Detection and representation for image understanding

Matthieu Cord

Laboratoire d'Informatique de Paris 6 (LIP6)
Université Pierre et Marie Curie (UPMC)

October, 2011

Outline

LIP6/DAPA/MALIRE

People

1. LIP6 lab in Paris
   - $\sim$ 200 permanent researchers, $\sim$ 250 Phd students
2. DAPA department: Databases and Machine learning
   - $\sim$ 35 permanent researchers, $\sim$ 50 Phd students
3. MALIRE team: MAchine Learning and Information REtrieval
   - $\sim$ 20 permanent researchers, $\sim$ 35 Phd students, $\sim$ 10 Post-docs
4. Our research group
   - 2 permanent researchers (M. Cord, N.Thome), $\sim$ 10 Phd/Post-docs

LIP6/DAPA/MALIRE

Research Topics: image representations and similarities

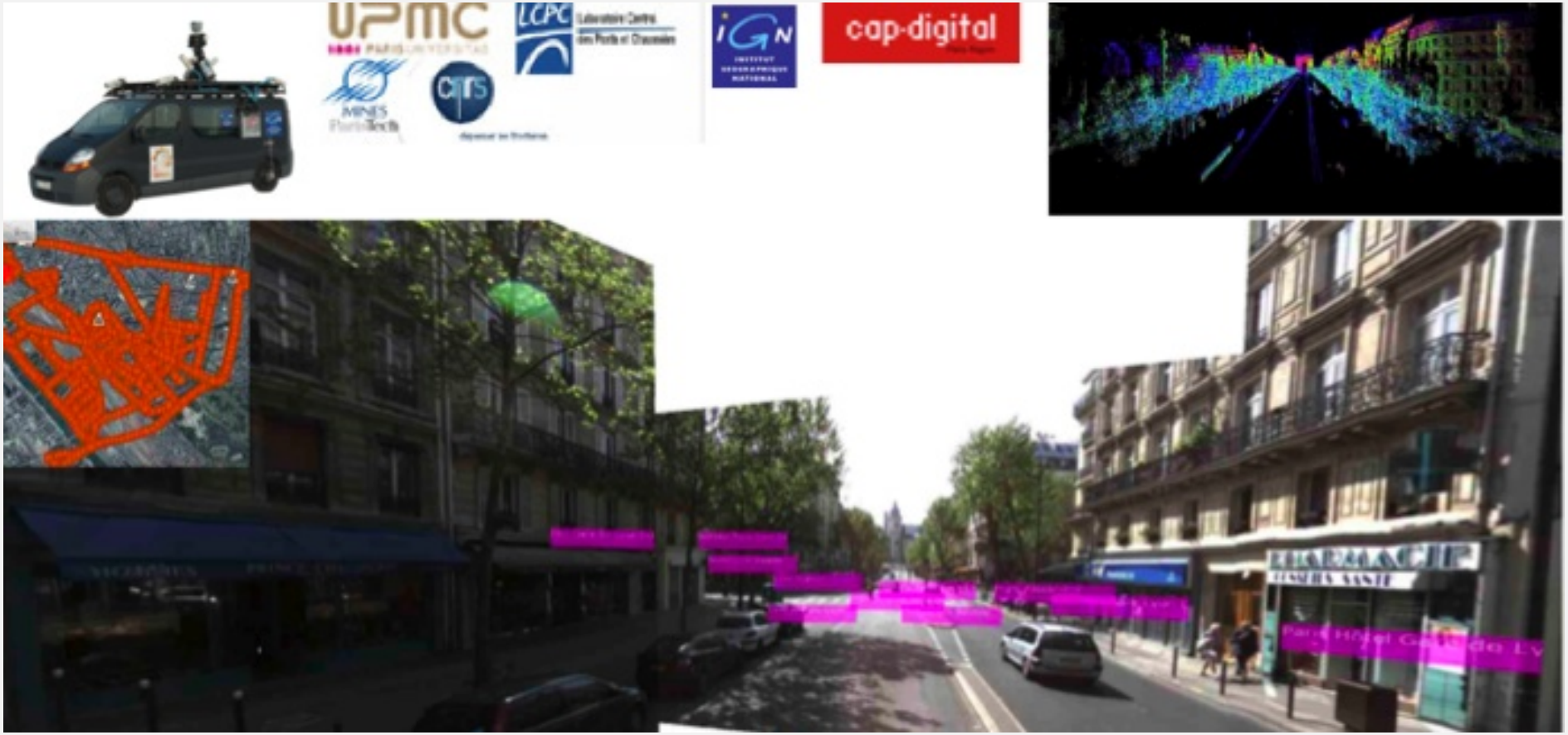
- Image Representations:
  - Deep learning for image understanding
  - Computer-vision-based image representations (beyond BoW)
- Kernel methods
  - Similarity design (image search, actor video retrieval)
  - Similarity learning (feature combination)
- Content-based image and video retrieval systems in huge multimedia databases
- Active/Interactive strategies for image and video retrieval
- Object (Text) Detection with context
- Hybrid strategies for web archiving

Outline

Main Projects

Retrieval, object detection

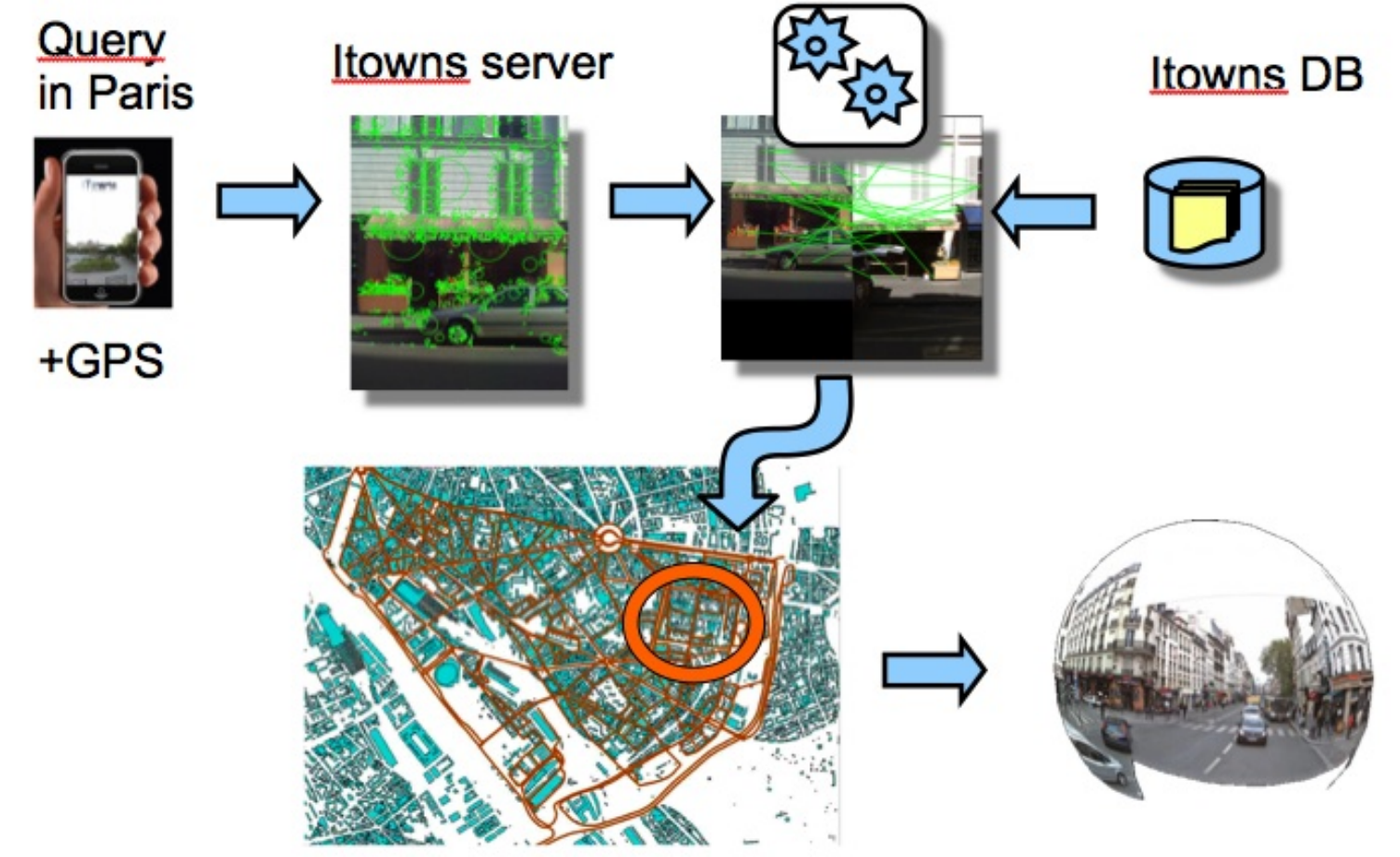
- ANR ITowns: street digitalization, visu., understand., search ...



Main Projects

Retrieval, object detection

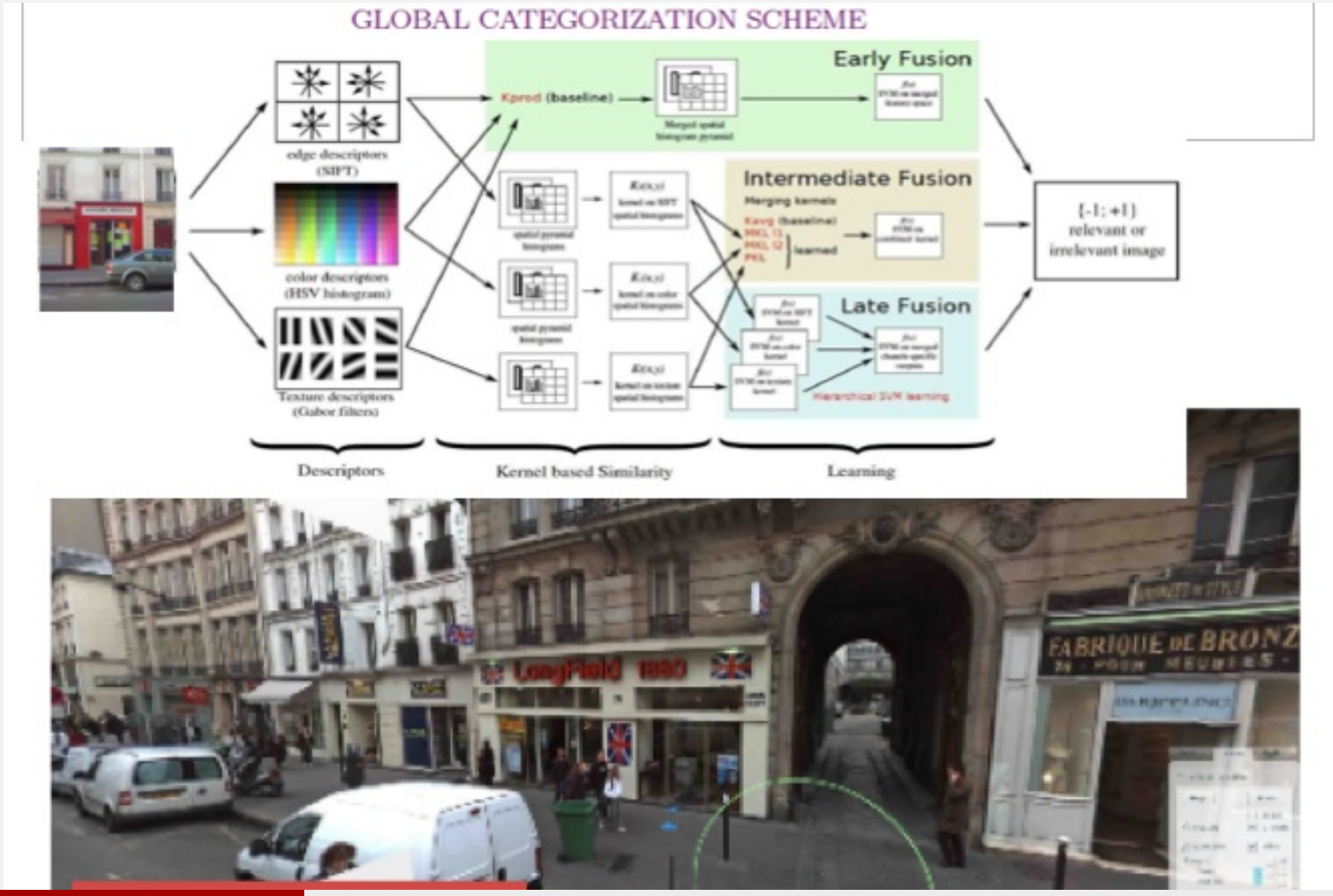
- ANR ITowns: street digitalization, visu., understand., search ...



Main Projects

Retrieval, object detection

- ANR ITowns: street digitalization, visu., understand., search ...



Main Projects

Retrieval, object detection

- ANR ITowns: street digitalization, visu., understand., search ...

Main Projects

Retrieval, object detection

- ANR ITowns: street digitalization, visu., understand., search ...

Deep Learning

- ANR ASAP: french consortium for learning deep representations
- Bilateral project MERLION: partnership with IPAL, Singapore

Main Projects

Retrieval, object detection

- ANR ITowns: street digitalization, visu., understand., search ...

Deep Learning

- ANR ASAP: french consortium for learning deep representations
- Bilateral project MERLION: partnership with IPAL, Singapore

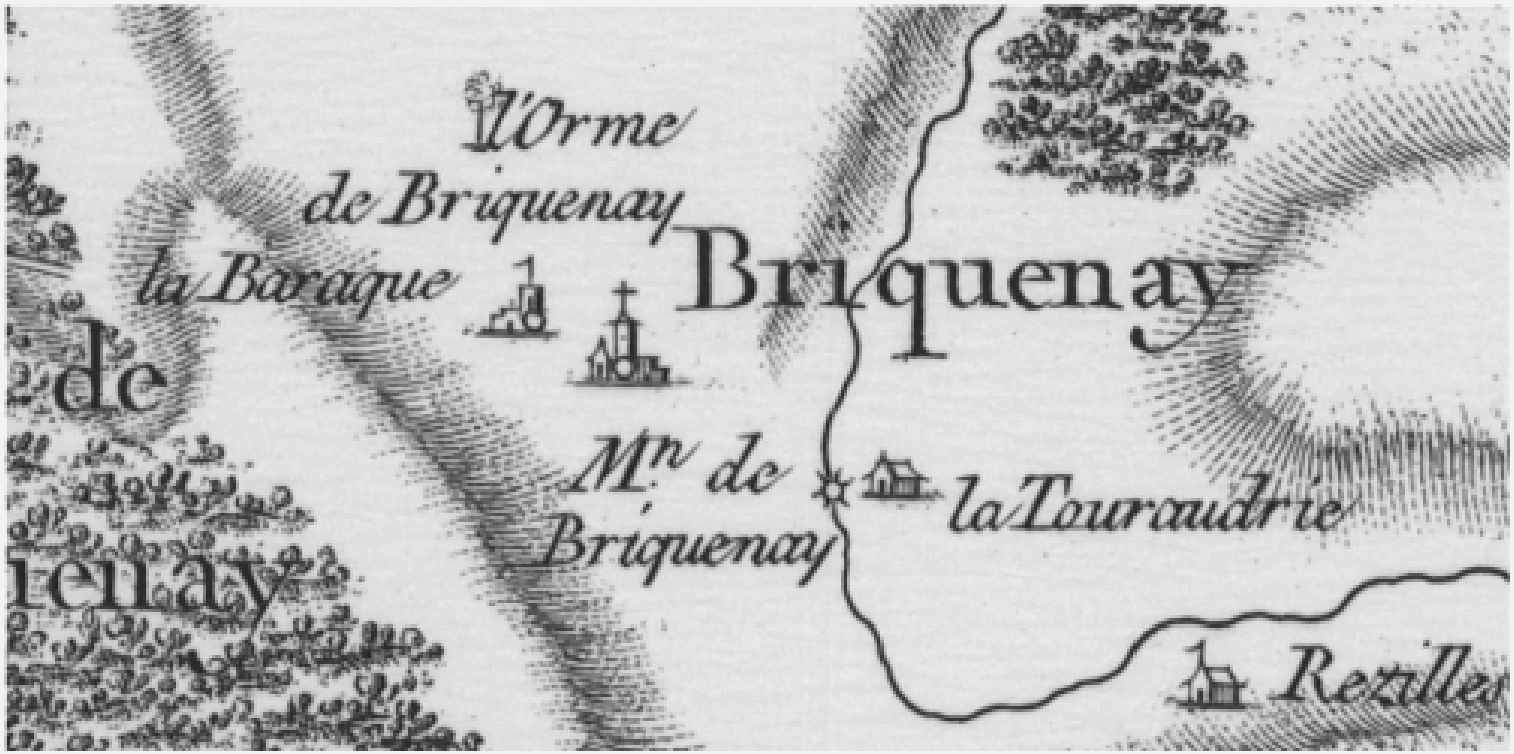
Image and video analysis

- French-Bresilian CAPES-COFECUB Project (S. Avila and R. Minetto's Phd)

Main Projects

Object detection

- ANR Geopeuple: old Maps Analysis (EHESS, COGIT/IGN)
- Interpreting Population Evolution: from 18th century → nowdays

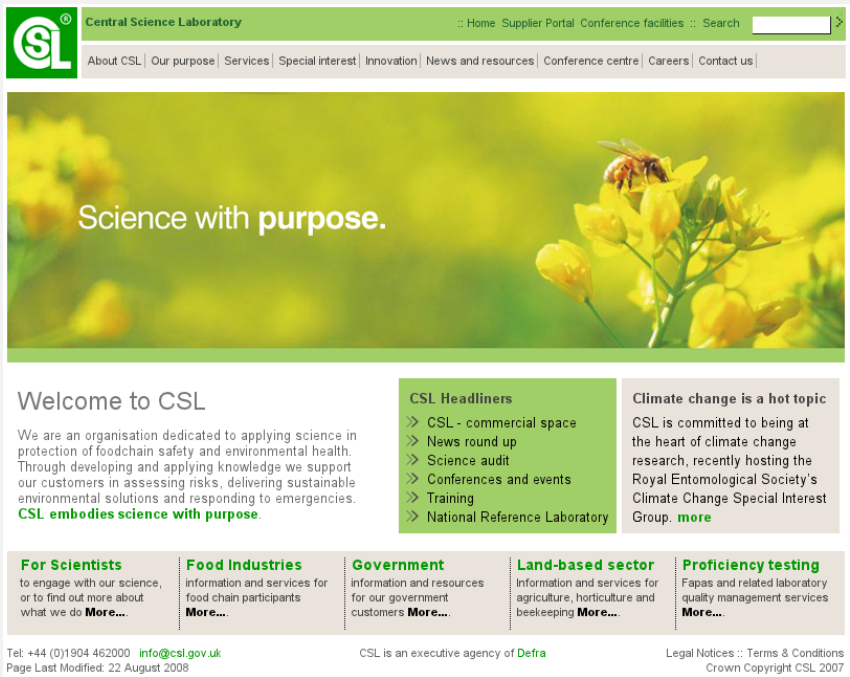


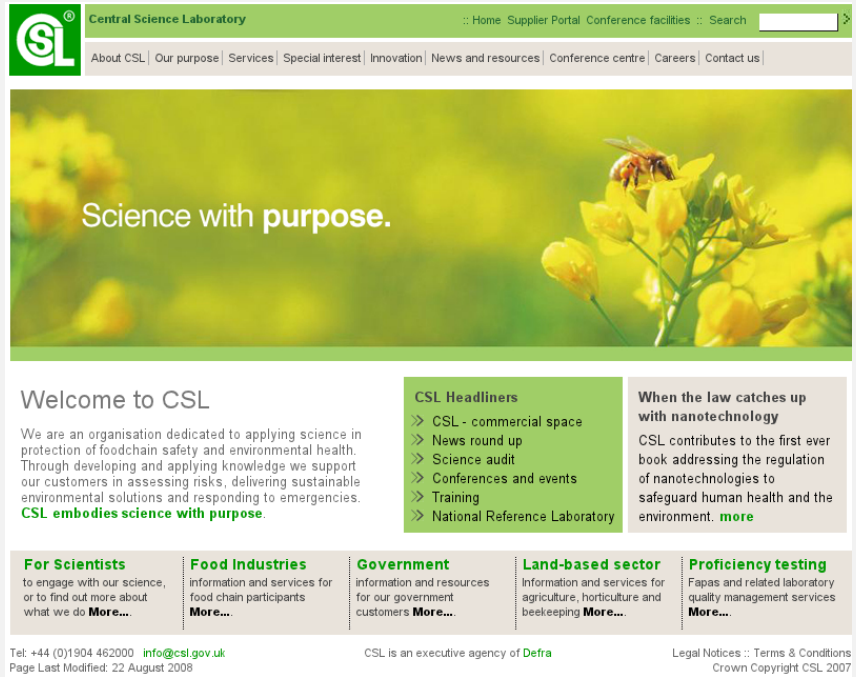
- Contextual object detection

Main Projects

Scalable data migration preservation

- European Project (IP) SCAPE: SCAlable Preservation Environments
- LIP6 (BD/MALIRE): automation on the web page versioning
  - Combining structural and visual feature

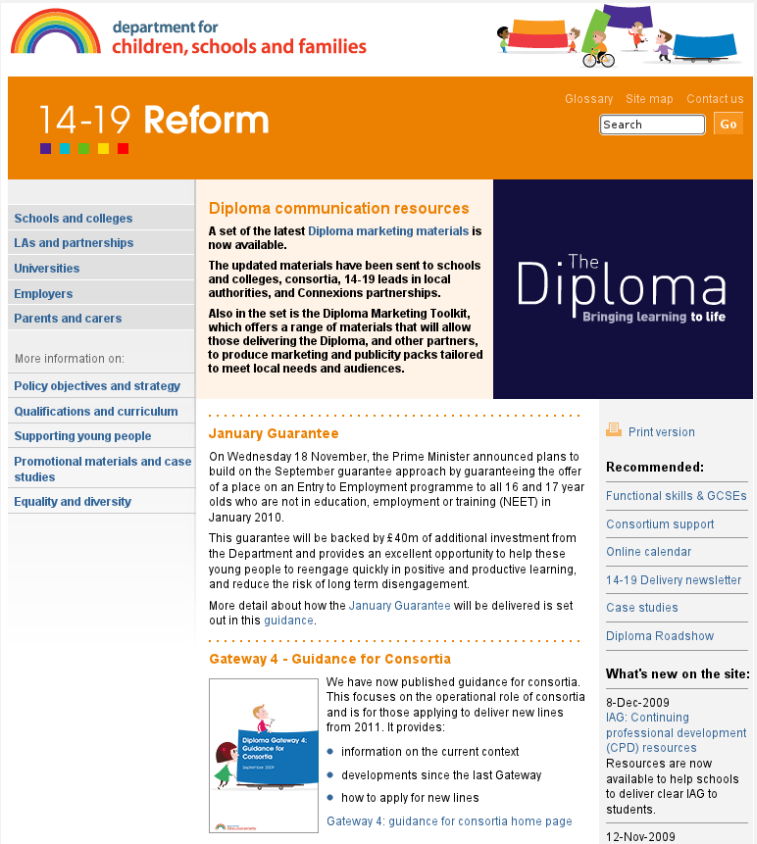




No significant change $\rightarrow$ same versionning

Main Projects

SCAPE: SCAlable Preservation Environments



significant change $\rightarrow$ generate a new versionning

- LIP6 (BD/MALIRE): Combining structural and visual feature
  - Metric/kernel learning (MS student + PhD)

Outline

Introduction

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

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

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;

Related Works

- Hinnerk Becker [1] (Bottom-up approach)
- Alex Chen et al. [1] (Top-down approach)
- Epshtein et al. [2] (Bottom-up approach)
- Chen et al. [3] (Bottom-up approach)

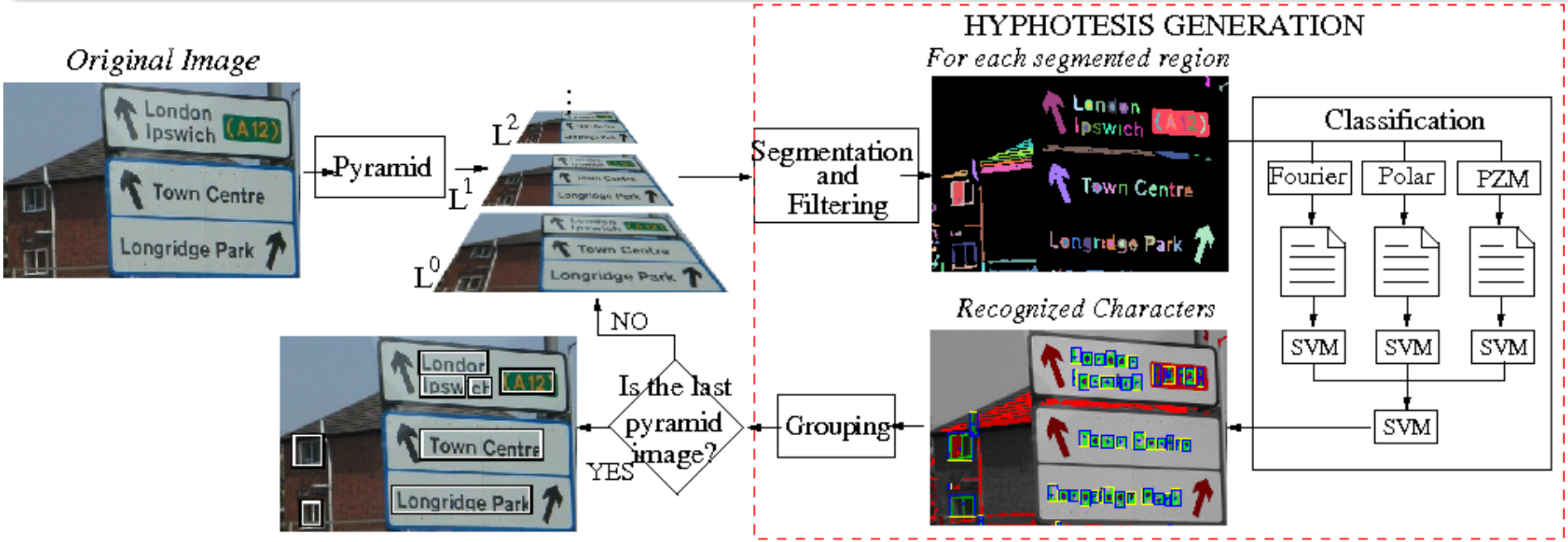
[1] S.M. Lucas. **Text Locating Competition Results**. ICDAR 2005.

[2] Boris Epshtein, Eyal Ofek and Yonatan Wexler. **Detecting Text in Natural Scenes with Stroke Width Transform**. CVPR 2010.

[3] 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 Hypotheses Generation

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



Bottom-Up Hypotheses 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

Bottom-Up Hypotheses Generation

a) Segmentation at $l = 2$

b) Segmentation at $l = 0$

c) Mono-resolution

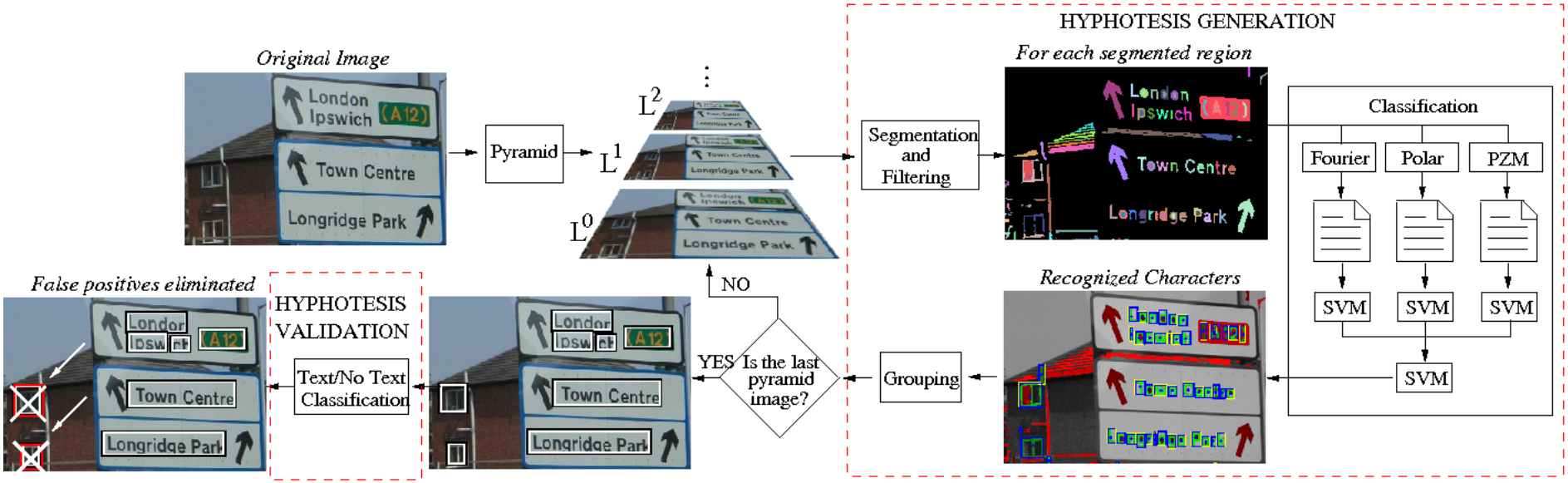
d) Multi-resolution

Bottom-Up Hypotheses Generation

Problem

- Analyzes the image content locally
  $\rightarrow$ Prone to false positives

Generation/validation process: SnooperText



- [*Minetto*2010] extends previous work [*Fabrizio*2009]
- Hybrid scheme: hypothesis generation/validation paradigm
  - Hypothesis generation: multiresolution bottom-up approach
    $\rightarrow$ improves segmentation robustness over [*Fabrizio*2009]
  - Hypothesis validation: top-down strategy
    $\rightarrow$ 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:
  $\rightarrow$ different gradient orientation distributions in the head, torso, legs, etc;

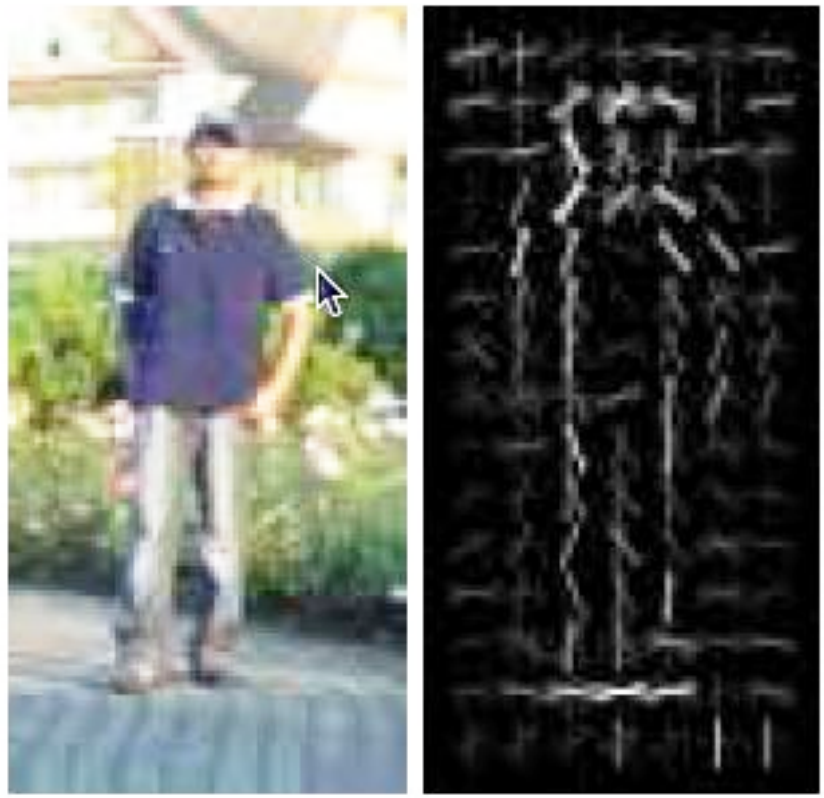
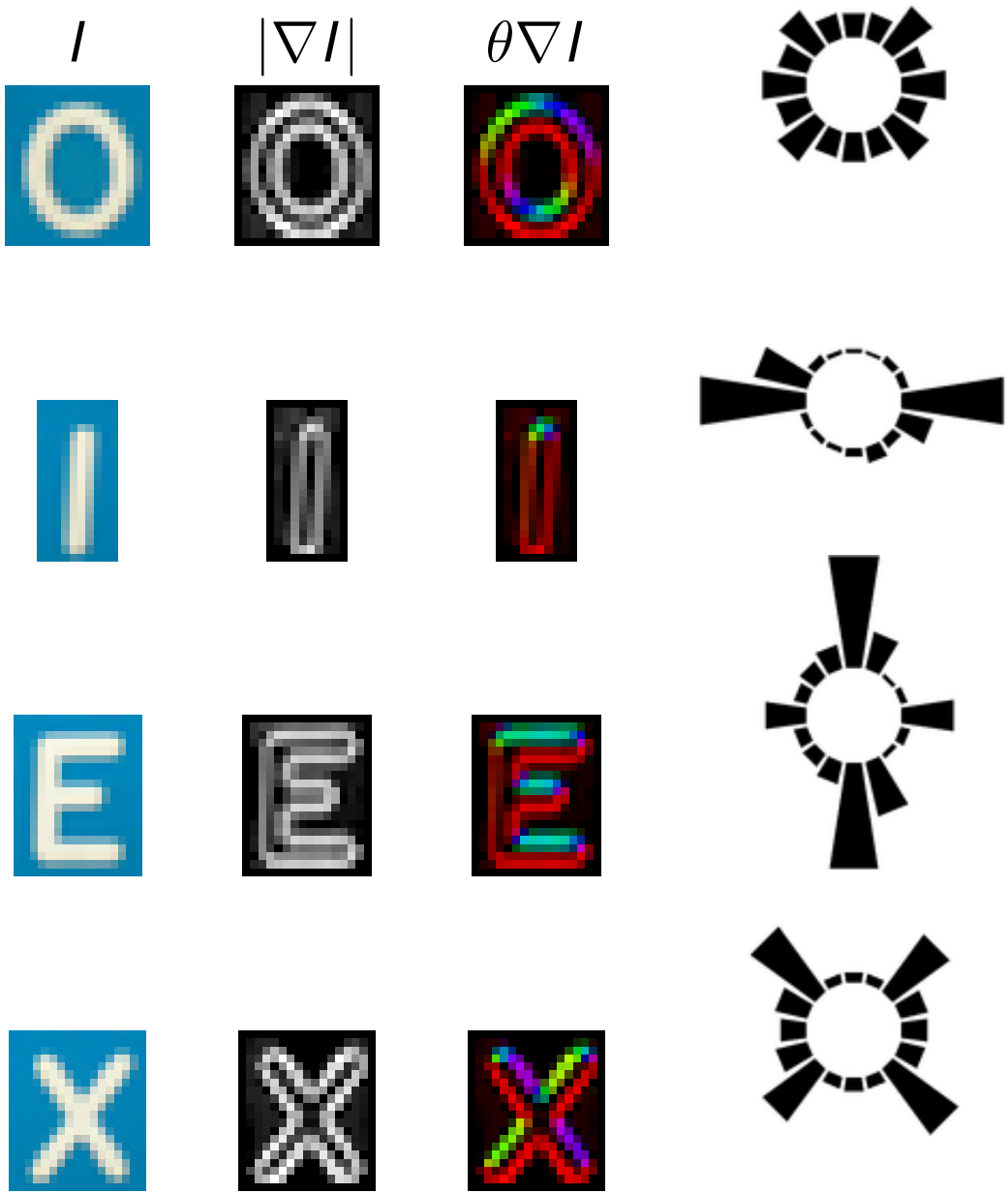


Figure: Image from: Histograms of Oriented Gradients for Human Detection. Navneet Dalal and Bill Triggs. CVPR 2004

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.

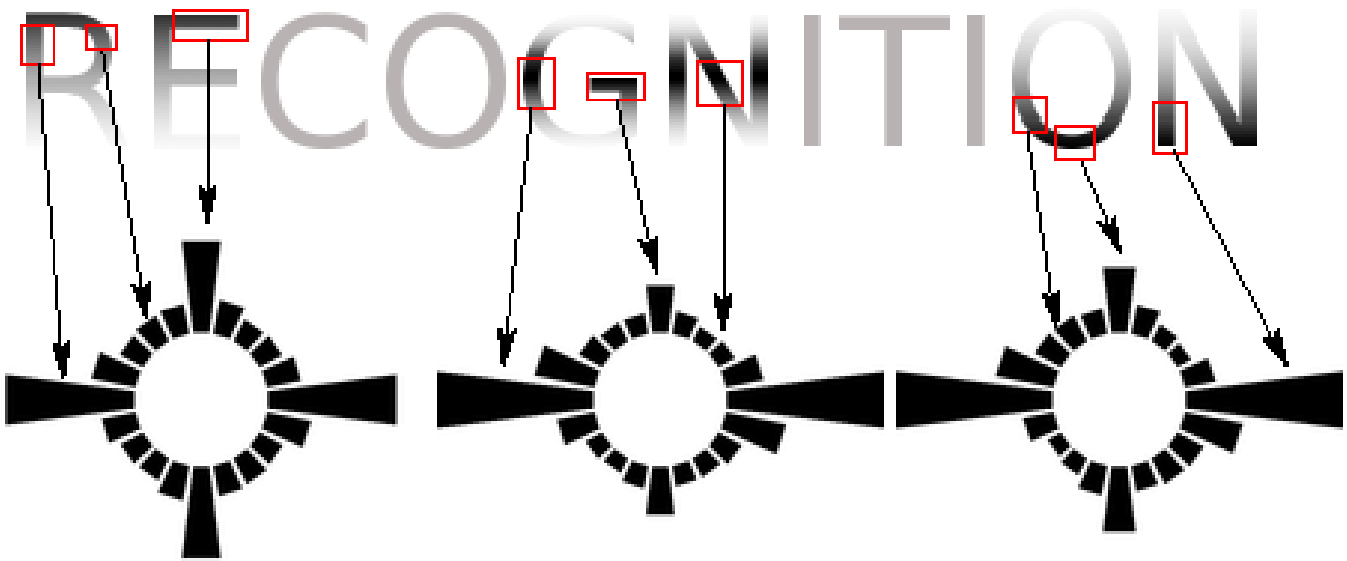
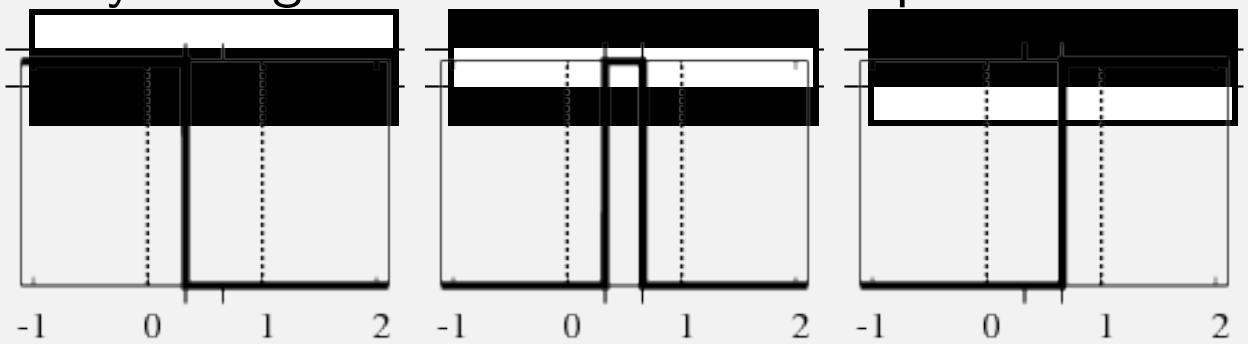


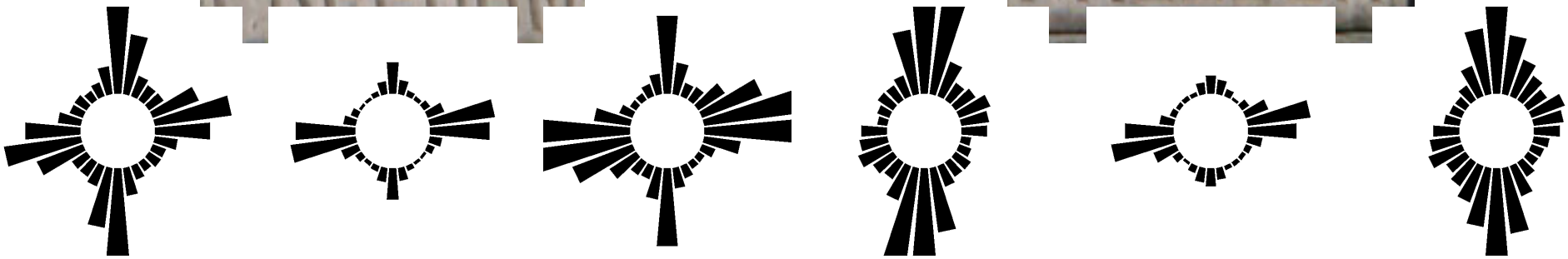
Figure: Top, middle and bottom HOGs for the text "RECOGNITION". The arrows show the contribution of specific letters strokes to the final descriptor.

Sharp cells

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

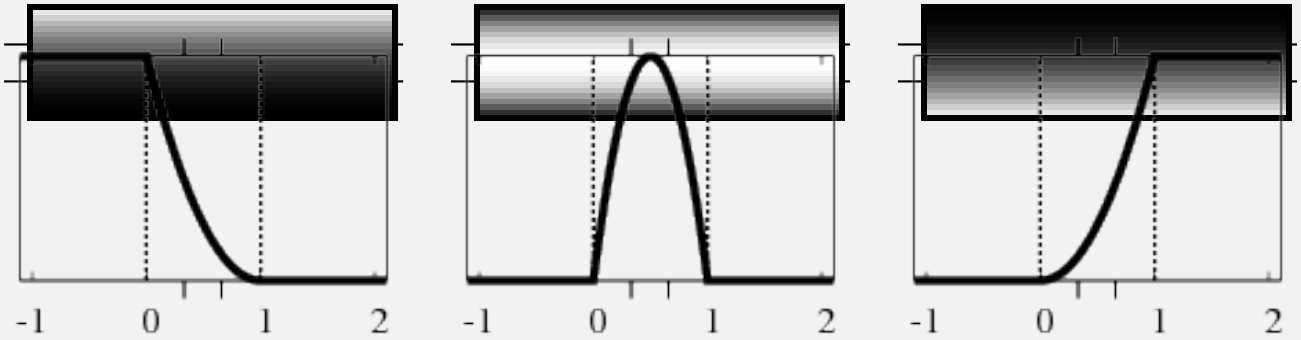


w_0 w_1 w_2

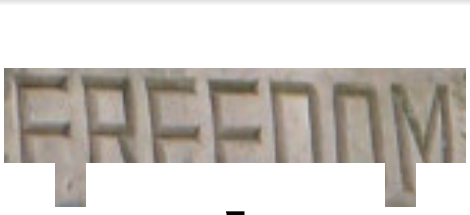


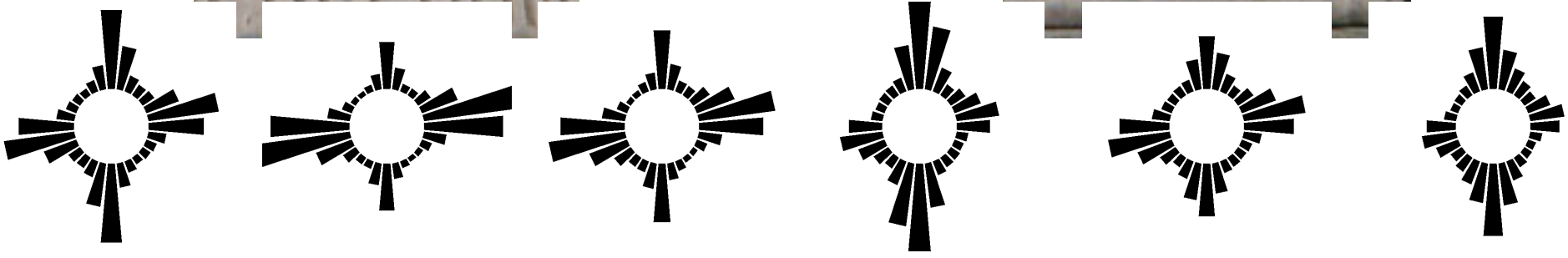
Fuzzy cells

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



w_0 w_1 w_2





Dalal et al. masks to human recognition

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

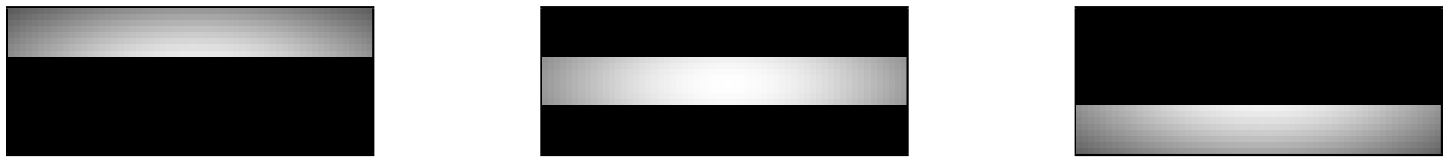


Figure: Weight functions for a single block of 1×3 cells ($\sigma_x = W/2$, $\sigma_y = H/2$).

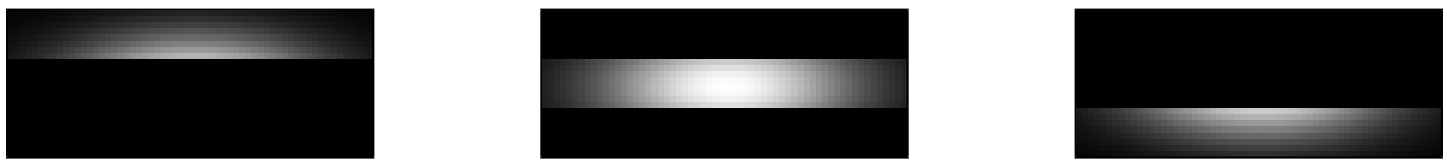


Figure: Weight functions for a single block of 1×3 cells ($\sigma_x = W/4$, $\sigma_y = H/4$).

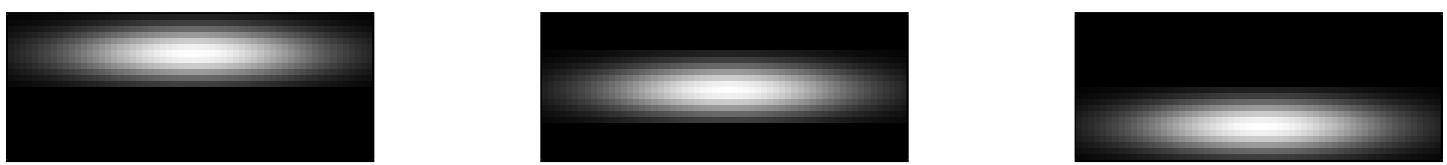
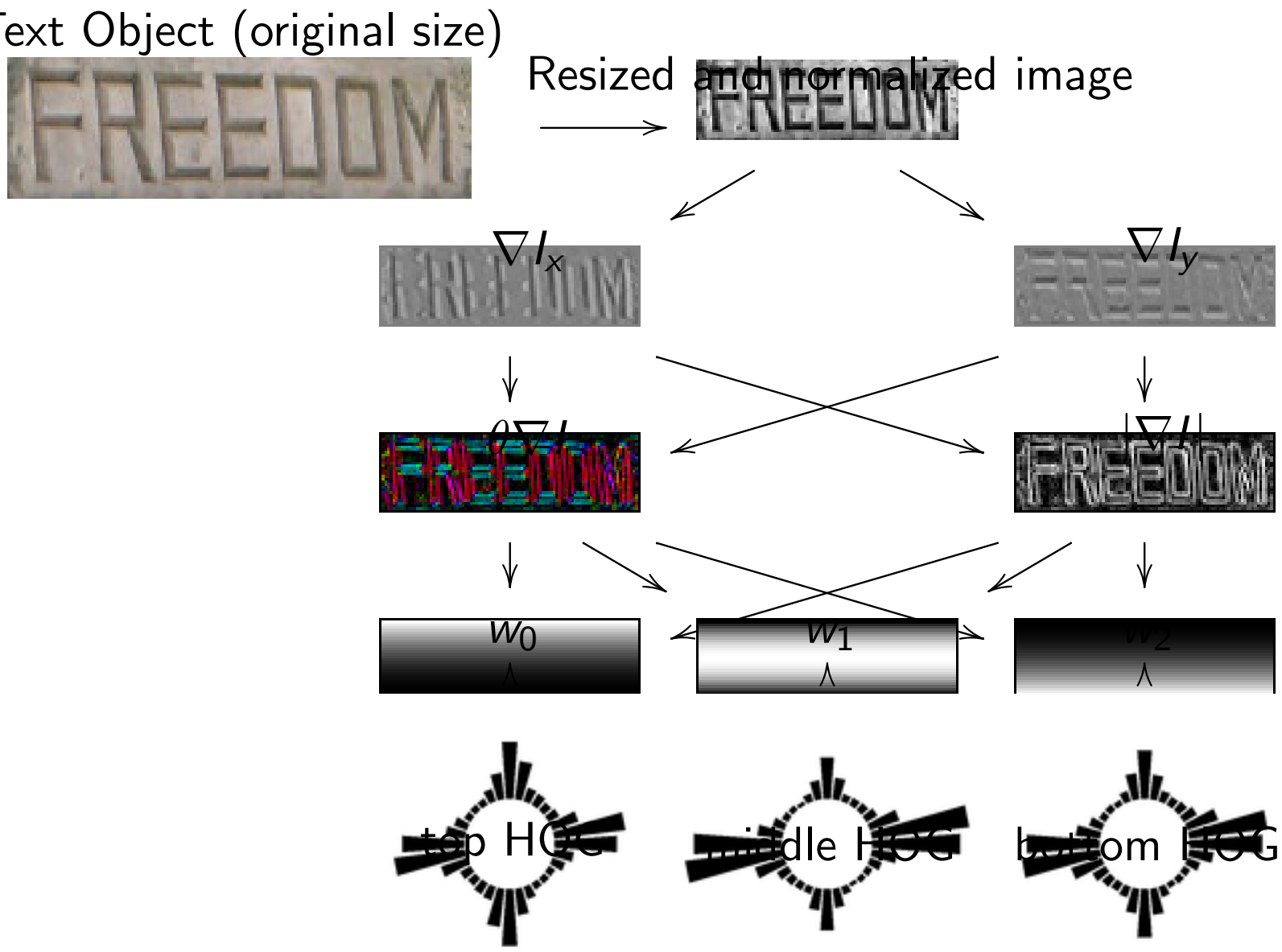
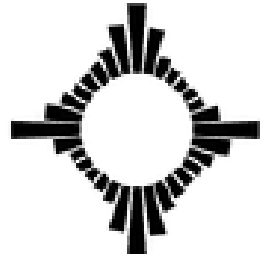


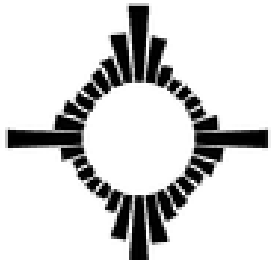
Figure: Weight functions for 1×3 single-cell blocks. Each with height $H/2$ and overlapped with stride $H/4$ ($\sigma_x = W/4$, $\sigma_y = H/8$).

Text HOG descriptor scheme

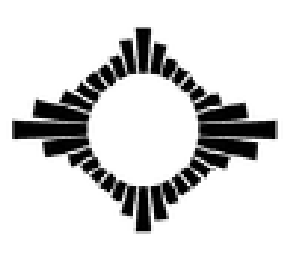


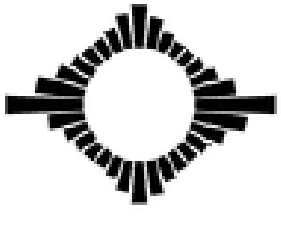
F-HOG of text and non-text regions



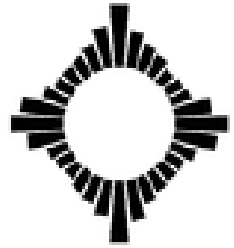


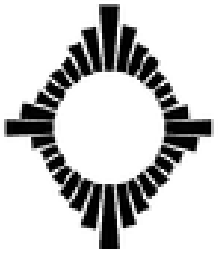
top



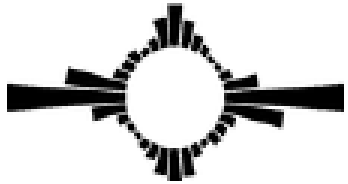


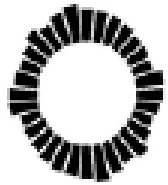
middle



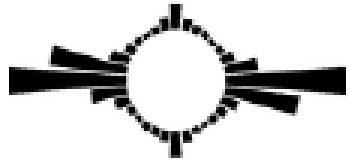


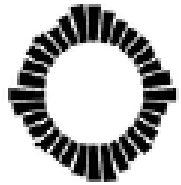
bottom



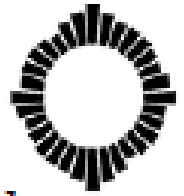


top





middle

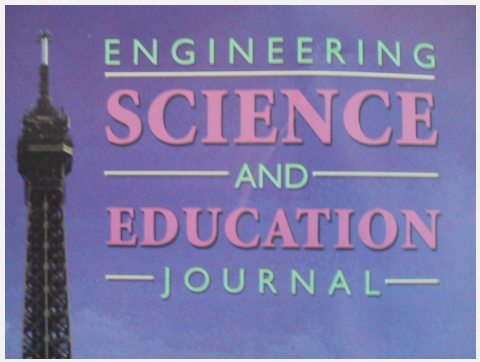


bottom

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

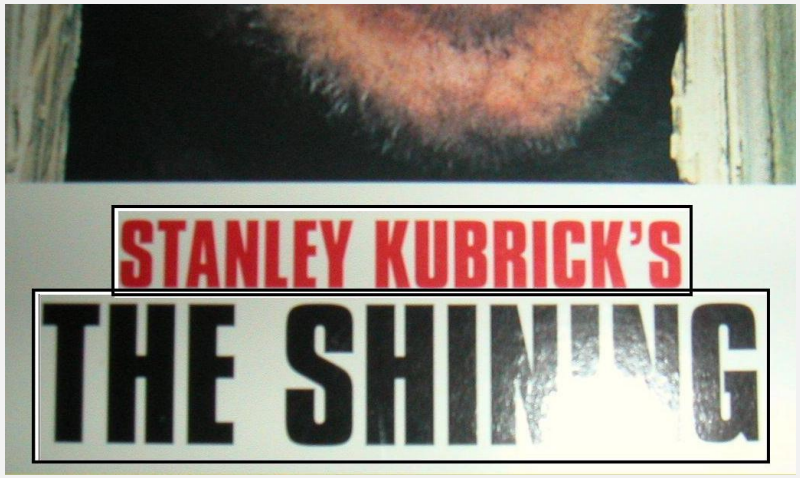
ICDAR

Successfull detections

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

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

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

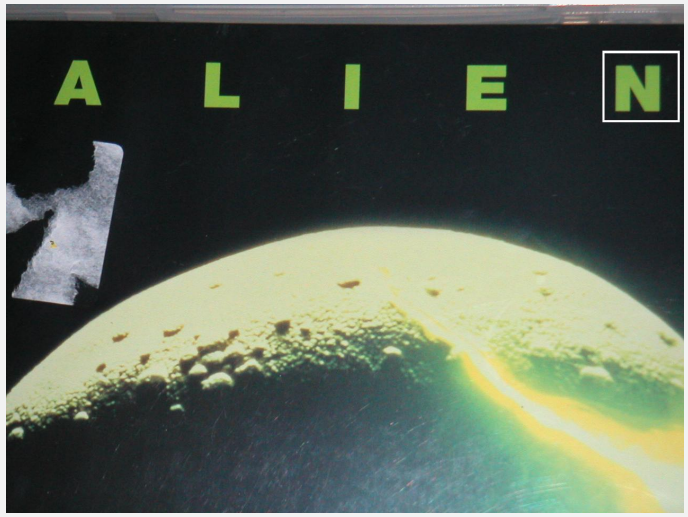


$p = 0.68, r = 0.56, f = 0.61$

ICDAR

Failures

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



$p = 0.17, r = 0.17, f = 0.17$

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

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

Performances

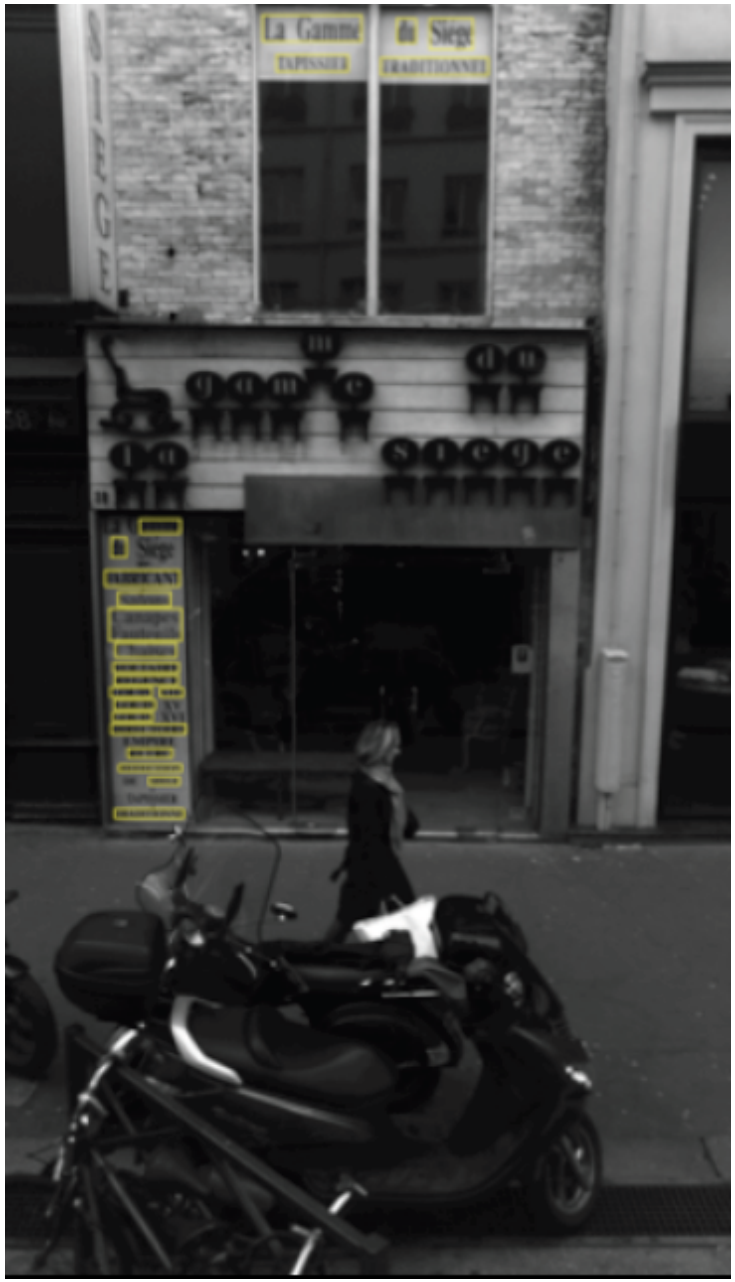
- ICDAR metrics;
- Text Detection + F-HOG: precision improvement of 23%

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

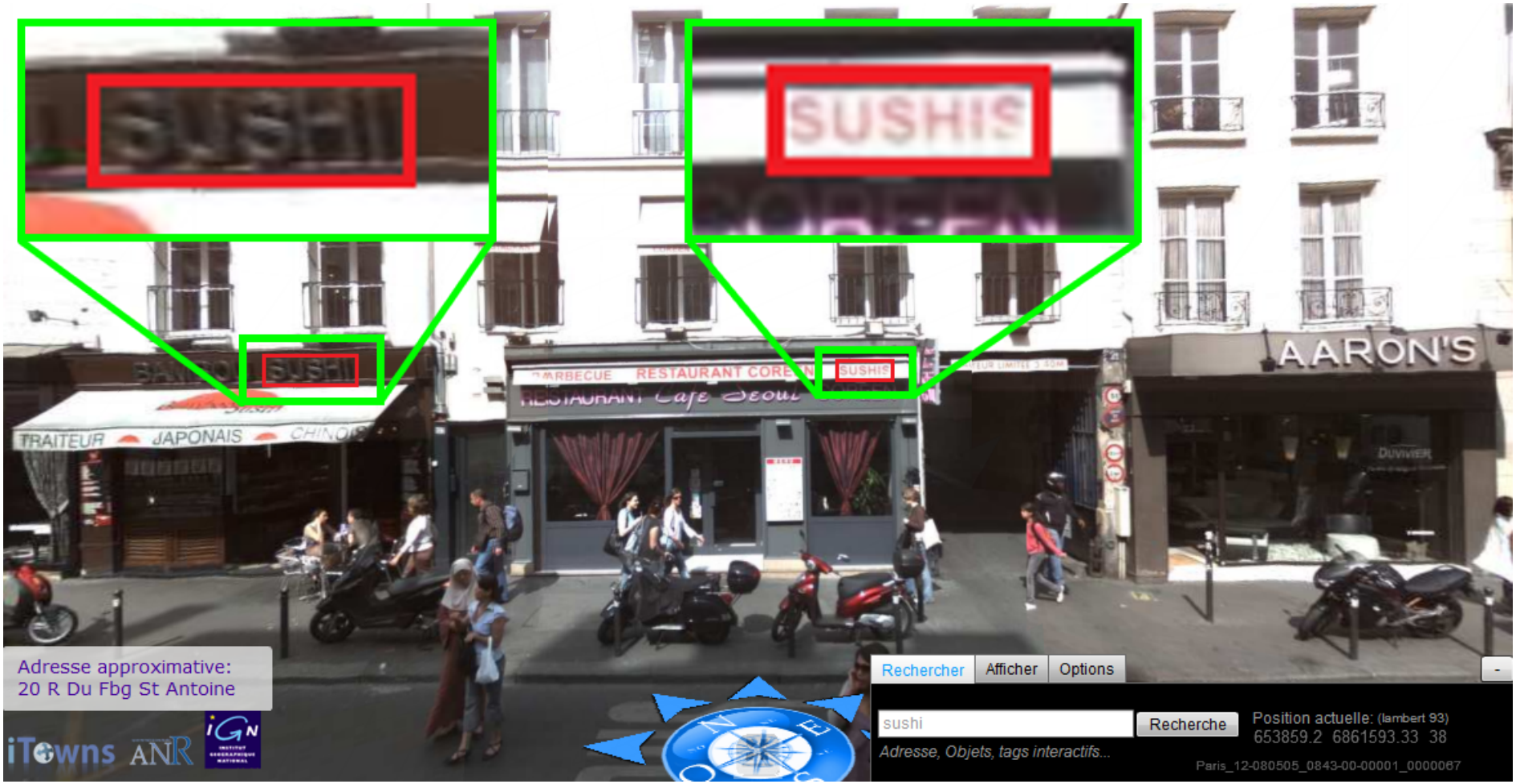
Jonathan Guyomard, Frederic Precioso, Nicolas Thome, Matthieu Cord

Text detection + OCR (Tesseract)

- Textual query to image retrieval;
- World matching by Edit distance.

itowns KeyWord Search

itowns KeyWord Search



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
	

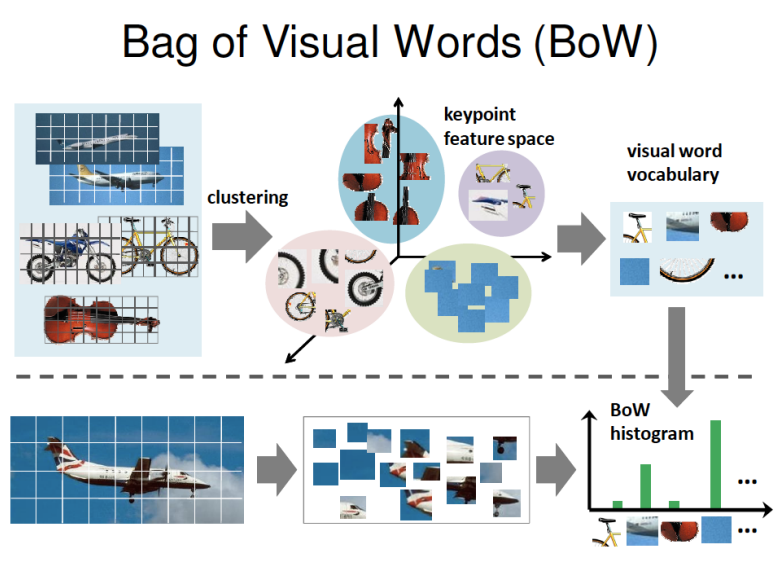
Loading ./images/text/trackingtext.avi

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

Outline

Image classification: state-of-the-art

State of the art Model: Bag-of-words (BOW)



Credit: Prof. Shih-Fu Chang

Some recent improvements:

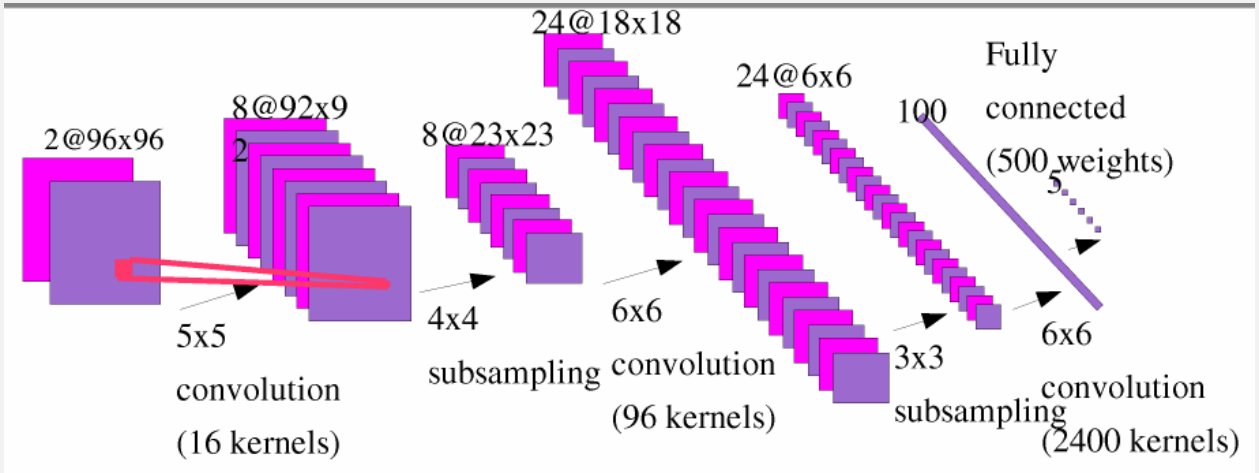
- Spatial Information [*Lazebnik*06]
- Finer coding of local descriptors: soft-assignement, sparse coding [*Wang*10]
- sum *v.s.* max pooling [*Boureau*10]
- gives state of the art classification performances

[*Lazebnik*06] P.Lazebnik.S, Schmid.C. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories CVPR2006.
[*Boureau*10]Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition CVPR2010.
[*Wang*10]J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong.Locality-constrained linear coding for image classification CVPR2010.

Image classification: state-of-the-art

Deep Networks

- Convolutional networks: [LeCun98], improvements [Jarrett09, Lee09]

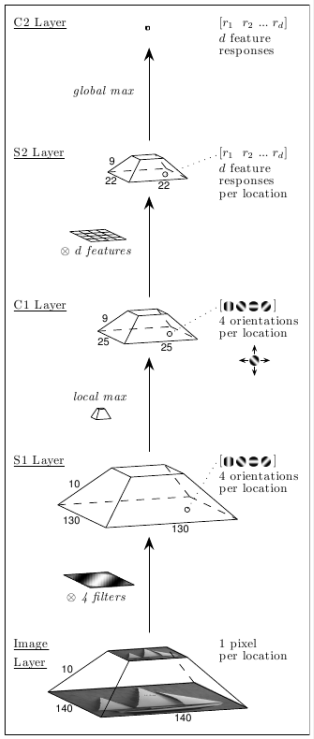


- $\oplus$ Deep models
- $\ominus$ Learning: training lower layers hard

[*Jarrett*09]K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun.What is the best multi-stage architecture for object recognition? In Proc ICCV2009.
[*Lee*09]H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. ICML2009

Image classification: state-of-the-art

Biologically-inspired Methods [*Fidler*08, *Serre*07, *Mutch*08]

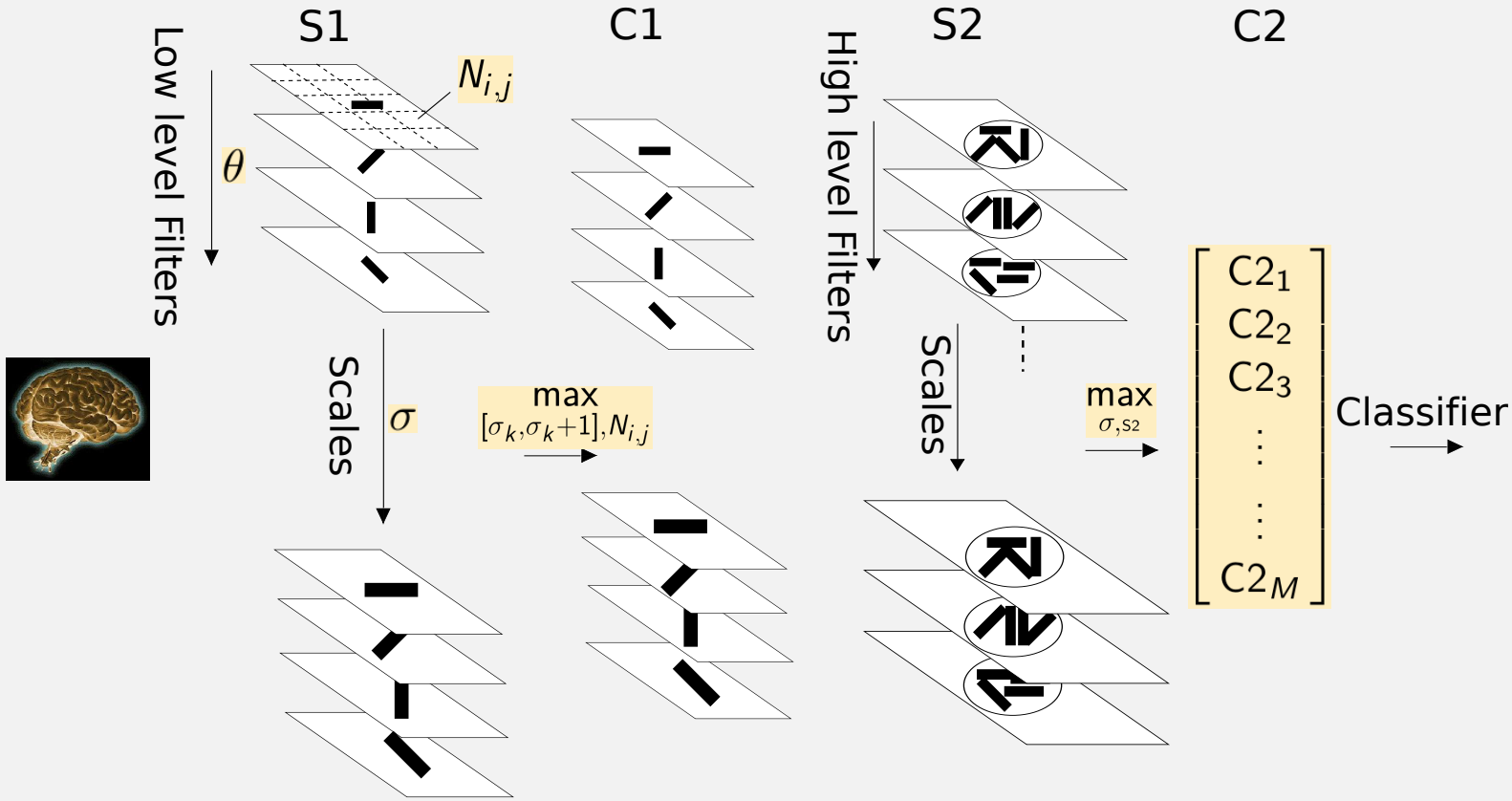


- Mimics feedforward properties of primate visual cortex V1 simple cells
- Based on the HMAX model [*Serre*07, *Mutch*08]
  - $\oplus$ Deep models
  - $\oplus$ Trainable with real images

[*Fidler*08]S. Fidler, B. Boben, and A. Leonardis. Similarity-based cross-layered hierarchical representation for object categorization CVPR2008.
[*Serre*07] T. Serre, *et.al*, Robust object recognition with cortex-like mechanisms, PAMI, 2007.
[*Mutch*08] Mutch.J and Lowe.D.G, Object class recognition and localization using sparse features with limited receptive fields, IJCV, 2008

Bio framework

HMAX-like architecture [*Serre*07, *Mutch*08]

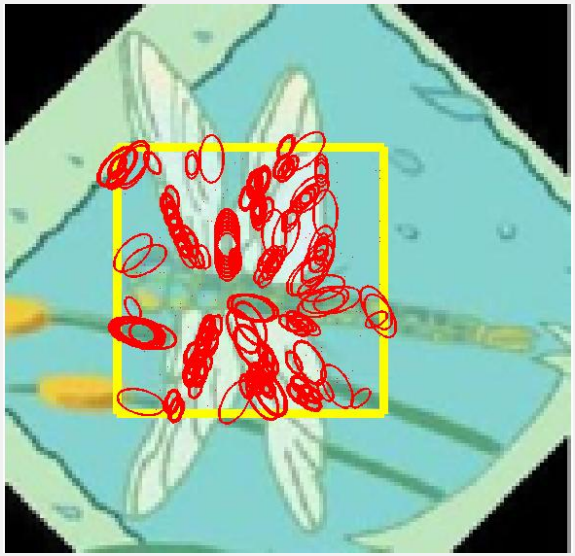


Max over positions and scales $\Longrightarrow$ Position and scale invariance

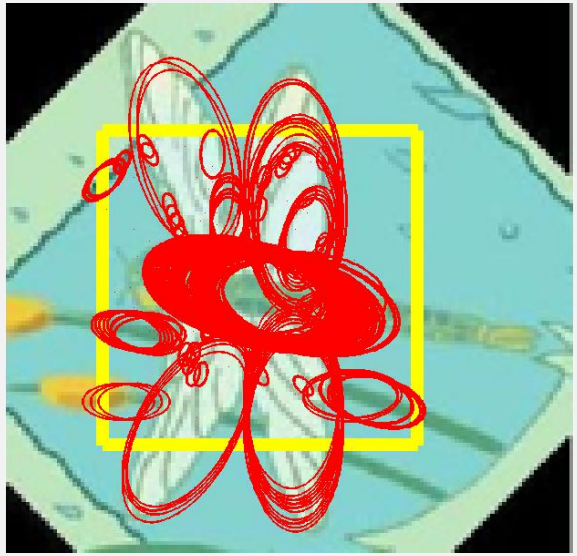
[*Serre*07] T. Serre, *et.al*, Robust object recognition with cortex-like mechanisms, PAMI, 2007.
[*Mutch*08] Mutch.J and Lowe.D.G, Object class recognition and localization using sparse features with limited receptive fields,

Contribution: Deep multiscale high level filters

- High level filters (yellow) should fit complex visual structures with multiple local scales (red)



Single scale [*Mutch*08]



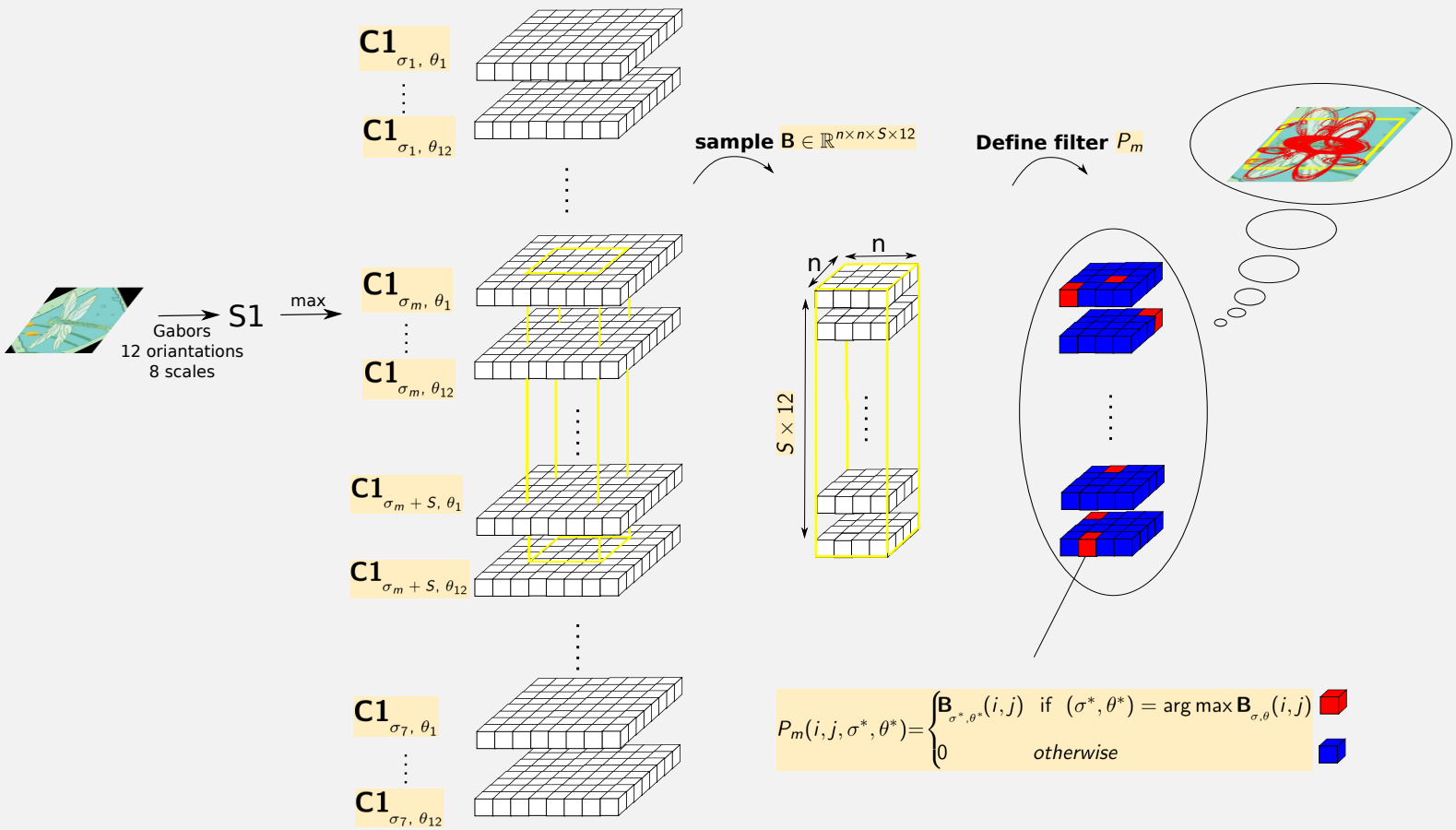
Our multiscale high level filter

Joint work with Christian Theriault (Post Doc) and Nicolas Thome LIP6

Proposed model

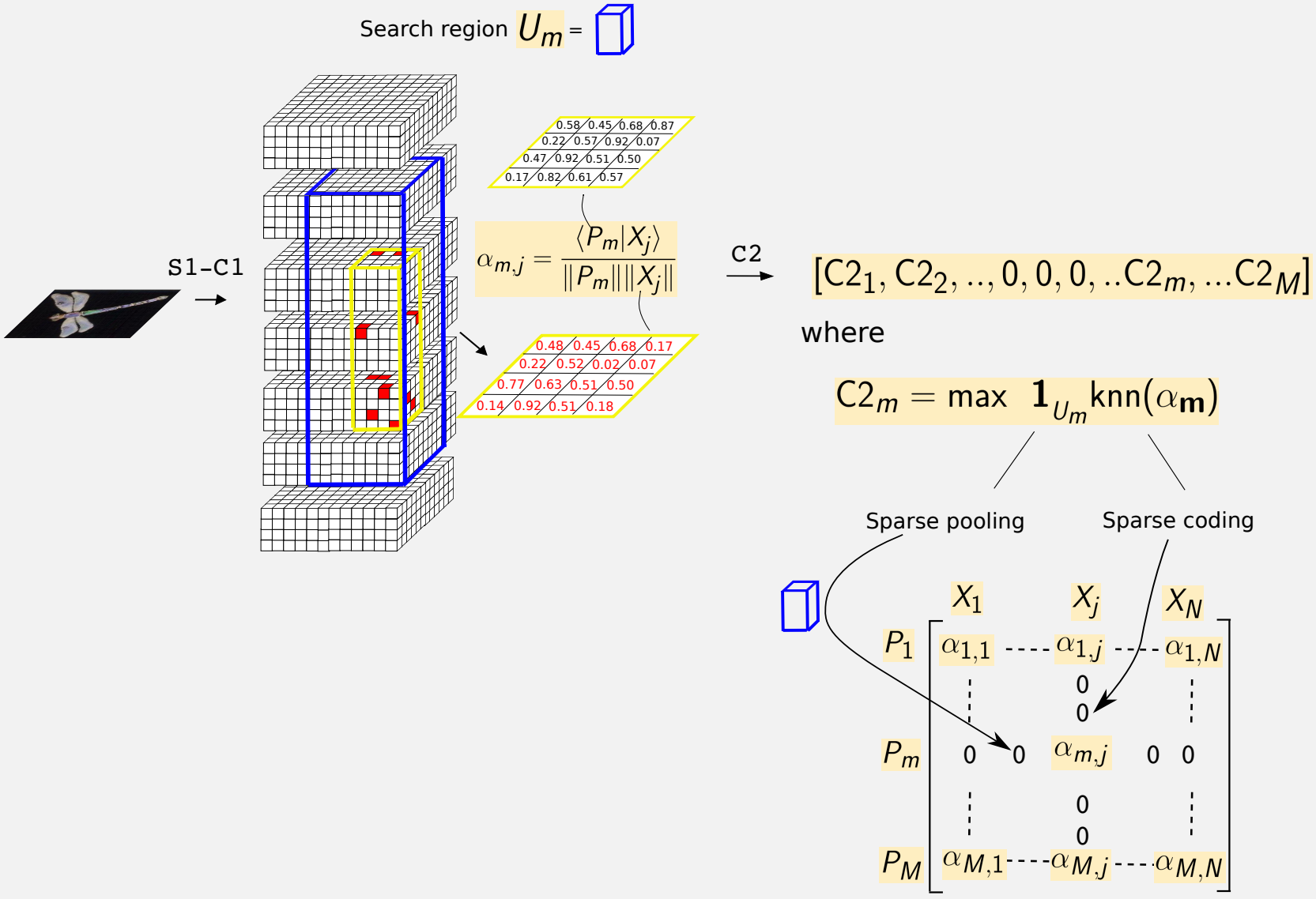
Learning of deep multiscale high level filters

Learning deep scale filters



Proposed model

Final Image Signature



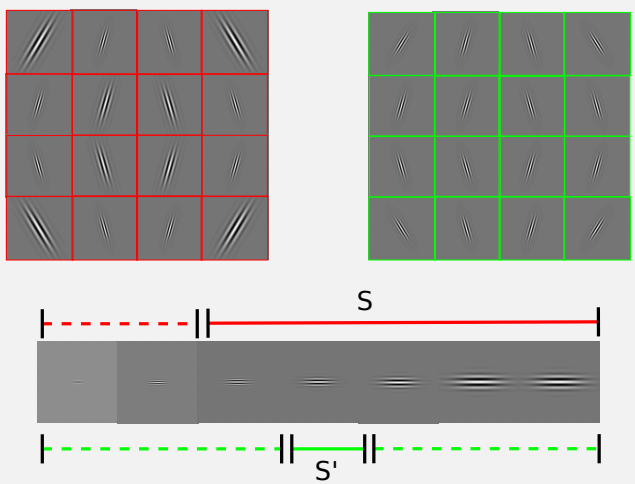
Proposed model

Main parametrization

The scale depth S of high level filter

$$S \uparrow \Longrightarrow \Delta \downarrow \quad \text{(More fitted, less scale invariance)}$$

$$S \downarrow \Longrightarrow \Delta \uparrow \quad \text{(Less fitted, more scale invariance)}$$

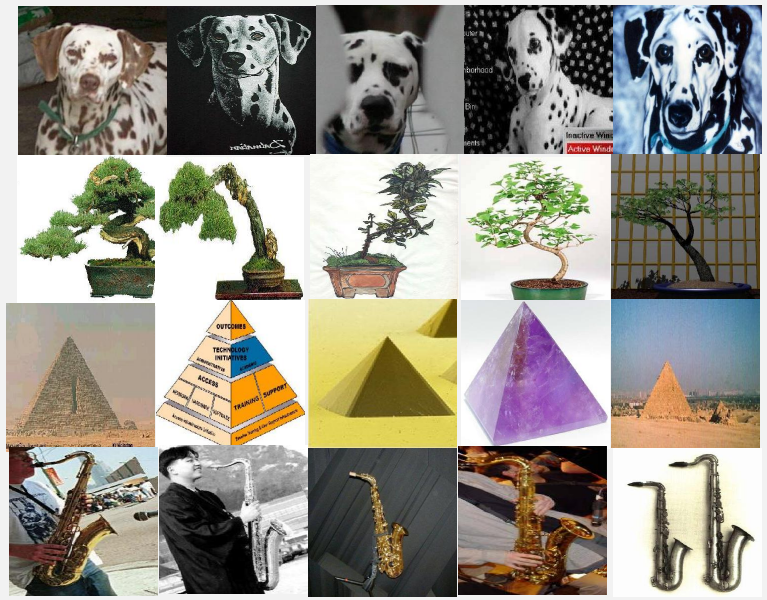


$\Longrightarrow$ Learn filters with different values of $S \in \{1, 2, 3, 4, 5, 6, 7\}$ to get both invariance and discrimination

Experiments

Dataset: Caltech101

- 9,144 images
- 102 categories
- 30 to 800 images per categorie

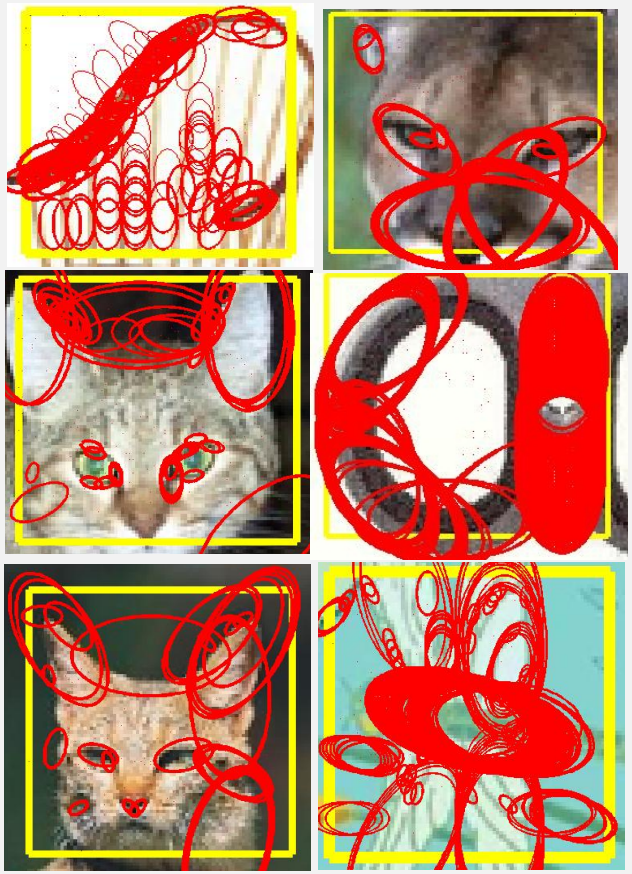


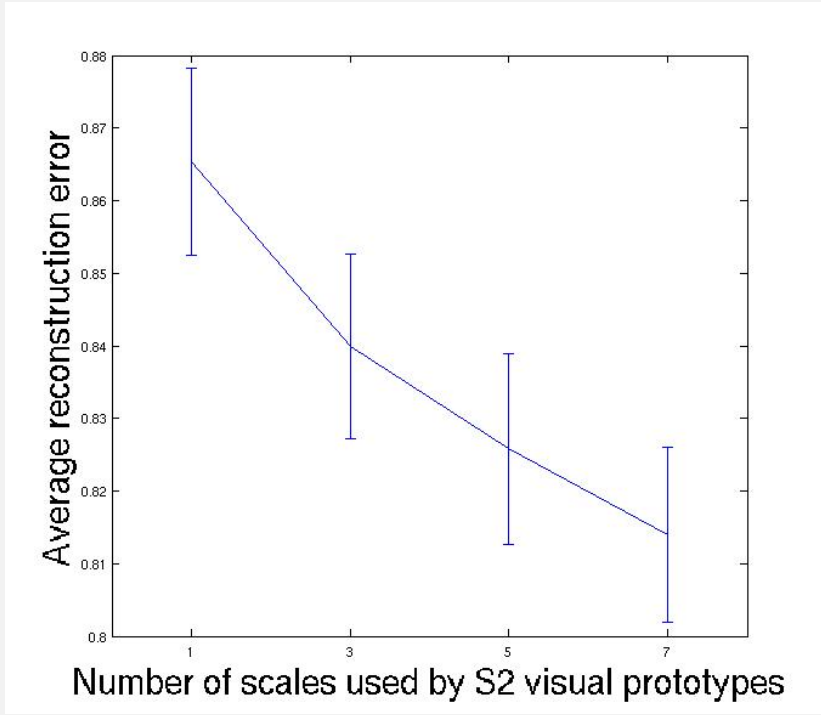
- Standard evaluation protocol for baseline comparison:
  - Train with 15-30 images / class
  - Test on the reamaining images
  - Metric: Multi-class Accuracy

Experiments

Caltech 101: Deep multiscale high level filters

Deep multiscale high level filters better fit visual content





Classification results

Average accuracy results

Setup:

- 8 scales, 12 orientations
- 4080 S2 filters
- One-Against-All gaussian SVM

Model	S	15 images	30 images
[*Mutch*08]	1	48	54
Our model	3	56.4%	62.5%
Our model	7	55.6 %	62.1%
Our model	1-7	**59.1 ± 0.2 %**	**66.9 ± 0.8%**

Average accuracy results

Model	15 images	30 images
Our model	**59.1 ± 0.2 %**	**66.9 ± 0.8%**

Deep biologically inspired architectures		
[*Mutch*08]	48	54
[Ranzato&alCVPR07]	-	54
[Kavukcuoglu&alNIPS10]	-	66.3
[*Mutch*09]	57.7	64.4
[*Jarrett*09]	-	65.6
[Zeiler&alCVPR2010]	58.6	66.9
[*Fidler*08]	60.5	66.5
Shallow architectures		
[*Lazebnik*06]	56.4	64.6
[Zhang&alCVPR06]	59.1	62.2
[*Wang*10]	64.43	73.44

Integrate sparse code learning into the architecture

- We proposed an extension of the HMAX model by learning high level filters with deep scale range
- Our classification results indicates that these more fitted filters, combined with more invariant shallow filters, increase classification scores by nearly 12%
- Our next step is to include filter learning with sparse constraints: Fergus (2010), Olshausen (1996,2009), Ranzato(2007), Kavukcuoglu&Lecun (2010)

Thank you for your attention !

QUESTIONS ?

People

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

- PhD students: Sandra Avila, Rodrigo Minetto, Hanlin Goh, Mar Law, Denis Pitzalis
- Post-Docs: Christian Theriault
- Research Inge. J. Guyomard, C. Sureda

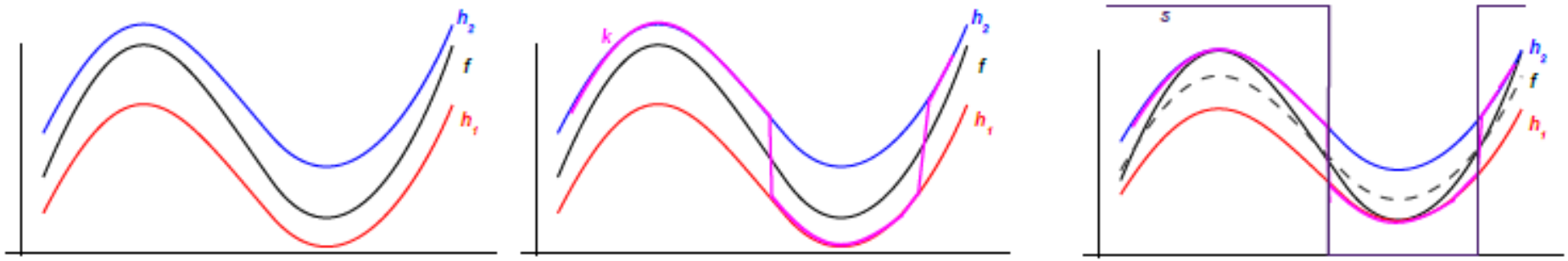
http://webia.lip6.fr/∼cord/

UPMC
SORBONNE UNIVERSITÉS

Text Detection: SnooperText

Hypothesis generation: Segmentation

- *Toggle Mapping* [*Serra*89]: morphological operator
- Efficient approach for segmenting characters

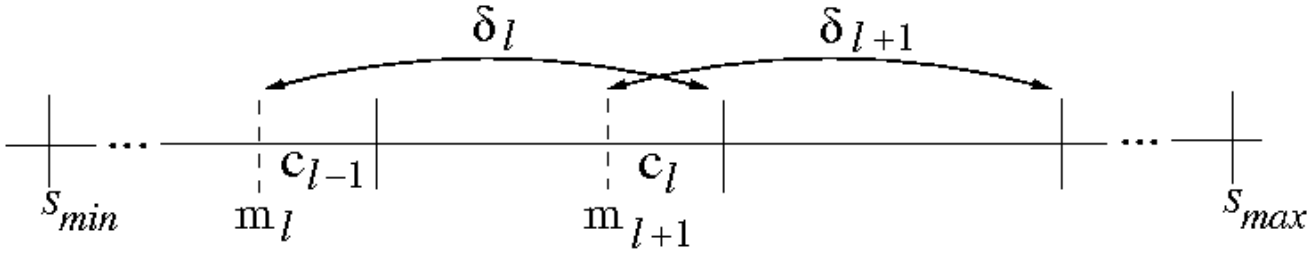


[*Serra*89] Jean Serra, Toggle mappings, From pixels to features,pp. 61-72, 1989, J.C. Simon (ed.), North-Holland, Elsevier.

SnooperText: Multi-Resolution Segmentation

- Difficult context for segmentation:
  $\rightarrow$ very small relevant regions
  $\rightarrow$ large textured text regions
- Multi-resolution goal:
  - Coarser levels: detect large regions and ignoring texture details (high frequencies)
  - Finer levels: detect smaller regions (analyzing accurately the local image content)

Region sizes managed at different resolution levels:

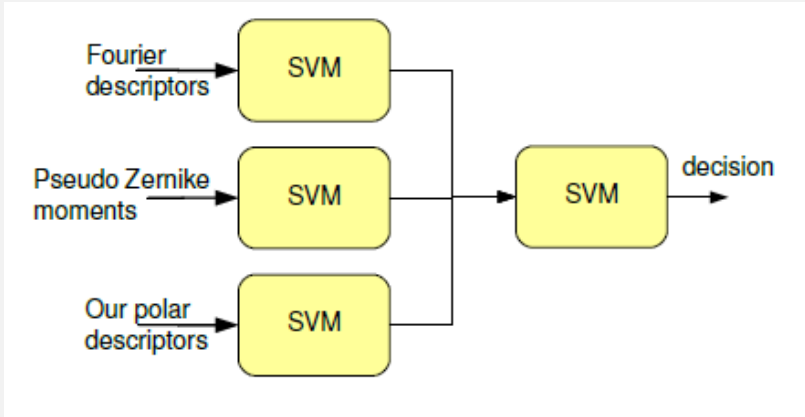


at level l, detecting text regions with size $s_l \in [m_l; m_l + \delta_l]$ (c_l : overlap)

SnooperText: Character extraction & grouping

Character extraction

- Each region described by 3 shape descriptors:
  - Pzeudo-Zernike Moments (PZM)
  - Fourier Descriptors
  - Polar Descriptor [*Fabrizzio*2010]
- Late fusion: Hierachical SVM classifier



Character grouping

- Each character is merged with neighboring characters
  - Constraints related to distance, relative size, *etc* (see [*Retornaz*07])

[*Fabrizzio*2010] J. Fabrizio, M. Cord, B. Marcotegui, Text extraction from street level images, ISPRS Workshop CMRT, 2009
[*Retornaz*07] Thomas Retornaz and Beatriz Marcotegui, Scene text localization based on the ultimate opening, ISMM, vol.1, pp. 177-188, 2007.