Top-Down and Bottom-up Cues for Scene Text Recognition, A. Mishra, K. Alahari, C. V. Jawahar

Vincent BODIN & Thomas MOREAU

December 17, 2013

Introduction

Figure: Task of the project (image from SVT)

Sommaire

Learning characters

We need to build classifiers to recognize characters in a natural picture:

1. Character databases: ICDAR 2003 [1], Chars74K [2];
2. features: Histogram Of Gradient (HOG) [3][1];
3. build K SVMs ($K = 62$) with RBF Kernel, Fig.(2):

$$\exp(-\gamma|x - x'|^2), \gamma > 0 \quad (1)$$

method one-versus-all, we used a Python library scikit-learn [4]: K coefficients γ and regularization C optimized by cross-validation => test error without optimization 25%, with optimization 16%.

[1] N. Dalal, B. Triggs, Histogram of oriented gradients for human detectection

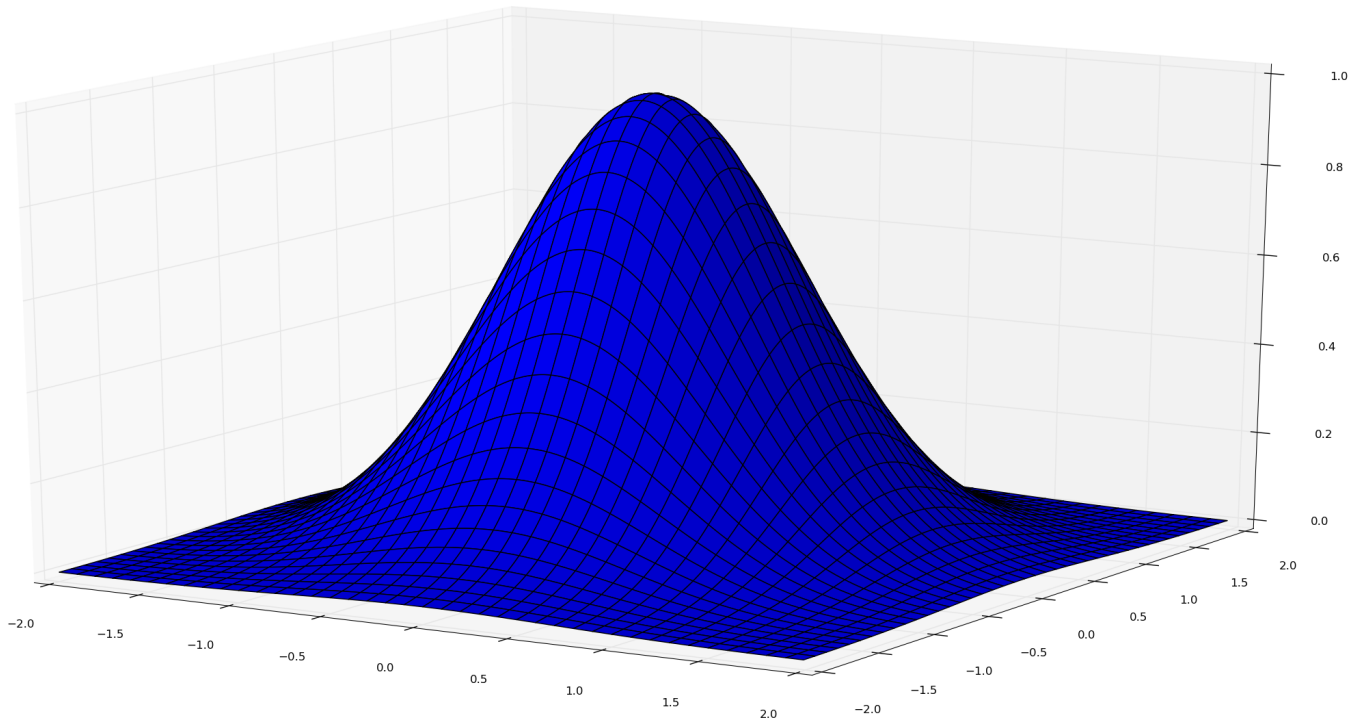

Figure: RBF Kernel

Learning words

We have to build a prior lexicon that contains how characters interact to create words.

1. Word database ICDAR 2003 word [5][2];
2. compute for each pair of characters (c_i, c_j) their frequency of occurrence $p(c_i, c_j)$ in the database (bi-gram model);
3. will be used as an energy term to be minimized afterward.

[2] http://algoval.essex.ac.uk/icdar/Datasets.html

Sommaire

Sliding window and pruning

- Scan the image with sliding windows;
- for each window l_i, compute ϕ_i, feature of HOG with 12 orientations;
- run the 62 SVMs on ϕ_i and compute the goodness score:

$$GS(l_i) = \max_c p(c|\phi_i) \exp\left(-\frac{(a_i - \mu_{a_j})^2}{2\sigma_{a_j}^2}\right) \quad (2)$$

a is the aspect ratio and j is the class that reaches the maximum.

Sliding window and pruning

- if $GS(l_i) > \texttt{threshold}$, keep the window **and** all the probabilities of each character.
- apply Non-Maximum Suppression algorithm (NMS) to merge overlapping windows:
  1. find l_i with maximum confidence.
  2. for all others l_j compute:

$$\text{criterion} = \frac{l_i \cap l_j}{l_i \cup l_j} \tag{3}$$

  3. if criterion $> \texttt{threshold}$ **and** highest character is the same then merge the windows.
  4. the ending merged window has the highest confidence and is located at the barycenter of all the windows merged.

Sommaire

Graph construction and retrieval of the word

PGM part

Technical part of graphical model implemented for PGM course, here is a (very) short summary.

Input. set of n windows with possible characters.

1. for each window, assign a node taking values in $\mathcal{K}_\epsilon$ (ϵ void label);
2. build edges if the windows are 'close enough';
3. assign unary energy to nodes, **and** pairwise energy for edges (overlapping terms + lexicon prior):

$$E(\mathbf{x}) = \sum_{i \in \mathcal{V}} E_i(x_i) + \sum_{(x_i, x_j) \in \mathcal{E}} E_{i,j}(x_i, x_j) \tag{4}$$

4. minimize the discrete energy on the graph (NP-hard) with TRW-S algorithm [6].

Sommaire

Implementation and first results

Figure: Image for testing the algorithm

Implementation and first results

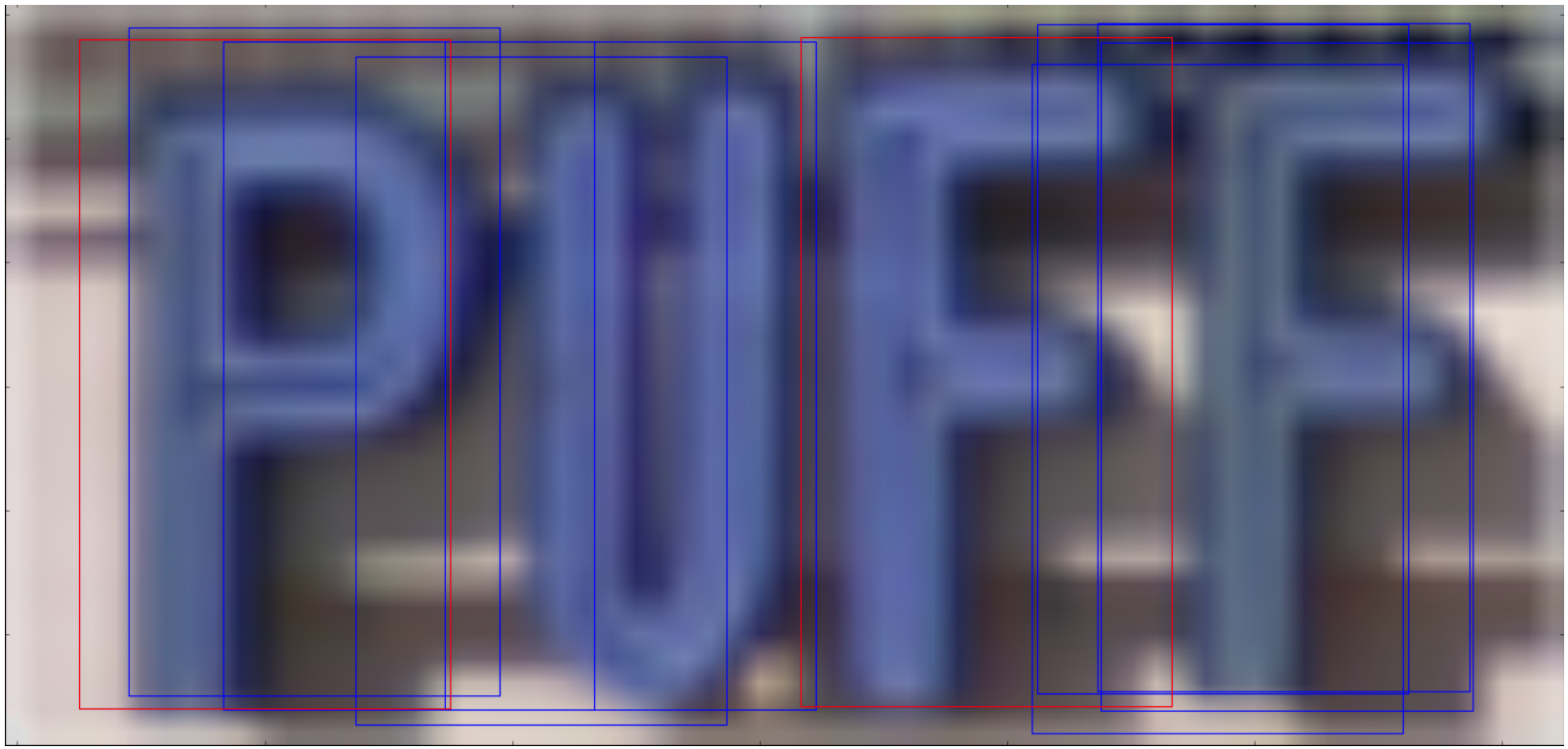

Figure: Word retrieved: 'PE'

Implementation and first results

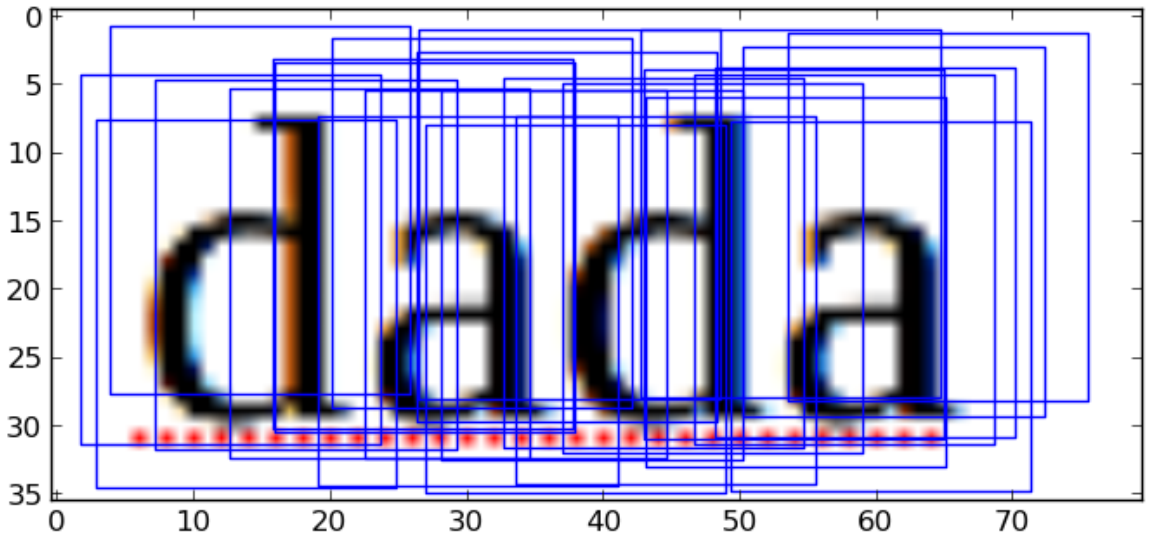

Figure: All the sliding windows kept to build graphical model

Implementation and first results

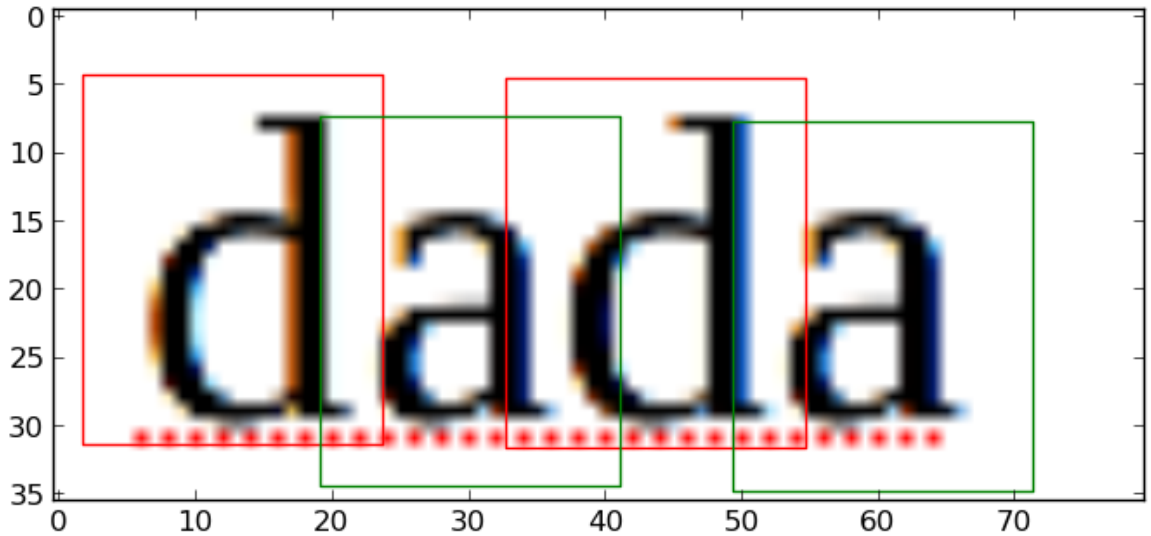

Figure: Presence of the letters 'a' (green) and 'd' (red)

Implementation and first results

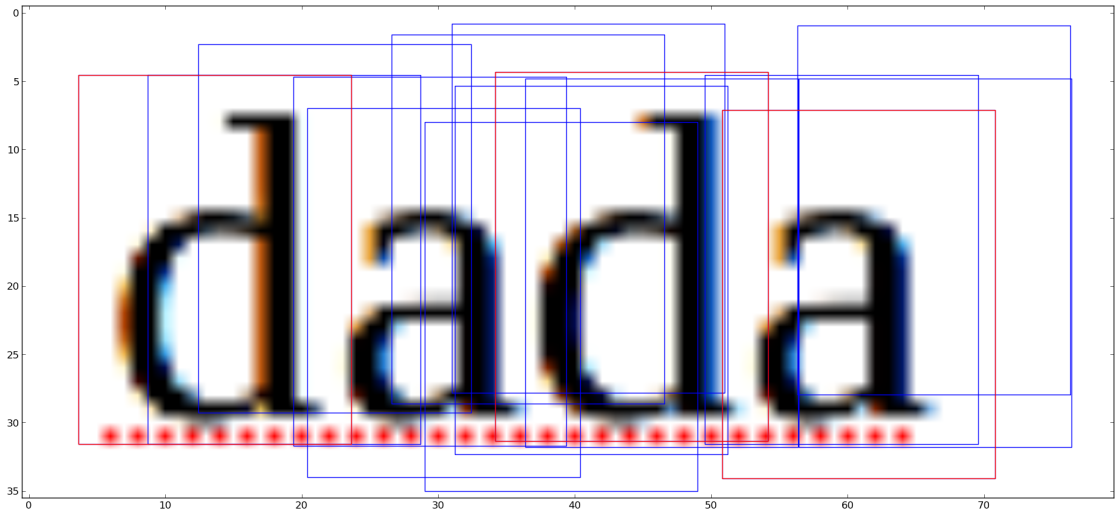

Figure: Word retrieved 'dda' -> the second a has a too big overlap (penalized by pairwise energy, PGM part)

Improvements

1. **Difficult problem**: for each window, classify K times.
2. SVM is not good enough (albeit a 14-hours cross-validation): error on the training set is still 16%.
3. parameters of the algorithm: $2 \times K$ for SVMs + 2 for HOG (bins, number of orientations) + threshold for GS + threshold in NMS + parameters of energy and of TRW-S => complex algorithm.
4. (PGM part) The energy tends to explode letters instead of joining them into words (example 'dda').

Improvements

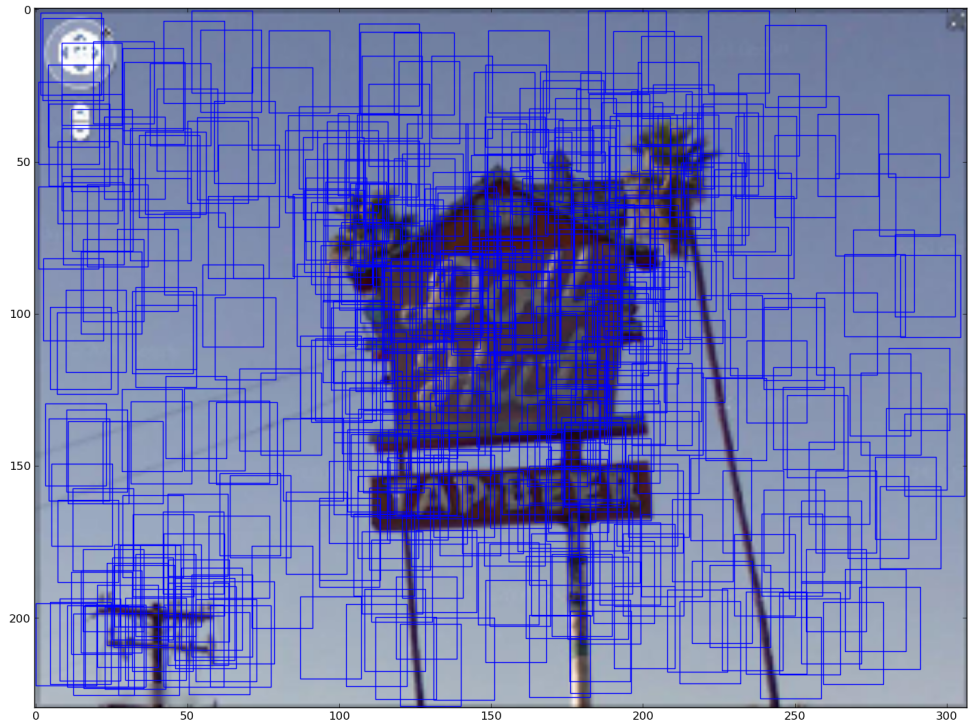

Figure: Flexible GS threshold... too many windows

References

Robust ocr.

http://algoval.essex.ac.uk/icdar/Datasets.html.
Accessed: 2013-12-16.

The chars74k dataset: Character recognition in natural images.

http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/.
Accessed: 2013-12-16.

B. Triggs N. Dalal.

Histogram of oriented gradients for human detectection.
CVPR, 2005.

Scikit-learn.

http://scikit-learn.org/stable/modules/svm.html#parameters-of-the-rbf-kernel.
Accessed: 2013-12-16.

Robust word recognition.

http://algoval.essex.ac.uk/icdar/Datasets.html.
Accessed: 2013-12-16.

V. Kolmogorov.

Convergent tree-reweighted message passing for energy minimization.
IEEE Transactions on Pattern Analysis and Machine Learning (PAMI), 2006.