Scene Text Recognition using Convolutional Neural Networks and an attribute-based word representation

Supervisor: Ernest Valveny (UAB)
e-mail: ernest@cvc.uab.es

Recognition and detection in scene images (figure 1) is an emerging topic in computer vision with multiple applications due to the exponential increase and availability of pictures and captured with digital cameras and mobile devices.

In a previous work [1] we have developed a method for word recognition in scene images (a generic overview of the method is shown in figure 2) that obtains very competitive results with current state-of-the-art based on learning an attribute-based representation of both images and strings. In this representation (figure 3) every attribute encodes the probability of appearance that a given character or bi-gram appears at a given position of the image or string. For strings, this representation can be obtained directly from the characters of the string. For images, this representation is obtained applying an SVM classifier for each attribute that is trained using a set of annotated training data using a Fisher Vector low-level representation of the image.
The strongest point of this approach is the attribute-based representation that permits to perform with the same framework text recognition and text retrieval (either using an image or a string as a query).

Recently, deep Convolutional Neuronal Networks [2] trained on synthetic data have been shown to obtain state-of-the-art results for text recognition. However, these methods are usually based on the composition of individual character recognition and can only perform text recognition, not retrieval.

In this project we propose to investigate the combination of the strong points of both approaches, analyzing the application of a CNN to learn the intermediate attribute-based representation of [1], and not a final character/word score as in [2]. It is expected that a CNN will yield to a better representation than the original approach and thus, also to an improved performance. Preserving the attribute-based representation should permit to keep the ability to perform indistinctly recognition and retrieval.

The main tasks to develop during the project should be:

- Collect/generate a large set of synthetic data
- Train CNN configurations using this synthetic data in order to learn the attribute-based representation
- Apply the trained CNNs to the problem of word recognition/retrieval
- Analyze the results

References:
