Efficient Exemplar Word Spotting

Jon Almazán
www.cvc.uab.es/~almazan
Albert Gordo
agordo@cvc.uab.es
Alicia Fornés
afornes@cvc.uab.es
Ernest Valveny
ernest@cvc.uab.es

Computer Vision Center
Departament de Ciències de la Computació
Universitat Autònoma de Barcelona
Barcelona, Spain

This work addresses the problem of word spotting: given a query word image, the goal is to retrieve locations in a set of document images where this word may be present. Traditionally, word spotting systems have followed a well defined flow. First, an initial layout analysis is performed to segment word candidates. Then, the extracted candidates are represented as sequences of features, and, by using a similarity measure – commonly a Dynamic Time Warping (DTW) or a Hidden Markov Model (HMM)-based similarity –, the query word is compared and candidates are ranked. An example of this framework is the work of Rath and Manmatha [4].

One of the main drawbacks of these systems is that they need to perform a costly and error prone segmentation step to select candidate windows. Any error introduced in this step will negatively affect the following stages, and so it is desirable to avoid segmenting the image whenever possible. Unfortunately, since the comparison of window regions, represented by sequences, is based on costly approaches such as a DTW or a HMM, it is not feasible to perform this comparison exhaustively with a sliding window approach over the whole document image. Another important drawback is that best performance techniques – such as HMMs or Neural Networks – need a large amount of annotated training images. However, this is not a common case in real scenarios.

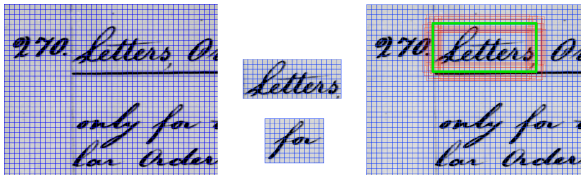
The recent [5] addresses the segmentation problem by representing regions with a fixed-length descriptor based on the bag of visual words framework. To further improve the system, unlabeled training data is used to learn a LSI space, where the distance between words is more meaningful than in the original space. We follow [5] and address the word spotting problem in an unsupervised, segmentation-free setting, and argue that the current methods can be improved in several ways.

First, they can be improved in the choice of low level features. We address these issues by using HOG descriptors, which have been shown to obtain excellent results when dealing with large variations in the difficult tasks of object detection and image retrieval. In our baseline system, the document images are divided in equal-sized cells (see Fig 1a) and represented with HOG histograms. Queries are represented analogously using cells of the same size in pixels (Fig 1(b)). The score of a document region is computed using the cosine similarity, *i.e.*, calculating the dot-product between the L2 normalized descriptors. Following this approach, we can compute the similarity of all the regions in the document image with respect to the query using a sliding window and rank the results.

Second, spotting methods can be improved in the learning of a more semantic space. We propose to perform this unsupervised learning once the query has been observed, and adapt the learning to the query. For this task, we propose to use a similar approach to the Exemplar SVM framework of [3]. We need first a set $\mathcal{P}$ of relevant regions to the query. This set is constructed by slightly shifting the window around the query word to produce many almost identical, shifted positive samples (see Fig 1(c)). Then we need a set of non-relevant regions. To produce this negative set $\mathcal{N}$, we sample random regions over all the documents. Given these sets, we can solve the following optimization problem

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||^2 + C \sum_{\mathbf{y}_p \in \mathcal{P}} L(\mathbf{w}^T \mathbf{y}_p) + C \sum_{\mathbf{y}_n \in \mathcal{N}} L(-\mathbf{w}^T \mathbf{y}_n), \qquad (1)$$

where $L(\mathbf{x}) = \max(0, 1-\mathbf{x})$ is the hinge loss and C is the cost parameter. Solving this optimization produces a weight vector $\mathbf{w}$, which can be seen as a new representation of the query. This new representation has been directly optimized to give a high positive score to relevant regions, and a high negative score to non-relevant regions when using the dot-product with L2 normalized regions.



(a) HOG grid. (b) (c)

Figure 1: a) Grid of HOG cells. b) Two random queries. The windows adjust around the HOG cells. c) A query (in green) and some positive samples (in red) used to learn the Exemplar SVM.

Finally, these methods can be improved in the cost of storing the descriptors of all the possible windows of all the dataset items. Assuming HOG descriptors of 31 dimensions represented with single-precision floats of 4 bytes each, and 50,000 cells per image, storing as few as 1,000 precomputed dataset images would require 5.8GB of RAM. Since documents will not fit in RAM memory when dealing with large collections, it would produce a huge performance drop in the speed at query time. To address this problem, we propose to encode the HOG descriptors using Product Quantization (PQ) [2]. Encoding the descriptors with PQ would allow us to preserve a much larger amount of images in RAM at the same time. As a side effect, computing the scores of the sliding window also becomes significantly faster. In our case, we can reduce the size of the HOG descriptor to one byte with minimal loss, and achieve a 10-fold improvement in computational time.

We evaluate our approach on two public datasets: The George Washington (GW) and the Lord Byron (LB). Figure 2 shows the mean Average Precision of our vanilla method and its improvements – EWS and EWS+PQ – on the GW dataset as a function of the HOG cell size and the sliding window step size.

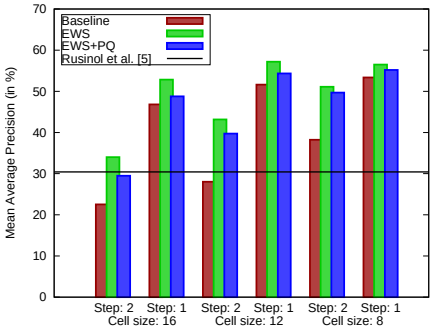


Figure 2: Mean Average Precision of our method for different setups.

We publish the MATLAB code implementation for training and testing the Exemplar Word Spotting on the web page [1].

[1] J. Almazán, A. Gordo, A. Fornés, and E. Valveny. Exemplar Word Spotting library. URL `http://almazan.github.com/ews/`.

[2] H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE TPAMI*, 2011.

[3] T. Malisiewicz, A. Gupta, and A. Efros. Ensemble of Exemplar-SVMs for object detection and beyond. In *ICCV*, 2011.

[4] T. Rath and R. Manmatha. Word spotting for historical documents. *IJDAR*, 2007.

[5] M. Rusiñol, D. Aldavert, R. Toledo, and J. Lladós. Browsing heterogeneous document collections by a segmentation-free word spotting method. In *ICDAR*, 2011.