Label embedding for text recognition

Jose A. Rodriguez-Serrano
jose-antonio.rodriguez@xrce.xerox.com

Florent Perronnin
Florent.Perronnin@xrce.xerox.com

Xerox Research Centre Europe
Meylan, France

Summary The standard approach to recognizing text in images consists in first classifying local image regions into candidate characters and then combining them with high-level word models such as conditional random fields (CRF). *This paper explores a new paradigm that departs from this bottom-up view.*

In our approach, every label from a lexicon is *embedded* to an Euclidean vector space. We refer to this step as *label embedding*. Each vector of image features is then projected to this space. To that end, we formulate the problem in a structured support vector machine (SSVM) framework [3] and learn the linear projection that optimizes a proximity criterion between word images and their corresponding labels: matching label-image pairs should be closer than non-matching pairs. In this space, the "compatibility" between a word image and a label is measured simply as the dot product between their representations. Therefore, given a new word image, recognition amounts to finding the closest label in the common space (Fig. 1).

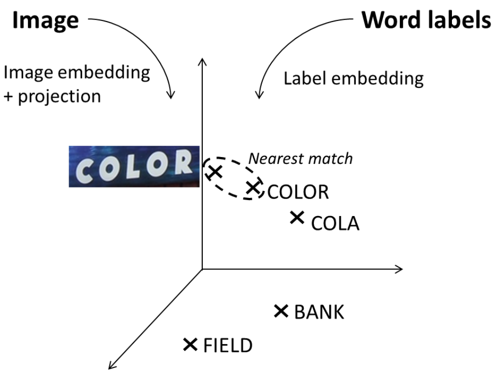

Figure 1: Illustration of recognition with label embedding.

This method presents the following advantages: (i) it does not require costly pre- or post-processing operations, (ii) it allows for the recognition of never-seen-before words, (iii) the recognition process is efficient.

Model Let $\theta : \mathcal{X} \rightarrow \mathbb{R}^D$ be a function that acts on the pixels of x and extracts a D-dimensional feature vector $\theta(x)$ (feature embedding).

Let $\varphi : \mathcal{Y} \rightarrow \mathbb{R}^E$ denote a function that computes a fixed-length feature vector from the label y (label embedding).

We use the following similarity function between the (projected) image embeddings and the label embeddings :

$$F(x,y;W) = \tilde{\theta}^T(x)\varphi(y) = \theta^T(x)W\varphi(y). \tag{1}$$

If the matrix W is known, recognizing the text in image x amounts to scanning the lexicon $\mathcal{Y}$ for a best match:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} F(x,y;W) \tag{2}$$

The goal of learning is to find the optimal matrix W.

Embeddings For the image embeddings, we use the widely adopted bag-of-patches framework. We choose to compute the patch statistics using the Fisher Vector (FV) principle [4].

For the image embeddings, we propose a Spatial Pyramid of Characters (SPOC). Given a text label, the SPOC counts the frequencies of character appearances at certain subdivisions of the text label, as illustrated in Fig. 2. This embedding is data-free (*i.e.* any label can be easily embedded on-the-fly), respects the lexical similarity between words and is expressed with a fixed-length feature vector.

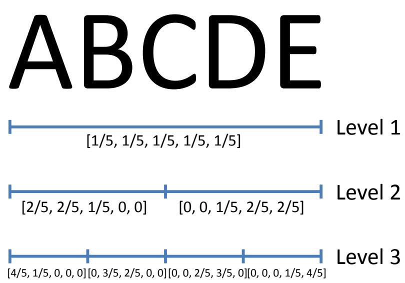

Figure 2: SPOC label embedding.

Learning We note that Eqs. (1) and (2) can be re-written in the form of a ranking SSVM with an objective of the form

$$w^* = \arg\min_w \frac{1}{N}\sum_{n=1}^{N} B_2(x_n, f(y_n)) + \frac{\lambda}{2}||w||^2, \tag{3}$$

where

$$B_2(y_n, f(x_n)) = \sum_{y \in \mathcal{Y}} \Delta(y_n, y) - F(x_n, y_n; w) + F(x_n, y; w), \tag{4}$$

which can be optimized with Stochastic Gradient Descent (SGD) [1].

Experiments Experiments are performed on a private license plate recognition dataset and on the IIIT-5K scene text dataset [2] show that the proposed method is competitive with standard bottom-up approaches to text recognition.

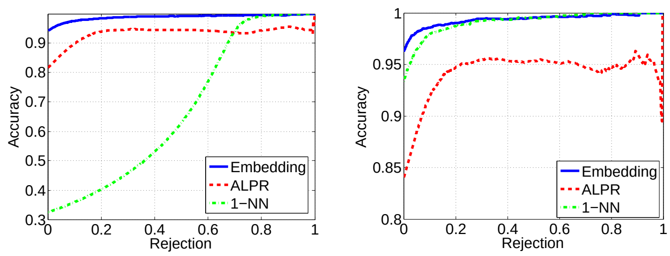

Figure 3: License plate results. Left: using the whole test set. Right: Using only the subset of images which have a true match in the database.

[1] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and Muller K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998.

[2] A. Mishra, K. Alahari, and C. V. Jawahar. Scene text recognition using higher order language priors. In *BMVC*, 2012.

[3] S. Nowozin and C. Lampert. Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 2011.

[4] F. Perronnin, J. Sánchez, and Thomas Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010.