Lecture 7:

Scene Text Detection and Recognition

Dr. Cong Yao

Megvii (Face++) Researcher

yaocong@megvii.com

Face++ 旷视

Outline

- Background and Introduction
- Conventional Methods
- Deep Learning Methods
- Datasets and Competitions
- Conclusion and Outlook

Outline

- Background and Introduction
- Conventional Methods
- Deep Learning Methods
- Datasets and Competitions
- Conclusion and Outlook

Text as a Hallmark of Civilization

Characteristics of Civilization

- Urban development
- Social stratification
- Symbolic systems of communication
- Perceived separation from natural environment

https://en.wikipedia.org/wiki/Civilization

Text as a Hallmark of Civilization

Characteristics of Civilization

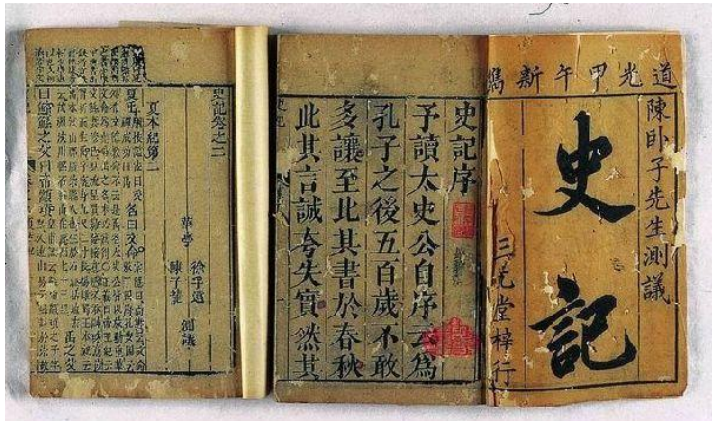
- Urban development
- Social stratification
- Symbolic systems of communication: **text**
- Perceived separation from natural environment

https://en.wikipedia.org/wiki/Civilization

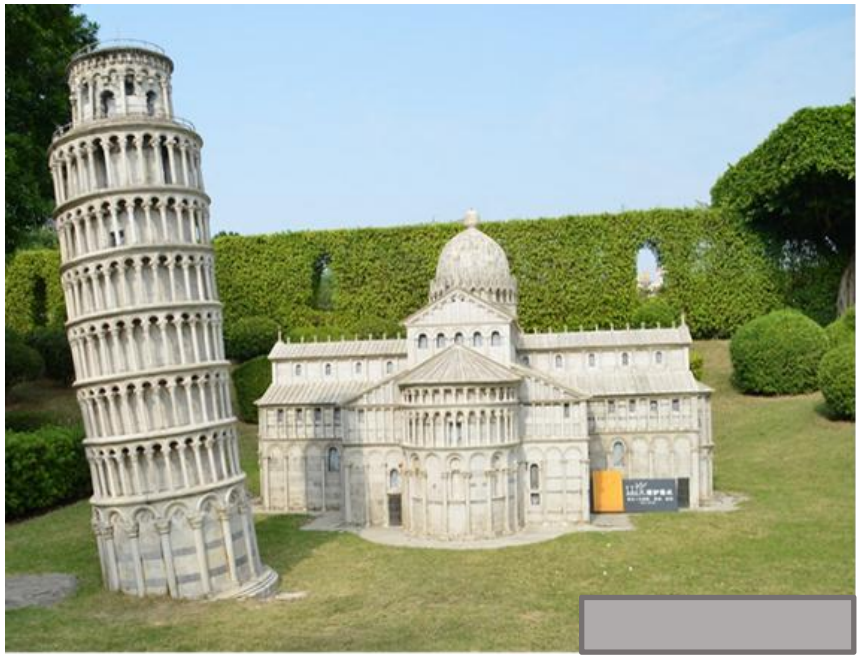
Text as a Carrier of High Level Semantics

Text is an invention of humankind that

- carries rich and precise high level semantics
- conveys human thoughts and emotions



Text as a Cue in Visual Recognition



(a)

(b)

Text as a Cue in Visual Recognition

Text is complementary to other visual cues, such as contour, color and texture

(a)

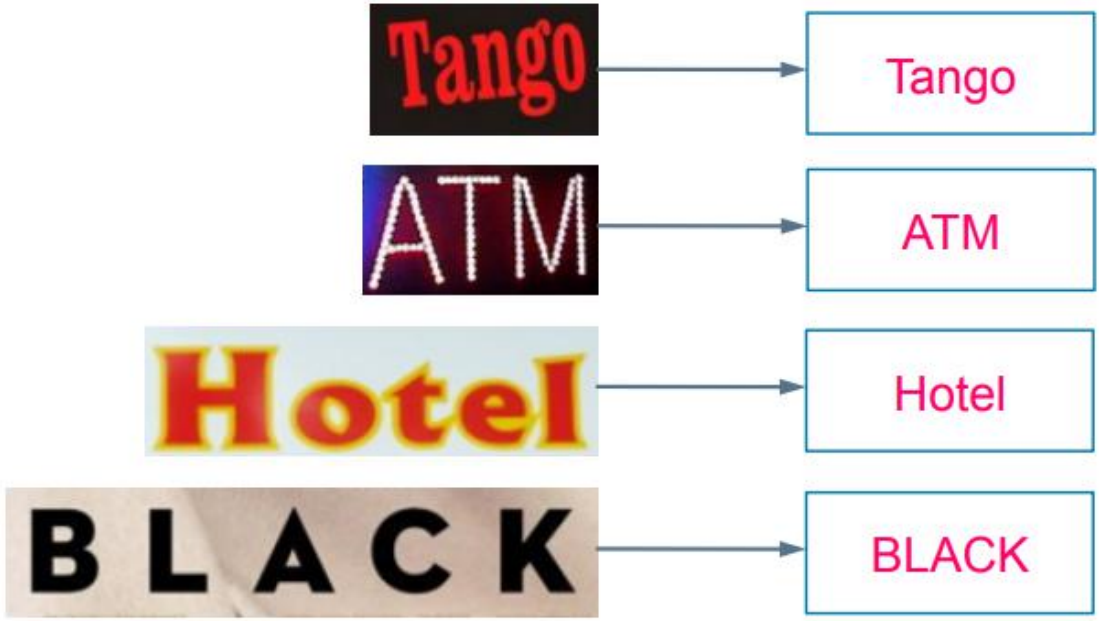
(b)

Face++ 旷视

Problem Definition

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

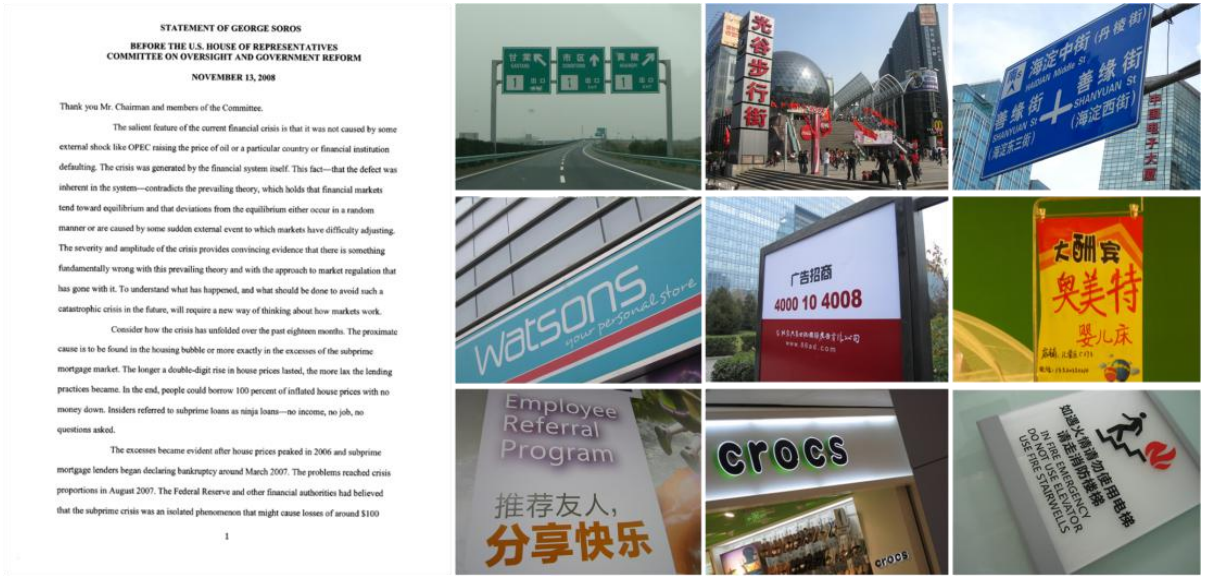
Problem Definition



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

Challenges

Traditional OCR *vs.* Scene Text Detection and Recognition



STATEMENT OF GEORGE SOROS

BEFORE THE U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON OVERSIGHT AND GOVERNMENT REFORM

NOVEMBER 13, 2008

Thank you Mr. Chairman and members of the Committee.

The salient feature of the current financial crisis is that it was not caused by some external shock like OPEC raising the price of oil or a particular country or financial institution defaulting. The crisis was generated by the financial system itself. This fact—that the defect was inherent in the system—contradicts the prevailing theory, which holds that financial markets tend toward equilibrium and that deviations from the equilibrium either occur in a random manner or are caused by some sudden external event to which markets have difficulty adjusting. The severity and amplitude of the crisis provides convincing evidence that there is something fundamentally wrong with this prevailing theory and with the approach to market regulation that has gone with it. To understand what has happened, and what should be done to avoid such a catastrophic crisis in the future, will require a new way of thinking about how markets work.

Consider how the crisis has unfolded over the past eighteen months. The proximate cause is to be found in the housing bubble or more exactly in the excesses of the subprime mortgage market. The longer a double-digit rise in house prices lasted, the more lax the lending practices became. In the end, people could borrow 100 percent of inflated house prices with no money down. Insiders referred to subprime loans as ninja loans—no income, no job, no questions asked.

The excesses became evident after house prices peaked in 2006 and subprime mortgage lenders began declaring bankruptcy around March 2007. The problems reached crisis proportions in August 2007. The Federal Reserve and other financial authorities had believed that the subprime crisis was an isolated phenomenon that might cause losses of around $100

1

- clean background vs. cluttered background
- regular font vs. various fonts
- plain layout vs. complex layouts
- monotone color vs. different colors

Face++ 旷视

Challenges

Diversity of scene text:

different colors, scales, orientations, fonts, languages…

Face++ 旷视

Challenges

Complexity of background:

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

Face++ 旷视

Challenges

Various interference factors:

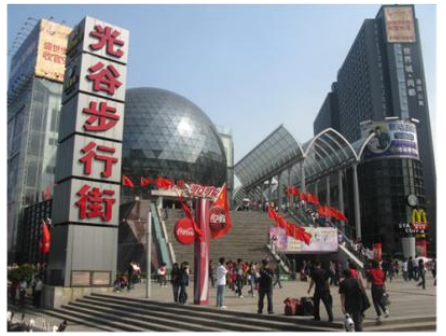
noise, blur, non-uniform illumination, low resolution, partial occlusion…

Face++ 旷视

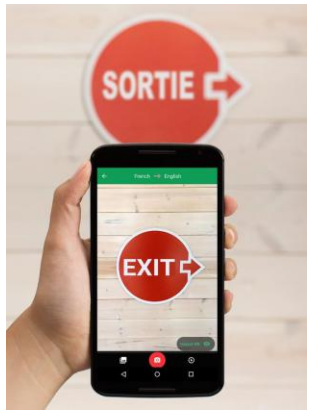
Applications

Card Recognition

Product Search



Geo-location



Instant Translation

Self-driving Car

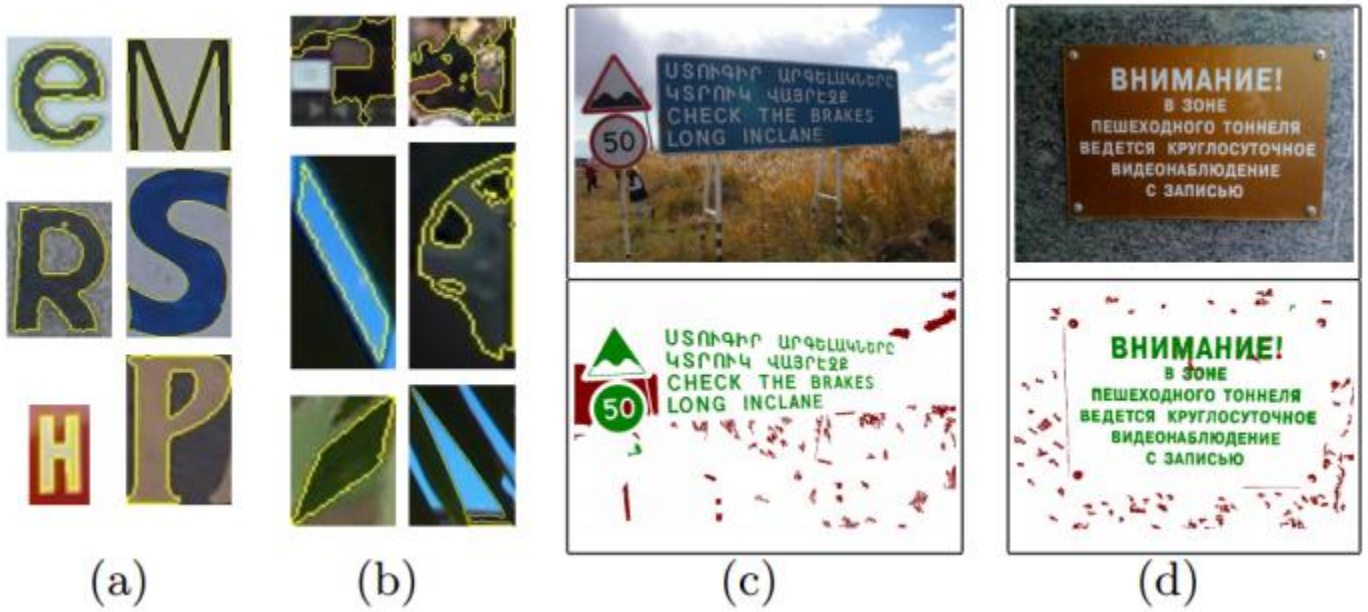
Industry Automation

Face++ 旷视

Outline

- Background and Introduction
- Conventional Methods
- Deep Learning Methods
- Conclusion and Outlook

Detection: MSER

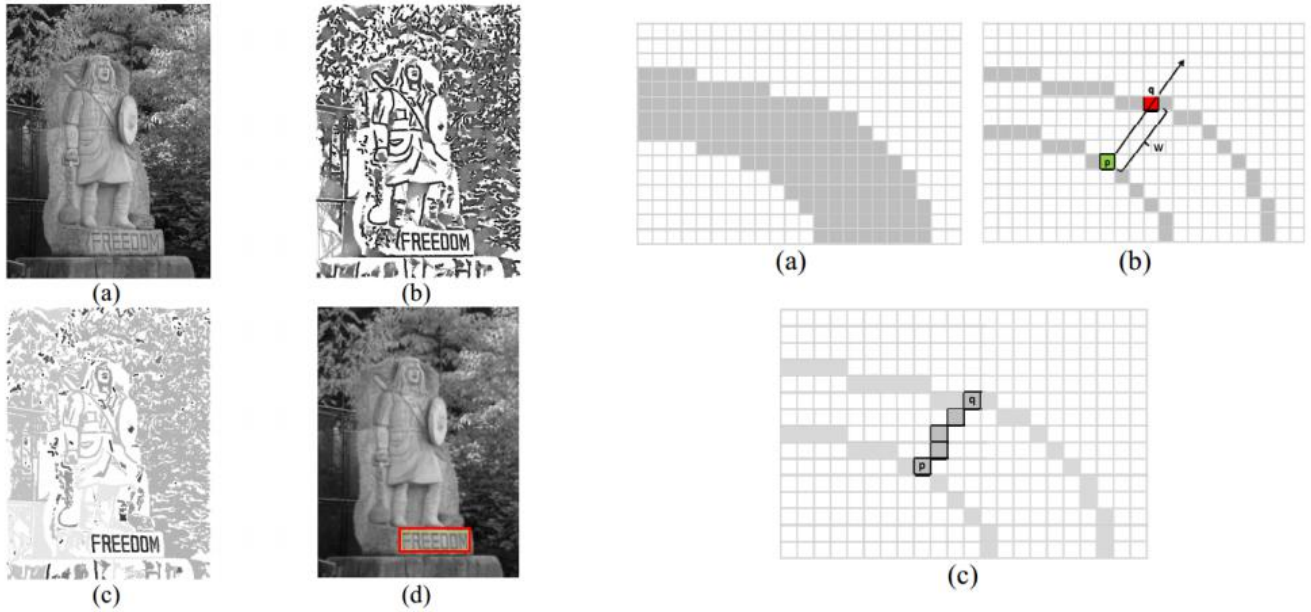


- extract character candidates using MSER (Maximally Stable Extremal Regions), assuming similar color within each character
- robust, fast to compute, independent of scale
- limitation: can only handle horizontal text, due to features and linking strategy

Face++ 旷视

Neumann and Matas. A method for text localization and recognition in real-world images. ACCV, 2010.

Detection: SWT



- extract character candidates with SWT (Stroke Width Transform), assuming consistent stroke width within each character
- robust, fast to compute, independent of scale
- limitation: can only handle horizontal text, due to features and linking strategy

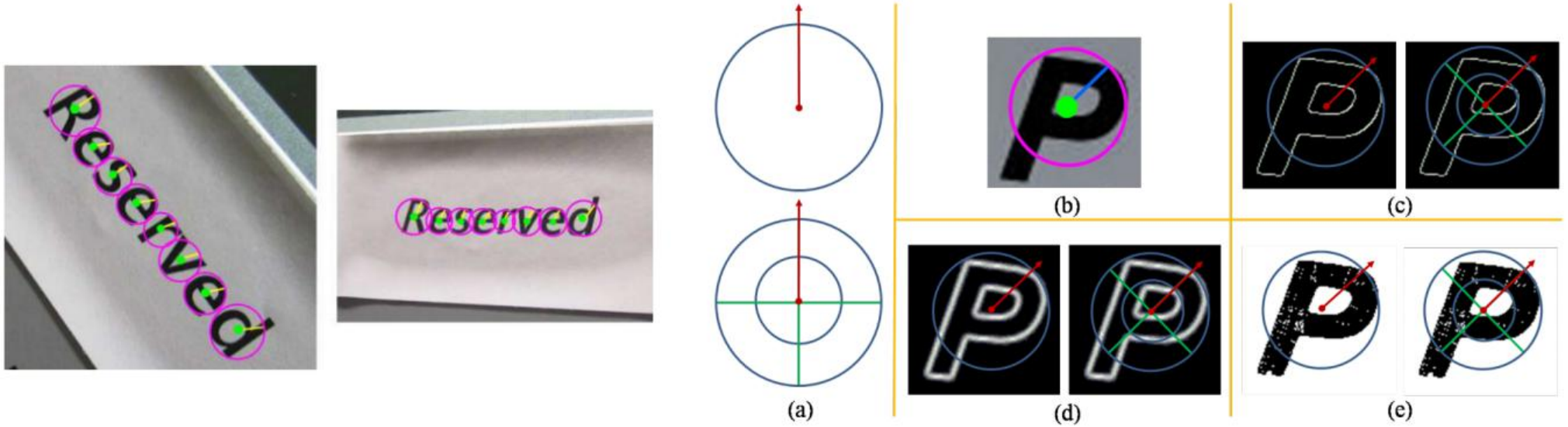
Epshtein et al.. Detecting Text in Natural Scenes with Stroke Width Transform. CVPR, 2010.

Detection: Multi-Oriented

- detect text instances of different orientations, not limited horizontal ones

Yao et al.. Detecting texts of arbitrary orientations in natural images. CVPR, 2012.

Detection: Multi-Oriented



- adopt SWT to hunt character candidates
- design rotation-invariant features that facilitate multi-oriented text detection
- propose a new dataset (MSRA-TD500) that contains text instances of different directions

Face++ 旷视

Yao et al.. Detecting texts of arbitrary orientations in natural images. CVPR, 2012.

Summary

- Role and status of MSER and SWT
  - two representative and dominant approaches before the era of deep learning
  - inspired a lot of subsequent works

[PDF] A method for text localization and recognition in real-world images
cmp.felk.cvut.cz/~matas/papers/neumann-text-accv10.pdf ▼ 翻译此页
作者：L Neumann - 被引用次数：311 - 相关文章
2010年11月8日 - Text localization and recognition in images of real-world scenes has ... Most published methods for text localization and recognition [1,6–8] are.

[PDF] Real-Time Scene Text Localization and Recognition - Center for ...
cmp.felk.cvut.cz/~matas/papers/neumann-2012-rt_text-cvpr.pdf ▼ 翻译此页
作者：L Neumann - 被引用次数：462 - 相关文章
An end-to-end real-time scene text localization and recognition method is presented. The real-time perfor- mance is achieved by posing the character detection ...

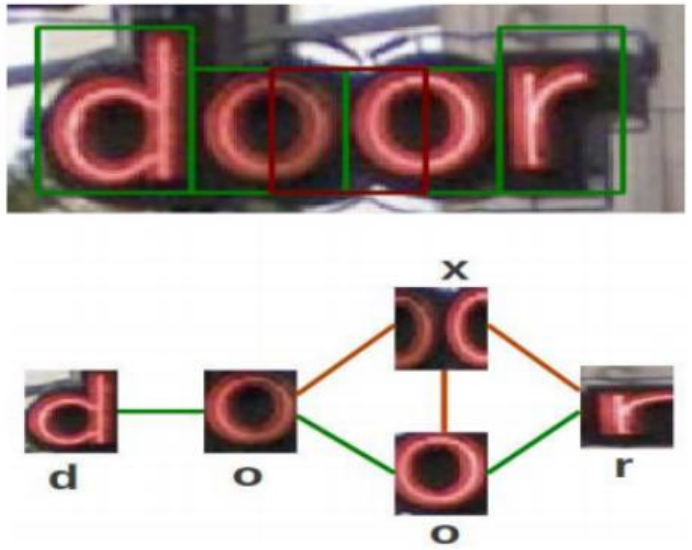
[PDF] Detecting Text in Natural Scenes with Stroke Width Transform
cmp.felk.cvut.cz/.../Detecting%20Text%20in%20Natural%20Scenes%20wit... ▼ 翻译此页
作者：B Epshtein - 被引用次数：1041 - 相关文章
We present a novel image operator that seeks to find the value of stroke width for each image pixel, and demonstrate its use on the task of text detection in ...

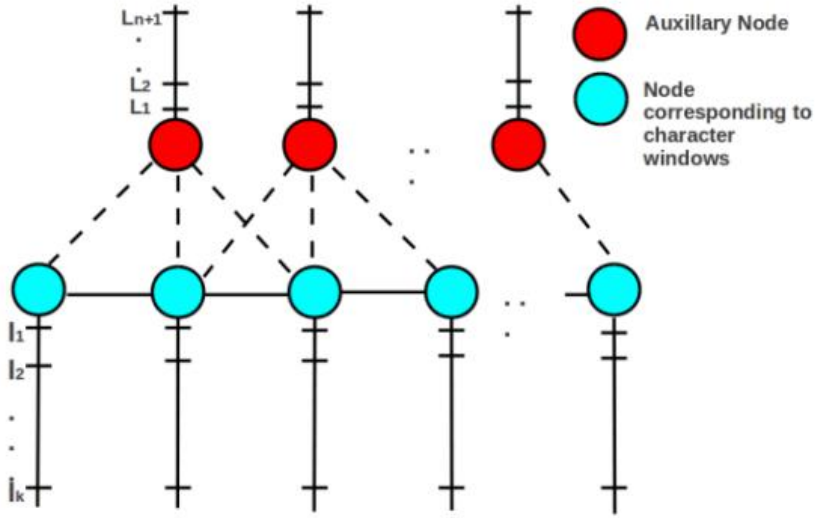
Face++ 旷视

Summary

- Common practices in scene text detection
  - extract character candidates by seeking connected components
  - eliminate non-text components using hand-crafted features (geometric features, gradient features) and strong classifiers (SVM ,Random Forest)
  - form words or text lines with pre-defined rules and parameters

Recognition: Top-Down and Bottom-Up Cues



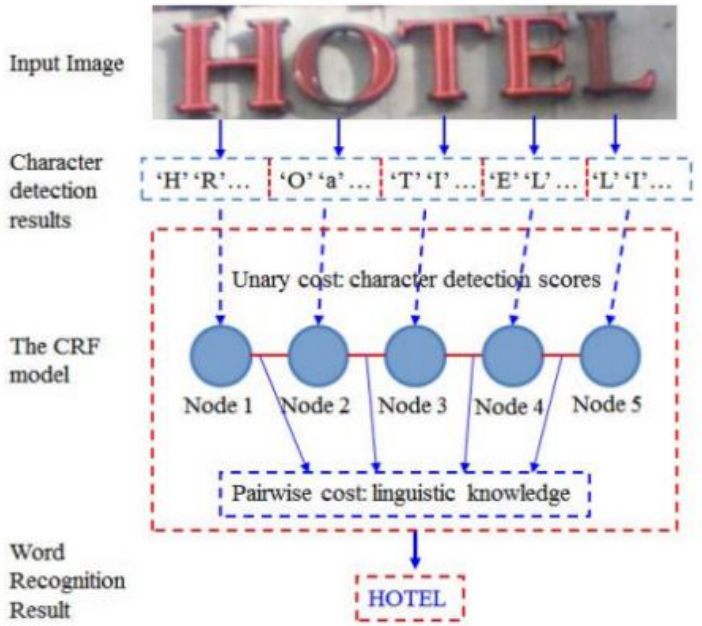


- seek character candidates using sliding window, instead of binarization
- construct a CRF model to impose both bottom-up (i.e. character detections) and top-down (i.e. language statistics) cues

Face++ 旷视

Mishra et al.. Top-down and bottom-up cues for scene text recognition. CVPR, 2012.

Recognition: Tree-Structured Model



- use 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

Face++ 旷视

Shi et al.. Scene Text Recognition using Part-Based Tree-Structured Character Detection. CVPR, 2013.

End-to-End Recognition: Lexicon Driven

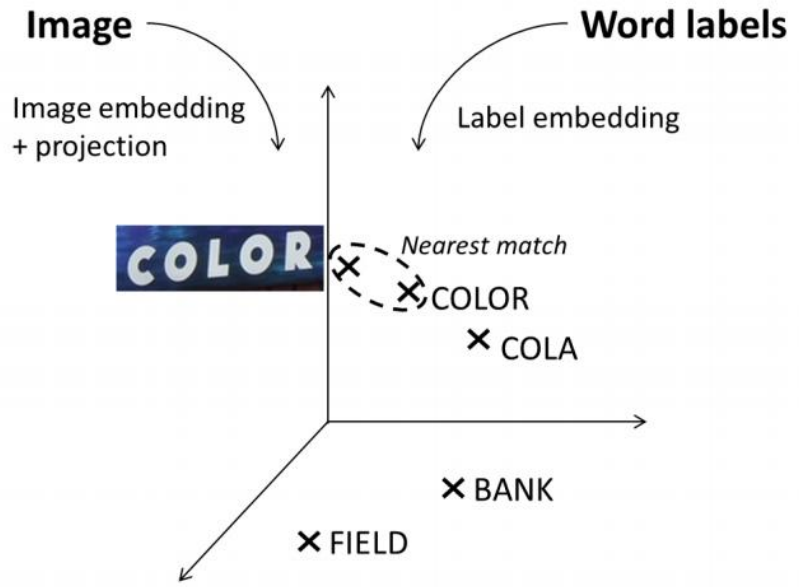
- end-to-end: perform both detection and recognition
- detect characters using Random Ferns + HOG
- find an optimal configuration of a particular word via Pictorial Structure with a Lexicon

Wang et al.. End-to-End Scene Text Recognition. ICCV, 2011.

Summary

- Common practices in scene text recognition
  - redundant character candidate extraction and recognition
  - high level model for error correction

Recognition: Label Embedding



- learn a common space for images and labels (words)
- given an image, text recognition is realized by retrieving the nearest word in the common space
- limitation: unable to handle out-of-lexicon words

Rodriguez-Serrano et al.. Label Embedding: A Frugal Baseline for Text Recognition. IJCV, 2015.

Outline

- Background and Introduction
- Conventional Methods
- Deep Learning Methods
- Datasets and Competitions
- Conclusion and Outlook

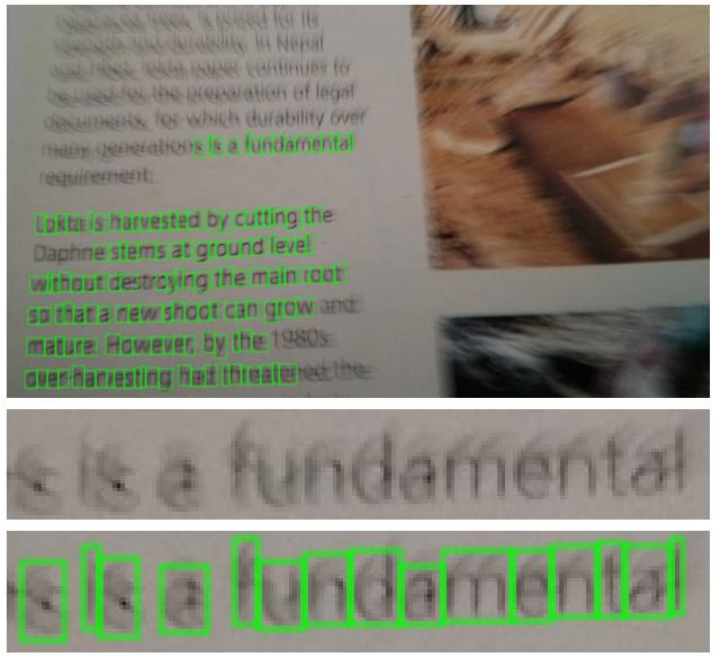
End-to-End Recognition: PhotoOCR

- 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 (in contrast to 2K training examples in the largest public dataset at that time)

Face++ 旷视

Bissacco et al.. PhotoOCR: Reading Text in Uncontrolled Conditions. ICCV, 2013.

End-to-End Recognition: PhotoOCR

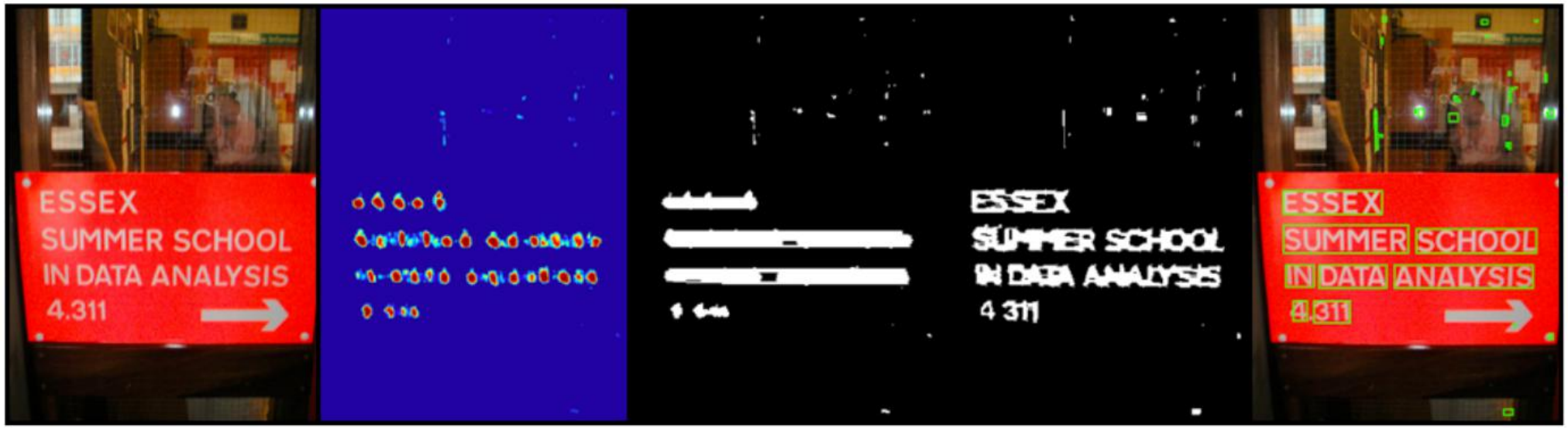


- also propose a mechanism for automatically generating training data
- perform OCR on web images using the trained system
- preliminary recognition results are verified and corrected by search engine

Face++ 旷视

Bissacco et al.. PhotoOCR: Reading Text in Uncontrolled Conditions. ICCV, 2013.

End-to-End Recognition: Deep Features

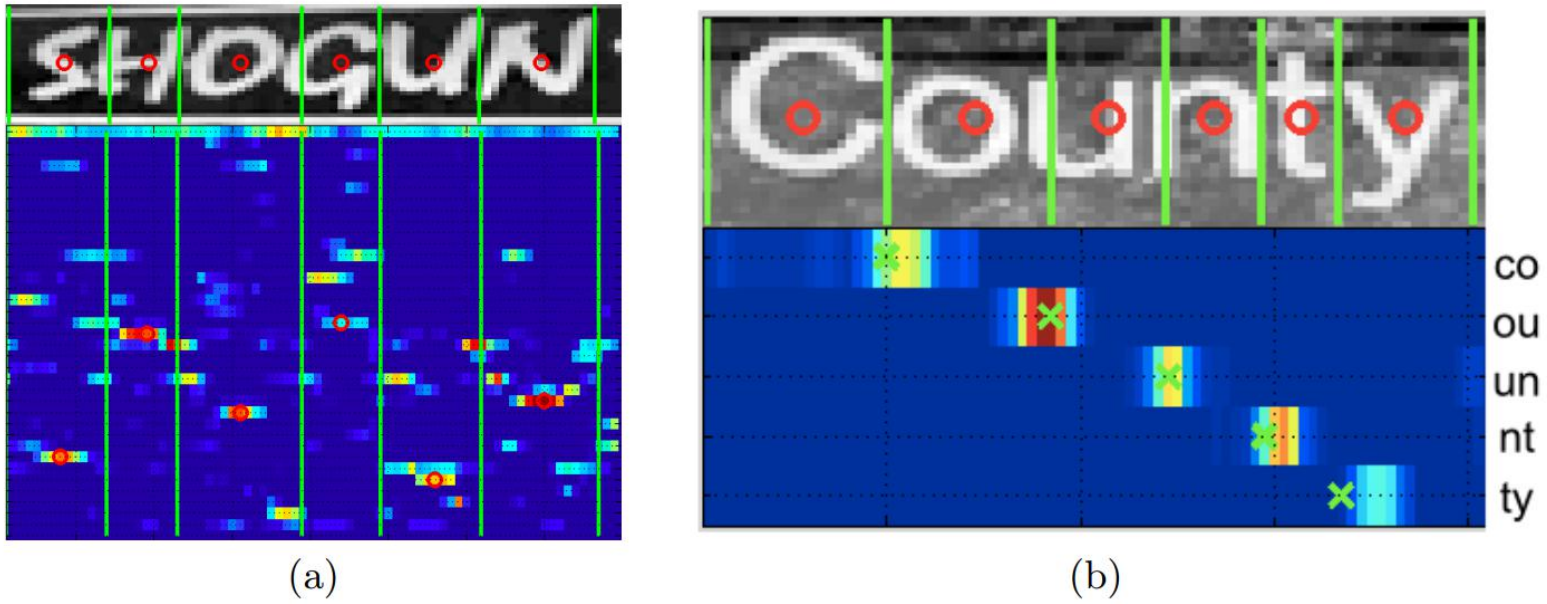


- propose a novel CNN architecture, enabling efficient feature sharing for text detection and character classification
- scan 16 different scales to handle text of different sizes

Face++ 旷视

Jaderberg et al.. Deep Features for Text Spotting . ECCV, 2014.

End-to-End Recognition: Deep Features



- generate a WxH map for each character hypothesis
- map reduced to Wx1 responses by averaging along each column
- breakpoints between characters are determined by dynamic programming

Jaderberg et al.. Deep Features for Text Spotting . ECCV, 2014.

End-to-End Recognition: Deep Features

- visualization of learned features

Face++ 旷视

Jaderberg et al.. Deep Features for Text Spotting . ECCV, 2014.

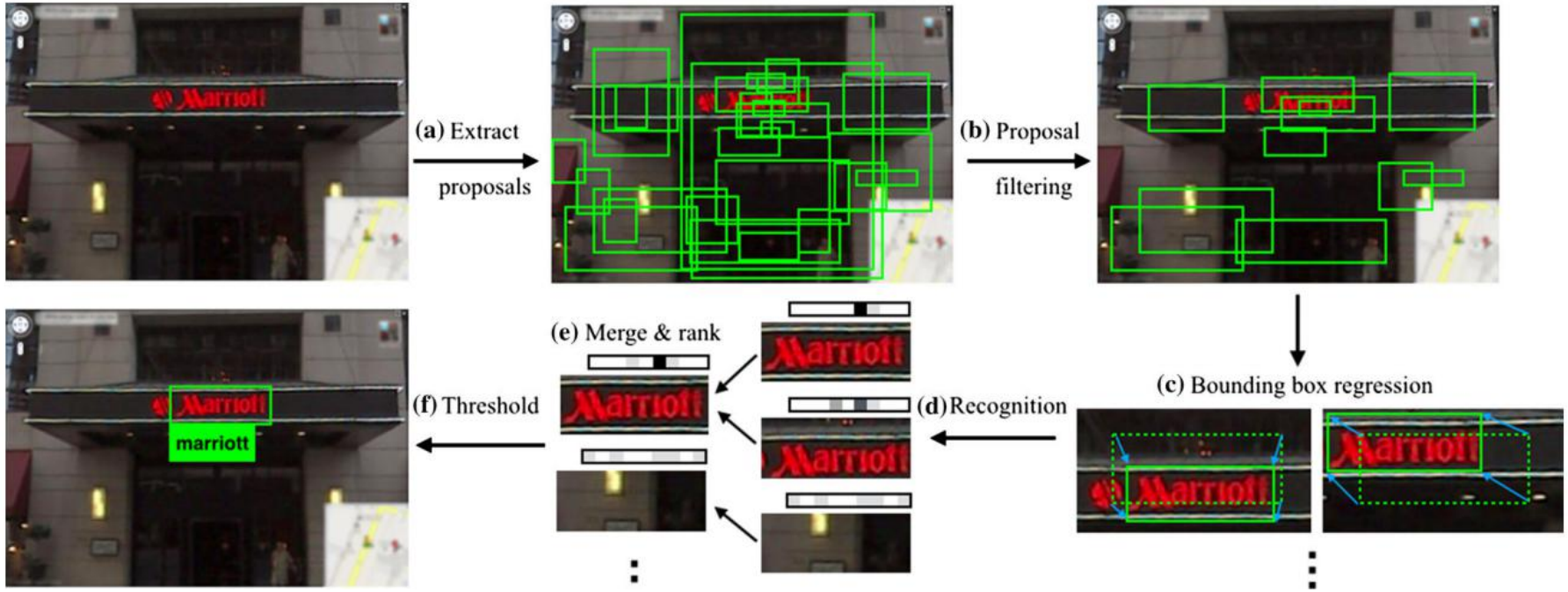
Detection: MSER Trees

- use MSER to seek character candidates
- utilize CNN classifiers to reject non-text candidates

Face++ 旷视

Huang et al.. Robust Scene Text Detection with Convolution Neural Network Induced MSER Trees. ECCV, 2014.

End-to-End Recognition: Reading Text



- seek word level candidates using multiple region proposal methods (EdgeBoxes, ACF detector)
- refine bounding boxes of words by regression
- perform word recognition using very large convolutional neural networks

Face++ 旷视

Jaderberg et al.. Reading Text in the Wild with Convolutional Neural Networks. IJCV, 2016.

Summary

- Common characteristics in early phase
  - pipelines with multiple stages
  - not purely deep learning based, adoption of conventional techniques and features (MSER, HOG, EdgeBoxes, etc.)

Detection: Holistic

local

- holistic *vs.* local
- text detection is casted as a semantic segmentation problem
- conceptionally and functionally different from previous sliding-window or connected component based approaches

Face++ 旷视

Yao et al.. Scene Text Detection via Holistic, Multi-Channel Prediction. 2016. arXiv preprint arXiv:1606.09002

Detection: Holistic

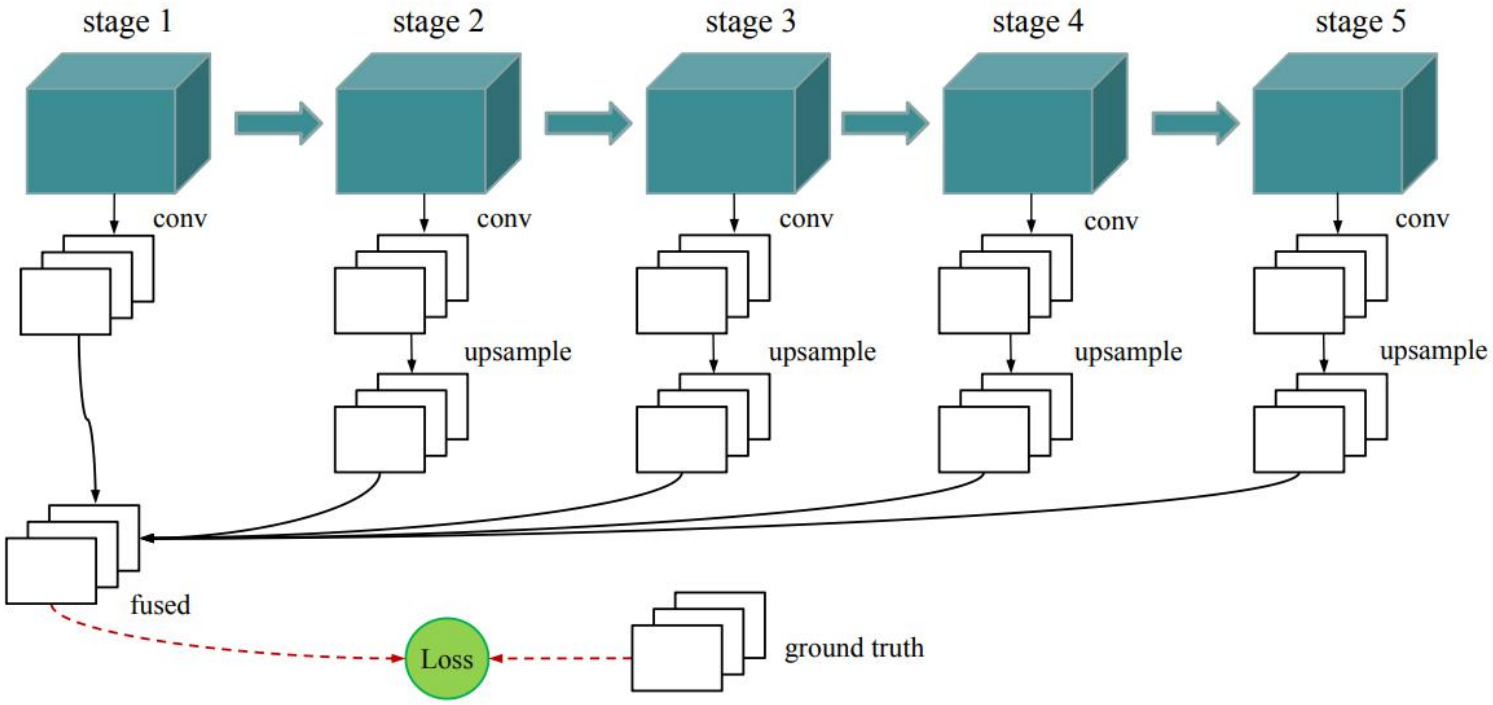
(a) (b) (c)

- holistic, pixel-wise predictions: text region map, character map and linking orientation map
- detections are formed using these three maps
- can simultaneously handle horizontal, multi-oriented and curved text in real-world natural images

Face++ 旷视

Yao et al.. Scene Text Detection via Holistic, Multi-Channel Prediction. 2016. arXiv preprint arXiv:1606.09002

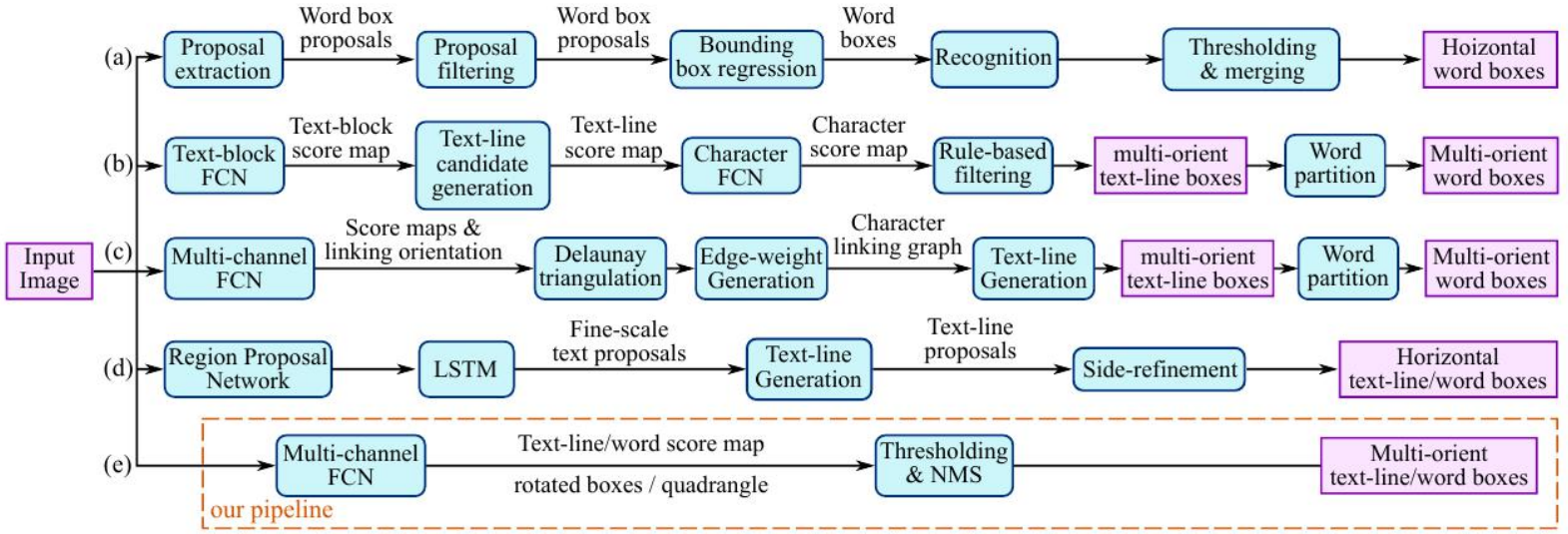
Detection: Holistic



- network architecture

Yao et al.. Scene Text Detection via Holistic, Multi-Channel Prediction. 2016. arXiv preprint arXiv:1606.09002

Detection: EAST (A Megvii work in CVPR 2017)

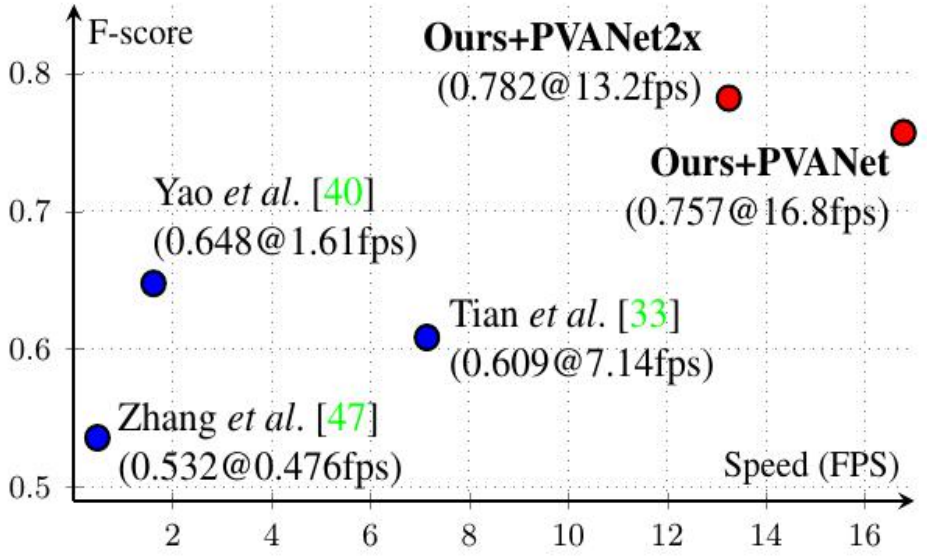


- highly simplified pipeline

Face++ 旷视

Zhou et al.. EAST: An Efficient and Accurate Scene Text Detector. CVPR, 2017.

Detection: EAST

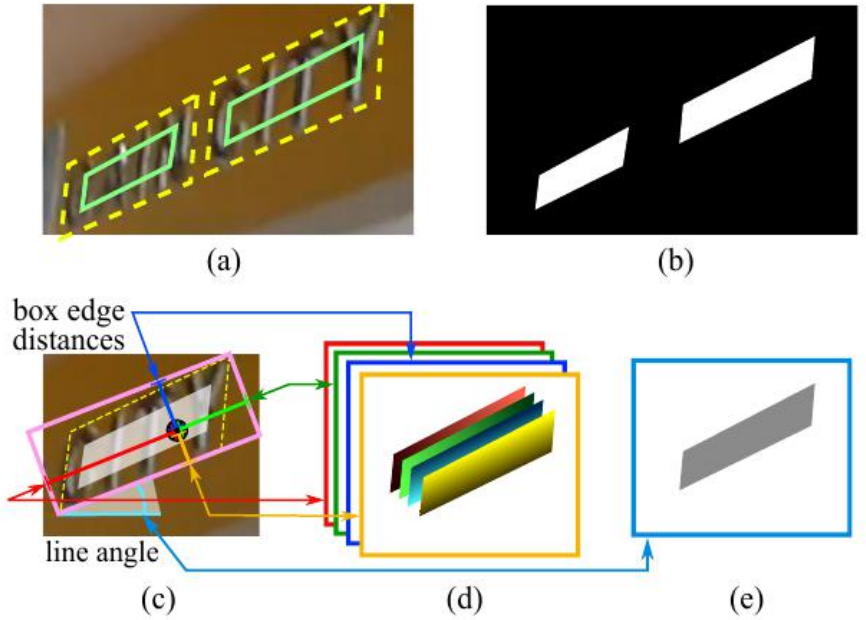


- strike a good balance between accuracy and speed
- code available at: https://github.com/argman/EAST (reimplemented by a student outside Megvii (Face++), credit goes to @argman)

Face++ 旷视

Zhou et al.. EAST: An Efficient and Accurate Scene Text Detector. CVPR, 2017.

Detection: EAST



- main idea: predict location, scale and orientation of text with a single model and multiple loss functions (multi-task training)
- advantages:
  (a). accuracy: allow for end-to-end training and optimization
  (b). efficiency: remove redundant stages and processings

Face++ 旷视

Zhou et al.. EAST: An Efficient and Accurate Scene Text Detector. CVPR, 2017.

Detection: EAST

Examples

Zhou et al.. EAST: An Efficient and Accurate Scene Text Detector. CVPR, 2017.

Detection: EAST

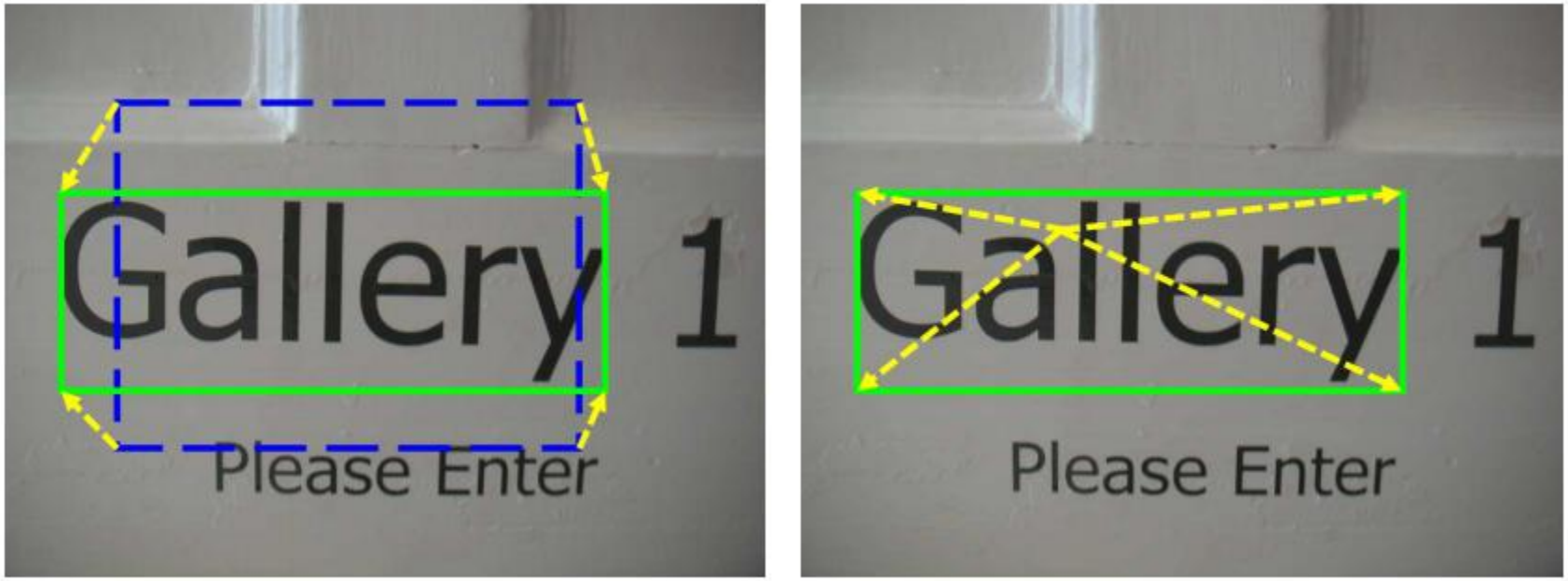
Demo Video

video also available at: https://www.youtube.com/watch?v=o5asMTdhmvA

Face++ 旷视

Zhou et al.. EAST: An Efficient and Accurate Scene Text Detector. CVPR, 2017.

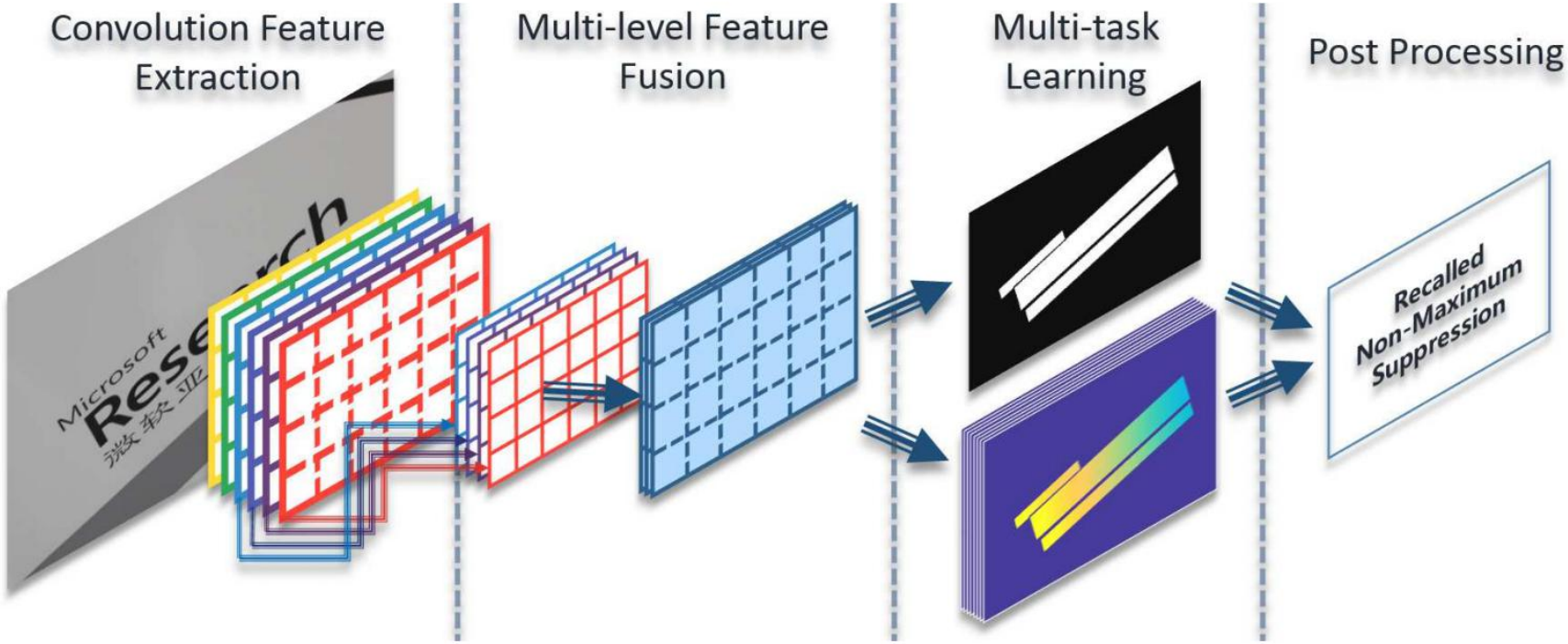
Detection: Deep Direct Regression



- directly regress the offsets from a point (as shown on the right), instead of predicting the offsets from bounding box proposals (on the left)

He et al.. Deep Direct Regression for Multi-Oriented Scene Text Detection. ICCV, 2017.

Detection: Deep Direct Regression



- produce maps representing properties of text instances via multi-task learning in a single model
- main idea is very similar to EAST

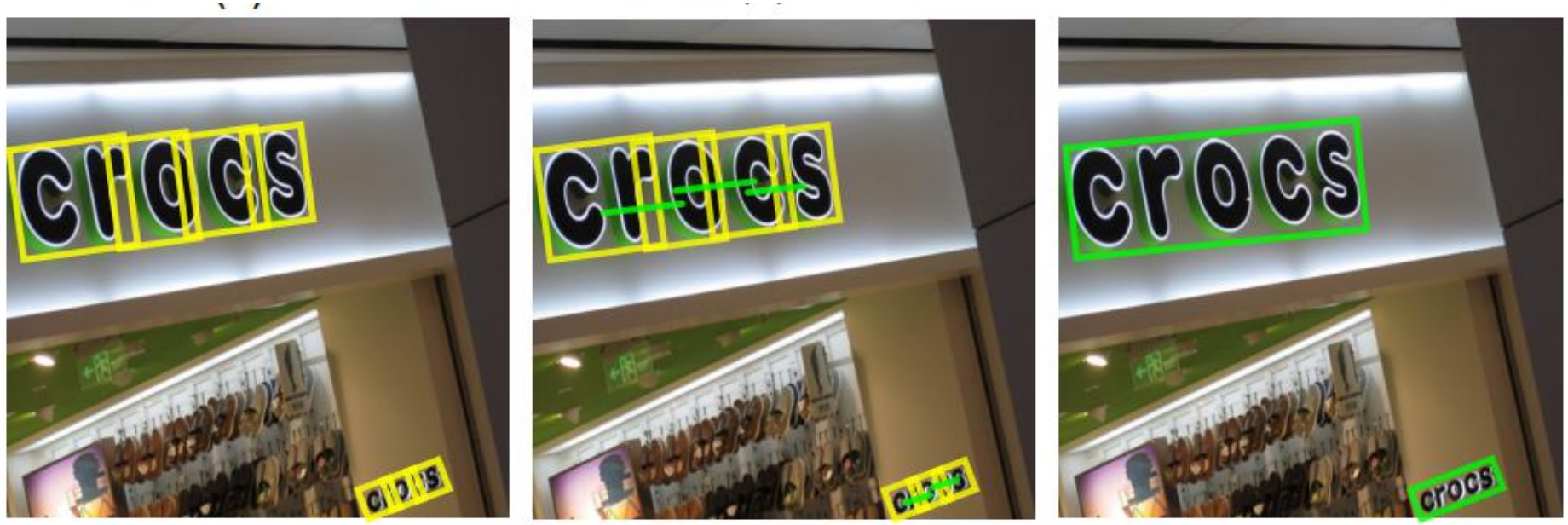
He et al.. Deep Direct Regression for Multi-Oriented Scene Text Detection. ICCV, 2017.

Detection: Deep Direct Regression

Examples

He et al.. Deep Direct Regression for Multi-Oriented Scene Text Detection. ICCV, 2017.

Detection: SegLink

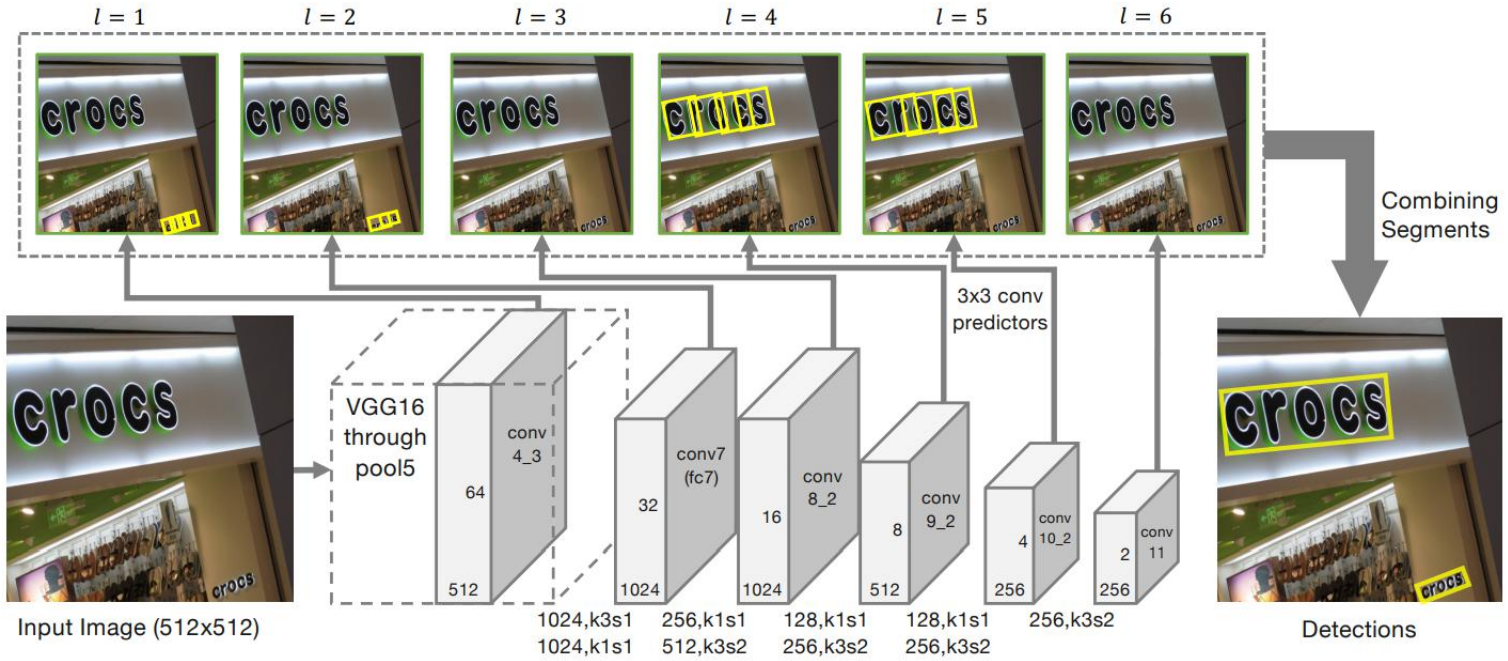


- decompose text into two locally detectable elements, namely segments and links
- segment is an oriented box covering a part of a word or text line
- link connects two adjacent segments

Face++ 旷视

Shi et al.. Detecting Oriented Text in Natural Images by Linking Segments. CVPR, 2017.

Detection: SegLink



- segments (yellow boxes) and links (not displayed) are detected by convolutional predictors on multiple feature layers
- detected segments and links are combined into whole words by a combining algorithm

Face++ 旷视

Shi et al.. Detecting Oriented Text in Natural Images by Linking Segments. CVPR, 2017.

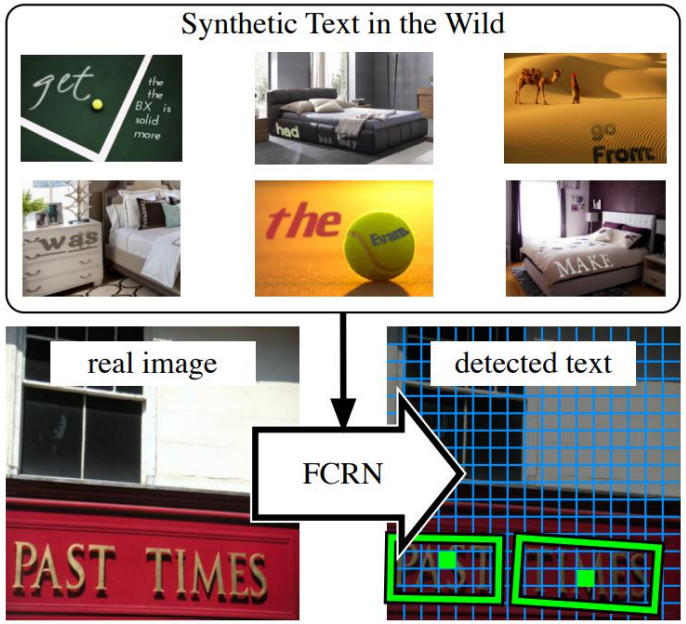
Detection: SegLink

Examples

- able to detect long lines of Latin and non-Latin text, such as Chinese

Shi et al.. Detecting Oriented Text in Natural Images by Linking Segments. CVPR, 2017.

Detection: Synthetic Data

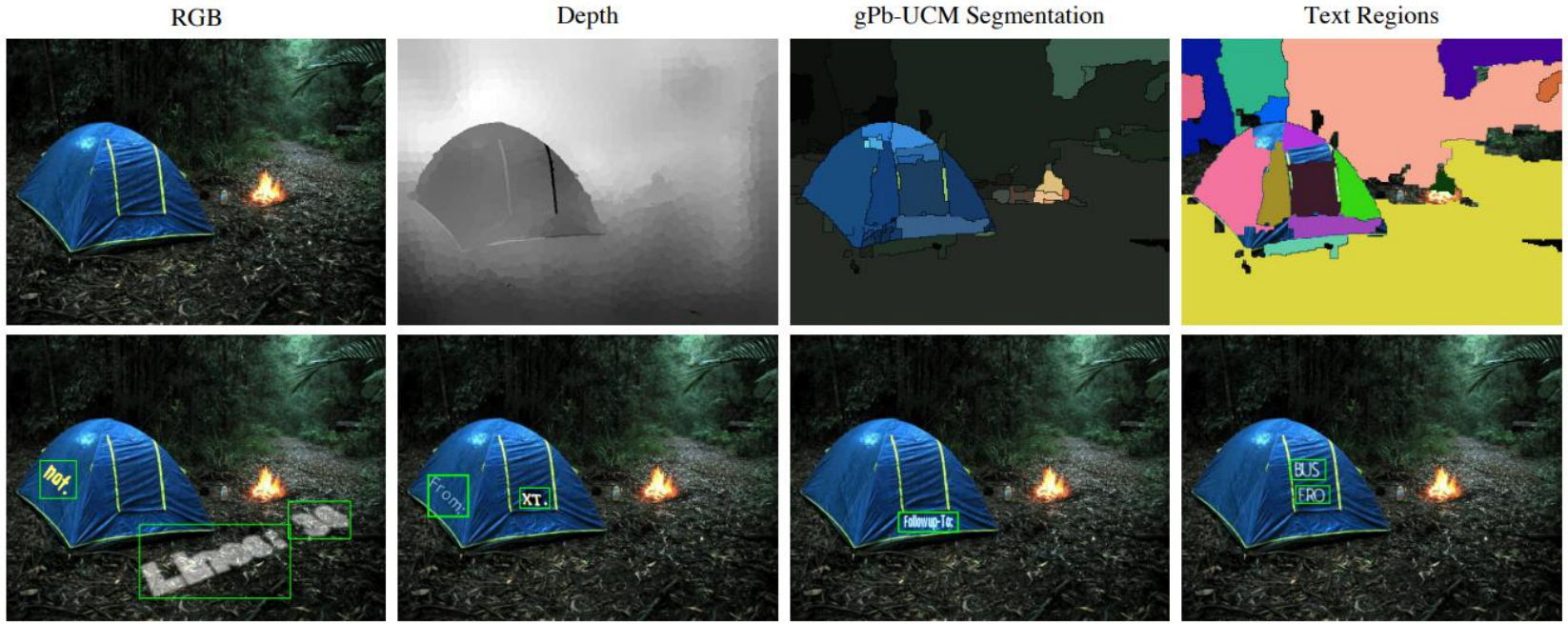


- present a fast and scalable engine to generate synthetic images of text in clutter
- propose a Fully-Convolutional Regression Network (FCRN) for high-performance text detection in natural scenes

Face++ 旷视

Gupta et al.. Synthetic Data for Text Localisation in Natural Images. CVPR, 2016.

Detection: Synthetic Data



Sample Scene-Text Images

- overlay synthetic text to existing background images in a natural way, accounting for the local 3D scene geometry

Gupta et al.. Synthetic Data for Text Localisation in Natural Images. CVPR, 2016.

Detection: Synthetic Data

- local colour/texture sensitive placement

Face++ 旷视

Gupta et al.. Synthetic Data for Text Localisation in Natural Images. CVPR, 2016.

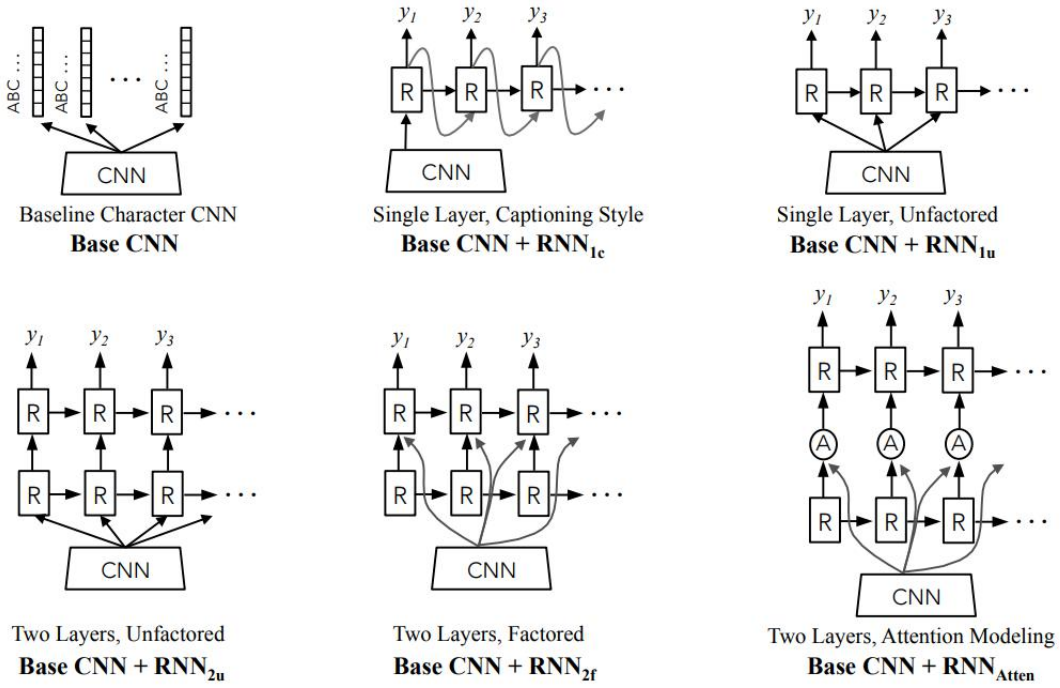
Detection: Synthetic Data

- a dataset consists of 800 thousand images with approximately 8 million synthetic word instances
- dataset available at: http://www.robots.ox.ac.uk/~vgg/data/scenetext/
- code available at: https://github.com/ankush-me/SynthText

Face++ 旷视

Gupta et al.. Synthetic Data for Text Localisation in Natural Images. CVPR, 2016.

Recognition: R^{2}AM

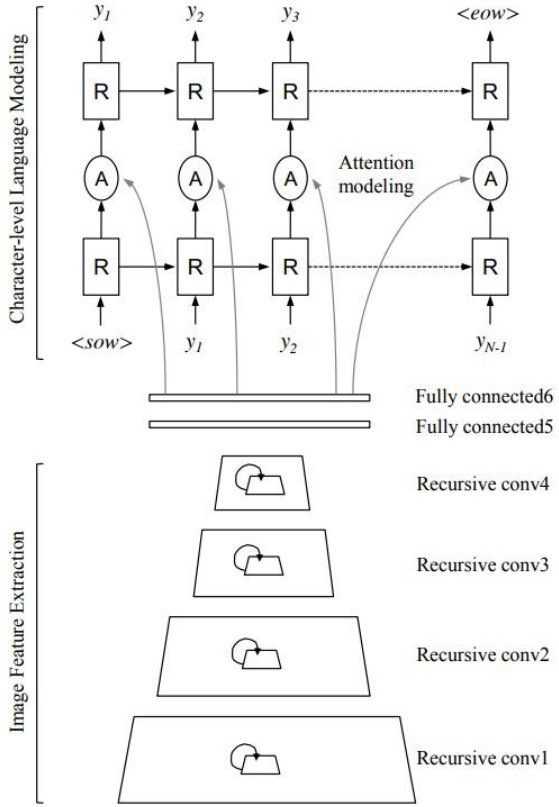


- explore five variations of the recurrent in time architecture for text recognition
- present recursive recurrent neural networks with attention modeling (R2AM) for lexicon-free text recognition

Face++ 旷视

Lee et al.. Recursive Recurrent Nets with Attention Modeling for OCR in the Wild. CVPR, 2016.

Recognition: R^{2}AM



- an implicitly learned character-level language model, embodied in a recurrent neural network
- use of a soft-attention mechanism, allowing the model to selectively exploit image features in a coordinated way

Face++ 旷视

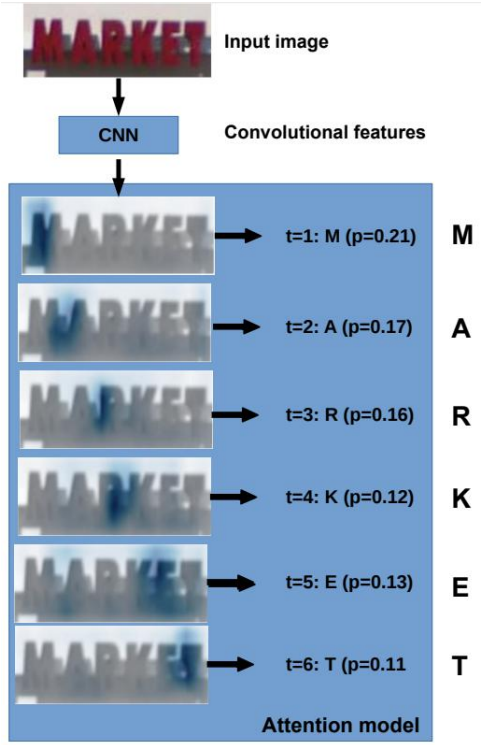
Lee et al.. Recursive Recurrent Nets with Attention Modeling for OCR in the Wild. CVPR, 2016.

Recognition:

Examples

Lee et al.. Recursive Recurrent Nets with Attention Modeling for OCR in the Wild. CVPR, 2016.

Recognition: Visual Attention

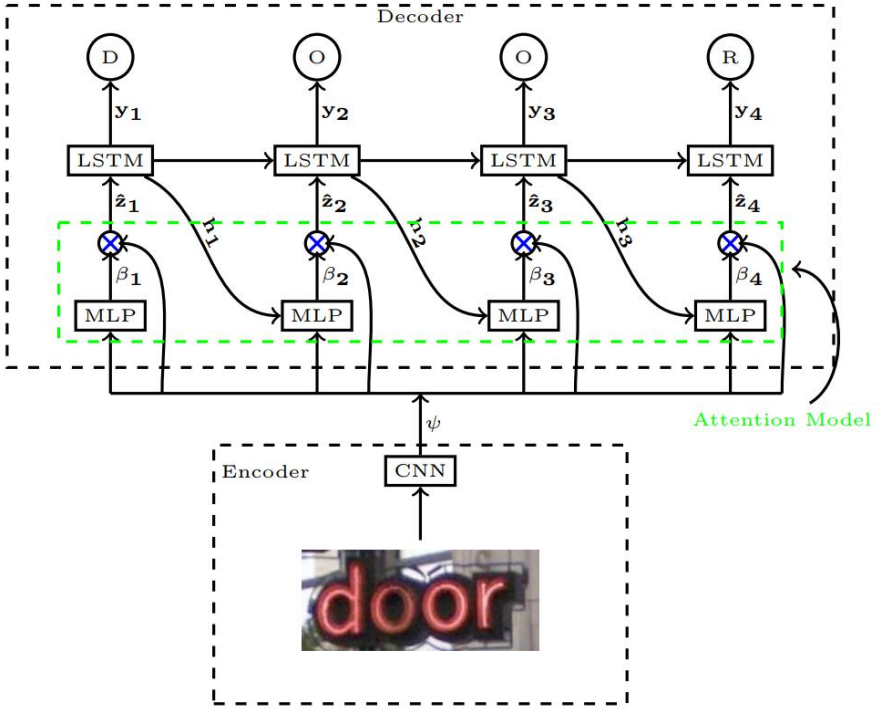


- a set of spatially localized features are obtained using a CNN
- at every time step the attention model weights the set of feature vectors to make the LSTM focus on a specific part of the image

Face++ 旷视

Ghosh et al.. Visual attention models for scene text recognition. 2017. arXiv:1706.01487

Recognition: Visual Attention



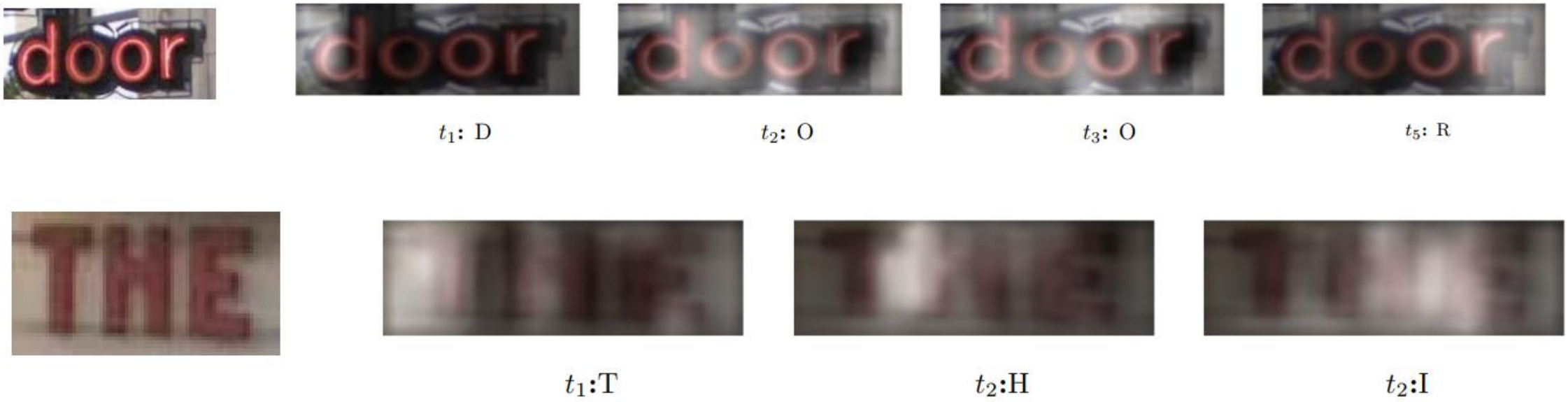
- encoder-decoder framework with attention model

Face++ 旷视

Ghosh et al.. Visual attention models for scene text recognition. 2017. arXiv:1706.01487

Recognition: Visual Attention

Examples



Face++ 旷视

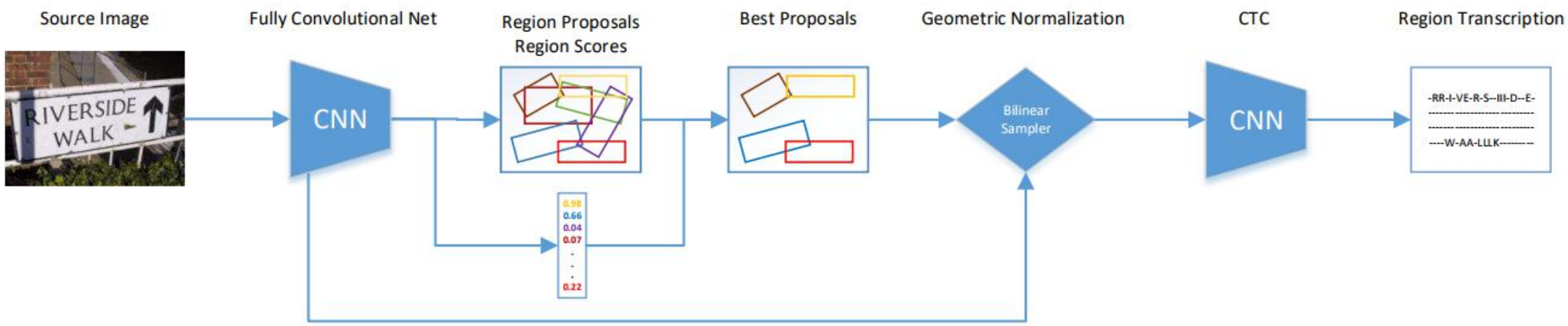
Ghosh et al.. Visual attention models for scene text recognition. 2017. arXiv:1706.01487

End-to-End Recognition: Deep TextSpotter

- achieve both text detection and recognition in a single end-to-end pass
- state-of-the-art accuracy in end-to-end recognition

Busta et al.. Deep TextSpotter: An End-To-End Trainable Scene Text Localization and Recognition Framework. ICCV, 2017.

End-to-End Recognition: Deep TextSpotter

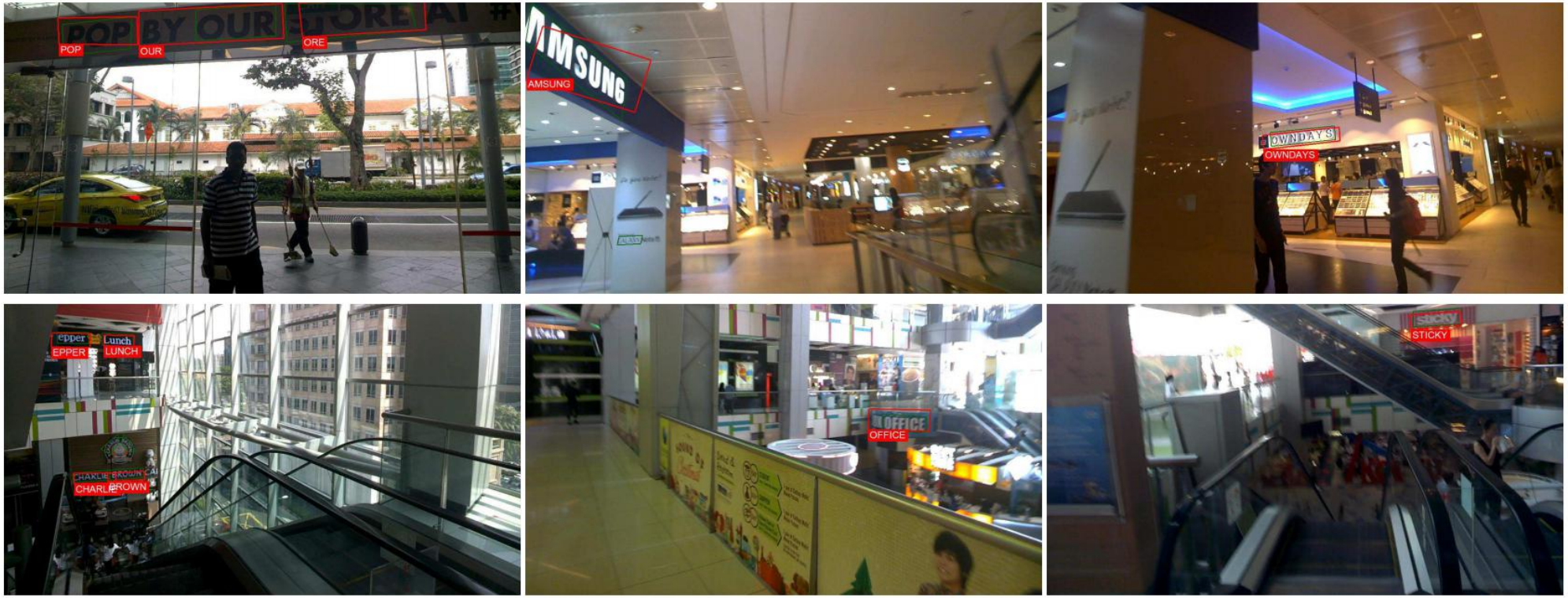


- text region proposals are generated by a Region Proposal Network (Faster-RCNN)
- each region is associated with a sequence of characters or rejected as not text
- model is jointly optimized for both text localization and recognition in an end-to-end training framework

Busta et al.. Deep TextSpotter: An End-To-End Trainable Scene Text Localization and Recognition Framework. ICCV, 2017.

End-to-End Recognition: Deep TextSpotter

Examples



code available at: https://github.com/MichalBusta/DeepTextSpotter

Face++ 旷视

Busta et al.. Deep TextSpotter: An End-To-End Trainable Scene Text Localization and Recognition Framework. ICCV, 2017.

Summary

- Common characteristics in recent phase
  - highly simplified pipelines, removing intermediate steps
  - deep learning based, hardly any conventional techniques and features
  - ideas borrowed from methods for semantic segmentation and object detection, like FCN, Faster-RCNN
  - generation and use of synthetic data, rather than real data

Outline

- Background and Introduction
- Conventional Methods
- Deep Learning Methods
- Datasets and Competitions
- Conclusion and Outlook

ICDAR 2013

- 485 images containing text in a variety of colors and fonts on different backgrounds
- mostly horizontal text

http://rrc.cvc.uab.es/?ch=2&com=introduction

MSRA-TD500

- 500 images in total, with text instances of different orientations
- both Chinese and English text
- adopted by IAPR as official dataset

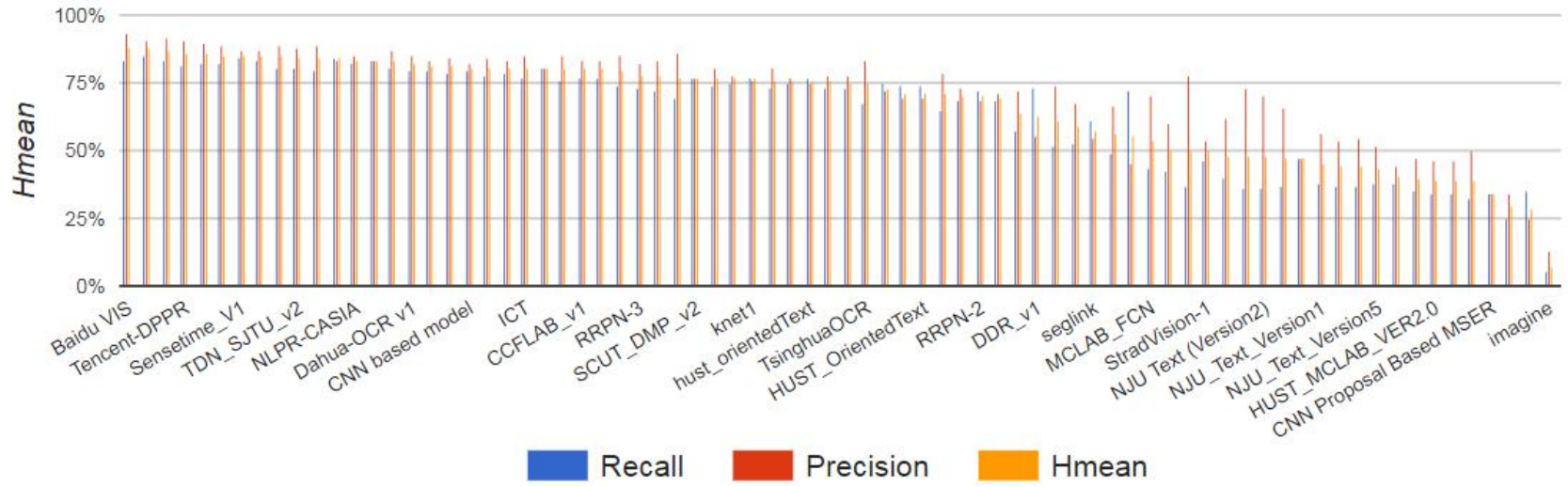
http://www.iapr-tc11.org/mediawiki/index.php/MSRA_Text_Detection_500_Database_(MSRA-TD500)

ICDAR 2015

- 1500 images in total, with text instances of different orientations
- incidental scene text: without the user having taken any specific prior action to cause its appearance or improve its positioning / quality in the frame
- only English text

http://rrc.cvc.uab.es/?ch=4&com=introduction

ICDAR 2015



- very popular benchmark
- about 50 submissions in 2017, about 80 submissions since 2015

http://rrc.cvc.uab.es/?ch=4&com=introduction

IIIT 5K-Word

- 5000 cropped word images from natural scene and born-digital images
- diversity in font, color, style, background, etc.
- used for cropped word recognition

Face++ 旷视

http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/IIIT5K.html

COCO-Text

- original images from the MS-COCO dataset
- 63,686 images, 145,859 text instances
- largest and most challenging dataset to date
- for both text detection and recognition

https://vision.cornell.edu/se3/coco-text-2/

MLT

- multilingual dataset, 9 languages: Chinese, Japanese, Korean, English, French, Arabic, Italian, German and Indian
- for text detection, script identification and recognition

http://rrc.cvc.uab.es/?ch=8&com=introduction

Total-Text (released on Oct. 31, 2017)

- 1555 images with different text orientations: Horizontal, Multi-Oriented, and Curved
- facilitate a new research direction for the scene text community

Face++ 旷视

https://github.com/cs-chan/Total-Text-Dataset

Outline

- Background and Introduction
- Conventional Methods
- Deep Learning Methods
- Datasets and Competitions
- Conclusion and Outlook

Conclusion and Outlook

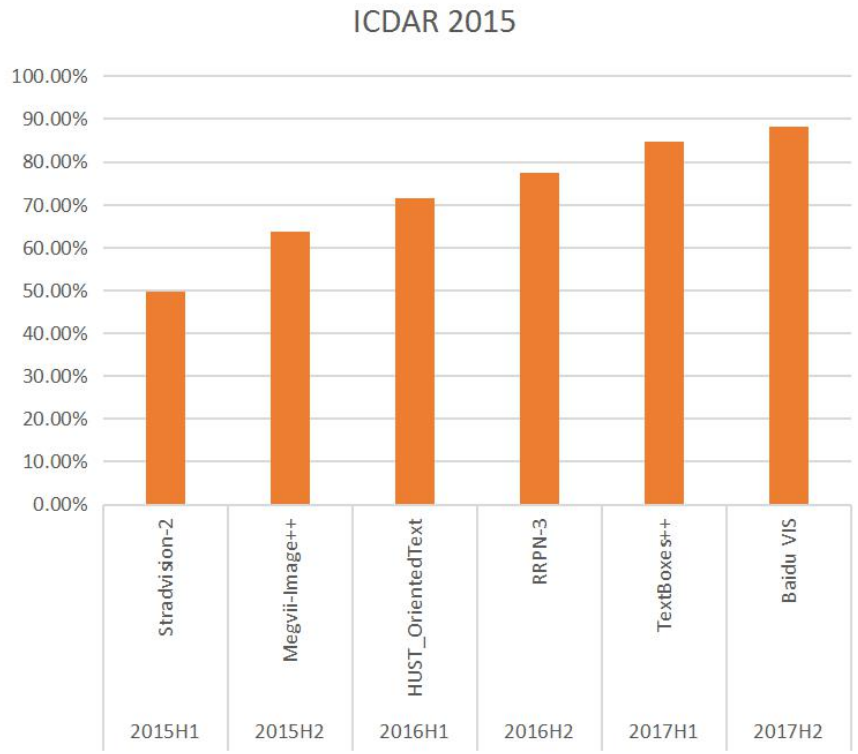
- Evolution path
  - Pre-deep-learning era [1914-2013]: conventional techniques and features
    - MSER [Neumann *et al.*, 2010;]
    - SWT [Epshtein *et al.*, 2010; Yao *et al.*, 2012]
    - HOG [Wang *et al.*, 2011]
    - CRF [Mishra *et al.*, 2011]
  - Transition period [2013-2015]: mixture of conventional techniques/features and deep models/features
    - HOG+DNN [Bissacco *et al.*, 2013]
    - MSER+CNN [Huang *et al.*, 2014; Zhang *et al.*, 2015]
    - HOG+LSTM [Su *et al.*, 2014]
  - Deep learning era [2015-now]: "pure" deep models/features
    - CNN [Gupta *et al.*, 2016]
    - RNN [Ghosh *et al.*, 2016]
    - FCN [Yao *et al.*, 2016; Zhou *et al.*, 2017]
    - Faster-RCNN [Busta *et al.*, 2017]

Face++ 旷视

https://en.wikipedia.org/wiki/Optical_character_recognition

Conclusion and Outlook

- Substantial progresses achieved
  - Two core factors: Deep Learning (CNN and RNN) and Data (real and synthetic)



source: http://rrc.cvc.uab.es/?ch=4&com=evaluation&task=1&gtv=1

Conclusion and Outlook

- Grand challenges remain
  - Diversity of text: language, font, scale, orientation, arrangement, etc.
  - Complexity of background: virtually indistinguishable elements (signs, fences, bricks and grasses, etc.)
  - Interferences: noise, blur, distortion, low resolution, nonuniform illumination, partial occlusion, etc.

Conclusion and Outlook

- Future Trends
  - Stronger models (accuracy, efficiency, **interpretability**)
  - Data synthesis
  - Muiti-oriented text
  - Curved text
  - Muiti-language text

Appendix: references

- Survey
  - Ye *et al.*. Text Detection and Recognition in Imagery: A Survey. TPAMI, 2015
  - Zhu *et al.*. Scene Text Detection and Recognition: Recent Advances and Future Trends. FCS, 2015

Appendix: references

- Conventional Methods
  - Epshtein *et al.*. Detecting Text in Natural Scenes with Stroke Width Transform. CVPR, 2010.
  - Neumann *et al.*. A method for text localization and recognition in real-world images. ACCV, 2010.
  - Yao *et al.*. Detecting Texts of Arbitrary Orientations in Natural Images. CVPR, 2012
  - Wang *et al*.. End-to-End Scene Text Recognition. ICCV, 2011.
  - Mishra *et al.*. Scene Text Recognition using Higher Order Language Priors. BMVC, 2012.
  - Busta *et al.*. FASText: Efficient Unconstrained Scene Text Detector. ICCV 2015

Appendix: references

- Deep Learning Methods
  - Bissacco *et al.*. PhotoOCR: Reading Text in Uncontrolled Conditions. ICCV, 2013.
  - Jaderberg *et al.*. Deep Features for Text Spotting. ECCV, 2014.
  - Gupta *et al*.. Synthetic Data for Text Localisation in Natural Images. CVPR, 2016.
  - Zhou *et al*.. EAST: An Efficient and Accurate Scene Text Detector. CVPR, 2017.
  - Busta *et al*.. Deep TextSpotter: An End-To-End Trainable Scene Text Localization and Recognition Framework. ICCV, 2017.
  - Ghosh *et al*.. Visual attention models for scene text recognition. 2017. arXiv:1706.01487
  - Cheng *et al*.. Focusing Attention: Towards Accurate Text Recognition in Natural Images. ICCV, 2017.

Face++ 旷视

Appendix: useful resources

- Laboratories and Papers
  - https://github.com/chongyangtao/Awesome-Scene-Text-Recognition
- Datasets and Codes
  - https://github.com/seungwooYoo/Curated-scene-text-recognition-analysis
- Projects and Products
  - https://github.com/wanghaisheng/awesome-ocr

Thank You!

Face++ 旷视