Object Reading: Text Recognition for Object Recognition

Sezer Karaoglu, Jan C. van Gemert, and Theo Gevers

Intelligent Systems Lab Amsterdam (ISLA), University of Amsterdam,
Science Park 904, 1098 XH, Amsterdam, The Netherlands

Abstract. We propose to use text recognition to aid in visual object class recognition. To this end we first propose a new algorithm for text detection in natural images. The proposed text detection is based on saliency cues and a context fusion step. The algorithm does not need any parameter tuning and can deal with varying imaging conditions. We evaluate three different tasks: 1. Scene text recognition, where we increase the state-of-the-art by 0.17 on the ICDAR 2003 dataset. 2. Saliency based object recognition, where we outperform other state-of-the-art saliency methods for object recognition on the PASCAL VOC 2011 dataset. 3. Object recognition with the aid of recognized text, where we are the first to report multi-modal results on the IMET set. Results show that text helps for object class recognition if the text is not uniquely coupled to individual object instances.

1 Introduction

Text in natural scenes typically adds semantic information to the scene. For example, text adds identification on the brand or type of a product, it specifies which buildings serve food, and gives directions by road signs. In this paper, we propose to exploit this semantic relationship between text and the scene to improve automatic object recognition where visual cues may not prove sufficient.

Traditional optical character recognition (OCR) systems are well-suited for documents however their performance drastically drops when applied to natural scene images. The challenges for OCR include an unknown background and varying text sizes, styles and orientations. Moreover, the imaging conditions are often unknown, which adds sensitivity to specular reflections, shadows, occlusion, (motion) blur and resolution. To overcome such issues, we propose to rely on visual saliency for detecting text regions in the image. Color invariants and curvature saliency allows robustness to imaging conditions, whereas context information gathered by text distribution statistics addresses the background. Unlike state-of-the-art text detection methods [1, 2] we do not rely on any heuristics nor on a learning phase. Therefore, the proposed method can detect a-priori unseen text styles at any orientation and scale. Moreover, we show that the proposed saliency method is not limited to text, and can also be applied for generic object recognition.

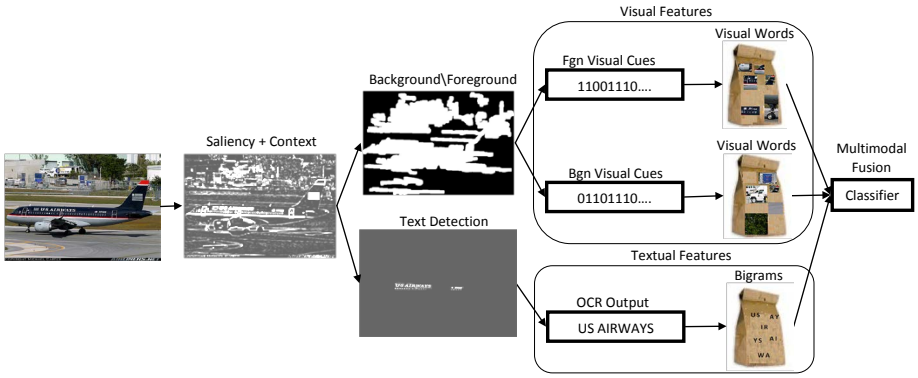

Fig. 1. Visual saliency is used for object recognition and for text detection. The detected text is converted to characters and words, which in turn are used in a bag-of-character-bigrams text classifier. This text classifier is combined with a bag-of-visual-words image representation

The contributions of this paper are twofold. First, we propose a novel saliency method with a context integration step that can be used both for text and object recognition. Second, we integrate saliency into a full end-to-end text detection and recognition pipeline that incorporates recognized text in the visual scene for image classification. In figure 1 we show an overview of our approach.

2 Related Work

Learning-based approaches for classifying text regions [1, 3, 4] typically use Histogram of Gradients (HOG), edge density or contrast features. Such features are computationally expensive since they require an exhaustive scale and spatial search. Moreover, the necessity of a learning phase makes these systems language dependent. They require a new classifier for each different language or text style and are therefore not robust for unknown sets of text structures. Our approach does not require any learning and is compatible with any a-priori unknown languages and text styles.

Other approaches without a learning phase use heuristics to exploit the geometric and structural properties of text. Such properties are typically based on connected components (CC) and include a stroke width transform [2], component size or distribution of gradients. A threshold is typically used to classify regions as text. However, relying on thresholds makes these approaches sensitive to changing conditions such as text font, text orientation and unknown languages. In this paper we introduce a novel contextual guidance approach to prevent using any predefined thresholds.

Texts in natural scenes are typically designed to attract attention. The work of Judd et al. [5] show that scene texts receive many eye fixations and psychophysical experiments done by Wang and Pomplun [6] confirm that text is

(a) (b) (c) (d) (e)

Fig. 2. An example of the saliency maps. (a) Original, (b) Color Boosting, (c) Curvature Shape Saliency (d) Context guided Saliency and (e) Final output.

salient. Other work [7] compares different saliency methods to show that such methods may be used to locate scene text. All this research evaluates existing visual attention models for text detection to show that scene text is salient. In this paper, we introduce a new visual attention model to improve text detection.

The work done by Zhu et al. [8] is the most similar to ours. The authors exploit information from text regions in natural scenes to improve object/scene classification accuracy. The authors combine visual features extracted from the full image with features extracted only from detected text regions. We, in contrast, truly recognize the text from images rather than investigating the visual features extracted from text regions.

3 Saliency and Context Cues for Text Detection

We propose text detection based on saliency cues. Such cues prevent an exhaustive spatial and scale search over all possible regions. We focus on color and curvature saliency cues, and rely on context information to bind them.

Color Boosting Saliency Texts often have a uniform distinct color that makes them stand out from their surroundings. The distinctiveness of the local color information is commonly ignored by most of the saliency detectors. Previous methods [7, 9] mainly focus on obtaining gradient information only from the luminance channel. To add color to general image gradients, Van de Weijer et al. [10] use information theory to boost color information by replacing gradient strength with information content. Let $f_x = (O_{1x}, O_{2x}, O_{3x})^T$ be the spatial image derivatives in the x dimension where O_1 stands for the first opponent color channel, then information theory relates the information content I as $I(f_x) = -\log(p(f_x))$, where $p(f_x)$ is the probability of the spatial derivative. We integrate color boosting to enhance the saliency of text color transitions. In figure 2b we show an example of color boosting applied to texts.

Curvature Saliency Color boosting saliency is not sensitive to colorless edge transitions, see the top image in figure 2b. Consequently, we aim for a shape-based saliency approach in addition to color boosting saliency. To this end, we employ curvature saliency (D) as used in the work by Valenti et al. [11], defined by $D = \sqrt{f_{I,xx}^2 + f_{I,xy}^2 + f_{I,yy}^2}$, where $f_{I,xx}$ and $f_{I,yy}$ stand for the second-order derivatives of the intensity image $f_I(x, y)$ in the x and y dimensions, respectively. Due to the contrast between text and its background, text regions strongly respond to curvature saliency.

Contextual Priors We remove non-text regions which are mistakenly detected by color boosting and curvature saliency maps by contextual priming. We analyzed the spatial occurrence probability of text in natural scene images from the ICDAR 2003[1] text locating competition training set. The occurrence probability for a given location is calculated by counting the presence of text for that location in the full training set. The spatial occurrence probability of text in the ICDAR dataset shows that the regions in the center of the image space are most likely to be text, see figure 3. This may be due to the fact that interesting objects are tend to be placed in the center of the images by human photographers.

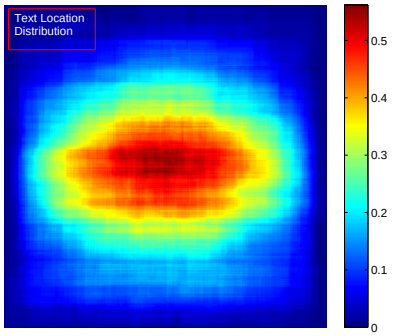

Fig. 3. Spatial occurrence probability of text in the ICDAR 2003 training dataset.

Since text is generally present in the image center, we base our contextual integration step on connectivity of the background. Specifically, we aim to connect all background regions with the help of the text occurrence probability map. We build a non-text map for each image, by connecting the non-text regions with conditional dilation. Conditional dilation (δ) is a morphological operation which can be defined as $\delta_B(\gamma|S) \triangleq (\gamma \oplus B) \cap S$, where B, γ and S are the structuring element, mask image and reference image (gray level input image), respectively.

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

Fig. 4. System output illustrating robustness to highlights and shadows.

The structuring dilation element is chosen as 3-by-3 square to preserve small characters. This operation is executed iteratively

$$\rho_B^-(\gamma|S) \triangleq \bigcup_{n=1}^{\infty} \delta_B^n(\gamma|S), \text{ until } \gamma_{n+1} = \delta_B(\gamma_n|S) \iff \gamma_n \neq \gamma_{n+1} \quad . \tag{1}$$

For the γ seed points we use the zero-probability spatial locations of the ICDAR 2003 training set as shown in figure 3. Conditional dilation uses simple subtraction operations to remove non-text regions from both the color and the curvature saliency maps. In this way, we avoid an exhaustive search for character recognition. In figure 2, we show an example of such contextual integration.

Feature Integration The color boosting and curvature maps are individually reconstructed with the help of conditional dilation. The maps are normalized to a fixed range $[0, 1]$ to make a linear combination possible. The linear weights of curvature and color boosting saliency for linear combination are empirically obtained as 3 and 1 respectively.

Noise Removal and Refinement For noise removal we follow Zhang and Kasturi [12] who divide the whole range of gradient orientation $[0, 2\pi)$ into 4 bins. The authors show that text is a closed region and contains all four gradient directions regardless of text type or style. The authors use only those pixels where the Canny edge detector finds an edge response, making it sensitive to the Canny edge detector response. To alleviate this, we use all CC regions. To extract CCs on the saliency maps, the saliency maps are converted into binary images by setting saliency values greater than zero to 1 and keep all others at 0.

The binarized and filtered saliency map is further processed to fill the holes within text regions and enlarged with morphological dilation. The conditional dilation is again applied to these enlarged regions to extract well-formed text masks. The zeros of our binarized saliency map are chosen as seeds of starting of equation 1 while the reference image S is chosen as gray level input image. At the end of the process we have one pair of text saliency map. One saliency map for text regions darker than their surroundings and another one for lighter text regions. The returned saliency maps are our final text detection output. For the system evaluation we feed the OCR system with these saliency maps. These outputs can be used for precise text localization or as input to other text recognition systems. In figure 4 we show an example of the proposed text detection method.

Method	Character	Word				
	Matched	Precision	Recall	F-Score	Time	Dictionary
TESS		0.35	0.18	0.27	**0.21s**	-
ABBYY	0.37	**0.71**	0.32	0.52	1.9s	-
Neumann and Matas'10 [3]	0.67	0.42	0.39	0.40	0.36s	-
Neumann and Matas'11 [13]		0.42	0.41	0.41	0.83s	-
Wang et al. [1]		0.45	0.54	0.51	15s	+
Our Method (Opponent)	0.21	0.47	0.24	0.36	7.2s	-
Our Method (Intensity)	0.62	0.64	0.52	**0.58**	7.2s	-
Our Method (Combination)	**0.68**	0.51	**0.59**	0.55	11.6s	-

Table 1. Character and word recognition results on ICDAR 2003 dataset

4 Results and Discussion

4.1 Character and Word Recognition Results

The text recognition is evaluated on the ICDAR Robust Reading Competition. We compare against five other pipelines: 1. Neumann and Matas [3]. 2. An extension of Neumann and Matas [13]. 3. A dictionary-based approach by Wang et al. [1]. 4. ABBYY FineReader[2], a leading commercial OCR engine. 5. An open source OCR engine Tessaract[3] (TESS). The output of our saliency map is given as input to an OCR system without any text localization. Although TESS can perform in real time, we choose ABBYY because of its superior accuracy. We follow [3] and evaluate precision, recall, f-score, time, and the use of a dictionary. In addition, we also compare character vs. word recognition.

We present our results in table 1. In this table we also evaluate the saliency maps in opponent color space and in intensity-only. The conversion from RGB color space to opponent color space produces low resolution at the chromatic channels. Therefore, connectivity within a character for opponent color space is sometimes broken by the conditional dilation. This may explain why the intensity based saliency method outperforms the opponent color based saliency map. Note that the combined results improve over the-state-of-the-art.

The method of Wang et al. [1] creates word list dictionaries for every image from the ground truth vocabulary and the authors refine their OCR results using the smallest edit distance to the created word lists. This helps them to reduce the number of false positives and also to increase the true positives for this system. However, creating these lists cannot be generalized beyond the training set. Our system does not use any word list, and improve precision and recall of [1] with 7% and 5%, respectively. The algorithm of Neumann and Matas [13] also does not use any word list. The results show that we increase their baseline score by 17%. Note that the accuracy of text recognition from our text segmentation

[2] http://finereader.abbyy.com
[3] http://code.google.com/p/tesseract-ocr/

masks is highly dependent on the OCR software. Even though the characters are precisely detected, the ABBYY OCR system could not always recognize the characters. Better recognition results can be obtained with a better OCR system.

4.2 Object Recognition Results

The proposed saliency method for text detection can be applied to generic object detection. To this end, the context guided saliency map is directly used for foreground/background separation. The scene borders are selected as seeds in equation 1. The full process is repeated in the same way as done for text detection, until the refinement step. We evaluate object recognition on the trainval classification set of the Pascal VOC[4] 2011. We report average precision scores as is commonly done for this set and repeat our experiment three times to obtain standard deviation scores to validate significance testing. As the classification baseline we use the popular bag of visual words (BOW) scheme. For features we use SIFT and the visual word vocabulary size is 4000 which is created per training set for each of the 3 splits with K-means on $500K$ randomly sampled SIFT features. For classification, we use the histogram intersection kernel in libsvm and use its built in cross-validation method to tune the C parameter.

We compare our saliency method in an object recognition task against three state of the art saliency detection algorithms. One is the curvature shape saliency method [11]. The second saliency detector is based on frequency, color and luminance and outperforms other approaches in an object segmentation task [14]. The last saliency method is a recent method based on multi-scale contrast, center surround histogram, and color spatial distributions [15].

To incorporate the saliency methods in the BOW scheme we follow the generalized spatial pyramid scheme by Van Gemert [16]. The traditional spatial pyramid [17] builds correspondences by quantizing the absolute position of an image feature in disjoint equivalence classes. The generalized spatial pyramid scheme of [16] allows us to use the same approach, only for more general equivalence classes based on saliency. We quantize the saliency-score in disjoint equivalence classes, and create correspondences between these equivalence classes. If saliency values range between $[0, \ldots, 1)$ then an example of two disjoint equivalence classes (low and high saliency) is given by $\left\{[0, \ldots, \frac{1}{2}), [\frac{1}{2}, \ldots, 1)\right\}$. Each image feature can be assigned to an equivalence class depending on its average amount of saliency. Note that we create equivalence classes based on saliency, keeping non-salient regions in a class of their own. Hence, our approach retains features that are non-salient. If they are consistently non-salient within an object category, such features will still aid in object classification.

We show the classification results per category in figure 5c. Since we use the scene borders as seeds our method does not work as well for object that touch the border, such as *bus*, *diningtable* and *chair*. We do best for categories where objects are strongly edged, and do not touch the border such as *aeroplane*, *bird*, *bicycle*, *cat*, *motorbike* and *tvmonitor*. The mean average precision scores

[4] http://pascallin.ecs.soton.ac.uk/challenges/VOC/

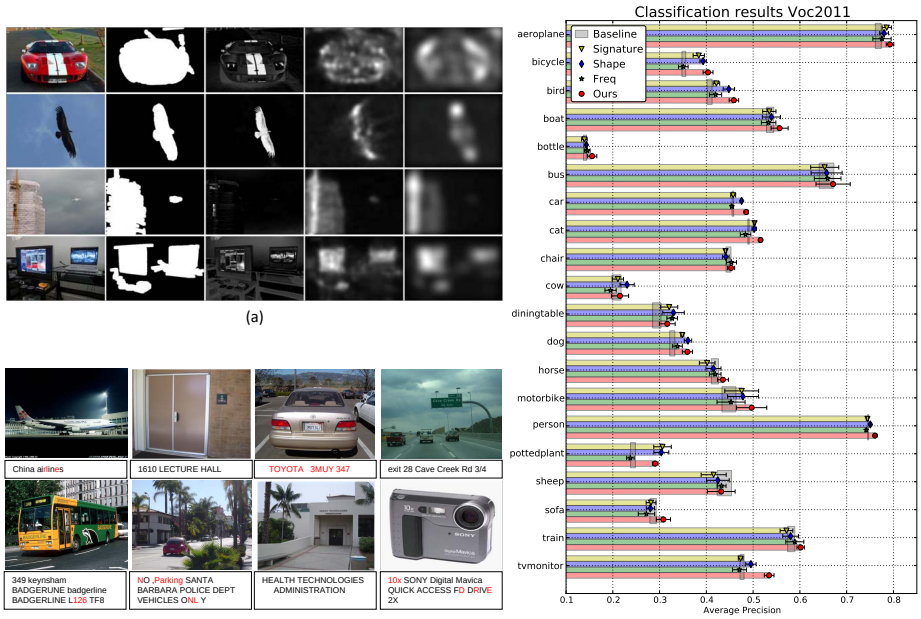

Fig. 5. (a) Original image and four saliency methods: 1. our proposed saliency, 2. Frequency based [14], Shape saliency [11] and Signature saliency [15]. (b) IMET dataset original images and OCR output; red chars are not recognized. (c) Classification results on the Pascal VOC 2011 for the BOW baseline and the four saliency methods.

in increasing order are: baseline: 0.429, Freq: 0.438, Signature: 0.443, Shape: 0.452, and ours: 0.462. Hence, our saliency method, which is designed with text detection in mind, outperforms other state-of-the-art saliency methods for object recognition, generalizing beyond the domain of text detection.

4.3 Multi-modal Object Recognition Results

We evaluate multi-modal object recognition based on visual features fused with text recognition on the IMET dataset [8]. This dataset consists of around 2,000 images, in 8 categories. The image set contains several natural images with text somewhere in them, see figure 5(b) for some examples. As the baseline we again use the BOW model with SIFT features. For the text recognition we use a bag-of-bigrams. We again use a histogram intersection kernel and tune the C parameter with cross-validation. We fuse the visual kernel with the OCR kernel by a simple addition. We identify two setups for text recognition: 1. the most realistic, where we feed our saliency output to the OCR; 2. feed the saliency output of the ground truth boxes to the OCR. This allows us to evaluate the effect of correct text localization.

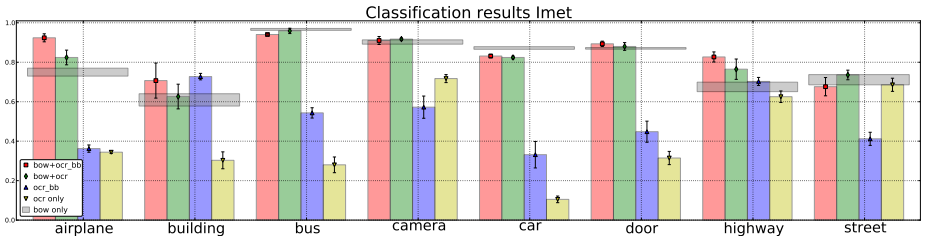

Fig. 6. Image classification results on the IMET dataset for the BOW baseline, and the four OCR methods; ocr: text only, with our saliency output direct to the OCR ocr_bb: text only, with ground truth bounding boxes over the saliency output, bow+ocr: visual features fused with ocr features, bow+ocr_bb: visual features fused with ocr_bb.

The classification scores per category are given in figure 6. The text-only results perform the worst. Of the text-only results the OCR with the ground truth bounding box over our saliency map is better than directly feeding the saliency map to the OCR. Only in the categories *camera* and *street* is the raw OCR better. This is because there are more actual text regions in the image than given by the ground truth. Generally, the results show that adding text to visual features helps. Only for *car* the results decrease, because the unique text on one license plate text does not generalize to other license plates. The MAP scores, in increasing accuracy are ocr: 0.422, ocr_bb: 0.512, bow: 0.779, bow+ocr: 0.816, bow+ocr_bb: 0.839. The dataset is well suited for visual information as indicated by the high baseline BOW score. However, we improve the visual-only score by 6% with fusing text information.
As a possible application we show in figure 7 our system's output on a Google image of the class *building*, where the text adds significant semantic information.

Fig. 7. Yellow dashes and boxes denote our output. The red boxes are not recognized.

5 Conclusion

We propose to use text in natural images to aid visual classification. To detect the text in a natural image, we propose a new saliency method that is based on low-level cues and novel contextual information integration. We show that this saliency method outperforms the state-of-the-art end-to-end scene text recognition scores with 0.17 on standard ICDAR 2003 datatset. Moreover, our saliency method outperforms other state-of-the art saliency methods for object recognition in the PASCAL VOC 2011 dataset. On the IMET dataset, we show that text recognition from natural scene images helps object classification if there is text in the image, and if the text is not too specific to a single object.

References

1. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: ICCV. (2011)
2. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: CVPR. (2010) 2963–2970
3. Neumann, L., Matas, J.: A method for text localization and recognition in real-world images. In: ACCV'10. (2011) 770–783
4. Wang, K., Rescorla, E., Shacham, H., Belongie, S.: Openscan: A fully transparent optical scan voting system. In: Electronic Voting Technology Workshop. (2010)
5. Judd, T., Ehinger, K.A., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV. (2009) 2106–2113
6. Wang, H.C., Pomplun, M.: The attraction of visual attention to texts in real-world scenes. In: CogSci2011. (2011)
7. Shahab, A., Shafait, F., Dengel, A., Uchida, S.: How salient is scene text? In: IAPR International Workshop on Document Analysis Systems. (2012)
8. Zhu, Q., Yeh, M.C., Cheng, K.T.: Multimodal fusion using learned text concepts for image categorization. In: ACM MM. (2006)
9. Uchida, S., Shigeyoshi, Y., Kunishige, Y., Feng, Y.: A keypoint-based approach toward scenery character detection. In: ICDAR. (2011) 819–823
10. van de Weijer, J., Gevers, T., Bagdanov, A.D.: Boosting color saliency in image feature detection. TPAMI **28** (2006) 150–156
11. Valenti, R., Sebe, N., Gevers, T.: Image saliency by isocentric curvedness and color. In: ICCV. (2009) 2185–2192
12. Zhang, J., Kasturi, R.: Text detection using edge gradient and graph spectrum. In: ICPR. (2010) 3979–3982
13. Neumann, L., Matas, J.: Text localization in real-world images using efficiently pruned exhaustive search. In: ICDAR. (2011) 687–691
14. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned Salient Region Detection. In: CVPR. (2009)
15. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. TPAMI **33** (2011)
16. van Gemert, J.C.: Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In: ICMR. (2011)
17. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)