Project

Vincent BODIN & Thomas MOREAU

Chosen topic. We decided to implement the paper by A. Mishra, K. Alahari and C.V. Jawahar called *'Top-Down and Bottom-up Cues for Scene Text Recognition'*[1] which deals with text recognition on typical street images. The idea is to build a model that detect characters from a street picture and determine the text contained in it.

Plan of the work. This article presents several methods that seem very interesting to implement by ourselves. It first begins with some character detection, then recognition of words and the there is an implementation part on two different databases of street images. Here is what we plan to do by ourselves:

- Character detection is performed by a sliding window detection method that retrieves if there is potentially a word at a given spatial position around this window. Windows are considered at multiple scales and move spatially. In each window, we compute Histogram of Gradient (HOG) as features ϕ_i. If $\mathcal{K} = \{c_1, \cdots, c_K\}$ is the set of character classes, we can compute the likelihood for this windows, represented by ϕ_i, to be one the c_i (formally $p(c_i|\phi_i)$) by a multi-class Support Vector Machine. We definitely want to do this part as this is the very basis of word recognition used in this paper. A score is also computed for every character window, defined by the so-called Goodness Score (GS). This part seems interesting to us as it permits to discriminate the choice of characters jointly with the choice of scale.

- A graphical model of the language is then introduced to perform word recognition. This part is central since it is shown in the article that it permits a big improvement for this type of task. It permits to define a adapted rule to add characters to create a word. An energy made of proper energy of each letter plus interaction with the nearest neighbor is introduced. Minimizing this energy over all the possible words that could be created with character detection is supposed to retrieve the 'real' word. Note that this assumes that we have a prior knowledge on character distribution in words, this knowledge being expressed by either a *bi-gram* model or a *node-specific prior*. As the results in the end seem to show that *node-specific prior* performs better, we hence decided to focus on this method and forget about the *bi-gram* method.

- Experiments are then performed on two databases. We will only test our code on one data set: the street view data set[2].

Share of the work. We do not plan specific separation in our work. As the subject is made as a whole, we will work together on each part.

[1] The paper can be downloaded at 'http://www.di.ens.fr/ alahari/papers/mishra12.pdf' and there is a project webpage 'http://cvit.iiit.ac.in/projects/SceneTextUnderstanding/'.

[2] This dataset is available at 'http://vision.ucsd.edu/ kai/svt/'