Supervised Mid-Level Features for Word Image Representation

Albert Gordo
Computer Vision Group, Xerox Research Centre Europe

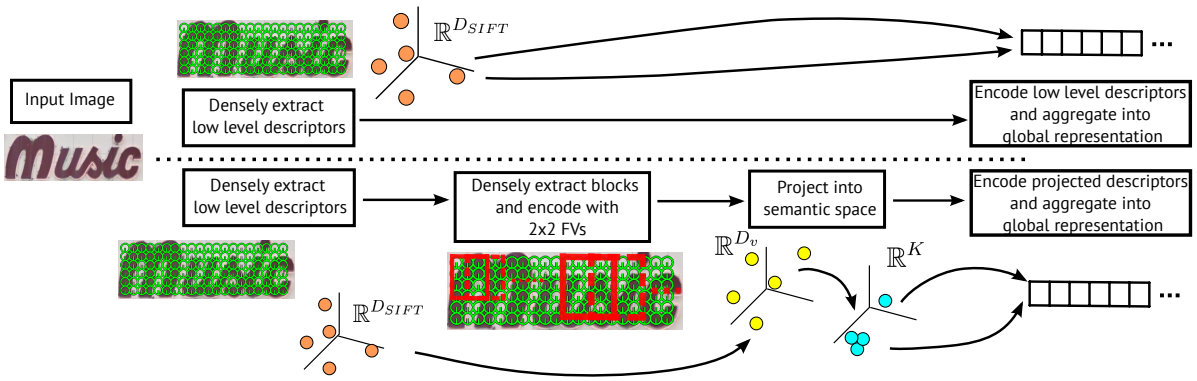

Figure 1

Top. Standard word image description flow: low-level descriptors (*e.g.* SIFT) are first densely extracted and then encoded and aggregated into a global representation using *e.g.* Fisher vectors (FV). Spatial pyramids may be used to add some weak geometry. **Bottom.** Proposed approach: we first densely extract low-level descriptors. Then we densely extract blocks of different sizes, and represent each block by aggregating the low-level descriptors it contains into a local FV with a 2×2 spatial pyramid. These local FV representations are then projected into a mid-level space correlated with characters. Finally, these mid-level features are aggregated into a global FV.

This paper addresses the problem of learning word image representations: given the cropped image of a word, we are interested in finding a descriptive, robust, and compact fixed-length representation. Machine learning techniques can then be supplied with these representations to produce models useful for word retrieval or recognition tasks. Although many works have focused on the machine learning aspect once a global representation has been produced, little work has been devoted to the construction of those base image representations: most works use standard coding and aggregation techniques directly on top of standard computer vision features such as SIFT or HOG.

We propose to learn local mid-level features suitable for building word image representations. These features are learnt by leveraging character bounding box annotations on a small set of training images. However, contrary to other approaches that use character bounding box information, our approach does not rely on detecting the individual characters explicitly at testing time.

At training time, blocks are randomly sampled from the training images and described using two types of representations: a visual representation, based on Fisher vectors on SIFT descriptors, and a character representation, based on the character annotations (see Figure 2).

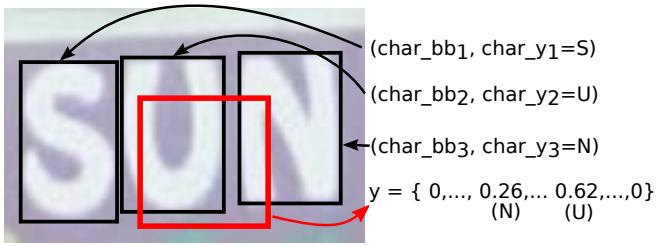

Figure 2: Example of annotated word and a sampled block with its label. The characters of the word contain bounding boxes (in black) and label annotations ('S','U','N'). The image also shows a sampled block (in red) with its respective computed label y. All the elements of y are set to 0 except the ones corresponding to the 'U' and 'N' characters.

The representation based on the character annotations is much more discriminative, but can only be obtained at training time on annotated images. Then, one can learn how to project the visual representation into a space correlated with the annotation space, and therefore with the characters contained in the word. This is achieved using canonical correlation analysis.

At testing time, one can extract all possible blocks in the image, represent them with visual features, and then project the visual representations in the space correlated with the character annotations with the learnt projections, with no need of character annotations at testing time. These local mid-level features can then be aggregated to produce a global word image signature (see Figure 1). Furthermore, these local mid-level features can be exploited by other frameworks that rely on low level features, such as the recent word attributes framework of Almazán et al [1].

We tested our representations on two standard benchmarks: Street view text (SVT) and IIIT5K. We show experimentally that using these mid-level features as a building step for word attributes improves over using SIFT descriptors directly, and outperforms recent state-of-the-art such as Google's PhotoOCR [2], although we do not match the accuracy of recent methods based on convolutional neural nets that use millions of training images [3].

Table 1: Recognition accuracy on the IIIT5K and SVT. Methods marked with an $*$ use several millions of training samples.

Method	SVT	IIIT5K
*PhotoOCR [2] *(in house training data)*	90.39	-
*Deep CNN [3] *(synthetic training data)*	**95.4**	-
[SIFT] + FV + Atts [1]	89.18	91.20
[**Prop. Mid-features**] + FV + Atts	89.49	92.67
[**Prop. Mid-features** + SIFT] + FV + Atts	90.73	**93.27**

[1] J. Almazán, A. Gordo, A. Fornés, and E. Valveny. Word spotting and recognition with embedded attributes. *TPAMI*, 2014.

[2] A. Bissacco, M. Cummins, Y. Netzer, and H. Neven. PhotoOCR: Reading Text in Uncontrolled Conditions. In *ICCV*, 2013.

[3] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition . *CoRR*, abs/1406.2227, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.