Text Detection and Recognition in Urban Scenes

R. Minetto[1,2], N. Thome[1], **Matthieu Cord**[1],
J. Stolfi[2], F. Précioso[1], J. Guyomard[1], N.J Leite[2]
matthieu.cord@lip6.fr

(1) LIP6)
Université Pierre et Marie Curie (UPMC) - PARIS 6 – Sorbonne Universities

(2) Institute of Computing (IC)
University of Campinas (UNICAMP)

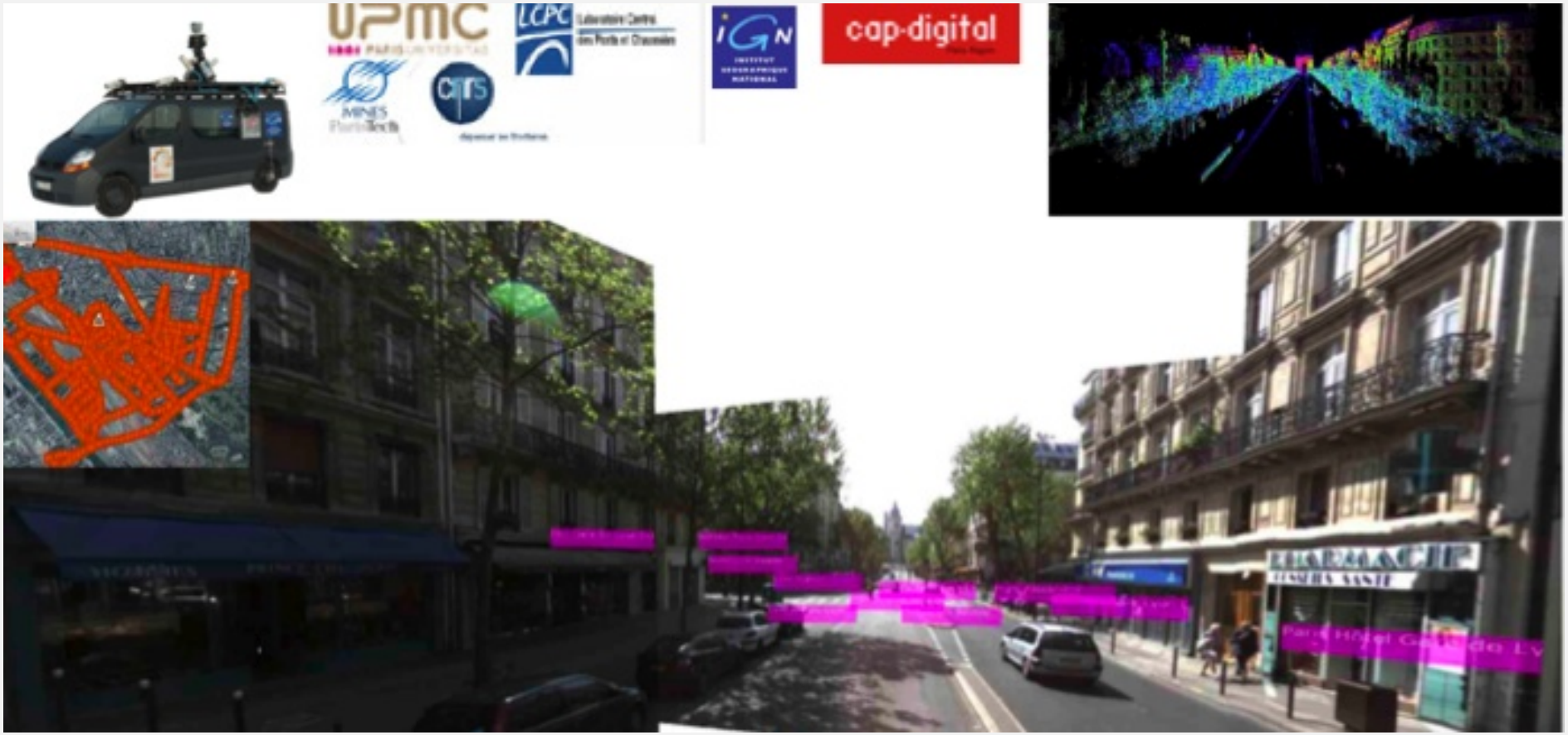
ICCV2011 Workshop
IEEE/ISPRS Workshop on Computer Vision for Remote Sensing of the Environment

7th of November 2011, Barcelona, Spain

Context of this work

ANR ITowns Project:

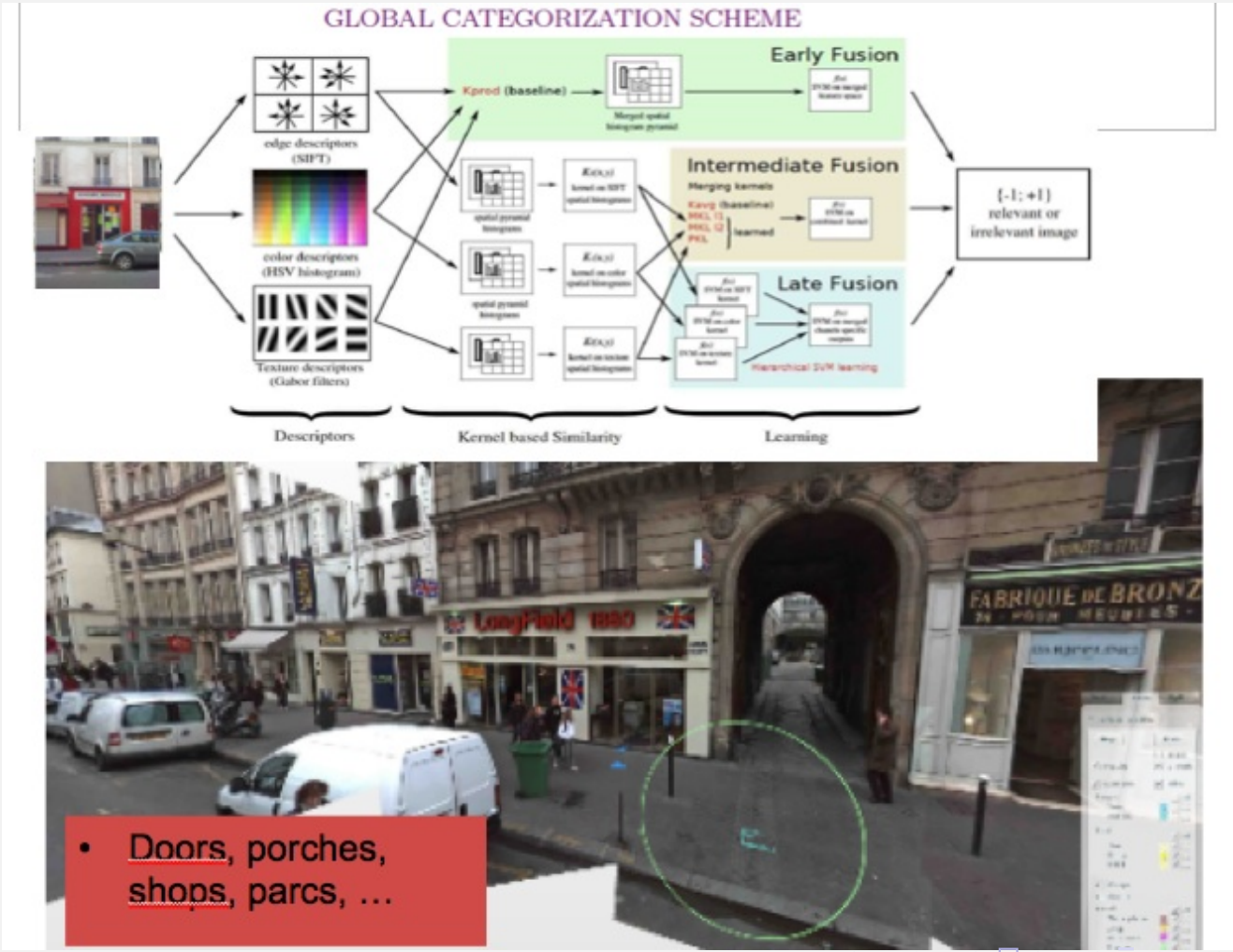
- Street digitalization, visualization, scene understanding, image-based search ...



Context of this work

ANR ITowns Project:

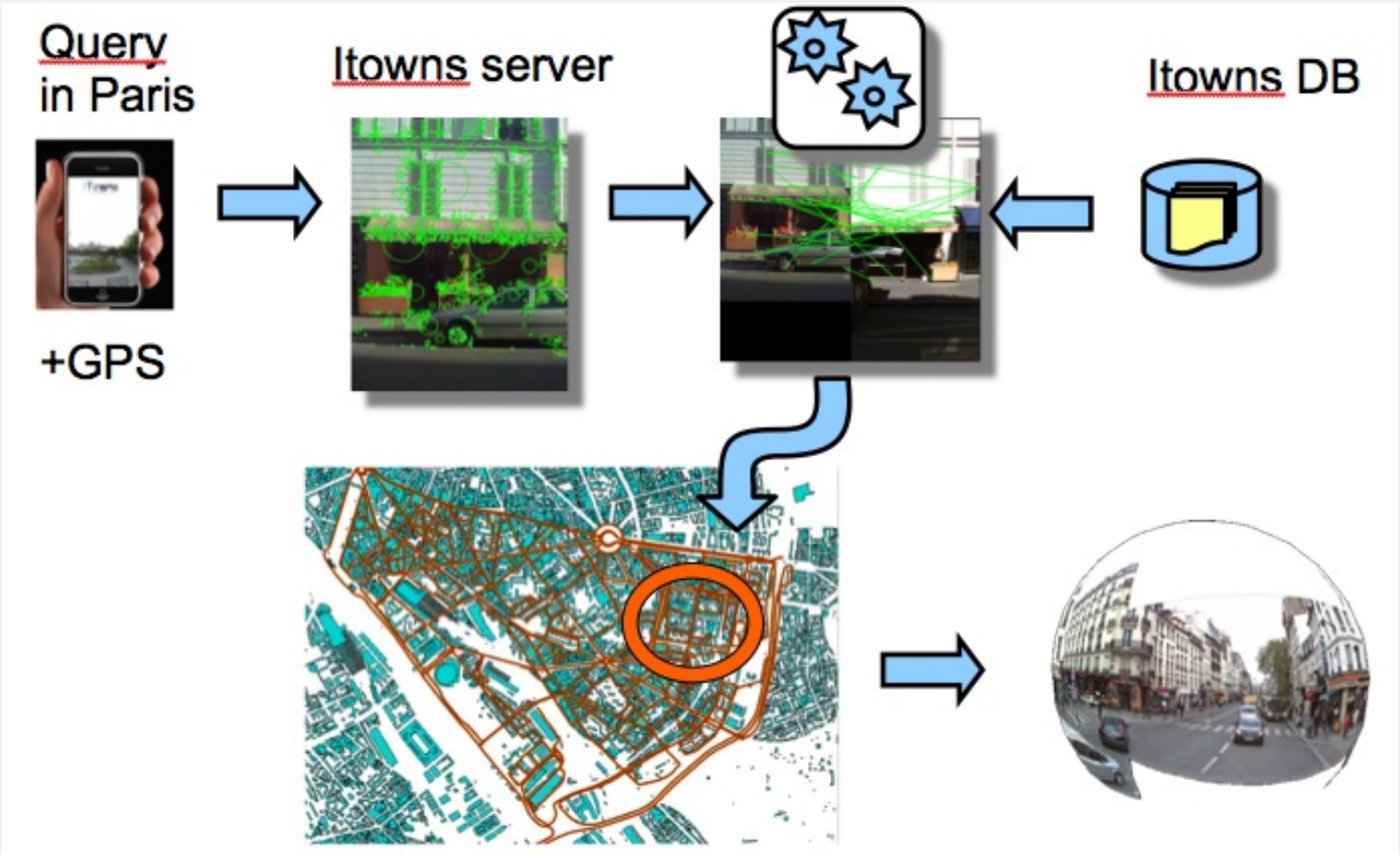
- Street digitalization, visualization, scene understanding, image-based search ...



Context of this work

ANR ITowns Project:

- Street digitalization, visualization, scene understanding, image-based search ...



Context of this work

ANR ITowns Project:

- Street digitalization, visualization, scene understanding, image-based search ...
- Object detection: Text detection in urban scenes
- Applications: Text detection is essential to build a GIS system

  - Text detection: a challenging task in computer vision;
  - Existing approaches dedicated to specific contexts;
  - Dificult in urban scenes:
    $\rightarrow$ Font variations;
    $\rightarrow$ Strong background clutter;
    $\rightarrow$ Natural noise;
    $\rightarrow$ Perspective distortion, blurring, illumination changes, etc;
  - State of the art OCR's fail in urban scenes images;

Motivation

Raw image

OCR (Tesseract)

```
It ;;'§`_L '"" '
Q ,,,¢,u,, Q ummm Q me
U Rh'
a ii ' E
» ~ !__ ,- !__f _ I _
IE ']; AVBAC- PM'
O Q I
Q* E' ~
~ U - " A _
~, » A ; v 3 f »
/ _ _ _ T ` T'
.  , Y ~- RA -\,  `
W -' 5 - gy ; :ww £-
N  ,_ _ ' _e
_ .¢~
» _ ' V nf' f
` * ` > '  7 Y, ,.
 _ ' . ' »
 rr:  »y; ~,_»;,» ' ;_ §:. ,  2
:_,-__-»a>;,a._..;'=-,.' A iz-
" ,, ,..-f` , /.
 if ;, _  'LMS-.s%,\@=I
Wm
```

Motivation

Text detection

Text regions

OCR (Tesseract)

LE RALLY

SALLE

BAR TABAC LE RALLY

un ¢*
›|:Klt-

num;

Contributions

- Development of a robust text detection scheme:
  - Bottom-up Text Hypothesis Generation
  - Hypothesis Validation (F-HOG):

    $\rightarrow$ HOG for text recognition (properties, shape, etc)

    $\rightarrow$ Horizontal slices

    $\rightarrow$ New layout of weight masks
- Integration of text detection into a GIS search engine application.

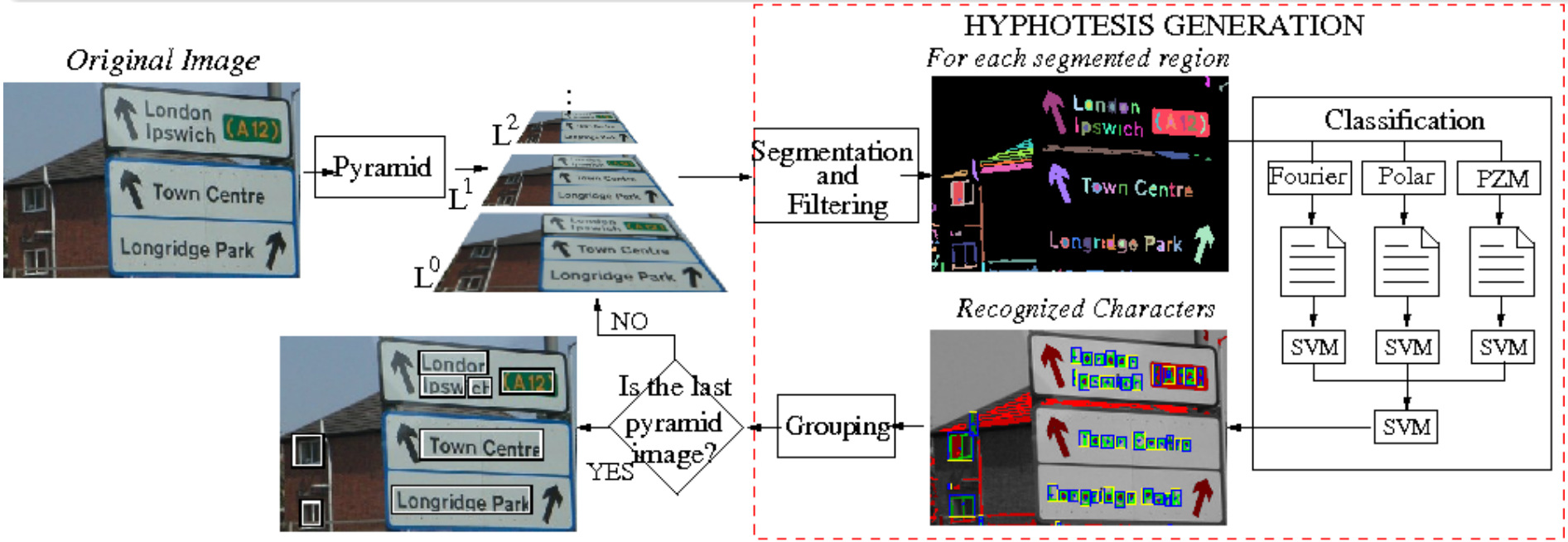
Outline

Related Works

- Texture-based approach (Top-down approach):
  (1) learning Text/NoText areas, (2) sliding window detection
  Alex Chen et al. (2d) from:
  – S.M. Lucas, **Text Locating Competition Results**, ICDAR 2005
- Connected-Component approach (Bottom-up approach):
  (1) character segmentation, (2) grouping
  Hinnerk Becker (winner of [1])
  – Boris Epshtein, Eyal Ofek and Yonatan Wexler, **Detecting Text in Natural Scenes with Stroke Width Transform**, CVPR 2010.
  – Huizhong Chen, Sam S. Tsai, Georg Schroth, David M. Chen, Radek Grzeszczuk and Bernd Girod, **Robust Text Detection in Natural Images with Edge-enhanced Maximally Stable Extremal Regions**, ICIP 2011

Bottom-Up Hypothesis Generation

- Image segmentation:
  → Toggle mapping
- Character classification:
  → Rotation invariant image descriptors
- Character grouping:
  → Geometric criteria
- Multi-resolution



Bottom-Up Hypothesis Generation

Mono-resolution v.s. Multi-resolution segmentation

- Coarser levels:
  → detects large text areas
  → ignores texture details
- Finer levels:
  → detects small regions
  → analyses more accurately the local image content

Multi-resolution

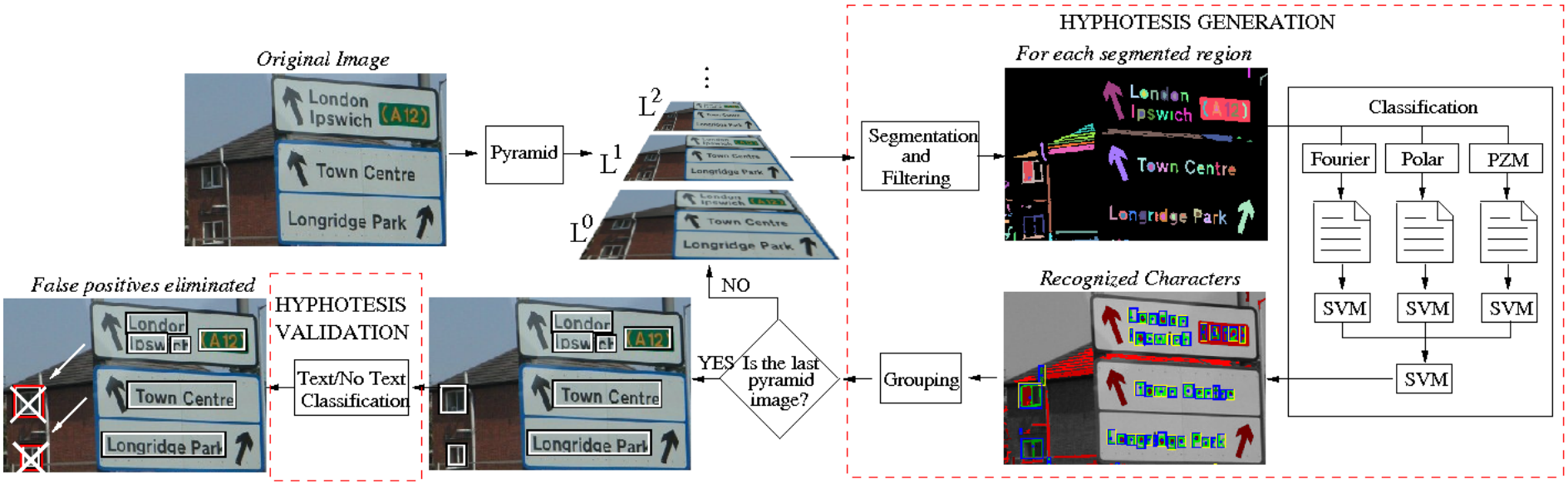
Bottom-Up Hypothesis Generation

Result

- Local analysis of image content
  $\longrightarrow$ Prone to false positives

Outline

Generation/validation process: SnooperText



- Hybrid scheme: hypothesis generation/validation paradigm
  - Hypothesis generation: multiresolution bottom-up approach
  - Hypothesis validation: top-down strategy
    $\longrightarrow$ To remove false positives by analyzing globally the window content

[*Minetto*2010] R. Minetto, N. Thome, M. Cord, J. Fabrizio, B. Marcotegui, SnooperText: A Multiresolution System for Text Detection in Complex Visual Scenes, ICIP 2010.
[*Fabrizzio*2009] J. Fabrizio, B. Marcotegui, and M. Cord, text segmentation in natural scenes using togglemapping, ICIP 09.

Hypothesis Validation

Fuzzy HOG

- Idea: analyze each candidate text region globally
- Fuzzy HOG: a global HOG descriptor with different weight masks
- Eliminate the regions with non "text-like" periodical patterns

Hypothesis Validation

Fuzzy HOG

- Idea: analyze each candidate text region globally
- Fuzzy HOG: a global HOG descriptor with different weight masks
- Eliminate the regions with non "text-like" periodical patterns

HOG idea

- Images of complex objects typically have different HOG's in different parts;
- Humans:
  $\longrightarrow$ different gradient orientation distributions in the head, torso, legs, etc;

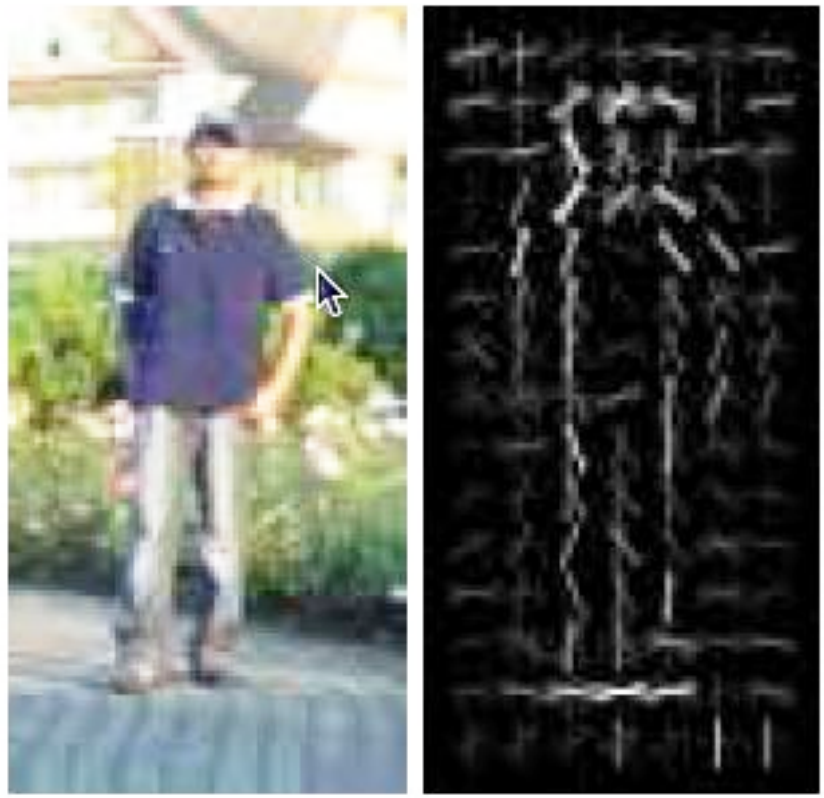
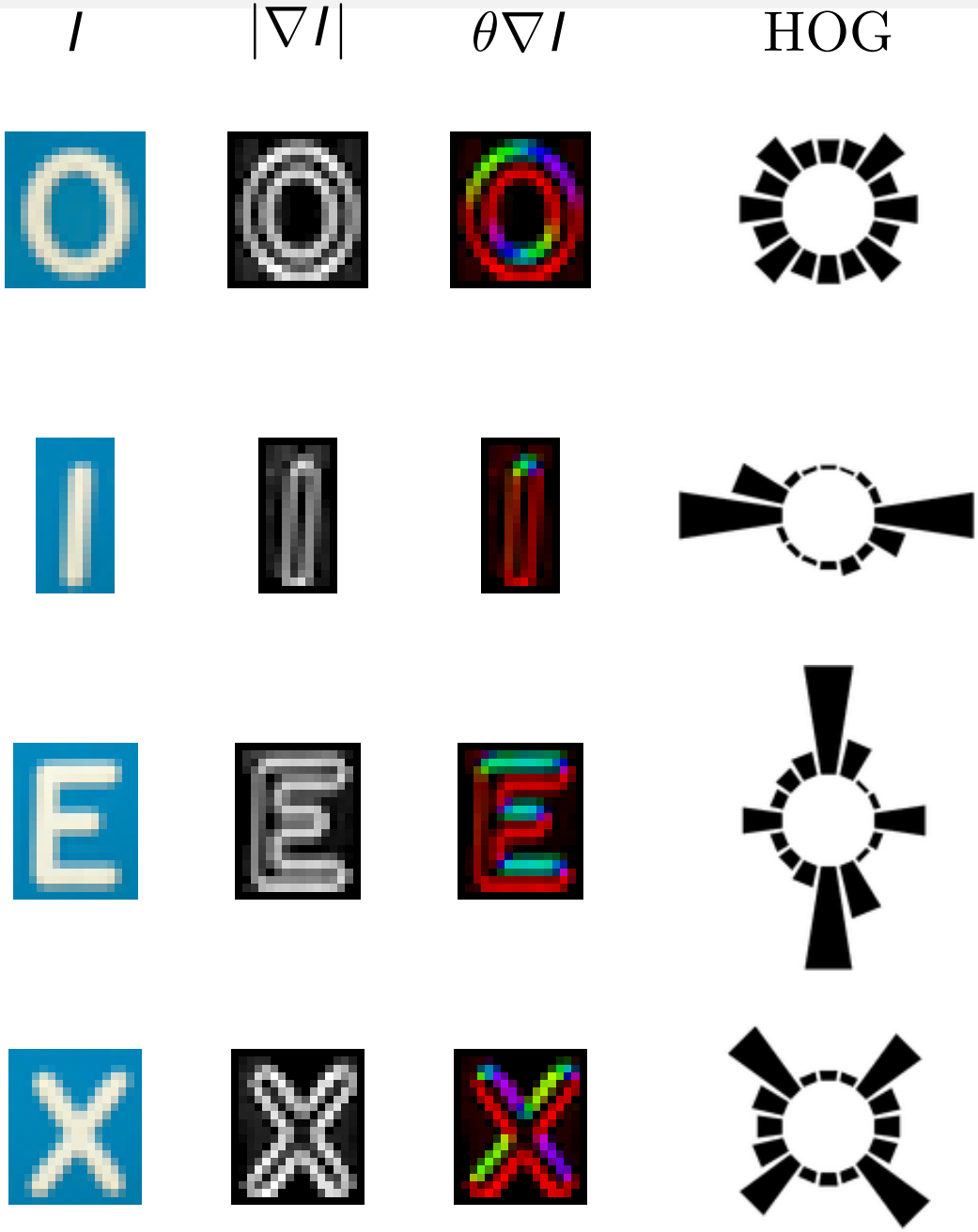


Figure: Image from: Histograms of Oriented Gradients for Human Detection. Navneet Dalal and Bill Triggs. CVPR 2005

HOG of some isolated letters



Text HOG idea

- Text-lines of Roman letters: $\neq$ HOG's in the top, middle and bottom parts:
  $\rightarrow$ The image is divided into an array of cells with one HOG to each cell;
- Top and bottom parts: **Large proportion of horizontal strokes**
  $\rightarrow$ gradients pointing mostly in the vertical direction;
- Middle part: **Large proportion of vertical strokes**
  $\rightarrow$ gradients pointing mostly in the horizontal direction;
- All parts: **Amall amount of diagonal strokes**
- The concanetation of the 3 HOG's is the descriptor of the full region.

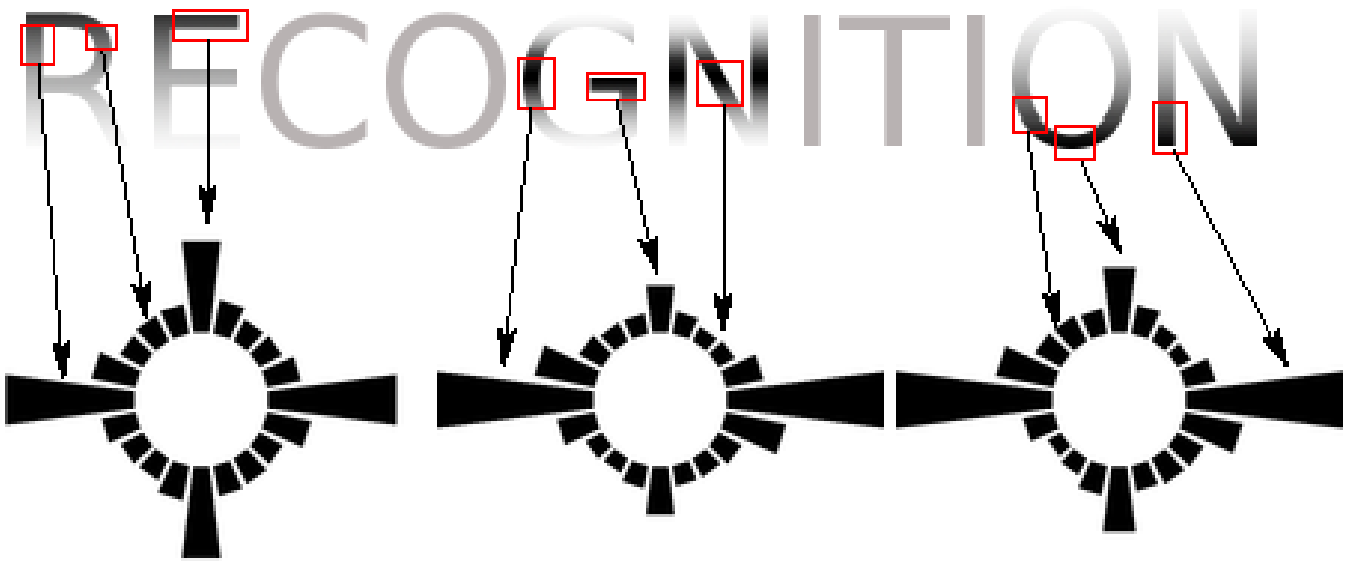
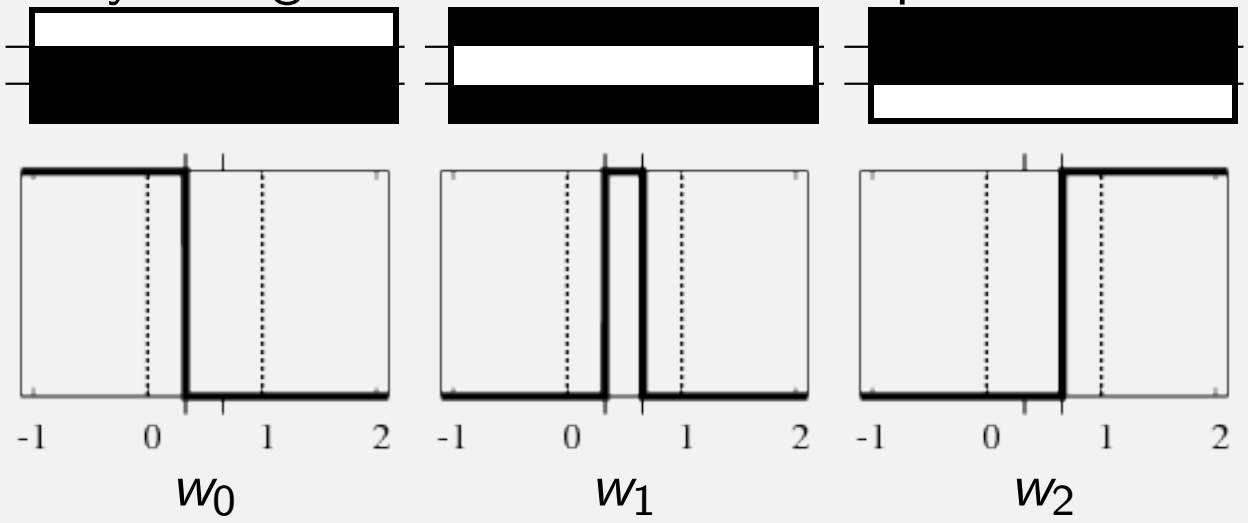
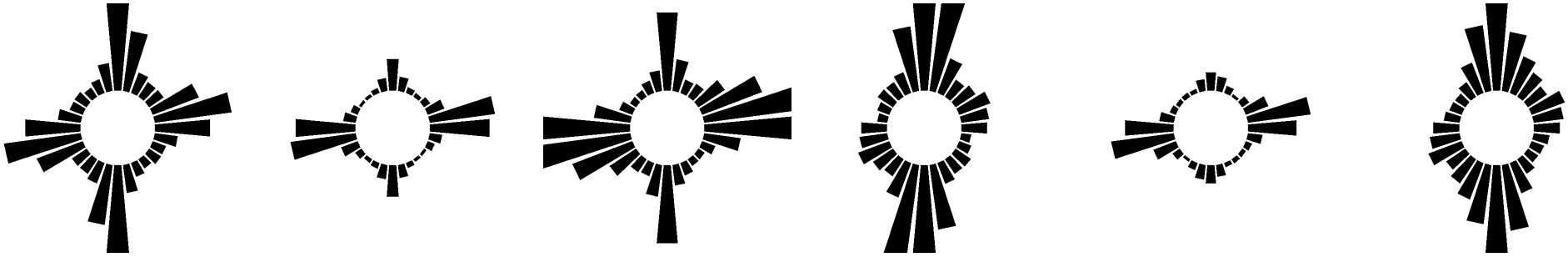


Figure: Top, middle and bottom HOGs for the text "RECOGNITION". The arrows show the contribution of specific letters strokes to the final descriptor.

Sharp cells

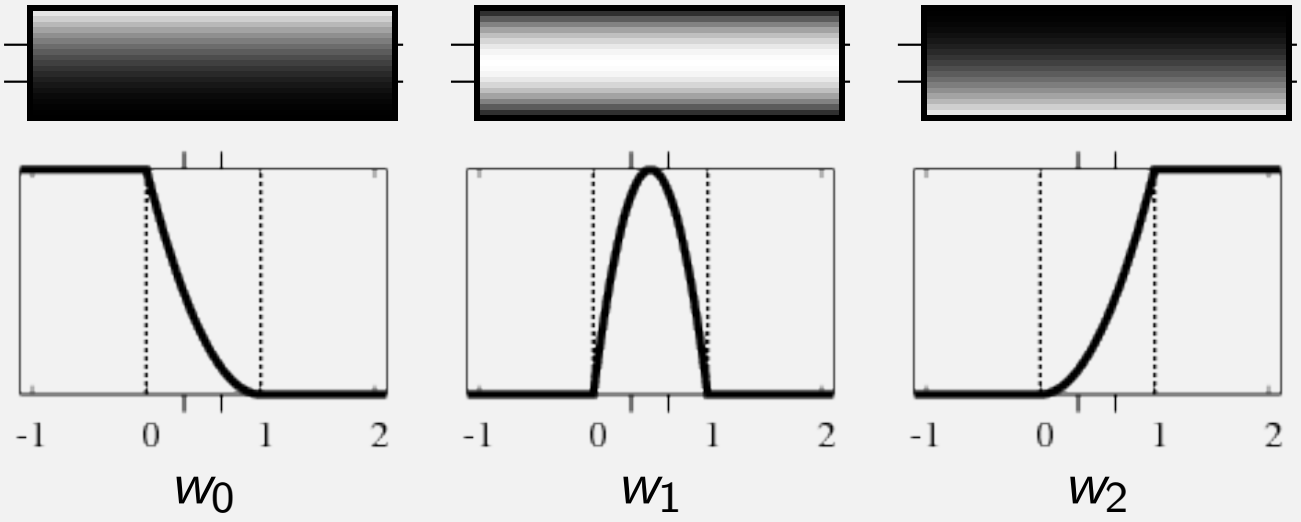
- Cells defined by sharp boundaries:
  $\longrightarrow$ HOG may change with small vertical displacements

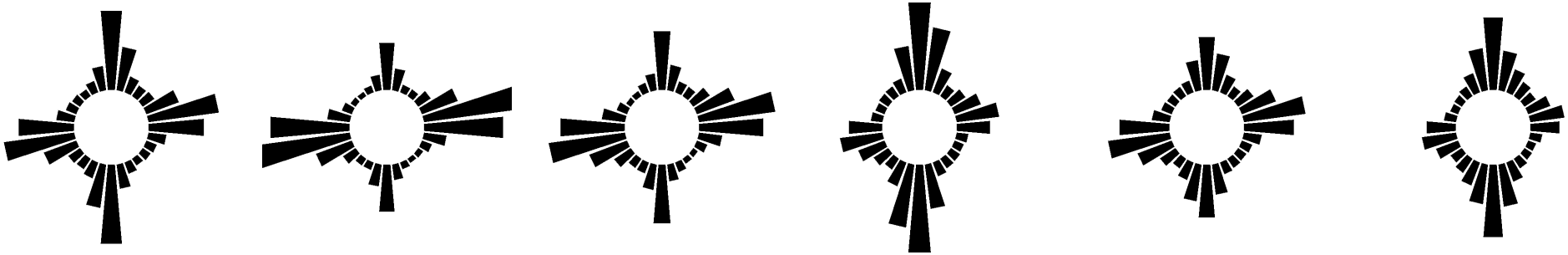




Fuzzy cells

- To avoid this problem, we used "fuzzy" cells :





Dalal et al. masks to human recognition

- Gaussian weight functions:
  $\longrightarrow$ Problem: Sharp boundaries.

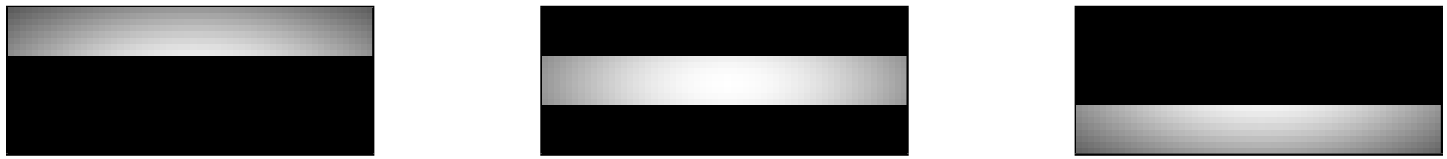


Figure: Weight functions for a single block of 1×3 cells ($\sigma_x = W/2$, $\sigma_y = H/2$).

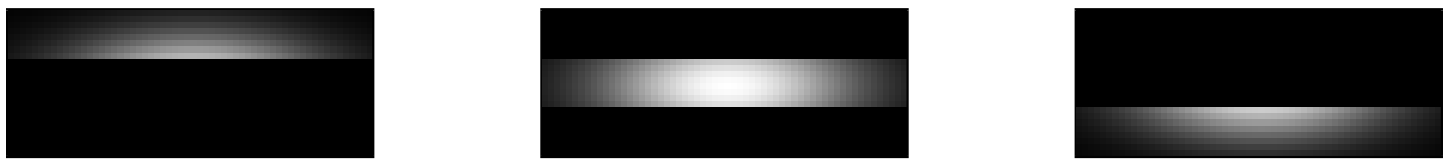


Figure: Weight functions for a single block of 1×3 cells ($\sigma_x = W/4$, $\sigma_y = H/4$).

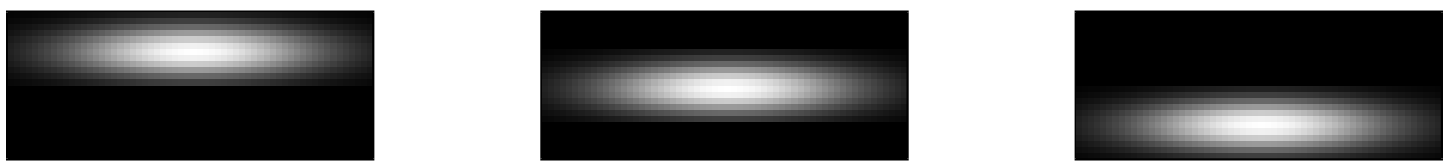
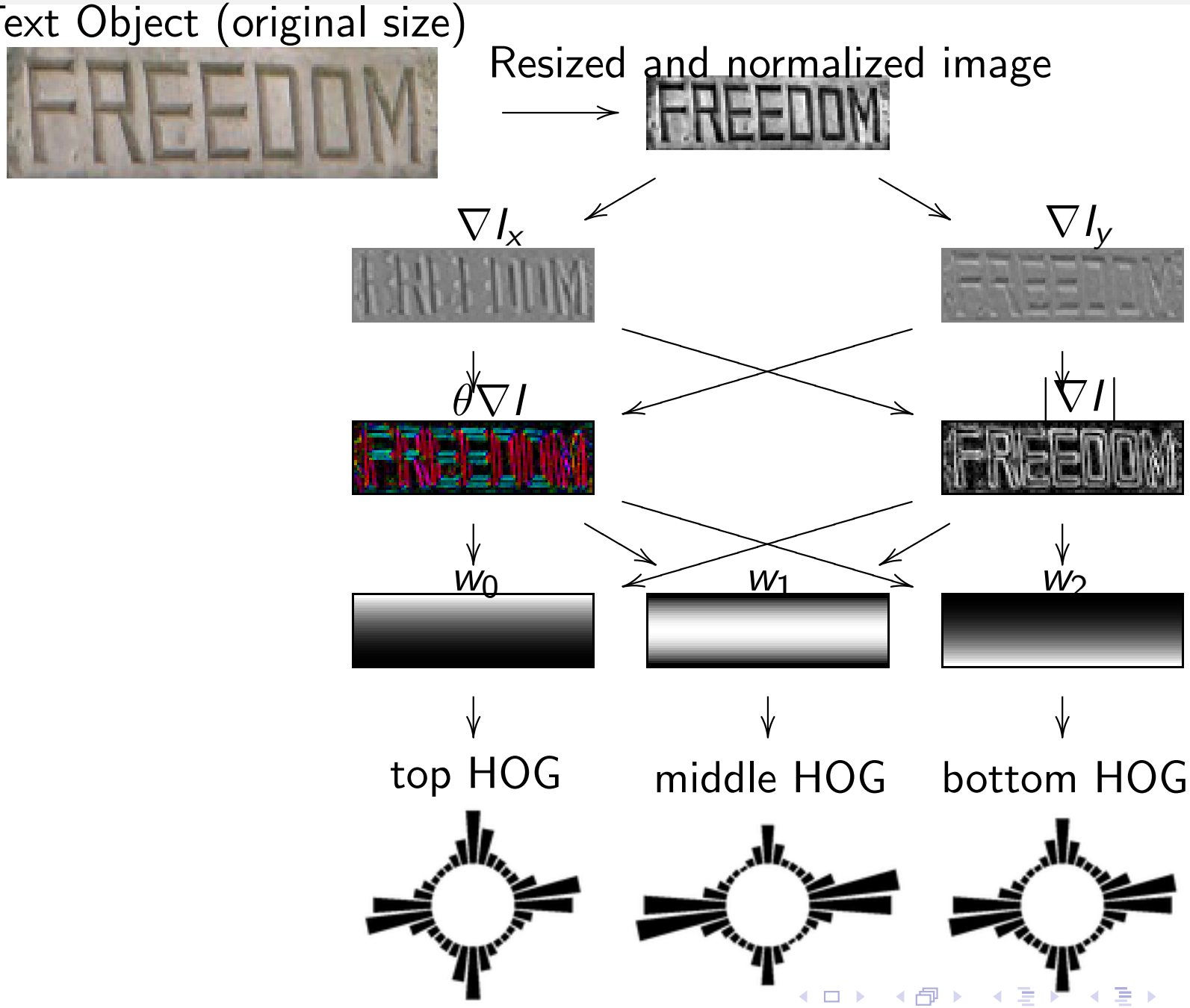
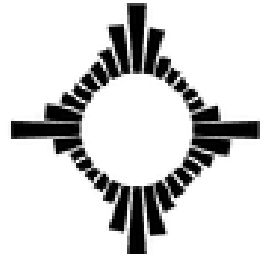


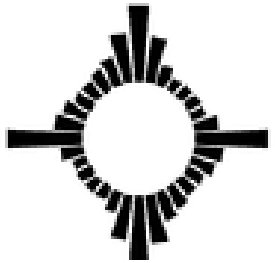
Figure: Weight functions for 1×3 single-cell blocks. Each with height $H/2$ and overlapped with stride $H/4$ ($\sigma_x = W/4$, $\sigma_y = H/8$).

Text HOG descriptor scheme

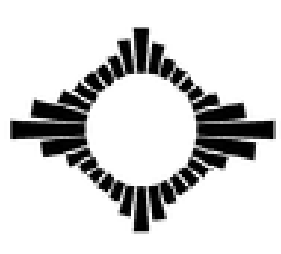


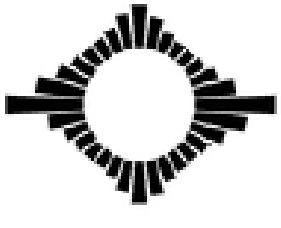
F-HOG of text and non-text regions



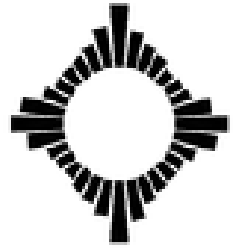


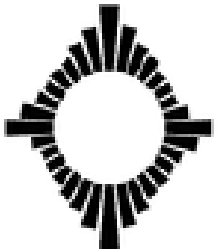
top



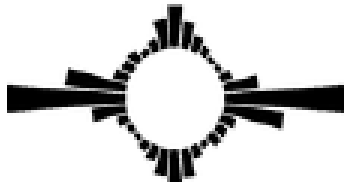


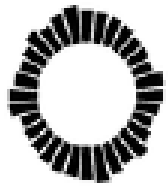
middle



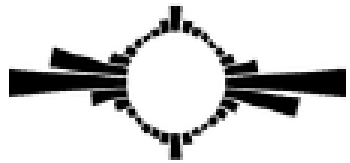


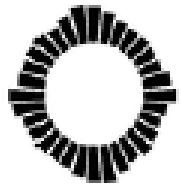
bottom



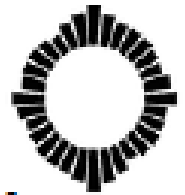


top





middle



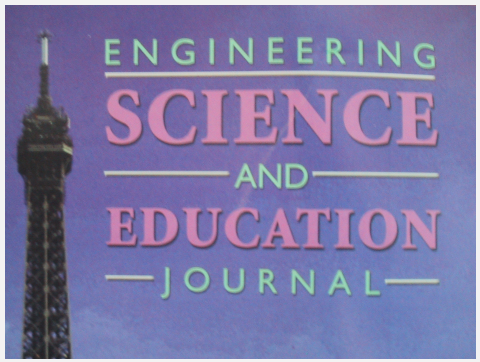
bottom

Outline

ICDAR

Dataset

- 499 color images (training/testing)
- Captured with different digital cameras and resolutions
- Images from indoor and outdoor scenes
- Groundtruth available (XML)



ICDAR

Metrics

Precision

$$p = \frac{\sum_{r_e \in E} m(r_e, T)}{|E|}$$

Recall

$$r = \frac{\sum_{r_t \in T} m(r_t, E)}{|T|}$$

Ranking

$$f = \frac{1}{\alpha/p + (1-\alpha)/r}$$

- $m(r, R)$: best match for a rectangle r in a set of rectangles R.
- T: set of manually identified text regions (groundtruth);
- E: set of text regions reported by the detector;
- f: harmonic mean of precision and recall ($\alpha = 0.5$)

ICDAR

Performances results

System	Precision (p)	Recall (r)	f
Our System	**0.73**	**0.61**	**0.67**
Epshtein et al. (CVPR 2010)	**0.73**	0.60	0.66
Chen et al. (ICIP 2011)	**0.73**	0.60	0.66
SnooperText (ICIP 2010)	0.63	0.61	0.61
Hinnerk Becker [1]	0.62	0.67	0.62
Alex Chen	0.60	0.60	0.58
Ashida	0.55	0.46	0.50
HWDavid	0.44	0.46	0.45
Wolf	0.30	0.44	0.35
Qiang Zhu	0.33	0.40	0.33
Jisoo Kim	0.22	0.28	0.22
Nobuo Ezaki	0.18	0.36	0.22
Todoran	0.19	0.18	0.18
Full	0.01	0.06	0.08

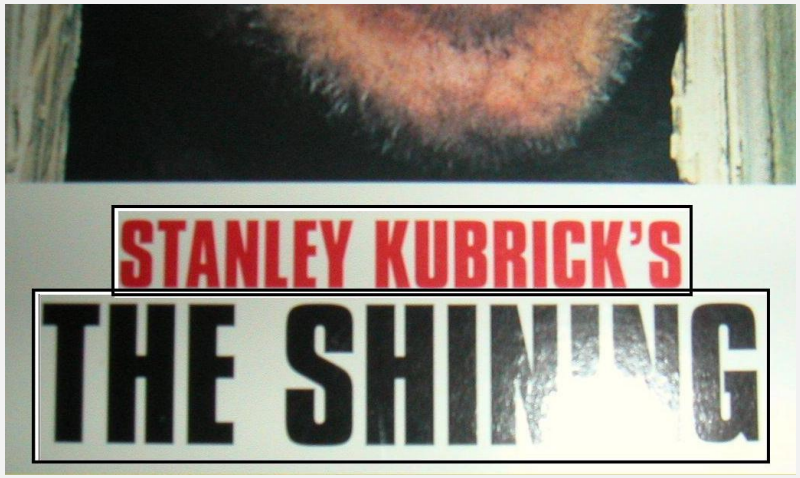
ICDAR

Successfull detections

$p = 0.96, r = 0.64, f = 0.77$

$p = 0.93, r = 0.93, f = 0.93$

$p = 0.90, r = 0.90, f = 0.90$



$p = 0.68, r = 0.56, f = 0.61$

ICDAR

Failures

$p = 0.00, r = 0.00, f = 0.00$

$p = 0.00, r = 0.00, f = 0.00$

$p = 0.71, r = 0.95, f = 0.81$

iTowns

Performances

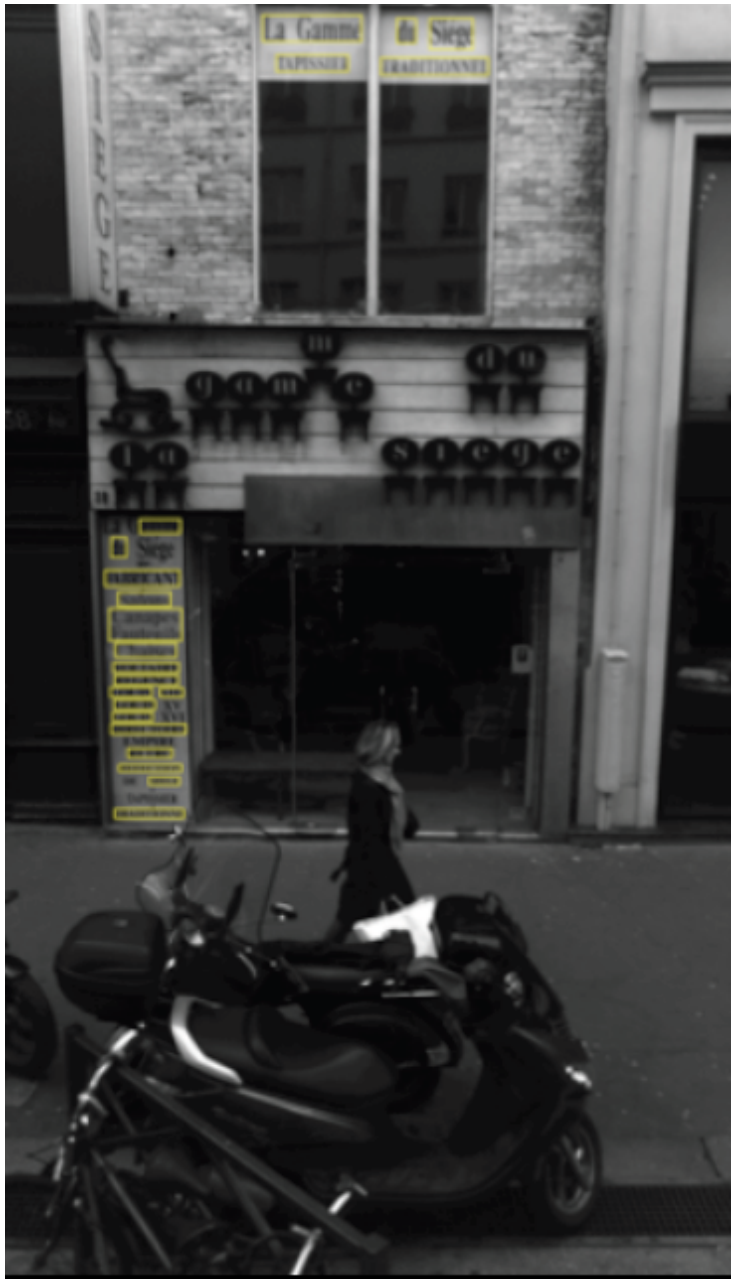
- ICDAR metrics;
- Text Detection + F-HOG: precision improvement of 23%

System	Precision (p)	Recall (r)	f
Our System	**0.69**	**0.49**	**0.55**
SnooperText (ICIP 2010)	0.46	0.49	0.48

iTowns - Detection results

iTowns - Detection results

iTowns - Detection results



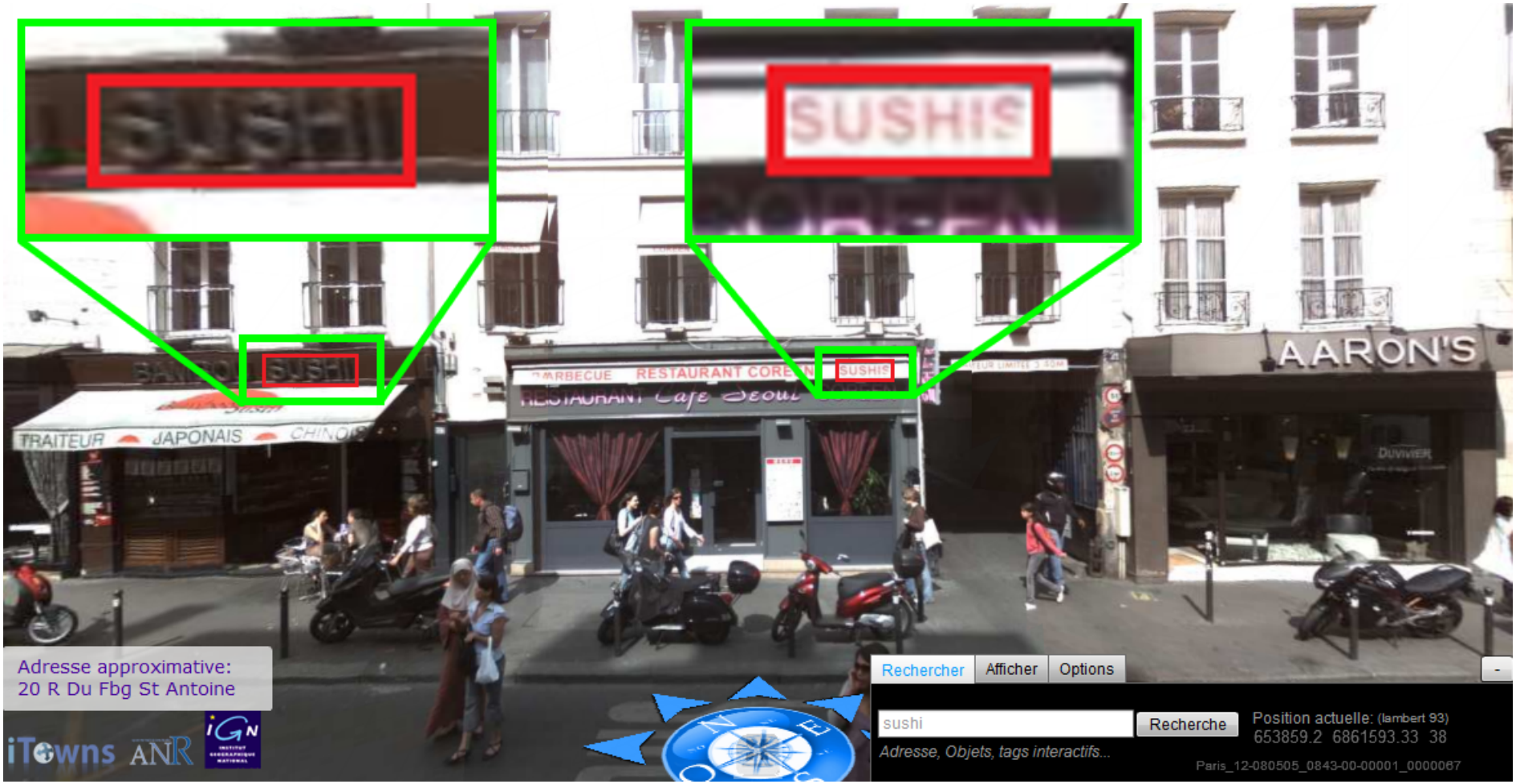
itowns KeyWord Search

Text detection + OCR (Tesseract)

- Textual query for image retrieval
- Word matching by Edit distance

itowns KeyWord Search

itowns KeyWord Search



Conclusion

- Robust text detection system for urban scenes
- Analyze of the HOG for text recognition (properties, shape, etc)
- Efficient way to use the HOG as a text recognizer (F-HOG)
- Very good results on the ICDAR dataset
- System integrated in iTowns for real street image databases
- Application with an OCR to build a GIS system

Thank you for your attention !

QUESTIONS ?

People

Matthieu Cord, Nicolas Thome
LIP6, Univ. UPMC-PARIS VI
matthieu.cord@lip6.fr

- Co-supervision PhD student R. Minetto with Prof. J. Stolfi, University of Campinas (UNICAMP), Brazil
- Continuation of a strategy developed with CMM in ANR itowns

http://webia.lip6.fr/~cord/

UPMC
SORBONNE UNIVERSITÉS

SnooperTrack: Extension to videos

SnooperText: Conclusion

- Combines bottom-up & top-down mechanisms
- Efficient in various contexts: urban images, standard databases
- Computational time may make approach difficult to scale up: 640×480 pixel images $\sim$ 1 minute

SnooperTrack: Motivations

- Combining detection & tracking:
  - Speedup text detection in image sequences
  - Discard false positives
  - Improves detection accuracy
- Detection: SnooperText
- Tracking: Particle Filtering (HoG)
- Merging detection & tracking with a combination of position, size and appearance features

SnooperTrack: Results

SnooperText: Detection only	SnooperTrack: Detection + Tracking
	

Loading ./images/text/trackingtext.avi

[*MinettoICIP*11] Rodrigo Minetto, Nicolas Thome, Matthieu Cord, Neucimar Leite, Jorge Stolfi, SnooperTrack: Text Detection and Tracking for Outdoor Videos, ICIP 2011