Exemplar Driven Character Recognition in the Wild

Karthik Sheshadri
sheshadri@cmu.edu
Santosh K. Divvala
santosh@ri.cmu.edu

The Robotics Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania
USA

Character recognition in natural scenes continues to represent a formidable challenge in computer vision. Traditional optical character recognition (OCR) methods fail to perform well on characters from scene text owing to a variety of difficulties in background clutter, binarisation, and arbitrary skew. Further, English characters group into only 62 classes whereas many of the world's languages have several hundred classes. In particular, most Indic script languages such as Kannada exhibit large intra class diversity, while the only difference between two classes may be in a minor contour above or below the character. These considerations motivate an exemplar approach to classification; one which does not seek intra class commonality among extreme examples which are essentially sub classes of their own. Exemplar SVM's have been recently introduced in the object recognition context. The essence of the exemplar approach is that rather than seeking to establish commonality within classes, a separate classifier is learnt for each exemplar in the dataset. To make individual classification simple, linear SVM's are used and each classifier is hence an exemplar specific weight vector. Each exemplar in the dataset is resized to standard dimensions, and thence HOG features are densely extracted to create a rigid template x_E. A set of negative samples N_E are created by the same process from classes not corresponding to the exemplar. Each classifier (w_E, b_E) maximizes the separation between x_E and every window in N_E. This is equivalent to optimizing the convex objective[4]:

$$\Omega_E(w,b) = \| w \|^2 + C_1 h(w^T x_E + b) + C_2 \sum_{x \in N_E} h(-w^T x - b), \qquad (1)$$

where $h(.)$ indicates the hinge loss function, and C_1, C_2 are constants.

1 Calibrating Exemplar SVM's for Character Recognition

In return for simpler classification at the level of each exemplar, we must now deal with the problem of decision calibration: combining decisions from independently trained and hence non compatible classifiers. In this work, we explore the following two calibration methods.

1.1 Calibration based on SVM scores

In the spirit of [4], we adopt an "on the fly" calibration method, using positives selected by each exemplar based on SVM scores. Exemplars which achieve low scores on ground truth labelled query images from the validation set are suppressed by moving the decision boundary in their requisite classifier towards the exemplar and well performing exemplars are boosted by moving the decision boundary in their classifier away from the dataset. Given a detection 'x' and the learned sigmoid parameters α_E,β_E, the calibrated detection score for each exemplar E is as follows: $f(x|\alpha_E, \beta_E, w_E) = \frac{1}{1+e^{-\alpha_E(w_E^T x - \beta_E)}}$. This rescaling and shifting of the decision boundary conditions each classifier to fire only on visually similar examples, and underlines the explicit correspondence offered by the exemplar SVM based approach.

1.2 Calibration based on affine motion estimation

This calibration approach is based on a simple observation: variations in font and shape essentially constitute small affine transformations. Characters from visual scenes are often affine warps of characters from normal text: they are oriented differently, different character contours are irregularly shaped, and are of different sizes, etc. Hence on thinned character images, one could compute affine motion between train and test characters, and minimize the sum of absolute differences to refine candidate choices obtained by simple max voting of the exemplar SVM's. Our proposed approach is summarized as follows: (i) count the number of positive votes in favour of each class, computed based on a preselected threshold (ii) extract the top k of these classes, and perform affine motion estimation $M_{E_C,Q}$ between the thinned binarized query image Q and every exemplar E_C, C being the class of the exemplar, in the training subset corresponding to the top k classes (iii) recognize the character as that class which minimizes the sum of absolute differences (SAD) between the test character and any exemplar in the training subset corresponding to top k classes. Equation (2) illustrates the approach:

$$B =_{C \in C_K} \{E_C - M_{E_C,Q} Q\} \qquad (2)$$

where B is the computed belief class of query image Q, and $M_{E_C,Q}$ is the affine transformation matrix which warps Q with respect to E_C.

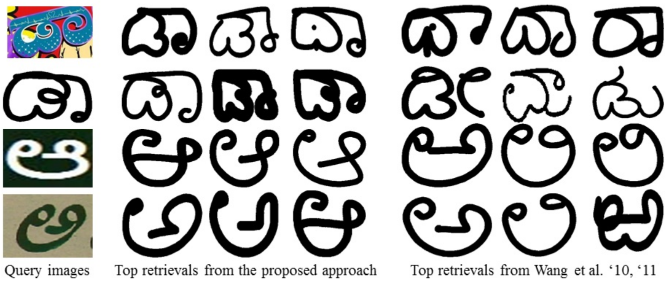
Query images
Top retrievals from the proposed approach
Top retrievals from Wang et al. '10, '11

Figure 1: The problem of fine grained character recognition in unconstrained visual scenes is addressed in this paper.

Table 1: Our results (English) on chars74k and ICDAR-CH, and comparison to baseline methods.

Model	Chars74k-5	Chars74k-15	ICDAR-CH
HOG+ESVM+AFF	**48.43 ± 2.40**	**69.66**	**70.53**
HOG+ESVM+on fly calib	27.76 ± 1.74	60.00	66.67
HOG+NN+AFF	47.61 ±0.81	64.22	63.59
HOG+ESVM	16.33 ± 2.33	44.68	41.44
HOG+NN[2]	45.33 ± 0.99	57.50	52
NATIVE+FERNS[3]	−−	54	64
MKL[1]	−−	55.26	−−

The proposed approach beats the existing state of the art on the chars74k and ICDAR datasets by over 10% for English, and around 24% for Kannada. Motivated by the performance on two languages ranging from conventional to extremely complex, we argue that leveraging fine grained categorization and generic object recognition approaches is a promising research direction for character recognition unconstrained by language or setting.

[1] T. de Campos, B. Babu, and M. Varma. Character recognition in natural images. In: VISAPP 2009.

[2] K. Wang and S. Belongie. Word Spotting in the Wild. In: ECCV 2010.

[3] Kai Wang, Boris Babenko, and Serge Belongie. End to End Scene Text Recognition. In: ICCV 2011.

[4] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A. Efros. Ensemble of Exemplar-SVMs for Object Detection and Beyond. In: ICCV 2011.