Detection and representation for image understanding

Matthieu Cord

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

October, 2011
Outline

1. Introduction
   - Multimedia Group
   - Projets

2. Text Detection in Natural Images
   - Motivation
   - Related Works
   - Hypotheses Generation
   - SnooperText
   - Hypothesis Validation
   - Experiments

3. Biologically inspired image classification
   - Image classification: state-of-the-art
   - Proposed model
   - Results
   - Conclusion
LIP6/DAPA/MALIRE

People

1. LIP6 lab in Paris
   - ~ 200 permanent researchers, ~ 250 Phd students

2. DAPA department: Databases and Machine learning
   - ~ 35 permanent researchers, ~ 50 Phd students

3. MALIRE team: MAchine Learning and Information REtrieval
   - ~ 20 permanent researchers, ~ 35 Phd students, ~ 10 Post-docs

4. Our research group
   - 2 permanent researchers (M. Cord, N. Thome), ~ 10 Phd/Post-docs

Matthieu.cord@lip6.fr
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

1. Introduction
   - Multimedia Group
   - Projets

2. Text Detection in Natural Images

3. Biologically inspired image classification
Main Projects

**Retrieval, object detection**

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

Retrieval, object detection

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

[Diagram of ITowns system with labels: Query in Paris, Itowns server, Itowns DB]
Main Projects

Retrieval, object detection

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

Retrieval, object detection

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

![Image of street scene with shops and signs]
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 → same versionning
**Main Projects**

### SCAPE: SCAlable Preservation Environments

**January Guarantee**
This guarantee will be backed by £40m of additional investment from the Department and provides an excellent opportunity to help these young people to re-engage quickly in positive and productive learning, and reduce the risk of long term disengagement.

**More detail about the January Guarantee will be delivered in due course.**

**Gateway 4: Guidance for Consortia**
We have now published guidance for consortia. This guides on the operational role of consortia and is for those applying to deliver new lines from 2011. It provides:
- information on the current context
- developments since the last Gateway
- how to apply for new lines

**RPA Moving Up day**
Moving Up day was successfully held on 6 December 2009. The purpose of the day was to celebrate the first anniversary of the passing of the Education and Skills Act, in which all young people will be required to continue in education or training until they are 18 by 2015.

**Quality Assurance System: overarching framework**
The Quality Assurance System (QAS) was published in 31 March 2010.

**Metrics and kernel learning (MS student + PhD)**

## Significant change → generate a new versioning

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

1. Introduction

2. Text Detection in Natural Images
   - Motivation
   - Related Works
   - Hypotheses Generation
   - SnooperText
   - Hypothesis Validation
   - Experiments

3. Biologically inspired image classification
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
Text Detection in Natural Images

Motivation

Raw image

OCR (Tesseract)

Matthieu.cord@lip6.fr

Detection and representation for image understanding
Motivation

Text detection

Text regions

OCR (Tesseract)

LE RALLY
SALLE
BAR TABAC LE RALLY
un ã
>: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)


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
  → Prone to false positives
Generation/validation process: SnooperText

[Minetto2010] extends previous work [Fabrizio2009]

Hybrid scheme: hypothesis generation/validation paradigm

- Hypothesis generation: multiresolution bottom-up approach
  → improves segmentation robustness over [Fabrizio2009]
- Hypothesis validation: top-down strategy
  → To remove false positives by analyzing globally the window content


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:
  → 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

$\nabla I$, $|\nabla I|$, $\theta \nabla I$
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 concatenation 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:

→ HOG may change with small vertical displacements

\[ w_0 \quad w_1 \quad w_2 \]
Fuzzy cells

To avoid this problem, we used “fuzzy” cells:

\[ w_0 \quad w_1 \quad w_2 \]
Dalal et al. masks to human recognition

- Gaussian weight functions:
  → Problem: Sharp boundaries.

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

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

**Figure:** Weight functions for $1 \times 3$ 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

Text Object (original size) → Resized and normalized image

$\nabla I_x$ $\nabla I_y$

$\theta \nabla I$

$|\nabla I|$

$w_0$ $w_1$ $w_2$

top HOG middle HOG bottom HOG
F-HOG of text and non-text regions
ICDAR

**Dataset**

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

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

\[
\begin{align*}
p &= 0.00, r = 0.00, f = 0.00 \\
p &= 0.17, r = 0.17, f = 0.17 \\
p &= 0.00, r = 0.00, f = 0.00 \\
p &= 0.71, r = 0.95, f = 0.81
\end{align*}
\]
iTowns

Performances

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

<table>
<thead>
<tr>
<th>System</th>
<th>Precision (p)</th>
<th>Recall (r)</th>
<th>f</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our System</td>
<td>0.69</td>
<td>0.49</td>
<td>0.55</td>
</tr>
<tr>
<td>SnooperText (ICIP 2010)</td>
<td>0.46</td>
<td>0.49</td>
<td>0.48</td>
</tr>
</tbody>
</table>
iTowns - Detection results
iTowns - Detection results
iTowns - Detection results
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 ~ 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

Outline

1 Introduction

2 Text Detection in Natural Images

3 Biologically inspired image classification
   - Image classification: state-of-the-art
   - Proposed model
   - Results
   - Conclusion
Biologically inspired image classification

Image classification: state-of-the-art

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

Some recent improvements:

- Spatial Information [Lazebnik06]
- Finer coding of local descriptors: soft-assignment, sparse coding [Wang10]
- sum v.s. max pooling [Boureau10]
- gives state of the art classification performances

Credit: Prof. Shih-Fu Chang

Deep Networks

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


Biologically inspired Methods [Fidler08, Serre07, Mutch08]

- Mimics feedforward properties of primate visual cortex V1 simple cells
- Based on the HMAX model [Serre07, Mutch08]
  - ⊕ Deep models
  - ⊕ Trainable with real images

Bio framework

HMAX-like architecture [Serre07, Mutch08]

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

Contribution: Deep multiscale high level filters

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

Single scale \([Mutch08]\)  

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

\[ C_{\sigma, \theta} \]

\[ \sigma_m, \theta_1 \]

\[ \sigma_m + S, \theta_1 \]

\[ \sigma_{12}, \theta_1 \]

Sample \( B \in \mathbb{R}^{n \times S \times 12} \)

Define filter \( P_m \)

\[ P_m(i, j, \sigma^*, \theta^*) = \begin{cases} B_{\sigma^*, \theta^*}(i, j) & \text{if } (\sigma^*, \theta^*) = \arg \max B_{\sigma, \theta}(i, j) \\ 0 & \text{otherwise} \end{cases} \]
Proposed model

Final Image Signature

Search region \( U_m = \)

\[
\alpha_{m,j} = \frac{(P_m|X_j)}{\|P_m\| \|X_j\|}
\]

\[
[C_{21}, C_{22}, \ldots, 0, 0, 0, \ldots C_{2m}, \ldots C_{2M}]
\]

where

\[
C_{2m} = \max \ 1_{U_m} \knn(\alpha_m)
\]

Sparse pooling

Sparse coding

\[
\begin{bmatrix}
X_1 & X_j & X_N \\
\alpha_{1,1} & \alpha_{1,j} & - & \alpha_{1,N} \\
\vdots & & & \ddots & \vdots \\
0 & 0 & \alpha_{m,j} & 0 & 0 \\
\vdots & & & & \ddots & \vdots \\
0 & 0 & 0 & \alpha_{M,j} & - & \alpha_{M,N}
\end{bmatrix}
\]
Proposed model

Main parametrization

The scale depth $S$ of high level filter

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

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

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 remaining images
- Metric: Multi-class Accuracy
Experiments

Caltech 101: Deep multiscale high level filters

Deep multiscale high level filters better fit visual content

![Graph depicting average reconstruction error versus number of scales used by S2 visual prototypes.](image-url)
Classification results

Average accuracy results

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

<table>
<thead>
<tr>
<th>Model</th>
<th>$S$</th>
<th>15 images</th>
<th>30 images</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Mutch08]</td>
<td>1</td>
<td>48</td>
<td>54</td>
</tr>
<tr>
<td>Our model</td>
<td>3</td>
<td>56.4%</td>
<td>62.5%</td>
</tr>
<tr>
<td>Our model</td>
<td>7</td>
<td>55.6%</td>
<td>62.1%</td>
</tr>
<tr>
<td>Our model</td>
<td>1-7</td>
<td>59.1 ± 0.2%</td>
<td>66.9 ± 0.8%</td>
</tr>
</tbody>
</table>
# Average accuracy results

<table>
<thead>
<tr>
<th>Model</th>
<th>15 images</th>
<th>30 images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our model</td>
<td>59.1 ± 0.2 %</td>
<td>66.9 ± 0.8 %</td>
</tr>
</tbody>
</table>

## Deep biologically inspired architectures

<table>
<thead>
<tr>
<th>[Mutch08]</th>
<th>48</th>
<th>54</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Ranzato&amp;alCVPR07]</td>
<td>-</td>
<td>54</td>
</tr>
<tr>
<td>[Kavukcuoglu&amp;alNIPS10]</td>
<td>-</td>
<td>66.3</td>
</tr>
<tr>
<td>[Mutch09]</td>
<td>57.7</td>
<td>64.4</td>
</tr>
<tr>
<td>[Jarrett09]</td>
<td>-</td>
<td>65.6</td>
</tr>
<tr>
<td>[Zeiler&amp;alCVPR2010]</td>
<td>58.6</td>
<td>66.9</td>
</tr>
<tr>
<td>[Fidler08]</td>
<td>60.5</td>
<td>66.5</td>
</tr>
</tbody>
</table>

## Shallow architectures

<table>
<thead>
<tr>
<th>[Lazebnik06]</th>
<th>56.4</th>
<th>64.6</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Zhang&amp;alCVPR06]</td>
<td>59.1</td>
<td>62.2</td>
</tr>
<tr>
<td>[Wang10]</td>
<td>64.43</td>
<td>73.44</td>
</tr>
</tbody>
</table>
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 indicate that these more fitted filters, combined with more invariant shallow filters, increase classification scores by nearly 12%.
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/
Hypothesis generation: Segmentation

- **Toggle Mapping** [Serra89]: morphological operator
- Efficient approach for segmenting characters

SnooperText: Multi-Resolution Segmentation

- Difficult context for segmentation:
  - very small relevant regions
  - 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:

\[
\begin{align*}
\delta_l & \quad \delta_{l+1} \\
\cdots & \quad \cdots \\
S_{\min} & \quad m_{l-1} \quad m_l \quad m_{l+1} \quad S_{\max}
\end{align*}
\]

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:
  - Pseudo-Zernike Moments (PZM)
  - Fourier Descriptors
  - Polar Descriptor [Fabrizio2010]
- Late fusion: Hierarchical SVM classifier

![Diagram of SVM classification process](image)

Character grouping

- Each character is merged with neighboring characters
  - Constraints related to distance, relative size, etc (see [Retornaz07])