076
CO-OPERATIVE
משטרת ישראל
ISRAEL POLICE
شرطة إسرائيل

Limitations

1.Distinguish ability of features is not good enough(especially appearance feature).
2.Axes sought is not robust enough in street view dataset.
3.High time consumption

Future works

1. To explore better feature representation
2. To explore better axes sought method.
3. To expand our works to multi orientations text detection.

Contents

- Problem definition
- Significance and challenges
- Previous works
- Our algorithms
- Conclusion

Conclusion

The common key to the success of the above surveyed text detection and recognition methods is representation, just as in many other vision problems

Conclusion

Conventional methods rely on human designed representations (MSER, SWT, HOG), while CNN based algorithms directly learn representations from data

Conclusion

Learning representation from data
is a future trend

Conclusion

But there is still a long way to go,
since challenges remain:
multi-scale,
multi-orientation,
multi-language,
…

A brief advertisement

Valse Webinar

QQ群号： 364188996

Valse 2015 (Chengdu):

http://www.uestcrobot.net/valse2015/index.html

Thank You!

Email: xbai@hust.edu.cn