Con-Text: Text Detection Using Background Connectivity for Fine-Grained Object Classification

Sezer Karaoglu, Jan C. van Gemert and Theo Gevers
Intelligent Systems Lab Amsterdam (ISLA), University of Amsterdam
Science Park 904 Amsterdam, The Netherlands

ABSTRACT

This paper focuses on fine-grained classification by detecting photographed text in images. We introduce a text detection method that does not try to detect all possible foreground text regions but instead aims to reconstruct the scene background to eliminate non-text regions. Object cues such as color, contrast, and objectiveness are used in corporation with a random forest classifier to detect background pixels in the scene. Results on two publicly available datasets ICDAR03 and a fine-grained *Building* subcategories of ImageNet shows the effectiveness of the proposed method.

Categories and Subject Descriptors

I.4.8 [**Scene Analysis**]: Object Recognition; I.5.4 [**Computer Vision**]: [Applications]

Keywords

Fine-grained classification; scene text recognition

1. INTRODUCTION

Automatic visual classification of very similar instances, a.k.a. fine-grained classification, is the problem of assigning images to classes where all instances differ only by minute details. Examples include flower types [8] or specific bird species [18]. In this paper, we propose to use recognized photographed texts images to aid in fine-grained classification. As an application, we focus on classification of *Buildings* into their sub-classes such as *Cafe*, *Tavern*, *Diner*, etc. This can be used to link images from Google Street View to textual business information as in the Yellow pages.

When text is present in natural scenes, the text is typically there to give semantic meaning beyond what is obvious from exclusively visual cues. For instance, in fig 1, the left and middle images share a very similar scene layout. However, if one wants to group these images based on the semantics of the scenes, the middle and the right images belong together because they share the same business name ("Starbucks").

MM'13, October 21–25, 2013, Barcelona, Spain.

http://dx.doi.org/10.1145/2502081.2502197.

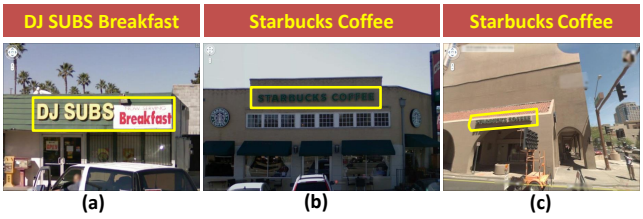

Figure 1: An example of fine-grained *Building* classification. Visual cues would group (a)-(b) whereas scene text reveals that and groups (b)-(c).

Traditional optical character recognition (OCR) systems work well for controlled documents, however their performance deteriorates in recognizing natural scene text. The challenges in natural images include non-linear illumination, occlusion, motion blur and variations in text size, style and orientation. To overcome such issues, rather than trying to detect all variations in text appearance, we propose to detect the background in the image using a Random Forest (RF). Then, text regions are detected by eliminating the scene background.

This paper has two main contributions. First, instead of trying to detect text directly, we propose to learn the background of the scene to infer the location of text. Second, we propose a system which incorporates the additional semantics in the scene text to infer the image content. We give results of multimodal fusion of text and visual cues for fine-grained classification on sub-classes of the ImageNet *building* and *place of business* dataset. The proposed method improves significantly over a visual-only classification.

2. RELATED WORK

Existing text detection methods extract features either from connected components [3, 4, 7] or bounding boxes [16] to decide whether the region contains text or not. These methods try to verify geometric, structural and appearance properties of text based on some heuristics or learning scheme. Since these methods rely on training or simple rules designed for specific text styles, it is hard to maintain robustness to a-priori unknown text styles. In contrast, our proposed method does not focus on extracting text features but tries to detect background pixels to eliminate non-text regions.

Our approach is inspired by Wei et al. [17] who propose to use scene background regions for salient object detection. We also use the background, but now for text detection. In contrast to the strong boundary dependency of [17] we use a careful background seed selection step by training a *RF*.

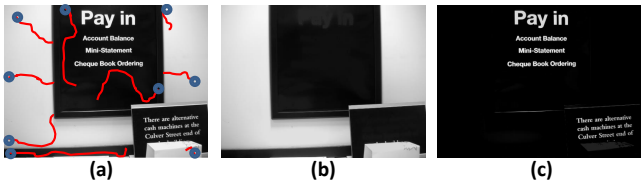

Figure 2: (a) Original image and selected background seeds (blue dots) (b) reconstructed background and (c) background removed image regions.

Text information is fused with visual cues for better scene classification by Wang et al. [14]. The authors use Flickr images and their associated social tags in an object recognition problem. Others [19] propose to combine visual features extracted from the surroundings of text regions with features from the full image. In contrast to these two methods, we propose text recognition in the image and use the recognized text to directly aid scene classification. The approaches done by Karaoglu et al. [5] and Tsai et al. [10] are the most similar in spirit to ours. Both approaches propose to use scene text information in combination with visual features to improve visual recognition. While Karaoglu et al. [5] combines text and visual features for generic object recognition such as *Cars* and *Aeroplanes*, Tsai et al. [10] combines text and visual features for book spine recognition problem.

3. BACKGROUND DETECTION

Instead of detecting text regions directly, we propose to detect the scene background to eliminate non-text regions. Intuitively, removing background clutter will reduce false text detections.

Texts are typically designed to attract attention [3, 7, 9, 11, 15] they typically contrasts strongly against the background. Hence, strong intensity changes can be expected on text boundaries. This give us the opportunity to distinguish between text and the surroundings by using an intensity-level pixel connectivity measure. Our method starts from background seeds and grow these seeds until the background is reconstructed. Background pixels can then easily be eliminated by subtracting them from the original image. An illustration of seed growing is shown in fig 2a. Selected background seeds are represented by blue dots and some of the background pixels obtained based on connectivity to those seeds are represented by red lines.

Mathematical morphology is a well known tool for defining connectivity between pixels. Therefore, we make use of morphological operations to define connectivity in our background detection problem. Morphological reconstruction uses conditional dilation (δ) which can be defined as

$$\delta_S(\gamma|I) \triangleq (\gamma \oplus S) \cap I \ , \tag{1}$$

where S is the structuring element (3-by-3 square), γ is the binary image where background seeds are ones and I is the gray level input image. This operation is executed iteratively until the stability is sustained,

$$\rho_S^-(\gamma|I) \triangleq \bigcup_{n=1}^{\infty} \delta_S^n(\gamma|I) \ , \tag{2}$$

$$\text{repeat } \gamma_{n+1} = \delta_S(\gamma_n|I_n) \text{ until } \gamma_n = \gamma_{n+1} \ . \tag{3}$$

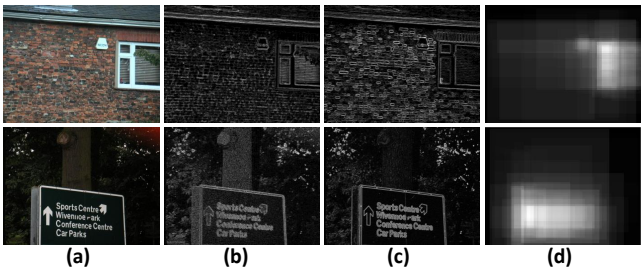

Figure 3: (a) Original image (b) color boosting (c) curvature and (d) objectness.

Careful selection of the seeds γ is essential for background reconstruction since the seeds will initialize the connectivity procedure. The proposed method selects the seeds based on learning scheme using color, contrast and objectiveness cues incorporation with *RF*.

3.1 Background Seed Selection

3.1.1 *Color and Contrast*

Color edges are useful to learn whether a region belongs to background or not. Color is homogeneous on most of the background scene such as on roads, sky, buildings and so on. Moreover, color edge responses correlate with high contrast texts. To exploit this, we use the color boosting algorithm proposed by Van de Weijer et al. [12] where the authors boost the color information in image gradients by replacing the gradient strength with information content carried by color. The response of the color boosting algorithm for each pixel is used as a feature in our RF, see fig 3b for an example. Moreover, we add RGB color values in addition to the color edge transition cues to help the classifier to distinguish unstructured regions such as sky, grass and trees.

In addition to color cues, we use the curvature (L) of colorless edges in the scene. Curvature is expressed by $L = \sqrt{I_{xx}^2 + I_{xy}^2 + I_{yy}^2}$, where I_{xx} and I_{yy} stand for the second-order derivatives of the intensity image $I(x, y)$ in the x and y dimensions, respectively. The response measured by curvature intensity for each pixel is used as a feature, see fig 3c for an example. Additionally, we extract two more features based on boundary priors which assumes that the border has a high likelihood to be background [5, 17]. We remove high responsive color boosting and curvature regions which are connected to the image border (equation 2).

3.1.2 *Objectness*

Generic object detection techniques have been widely used for text detection and recognition problem [6, 16]. In contrast, our goal here is not to detect text windows but to detect unstructured background context such as grass, bricks, fences and trees. The motivation to make use of object detectors is that they have low response on unstructured image regions and high response for structured image regions, as illustrated in fig 3d. We use a generic object detector called 'objectness' [1]. The score and coordinates of first 100 objectness windows are summed, normalized to the range [0,1] and mapped into image coordinates.

3.1.3 *Learning with Random Forest*

The described responses for color, contrast and objectness are extracted pixel-wise then smoothed and subsam-

Color [12]	Curvature (L)
(Boundary Prior + Color) [5]	(Boundary Prior + L) [5]
Objectness [1]	RGB values

Table 1: Features used for background detection.

Figure 4: Original (on top) and background removed (down) images using proposed algorithm.

pled (10x) to incorporate neighbor pixel information. For each pixel in the subsampled image we have six descriptive features for training a classifier to detect the background (see table 1). These features are used with a *RF* to classify background seeds. A *RF* is a combination of many un-pruned decision tree predictors [2]. In the training phase, each decision tree is constructed from randomly selected features at each split with bagging. When a new instance arrives to the forest for classification, the instance is put down each of the trees in the forest. Each tree gives a vote and the forest chooses the classification having the most votes.

We use 100 trees in the forest based on the out-of-bag error (*oob*) curve obtained for a small holdout set of the ICDAR03 dataset[1]. Each tree is constructed with three random features as suggested by Breiman [2] and is grown to the largest extent possible. The splitting of the nodes is made based on the GINI criterion [2]. The training for all experiments is done on the ICDAR03 training set where text bounding boxes are provided. We denote the pixels in the groundtruth text boxes positive and outside the boxes as negative.

4. EXPERIMENTS

To evaluate our system, the leading commercial OCR engine, ABBYY FineReader, is used for text recognition. After the background is detected and removed from the original image, ABBYY is fed with this output that retains the most likely text regions, see fig 4 for example outputs.

4.1 Character Recognition

The system takes roughly 5.2 seconds for objectness and 0.01 seconds for each color and curvature feature extraction processes to run on a 480×640 resolution image. The *RF* classification takes 0.01 seconds while conditional dilation to build connectivity takes 0.012 seconds.

The character recognition performance of the proposed method is evaluated on the publicly available ICDAR03 dataset. We present our results against ABBYY alone where ABBYY is directly fed with input image and with Karaoglu et al. [5] with their reported results in table 2. The results show that removing the background from the scene substantially increases the OCR performance. This is because traditional OCR systems only work well with controlled documents. We improve slightly (1%) over [5] which corresponds to an improvement of ± 100 characters.

Method	Cl. Rate (%)
ABBYY alone	37
Karaoglu et al [5]	62
Proposed Method	**63**

Table 2: Character recognition rates for different methods on ICDAR03.

We also evaluated how much of the true background is removed by using the ICDAR03 text localization ground truth. We counted the pixels which are correctly returned as a text region and the pixels which are not. Our proposed method removes 87% of the background regions. Since the dataset is not annotated with characters but with word bounding boxes only, we cannot give a precise estimation of the scene text amount that our method retains.

4.2 Fine Grained Classification

We use sub-classes of the ImageNet[2] *building* and *place of business* sets to evaluate fine-grained classification. The dataset consists of 28 categories with 24,255 images, see fig 5 for the category list. In the experiments, we use all images from these categories, i.e., many images do not necessarily contain photographed text. The number of images with text in the categories varies from 8% (*massage center*) to 30% (*bakery*).

We use average precision as the performance measure, and we repeat all experiments three times to obtain standard deviation scores to validate significance testing. We use a standard bag of visual words (BOW) approach with SIFT using 4000 words with 1×3 and 2×2 spatial pyramid as the visual classification baseline. We use the histogram intersection kernel in libsvm and use its built in cross-validation method to tune the C parameter. For text classification we use a bag-of-bigrams with the histogram intersection kernel. The two kernels are additive, and are thus fused by a simple addition. We compare results of visual-only, text-only and fused as our proposed method. The classification scores per category are given in fig 5.

The low accuracy of text cues over visual cues can be explained by the lack of photographed text in many images, as is the case for the *Massage Center* and *Bistro* classes. Moreover, this dataset is more realistic than existing text recognition datasets and therefore not all of the text in the scene is recognized. Nevertheless, text cues outperform visual cues for the *Discount House*, *Steak House*, *PawnShop*, *Cafe* and *Dry Cleaner* classes.

The images in *Discount House* and *Steak House* are visually not similar and are hard even for humans if there was no text information on them. The very low performance of the visual cues for *Discount House* cause it to be the only class where adding visual cues reduced the accuracy of the text-only score.

[1] `http://algoval.essex.ac.uk/icdar/`

[2] `http://image-net.org/`

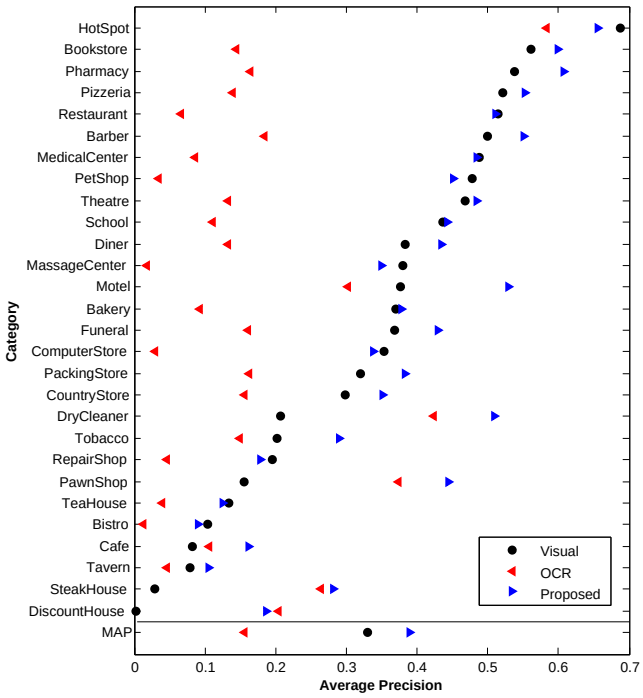

Figure 5: Fine-grained classification results for visual-only, text-only and proposed method. The visual MAP is 32.9 ± 1.7, text is 15.6 ± 0.4 and proposed is 39.0 ± 2.6.

Adding text information to visual cues significantly improved the accuracy of 19 out of the 28 classes. Note that the performance of each individual cue for classes *Cafe*, *Motel* and *Tobacco* are very close. However, the combination of visual and textual cues increases performance remarkably. This can be explained by reinforcing textual and visual cues where each modality brings new discriminative information. Moreover, when text is informative for a class (i.e. >0.1) combining visual and text cues always increases accuracy.

We show some example images in fig 6. The figure illustrates that the best cues can be learned per-category.

5. CONCLUSION

We have shown that background removal is a suitable approach for photographed text detection in natural scenes. Color, curvature and objectness prove valuable cues for background modeling.

We introduce a new fine-grained classification problem based on ImageNet subcategories and build a baseline for further research. We have shown that multimodal information fusion of visual and textual cues improves fine-grained classification on this dataset by 6%. For future work, we aim to build equivalence classes as proposed in [13] for text and non-text regions to exploit correspondences between text related visual features to help the system where OCR fails to recognize characters.

6. ACKNOWLEDGEMENT

This publication was supported by the Dutch national program COMMIT.

Figure 6: Result for *Discount House*.

7. REFERENCES

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *TPAMI*, 34(11), 2012.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1), 2001.

[3] B. Epshtein, E. Ofek, and Y. Wexler. Detecting text in natural scenes with stroke width transform. In *CVPR*, 2010.

[4] S. Karaoglu, B. Fernando, and A. Trémeau. A novel algorithm for text detection and localization in natural scene images. In *DICTA*, 2010.

[5] S. Karaoglu, J. van Gemert, and T. Gevers. Object reading: Text recognition for object recognition. In *ECCV Workshops (3)*, 2012.

[6] A. Mishra, K. Alahari, and C. V. Jawahar. Top-down and bottom-up cues for scene text recognition. In *CVPR*, 2012.

[7] L. Neumann and J. Matas. Real-time scene text localization and recognition. In *CVPR*, 2012.

[8] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, 2006.

[9] A. Trémeau, B. Fernando, S. Karaoglu, and D. Muselet. Detecting text in natural scenes based on a reduction of photometric effects: Problem of text detection. In *CCIW*, 2011.

[10] S. S. Tsai, D. Chen, H. Chen, C.-H. Hsu, K.-H. Kim, J. P. Singh, and B. Girod. Combining image and text features: a hybrid approach to mobile book spine recognition. In *ACM Multimedia*, 2011.

[11] S. Uchida, Y. Shigeyoshi, Y. Kunishige, and Y. Feng. A keypoint-based approach toward scenery character detection. In *ICDAR*, 2011.

[12] J. van de Weijer, T. Gevers, and A. D. Bagdanov. Boosting color saliency in image feature detection. *TPAMI*, 2006.

[13] J. C. van Gemert. Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In *ICMR*, 2011.

[14] G. Wang, D. Hoiem, and D. A. Forsyth. Building text features for object image classification. In *CVPR*, 2009.

[15] H. C. Wang and M. Pomplun. The attraction of visual attention to texts in real-world scenes. In *CogSci*, 2011.

[16] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In *ICCV*, 2011.

[17] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV*, 2012.

[18] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *CVPR*, 2012.

[19] Q. Zhu, M.-C. Yeh, and K.-T. Cheng. Multimodal fusion using learned text concepts for image categorization. In *ACM Multimedia*, 2006.