Rozpoznávání textu ve fotografiích a videu

Ing. Lukáš Neumann
Prof. Dr. Ing. Jiří Matas

Problem Introduction

Classical formulation


```
topLeft=(240;1428)
bottomRight=(391;1770)
text="TESCO"
```

- **Input**: Digital image (BMP, JPG, PNG) / video (AVI)
- **Output:** Set of words in the image
  word = horizontal rectangular bounding box + text content

Problem Introduction

Our extended formulation

"Amoeba Music"
"CD RECORD POSTER VIDEO"
"FREE PARKING"
"BUY HERE!"

- **Input**: Digital image (BMP, JPG, PNG) / video (AVI)
- **Output:** Set of displays in the image
  display = ordered set of words
  word = straight/curved baseline with letter height+ text content

Problem Introduction

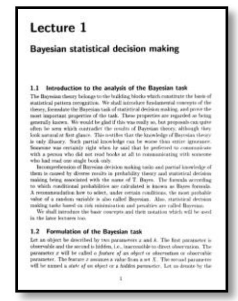
Real scene image

Recognition rate ~54% [2]

- **Text localization**
- Varying background
- Low text density, irregular layout
- Shadows, reflections, occlusions, perspective distortion, …
- Many different fonts

Printed documents (OCR)

Recognition rate >99% [1]


Lecture 1

Bayesian statistical decision making

1.1 Introduction to the analysis of the Bayesian task

1.2 Formulation of the Bayesian task

- High-contrast solid background
- High text density, structured text
- Only rotation and brightness adjustment

1. X. Lin. Reliable OCR solution for digital content remastering, 2001
2. T. E. de Campos, B. R. Babu, and M. Varma. Character recognition in natural images, 2009

Text Localization

- Can be computationally very expensive, generally in an image of N pixels generally any of the 2^N subsets can correspond to text
- Two different approaches widely used in the literature
  - Sliding Window based methods
  - Connected-Component (CC) based methods

Sliding Window methods

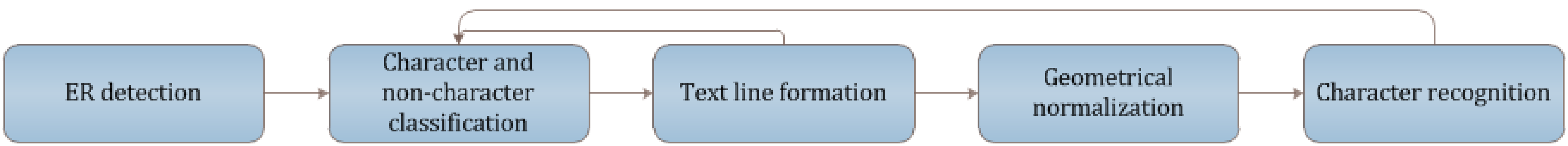
- Slide a window (of different sizes) across the image and let a classifier decide for each position whether the window contains the desired object (in our case either a character or a whole word)
- Successfully applied in many detection tasks (faces, pedestrians,…) for real-time detection
- BUT for text detection, there are much more window parameters to consider (aspect, skew, rotation) which makes such methods **very slow** (between 10^6 and 10^8 windows to **classify for each image**)

Connected-Component methods

- Recently more popular approach where individual characters are localized as connected components (CCs) based on local properties – color, intensity, stability (MSER), stroke width (SWT), “characterness” (CSER)
- Very fast because number of CCs is linear in the number of pixels and characters of all scales and orientations can be detected in a single pass
- BUT the assumption that a character is a connected component is very brittle and prone to noise – a change of intensity of a single pixel can disconnect a perfectly “nice” character, causing its disposal as clutter (now there are two connected components where neither of them look like a character)

Method Description

Stages Overview

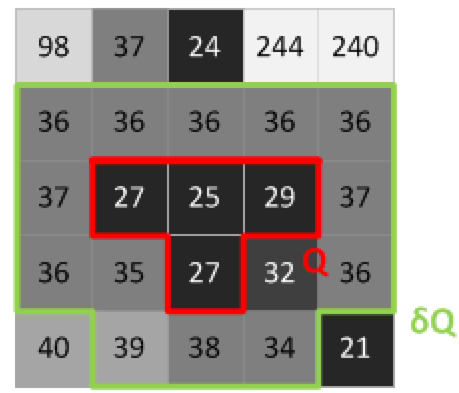


Method Description

Extremal Region

- Let image I be a mapping $I: Z^2 \rightarrow S$
- Let S be a totally ordered set, e.g. <0, 255>
- Let A be an adjacency relation (e.g. 4-neigbourhood)
- Region Q is a contiguous subset w.r.t. A
- (Outer) Region Boundary δQ is set of pixels adjacent but not belonging to Q
- Extremal Region is a region where there exists a threshold θ that separates the region and its boundary

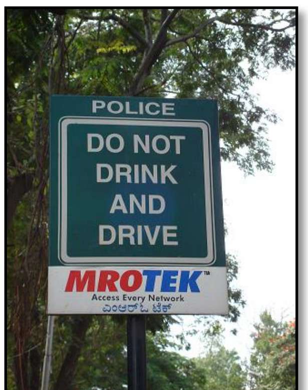
$$\exists \theta : \forall p \in Q, \forall q \in \delta Q : I(p) < \theta \leq I(q)$$



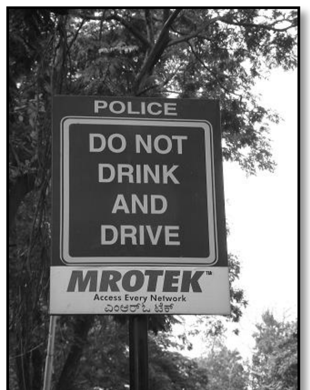
$\theta = 32$

Method Description

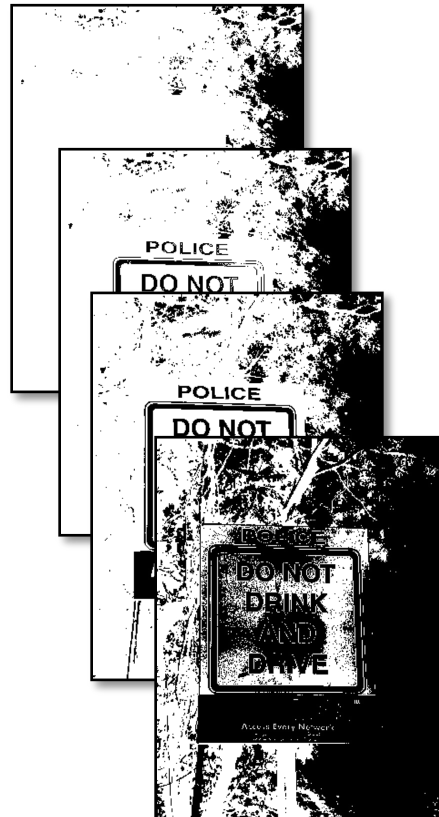
ER Detection



Input image
(PNG, JPEG, BMP)



1D projection
<0;255>
(grey scale, hue,…)

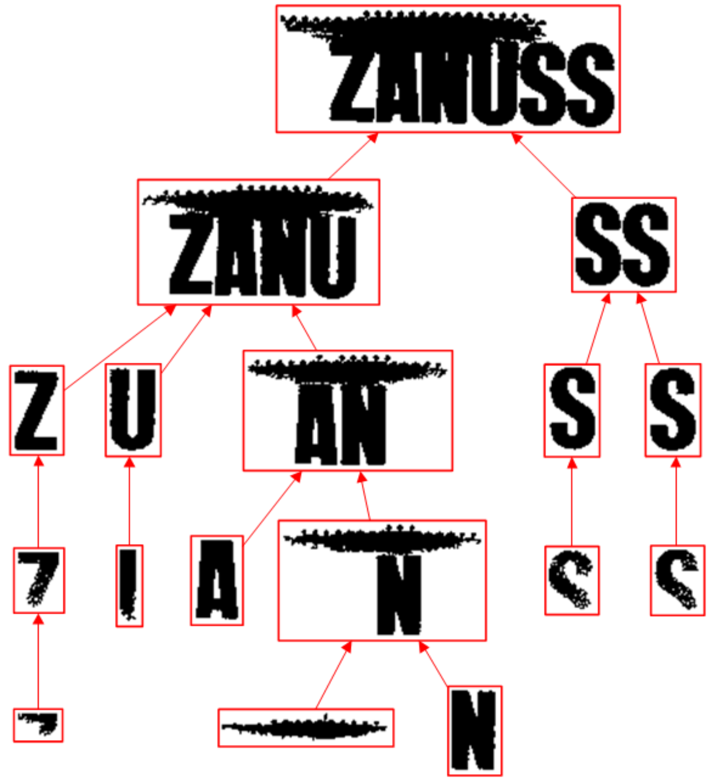


Extremal regions with threshold t
(t=50, 100, 150, 200)

ER Detection
↓
Region Classification
↓
Line Formation
↓
Geometrical Normalization
↓
Character Recognition

Method Description

ER Inclusion



ER Detection

Region Classification

Line Formation

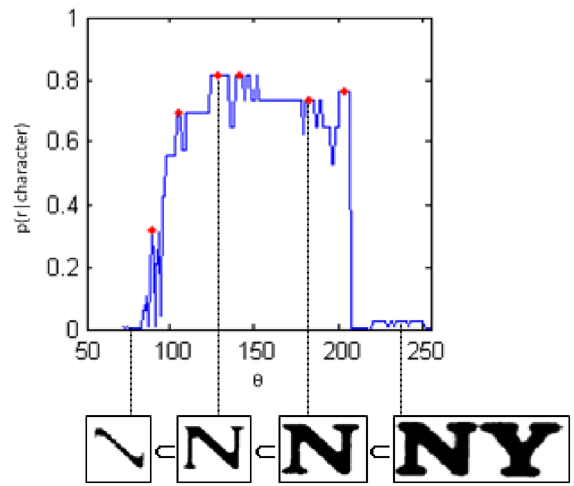
Geometrical Normalization

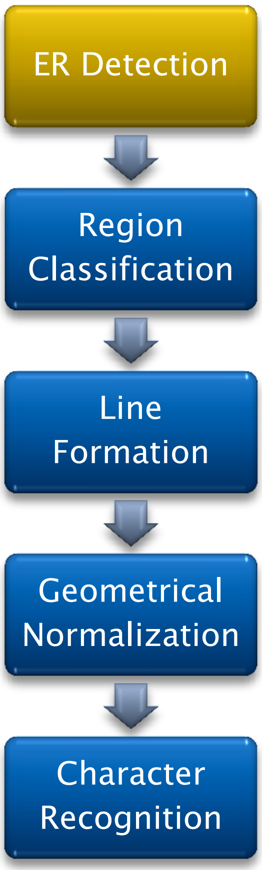
Character Recognition

Method Description

ER Detection

- p(r|character) estimated at each threshold for each region
- Only regions corresponding to local maxima selected by the detector
- Incrementally computed descriptors used for classification [3]
  - Aspect ratio
  - Compactness
  - Number of holes
  - Horizontal crossings
- Trained AdaBoost classifier with decision trees calibrated to output probabilities
- Real-time performance (300ms on an 800x600px image)

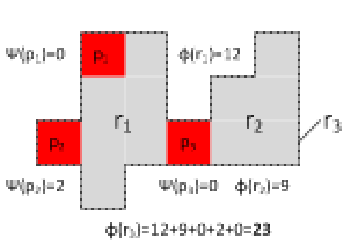




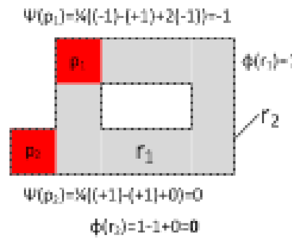
3. J. Matas and K. Zimmermann. A new class of learnable detectors for categorisation. In Image Analysis, volume 3540 of LNCS, pages 541-550. 2005.

Method Description

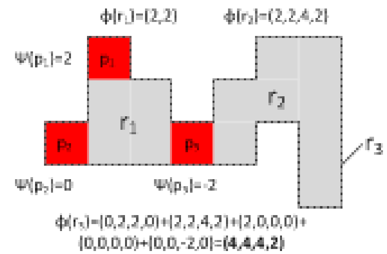
Incrementally Computed Descriptors



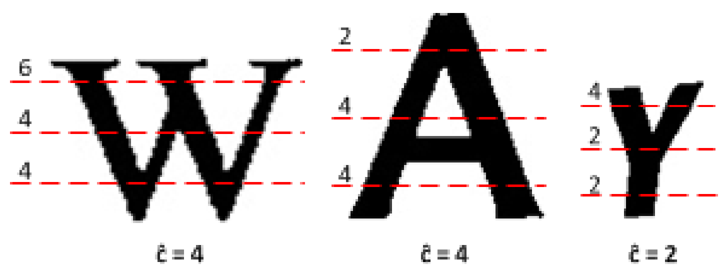
Perimeter



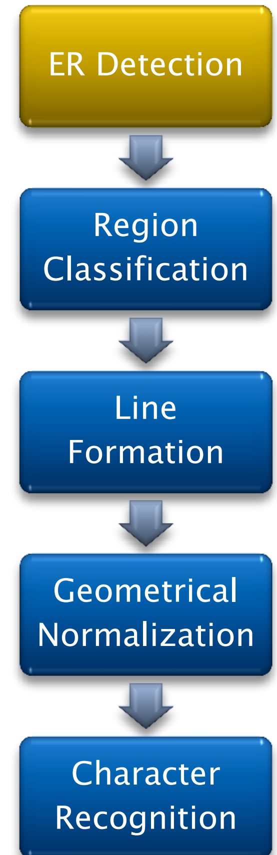
Euler Number



Horizontal Crossings

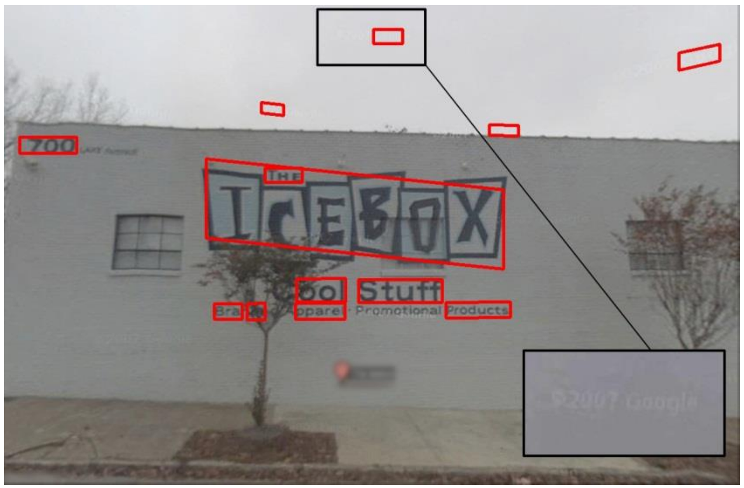


Horizontal Crossings Examples



Method Description

Robustness to blur, noise and low contrast



Examples from Street view dataset. All “false positives” in the images are caused by embedded watermarks

ER Detection → Region Classification → Line Formation → Geometrical Normalization → Character Recognition

Method Description

Multiple projections (channels)

- Multiple projections can be used
- Trade-off between recall and speed (although can be easily parallelized)
- Standard channels (R, G, B, H, S, I) of RGB / HSI color space
- 85,6% characters detected in the Intensity channel, combining all channels increases the recall to 94,8%

Source Image

Intensity Channel

Red Channel

ER Detection → Region Classification → Line Formation → Geometrical Normalization → Character Recognition

Method Description

Multiple projections (channels)

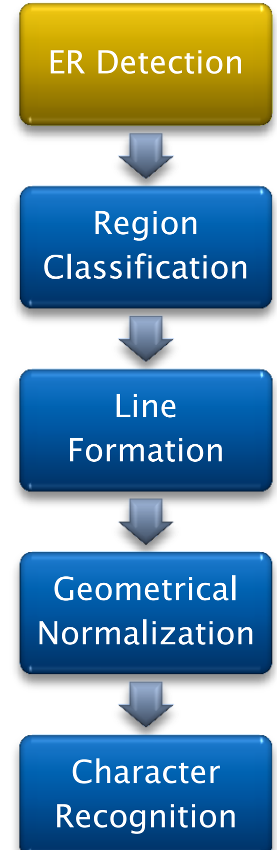
- Gradient projections can be used $\rightarrow$ edges induce ERs
- Intensity gradient projection ∇ appears to be orthogonal to standard channels (combination of Intensity, Hue and ∇ yields 93,7% recall, only 1% lower compared to all 6 standard channels combined)

Source Image

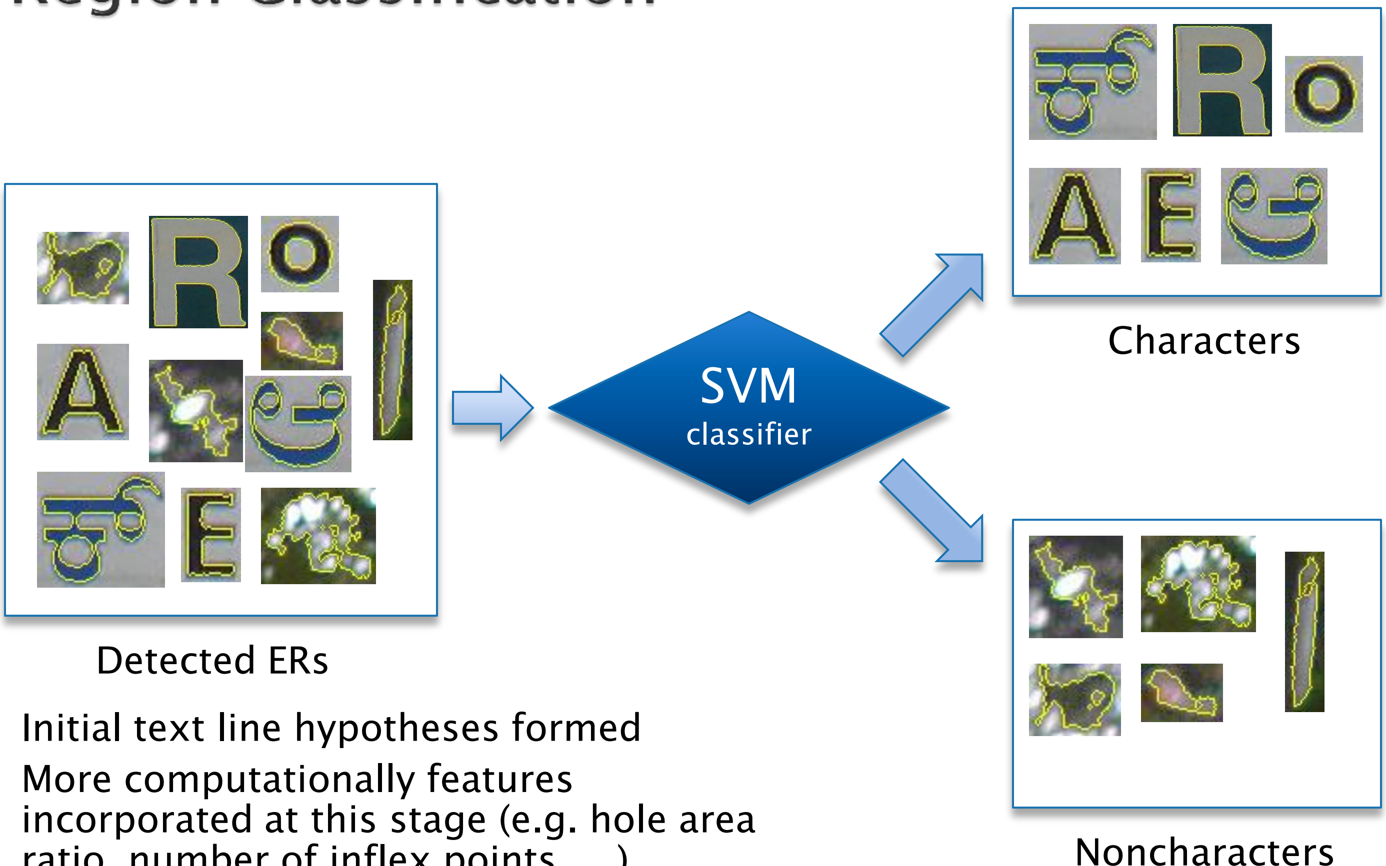
Intensity projection
(no threshold for letters “OW”)

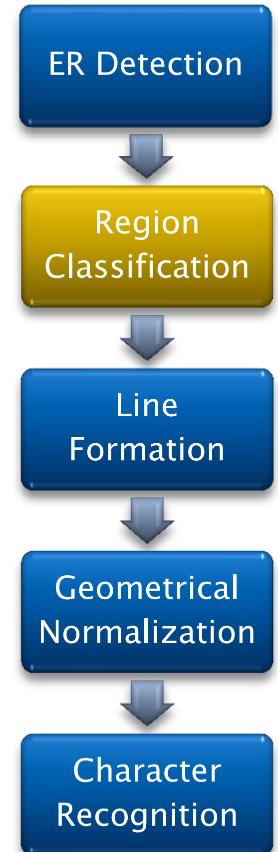
Intensity gradient
Projection ∇

Detected ERs in
Intensity gradient
projection



Method Description
Region Classification



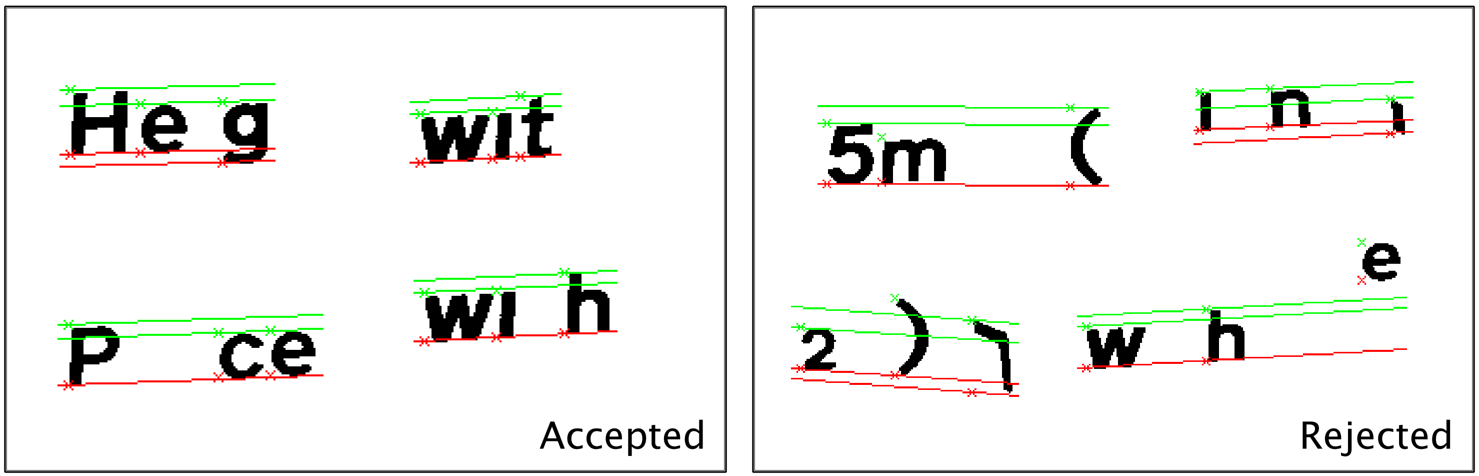


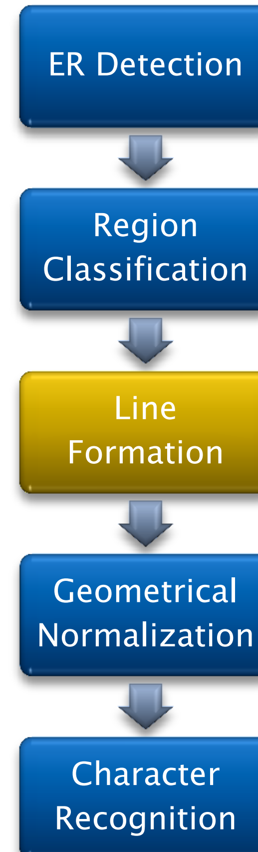
- Initial text line hypotheses formed
- More computationally features incorporated at this stage (e.g. hole area ratio, number of inflex points, …)
- Classification can be rectified later using feedback loops

Method Description

Line formation

- All neighboring region triplets exhaustively enumerated
- Text line typographical parameters (top line, middle, line, base line, bottom line) estimated by RANSAC
- Invalid triplets disregarded

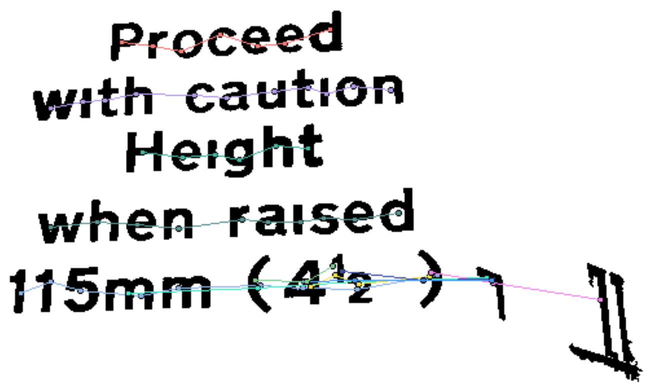




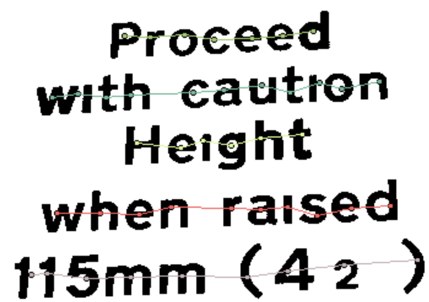
Method Description

Line formation

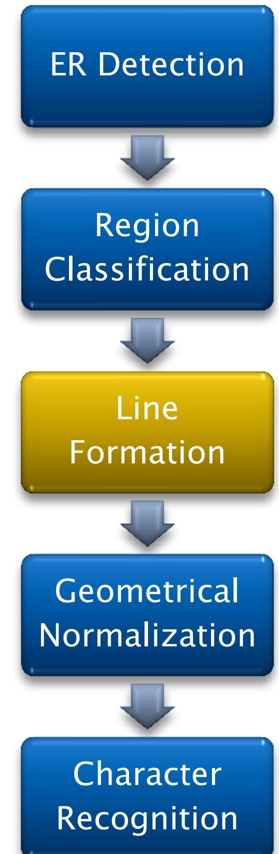
- Triplets are clustered by agglomerative grouping into text lines
- Conflicting text lines removed (by preference for longer text in given direction)
- Feedback loop to revisit initial region classification



Final stage of agglomerative grouping



Conflicting text lines removed



Method Description

Geometrical Normalization

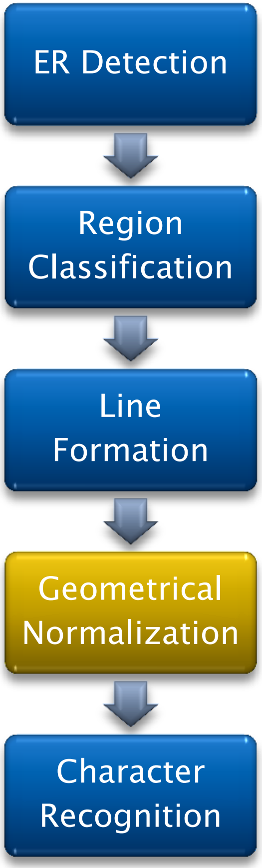
Input image

Detected text area

Top and bottom line

Normalized text area

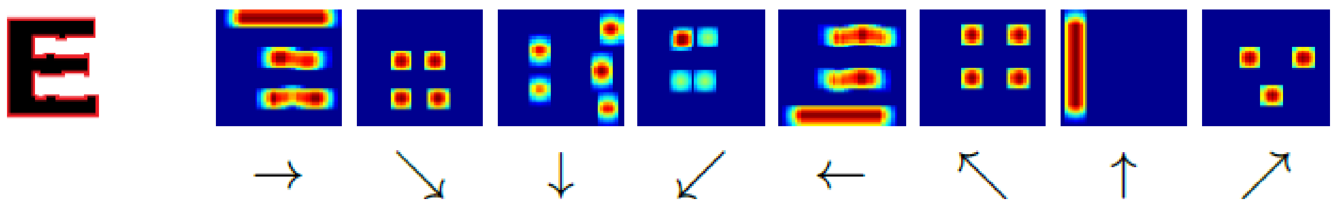
- Fitting top and bottom line to find horizontal vanishing point
- Creating inverse projection matrix

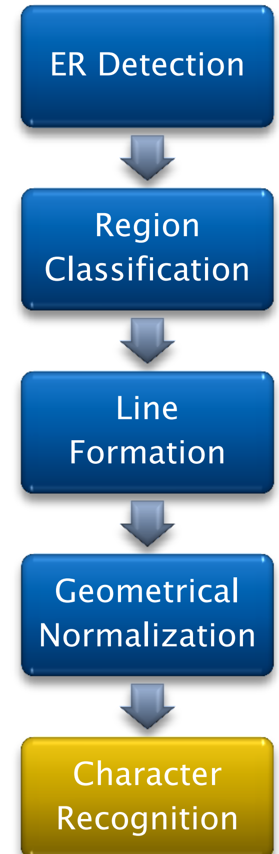


Method Description

Character Recognition

- Each region is normalized to a 20x20px matrix (while preserving aspect ratio)
- Chain code is generated on the region boundary
- Chain code direction bitmaps created for each direction (smoothed by Gaussian blur)





Method Description

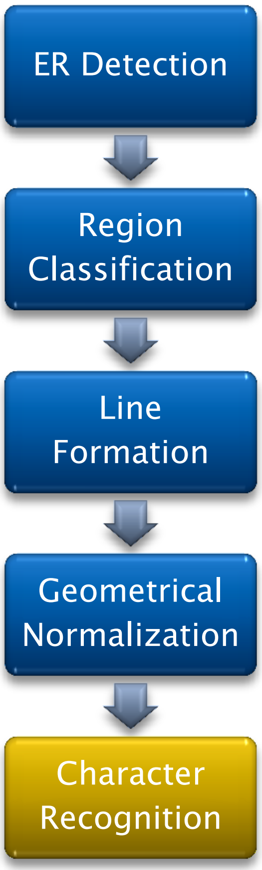
Character Recognition

- Approximate Nearest Neighbor classifier (namely FLANN) assigns (several) character labels to each region by finding K neighbors
- Recognition confidence given by ratio of equal labels in K neighbors
- Trained using synthetic data (Windows fonts)

0 1 2 3 4 5 6 7 8 9 ( )
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 ( )
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

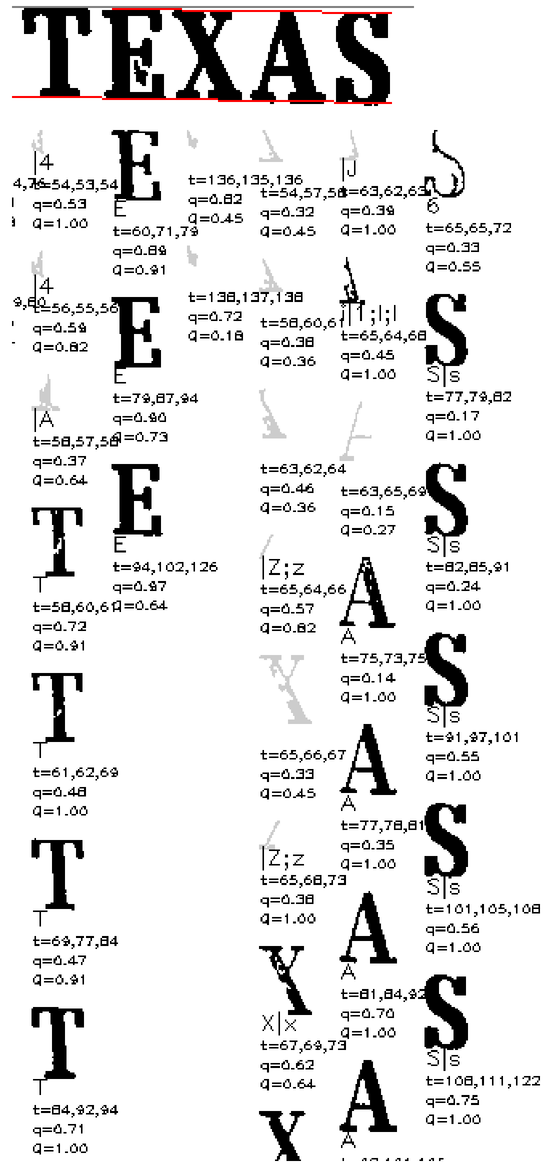
0 1 2 3 4 5 6 7 8 9 ( )
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
a b c d e f g h i j k l m n o p q r s t u v w x y z

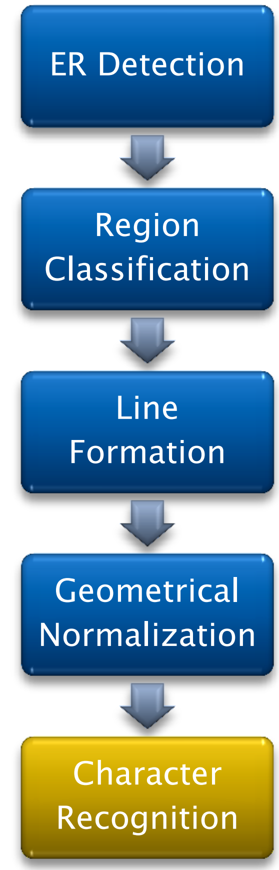


Method Description

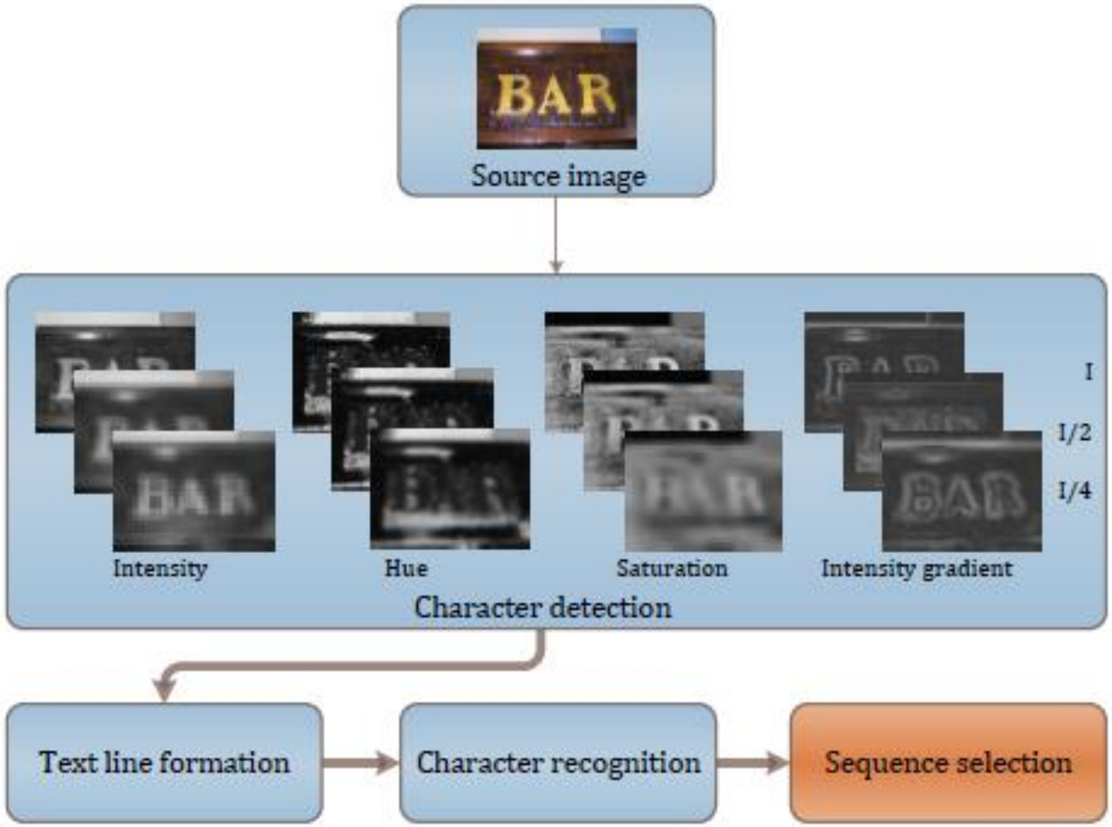
Character Recognition

- Multiple segmentations & label hypotheses for text line
- Cost function combines unary (OCR confidence) and pair-wise terms (threshold overlap, character pair frequency from a language model)
- Lowest cost found by Dynamic Programming
- Optimal weight setting / normalization still an open problem

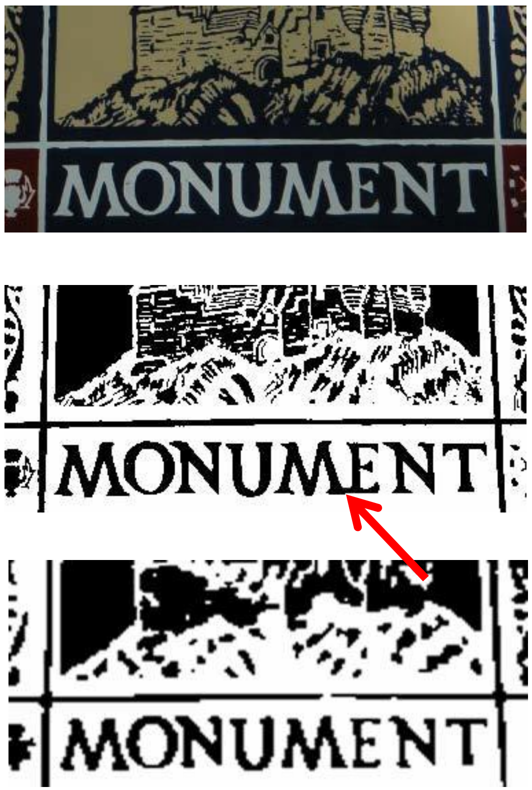




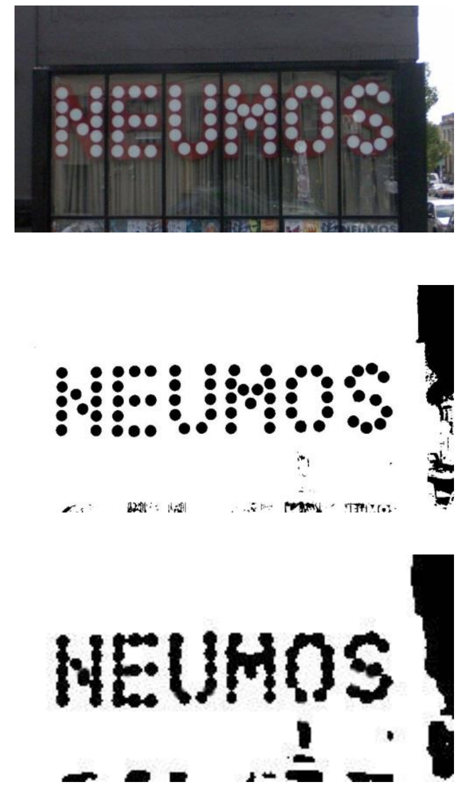
Gaussian Scale Space
Combining multiple segmentations



Gaussian Scale Space



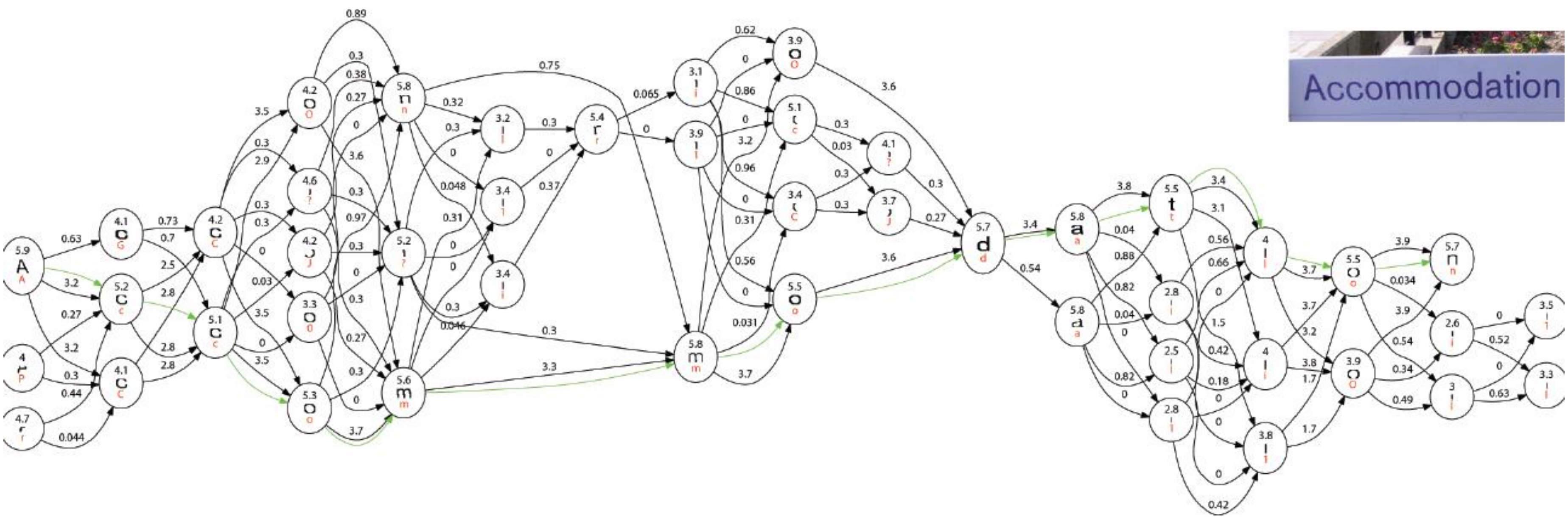
Multiple characters joint together



Characters formed of multiple small regions

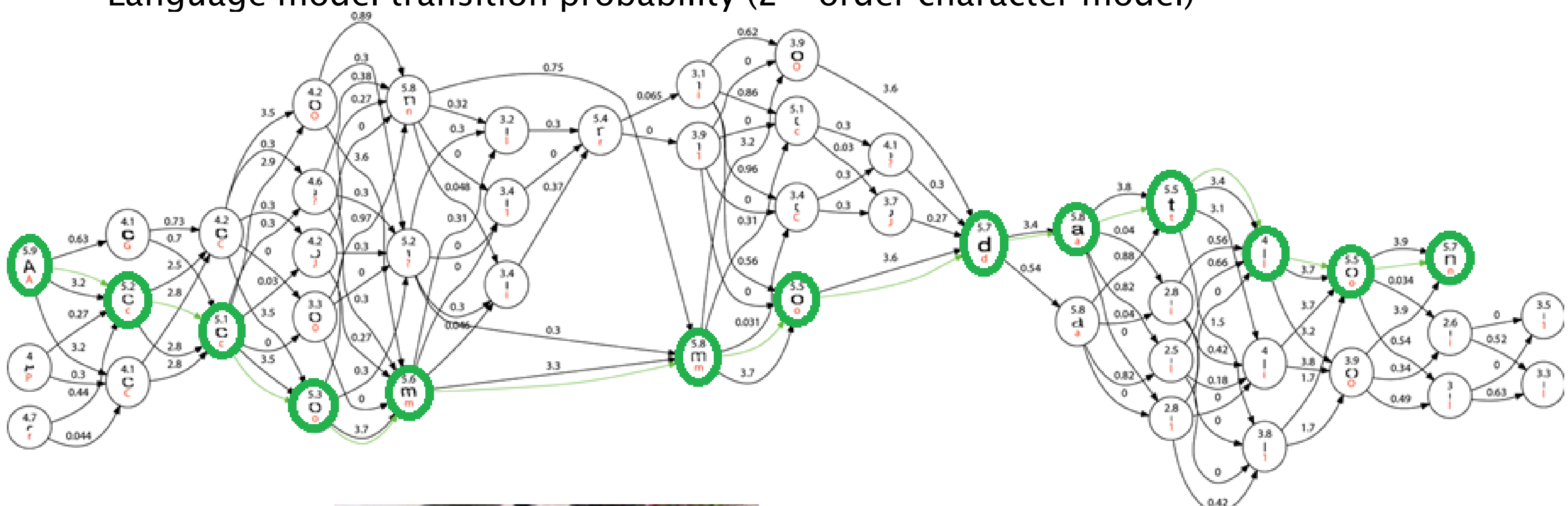
Method Description

Optimal sequence selection



Optimal Sequence Selection

- The final region sequence of each text line is selected as an optimal path in the graph, maximizing the total score
- Unary terms
  - Text line positioning (prefers regions which "sit nicely" in the text line)
  - Character recognition confidence
- Binary terms (regions pair compatibility score)
  - Threshold interval overlap (prefers that neighboring regions have similar threshold)
  - Language model transition probability (2nd order character model)



Accommodation

Sample Results

ICDAR 2011 Dataset

Osborne
Garages
AKES

Campus
Shop

ROUTE

Sample Results

ICDAR 2011 Dataset

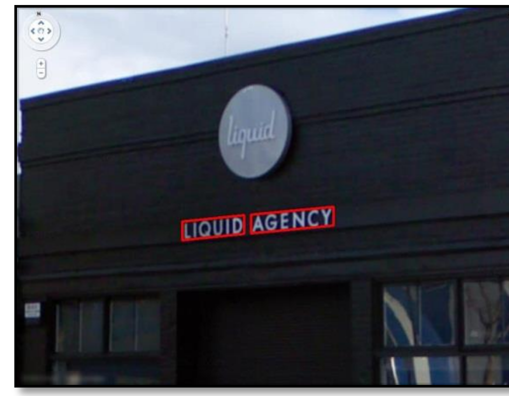
SALOON

TAX

Argo

Sample Results

Street View Dataset



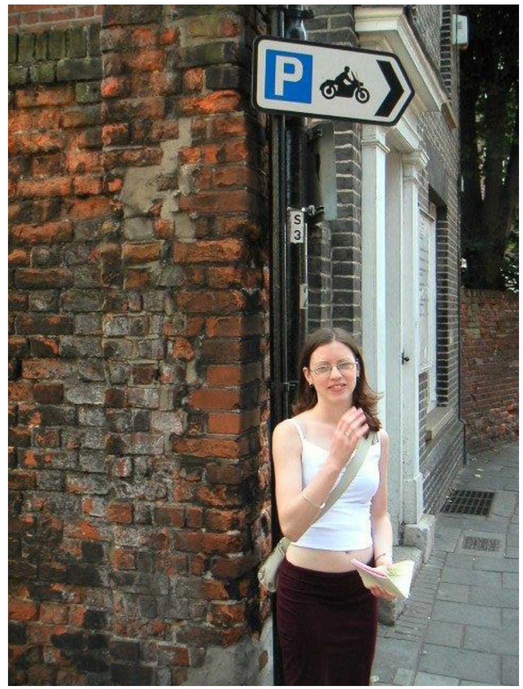
KFC
131

TADA
RESTAURANT

LIQUID
AGENCY

Limitations
Straight base line

Limitations
At least 3 characters in a text line



Sample Applications

Automatic Translator

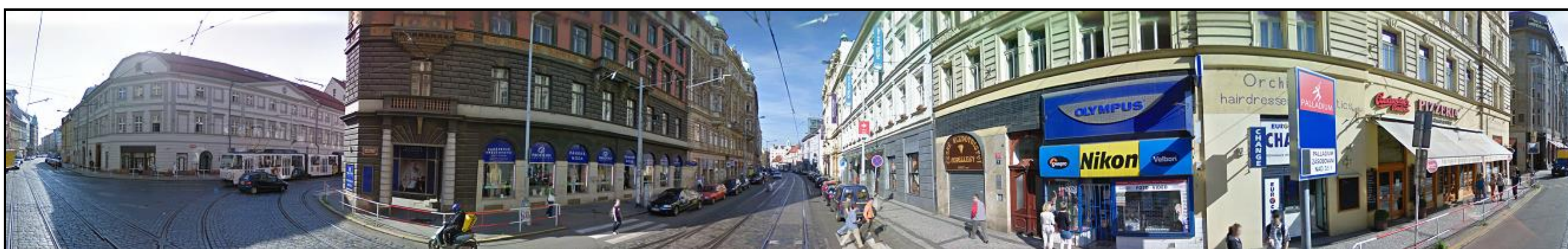
Photos taken by standard camera, downloaded from *http://www.flickr.com* and translated using *http://translate.google.com*; trained using synthetic font

DANGER	NEBEZPEČÍ
FORTS	Silný
COURANTS	Proudy
BAIGNADE	KOUPALIŠTĚ
TRAVERSEE	PŘECHOD
INTERDITE	ZAKÁZÁN

ВНИМАНИЕ T	UPOZORNĚNÍ T
В ЗОНЕ	ZÓNA
ПЕШЕХОДНОГО ТОННЕЛЯ	Pěší tunely
ВЕДЕТСЯ КРУГЛОСУТОЧНОЕ	Nonstop
ВИДЕОНАБЛЮДЕНИЕ	Video pozorování
С ЗАПИСЬЮ	S záznam

Sample Applications
Searching in image databases

Google Street View Application



Input image

Sample Applications

Assistant to visually impaired

Photos taken using standard mobile phone (5Mpix camera)

Tamiflu 75 mg
tvrde tobolky
Oseltamivirurm

TWININGS
EARL GREY
rfA BA CS

Achieved Results

- State-of-the-art results on most cited datasets (Chars74k, ICDAR 2011)
- Real-time processing
- Publications
  - Neumann L., Matas J.: Scene Text Localization and Recognition with Oriented Stroke Detection, IEEE International Conference on Computer Vision (ICCV 2013), 2013, Sydney, Australia
  - Neumann L., Matas J.: On Combining Multiple Segmentations in Scene Text Recognition, ICDAR 2013 (Washington D.C., USA) [Best Student Paper Award]
  - Neumann L., Matas J.: Real-Time Scene Text Localization and Recognition, CVPR 2012 (Providence, Rhode Island, USA)

LIVE DEMO