Con-Text: Text Detection for Fine-Grained Object Classification

Sezer Karaoglu, Ran Tao, Jan C. van Gemert, and Theo Gevers, *Member, IEEE*

***Abstract*— This paper focuses on fine-grained object classification using recognized scene text in natural images. While the state-of-the-art relies on visual cues only, this paper is the first work which proposes to combine textual and visual cues. Another novelty is the textual cue extraction. Unlike the state-of-the-art text detection methods, we focus more on the background instead of text regions. Once text regions are detected, they are further processed by two methods to perform text recognition, i.e., ABBYY commercial OCR engine and a state-of-the-art character recognition algorithm. Then, to perform textual cue encoding, bi- and trigrams are formed between the recognized characters by considering the proposed spatial pairwise constraints. Finally, extracted visual and textual cues are combined for fine-grained classification. The proposed method is validated on four publicly available data sets: ICDAR03, ICDAR13, *Con-Text*, and *Flickr-logo*. We improve the state-of-the-art end-to-end character recognition by a large margin of 15% on ICDAR03. We show that textual cues are useful in addition to visual cues for fine-grained classification. We show that textual cues are also useful for logo retrieval. Adding textual cues outperforms visual- and textual-only in fine-grained classification (70.7% to 60.3%) and logo retrieval (57.4% to 54.8%).**

***Index Terms*— Multimodal fusion, fine-grained classification, logo-retrieval, text detection, text saliency.**

I. Introduction

MANY existing object recognition methods are focused on distinguishing definite objects such as horses, bicycles and cars [6]. Object recognition results obtained for different benchmarks, e.g., Pascal VOC, show that there has been a significant progress to recognize these "distinct" object categories. However, the performance of these methods may deteriorate to distinguish categories of objects that only slightly differ in appearance, such tasks include fine-grained classification. Fine-grained classification is the problem of assigning images to sub-ordinate classes in which objects differ only in (subtle) details (e.g. flower types [28], bird species [59]). Although, visual cues (e.g. color, texture and shape) can be used to distinguish visually distinct objects, the same visual cues may lack discriminative power to differentiate object types of similar appearance. Therefore, existing fine-grained classification approaches increase the discriminative power of these visual cues by exploiting part information or, implying geometrical constraints [8], [9].

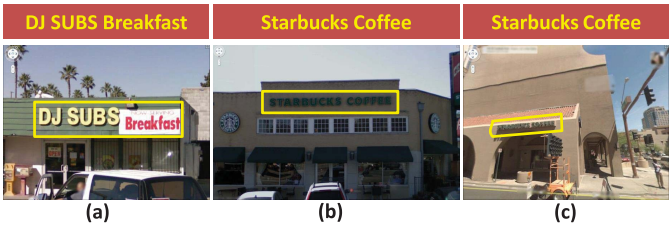


Fig. 1. An example of fine-grained *Building* classification [2]. Visual cues would group (a)-(b) whereas scene text reveals the semantics and clusters (b)-(c).

In this paper, we address the problem of fine-grained object classification by combining textual and visual cues. In particular, we focus on the classification of *Buildings* into their sub-classes such as *Cafe*, *Tavern*, *Diner*, etc. The reason to use textual cues for such task is that text adds semantics beyond visual cues. For instance, in Fig. 1, the aim is to classify the three images based on their semantics. In this case, visual cues are not sufficient or even misleading as the first two images have similar scene appearances. Textual cues are useful to recognize that the two (right) images belong to the same category since they contain the same brand name *Starbucks*. Therefore, we propose to use both textual and visual cues. Further, we propose a method for textual cue extraction. The success of the proposed fine-grained object classification method (fusion of visual and text modalities) highly depends on the completeness of the extracted textual image cues. Therefore, a robust character localization and a textual cue encoding method is proposed.

The state-of-the-art text detection methods [3], [4], [22], [39], [43], [54], [55], [66], [69] extract geometric, structural and appearance properties from candidate text regions which are obtained using a connected component or sliding window approach. These regions are further verified using the extracted features if they contain text or not. In contrast, we focus on the scene background (non-text regions) rather than text regions. Our motivation is that the majority of the scenes consist of background pixels e.g. 93% of ICDAR13.

Manuscript received February 20, 2016; revised October 18, 2016; accepted May 11, 2017. Date of publication May 24, 2017; date of current version June 13, 2017. This work was supported by the Dutch National Program and in part by COMMIT. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Weisi Lin. *(Corresponding author: Sezer Karaoglu.)*

R. Tao is with the Intelligent Sensory Information Systems Lab, University of Amsterdam, 1098 XH Amsterdam, The Netherlands (e-mail: r.tao@uva.nl).

S. Karaoglu is with the Computer Vision Laboratory, University of Amsterdam, 1098 XH Amsterdam, The Netherlands (e-mail: s.karaoglu@uva.nl).

J. C. van Gemert is with Computer Vision Laboratory, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: j.c.vangemert@tudelft.nl).

T. Gevers is with Computer Vision Laboratory, University of Amsterdam, 1098 XH Amsterdam, The Netherlands, and also with the Computer Vision Center, Universitat Autònoma de Barcelona, 08193 Barcelona, Spain (e-mail: th.gevers@uva.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2707805

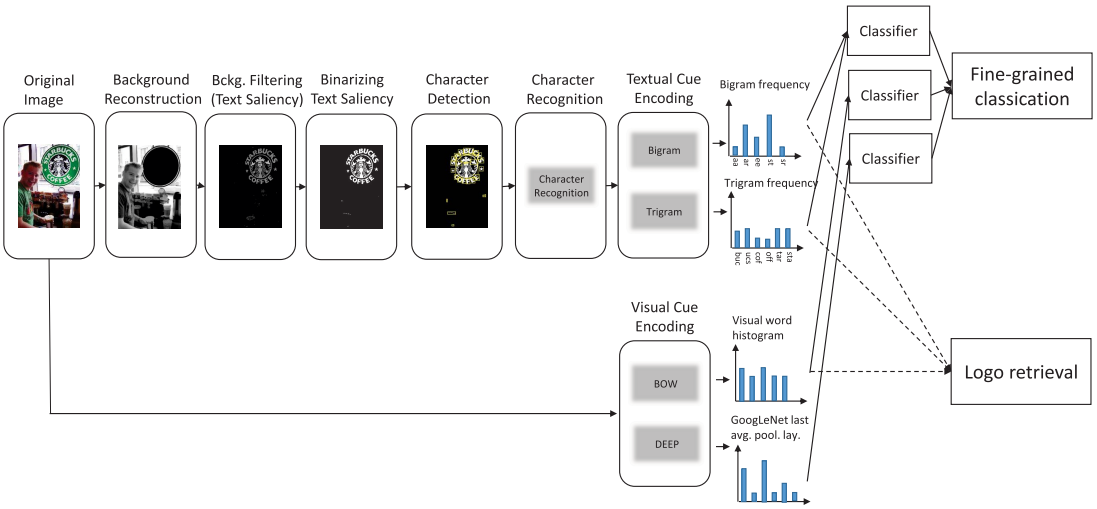


Fig. 2. The flow of the proposed method. We propose a generic, efficient and fully unsupervised text detection algorithm by eliminating scene background. Detected text regions are processed by a state-of-the-art character recognition algorithm. Then, bi- and trigrams (textual cues) are formed between the recognized characters by considering spatial pairwise constraints. Extracted textual and visual cues are combined for fine-grained classification and logo retrieval.

Moreover, background pixels are mostly homogenous within themselves e.g. fences, boards, roads, buildings, windows etc and highly contrasted with text regions. Focusing on eliminating background pixels using background connectivity rather than directly detecting text regions benefits from eliminating larger number of non-text regions at a reduced risk of eliminating true text regions (due to contrast). Moreover, since the proposed method does not extract text specific features, it does not require any tuning for varying text size, style and orientation. Furthermore, the proposed method is complementary to previous approaches, since it benefits more from background while previous approaches do not.

Text is designed to attract human attention. This has been verified by Judd et al. [50], and Wang et al. [45] who show that text features are more discriminative than generic object features and scene text receives human attention more than other generic objects. Accordingly, in this paper we consider salient and non-salient regions as text and background regions respectively. The proposed method initially selects background (non-text regions) seeds using color and curvature saliency and spatial context information. Then, it detects the background from these seeds based on background connectivity. Once the background has been detected and eliminated, text regions are identified. The detected character candidates are further processed by two methods to perform text recognition i.e. ABBYY commercial OCR machine and state-of-the-art character recognition algorithm [54]. Then, spatial pairwise constraints between character candidates are used to obtain textual cue representations. Finally, extracted textual cues are used in combination with visual cues for fine-grained classification and logo retrieval. The pipeline of the proposed method is summarized in Fig. 2.

The paper has six main contributions:

- We propose a generic and computationally efficient character detection algorithm without any training involved. Unlike the state-of-the-art text detection methods which try to detect scene text directly, the proposed method detects the background to infer the location of text.
- We experimentally show that removing scene background reduces clutter and subsequently improves the character recognition performance of standard OCR systems. Moreover, removing the background reduces the search space allowing the extraction of computationally expensive features for character recognition.
- We propose a fine-grained classification approach which combines textual and visual cues to distinguish objects. To the best of our knowledge, this is the first approach to combine textual and visual cues of objects in images.
- We are the first to combine textual and visual cues for logo retrieval in natural scene images.
- We propose to constrain textual cues by spatial information. We show that encoding textual cues with proposed constraints is superior than without these constraints.
- The introduced dataset, extracted features (textual and visual) and text detection code are publicly available.[1]

The previous versions of this paper appeared in [1] and [2]. We extended our previous studies in different aspects: 1) In [1] and [2], there are no spatial constraints to form textual cues. Characters, ‘P’, ‘E’, ‘A’, ‘T’, can be combined into ‘TAPE’, ‘PATE’, or ‘PEAT’. It is not possible to distinguish between ‘TAPE’, ‘PATE’ and ‘PEAT’ by considering all possible bigram combinations. Hence, we propose to encode textual cues using spatial constraints. 2) Location information of the recognized characters are essential to perform spatially constrained encoding. However, ABBYY is used as a character recognition system in [1] and [2]. Character location information is not provided by ABBYY. This restricts to perform proposed encoding. Therefore, a different character recognition algorithm [54] is used. A method is proposed to generate (locate) character candidates which are used as an input to [54]. in this way, we have the spatial information

[1] https://staff.fnwi.uva.nl/s.karaoglu/datasetWeb/Dataset.html

of the recognized characters. 3) Our previous work considers only bigrams to encode textual cues. In this paper, various layers of textual cue encoding is performed (bi- and trigrams) and their effects are compared. 4) New analysis on the background removal and text saliency is provided on ICDAR13. 5) In depth analysis on textual and visual cues for fine-grained classification is provided. 6) Our previous work uses only BOW as visual features. Visual baseline is substantially improved using GoogLeNet features [23]. 7) Various fusion techniques are used to combine different modalities and their influences are discussed. 8) The proposed method is compared against state-of-the-art text detection methods [15], [22], [55] on text saliency, end-to-end character recognition and fine-grained classification. 9) The proposed method is applied on a new application 'Logo retrieval'.

II. Related Work

A. Text Detection

Text detection methods aim at automatically detecting and generating bounding boxes of words in natural scene images. Text detection methods can be categorized into two classes based on how they search character regions: a connected component [3], [22], [39], [66] and a sliding window approach [4], [43], [54]. Connected component approaches aim at segmenting characters using pixel similarities, e.g. contrast [1], stroke width [3] and intensity [41] whereas sliding window based approaches search the image over different scales and window sizes to locate character regions. For both methods, word candidates are detected by further verifying and combining the generated character candidates. To verify and combine character candidates, geometric, structural and appearance properties of text are derived from hand-crafted rules [3] or obtained by learning [4], [27], [40], [54], [61], [69].

All these methods extract geometric, structural and appearance features from candidate regions to verify if a region contains text or not. It is difficult to have one global parameter setting which would accommodate for all possible text variations in natural images [62]. Therefore, it is necessary to tune these parameters for every new alphabet, text style and size. In contrast, our approach focuses more on background connectivity rather than text regions. The proposed method does not extract text specific features. Therefore, it does not require any tuning for varying text size, style and orientation. In [73], background information is used to avoid parameter tuning for text binarization in document images. Additionally, state-of-the-art methods combine characters into words by a learning or a rule based approach. However, the information loss at these steps are irreversible. In contrast, we use characters instead of words to represent textual information in the images. Recently, [70] and [71] use similar ideas as in object proposals [29] but this time to generate a small set of word candidates. A reduced number of word candidates makes it possible to use more complex classifiers for word recognition. Such work can highly benefit from the proposed text detection method to reduce word box proposals even further.

B. Visual Saliency

The aim of visual saliency detection is to separate attention-driven regions and other regions (e.g. background) [30], [31]. In this way, the vast amount of incoming visual data (background) is eliminated. This helps to extract more reliable information because the background is eliminated. Therefore, it is widely used in image processing, for scene classification [47], [48], object recognition [1] and visual search [25].

Saliency for text detection has only recently received some attention [33], [34], [42], [44]. Text in natural scenes is typically designed to attract attention. In the experiment conducted by Judd et al. [50], it is shown that scene text fragments receive a high number of eye fixations (i.e. attention). Psychophysical experiments conducted by Wang et al. [45] show that regardless of the text position, text features are more discriminative than generic saliency features. Recently, Jiang et al. [35] provide a large scale dataset for visual saliency. The authors measure saliency by following the mouse-tracking behavior of users. Interestingly, they observe that even though scene text is not explicitly defined as a category in MS COCO [36], scene text consistently attracts human attention. Other work by Shahab et al. [33] compares different saliency methods and concludes that scene text is the most salient. Recently, [55] proposes a text detection method which relies on text saliency. The method uses a Bayesian framework and integrates visual cues tailored for text detection to obtain text saliency. All this research shows that text in natural images is salient and therefore we rely on non-salient regions as our primary cue for background detection.

Existing methods for text detection using visual saliency [33] mainly focus on bottom-up information such as edges, corners, color distinctiveness and lines of symmetry. Top-down models are task dependent and use saliency and context to steer the search for objects in images. This process is inspired by human focus-of-attention mechanisms. For instance, while searching for text, humans will focus on road signs, commercials and billboards rather than other areas [50]. Torralba [26] use context information to fixate locations of targeted objects (e.g. pedestrians). Our approach is inspired by Wei et al. [38]. The authors use boundary priors to steer the search for salient object detection. In our approach, background information is used for text detection.

C. Multimodal Fusion

The use of textual (e.g. captions, video OCR) and visual information for video classification has extensively been studied [74]–[76]. We refer to [13] and [14] for an overview. Moreover, textual in combination with visual cues have also been used for document image analysis [16]–[18]. Others [19], [20] combine visual and textual information to recognize logos and stamps in documents. However, the use of automatically extracted text information from natural images for scene classification has largely been ignored. Text is fused with visual cues for scene classification by Wang et al. [24]. The method uses Flickr images and their associated social tags. Others [32], propose to combine visual features extracted from the surroundings of text regions with visual features from the

full image. In contrast, we propose to use recognized text from images in combination with visual features.

Similar to our approach, [21] proposes to use scene text in combination with visual features to improve book spine recognition. However, our method focuses on combining textual and visual features for fine-grained classification and logo retrieval.

D. Fine-Grained Classification

Similar to our work, [79] focuses on fine-grained classification of street view storefronts. Although the authors emphasize on the importance of textual information for visual classification, they do not explicitly detect and recognize the text in the images to help classification. In hindsight, [79] shows that, with a few qualitative results, the network is able to make use of the text in the images for the classification task. Different from that work, we explicitly detect and encode textual information to aid visual-only classification.

E. Context for Object Categorization

Contextual information is used for object categorization when visual cues are not sufficient [10], [11]. Semantic, spatial and scale context are the most common types of context information. In this paper, we use semantic and spatial context. Text usually gives more information beyond what is obvious. It can be used to indicate directions (e.g. bus), warnings, types of service (e.g. cleaning) etc. The proposed method exploits this (semantic) information in addition to visual cues. The proposed method uses non-salient image regions as scene background. However, some of the background pixels may also be detected as salient. To eliminate these false detections, the proposed method restricts the search space of scene text with the use of spatial context (i.e. background priors).

F. Problem Statement

In this paper, we address the problem of fine-grained classification for objects which slightly differ in appearance by combining textual and visual cues. Textual cues are important for the success of the approach. The state-of-the-art text detection methods perform an explicit word detection and non-text filtering steps to obtain high precision. These steps cause information loss. Unlike other text detection methods, where f-score is considered more important, we aim for high recall character detection because missed characters cannot be recovered. However, we still need to limit the number of false detections because increasing the number of text candidates increases the computational time for character localization and textual cue encoding. Therefore, we propose a text saliency method which reduces the search space for character candidates by using background removal. With the proposed text saliency method, we reduce the search space for character candidates at a reduce risk of recall decrease.

III. Background Removal

Text can appear on unknown background with unknown text size, style and orientation in natural scene images. It is difficult to have one global parameter setting which would accommodate for all these variations in text [62]. Therefore, we tackle the problem of detecting text from a different point of view. Rather than asking "what is the property of scene text", we ask the question "what is the property of scene background". Keeping in mind that scene text is highly contrasted with background, answering this question would also reveal the location of scene text. As a result, the proposed method would not require any tuning for varying text size, style and orientation. Moreover, eliminating background to infer text location has additional benefits, *(i)* the search space is reduced allowing the extraction of computationally intensive features for character recognition, *(ii)* background clutter is removed reducing false text detections/recognitions.

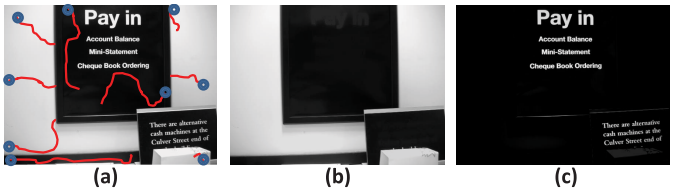


Fig. 3. (a) Original image and selected initial background seeds (blue dots). The connectivity path formed by these initial seeds is represented by the red dots (b) recovered background and (c) background removed image regions [2].

Background pixels are mostly homogeneous within themselves e.g. fences, boards, roads, buildings, windows etc.. This homogeneity allows defining connectivity between background pixels. Moreover, text is designed to attract attention [3], [22], [42], [45] and usually strongly contrasts with background. Hence, text boundaries usually correspond to strong intensity changes. Therefore, we propose to select initial background seeds from non-salient pixels and grow these seeds using connectivity of background. These seeds will grow until strong intensity changes are reached e.g., text/background transitions. Background seeds form connectivity between all pixels except those that belong to the text regions. An illustration of seed growing is shown in Fig. 3a. Blue dots represent initial background seeds whereas red lines represent the connectivity path formed by these initial seeds (blue dots).

To form connectivity between background pixels, we use conditional dilation (δ). Conditional dilation is a morphological operation where the dilation of a marker image is conditioned by a mask image (I). In this work, we use the image consisting of only background seeds as marker image γ and gray-level image as mask image. The conditioning is performed by intersecting dilated marker image with mask image, described as:

$$\delta_I(\gamma) = (\gamma \oplus S) \wedge I, \tag{1}$$

where $\oplus$, S and $\wedge$ denote for the dilation operation, the structuring element (3-by-3 square), and the element-wise minimum respectively. The initial background seeds, γ_0, are used to reconstruct the background image (ρ) of image I. The steps are described in algorithm 1.

The selection of background seeds γ is essential for background detection process. The proposed method selects the seeds based on saliency and scene text priors as explained in the next section.

Algorithm 1 Background Detection ρ

Input: Gray-level image I, the image consisting of the initial background seeds γ_0
Output: the reconstructed background image ρ

```
n = 1;
while true do
    γ_n = δ_I(γ_{n-1}) (Note that δ is the conditional dilation, which is explained in eq. 1);
    if γ_n == γ_{n-1} then
        break;
    end if
    n = n + 1;
end while
ρ = γ_n;
```

A. Background Seed Selection

1) Color Saliency: In general, it can be assumed that color is homogeneous for many background regions such as roads, sky, buildings and so on. Moreover, color edges correlate with high contrasted text fragments. To exploit this, we propose to detect color edges by color boosting algorithm [51]. The method uses information theory to correlate gradient strength with information content.

To be precise, let $f_{o,x} = (O_{1x}, O_{2x}, O_{3x})^T$ be the spatial image derivatives in the x dimension where O_1, O_2, and O_3 stand for the opponent color channels. The information theory relates the information content of an event to its frequency or probability

$$I(f_{o,x}) = -\log(p(f_{o,x})), \tag{2}$$

where I is the amount of information and $p(f_x)$ is the probability of the spatial x-derivative. According to information theory, rare events are more informative (i.e. higher information content). Consequently, [51] proposes to focus on rare color derivatives. A color saliency boosting function g is used to transform the vectors with equal information content to have equal influence on the saliency map. The distribution of image derivatives in opponent color space is characterized by a covariance matrix M. Eigenvector matrix U and an eigenvalue matrix V are obtained by the decomposition of matrix M. Then, the color saliency boosting function g is obtained by

$$g(f_{o,x}) = V^{-1}U^T f_{o,x}. \tag{3}$$

Once g is determined, the color boosting saliency (S_c) is expressed by

$$S_c = H(g(f_{o,x}), g(f_{o,y})), \tag{4}$$

where H is the saliency function. Color boosting approach is used to enhance the saliency of colorful text/background transitions and to suppress background regions. In Fig. 4b, an example is shown of applying color boosting.

2) Curvature Saliency: Obviously, the color saliency measure is inappropriate for colorless edge transitions, see the top image in Fig. 4b. Consequently, in addition to color saliency, we aim for a shape-based saliency measure. To this end, we define curvature saliency (L) by

$$L = \sqrt{f_{I,xx}^2 + f_{I,xy}^2 + f_{I,yy}^2}, \tag{5}$$

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 contrast between text and its background, text regions result in high responses to curvature saliency even for colorless edge transitions.

3) Spatial Context: Spatial context is described by the likelihood of finding an object in a certain position [10]. It has been shown to be beneficial to distinguish objects in the scene [10]. The proposed method also uses spatial context. To this end, text location priors are used to obtain background location priors. The proposed method treats background and text pixels as figure/ground pixels. To this end, the location occurrence probability of text is computed for the ICDAR03 [57] image training set. The occurrence probability for a given location is computed by counting the frequency of text for that particular location for the full training set. The text-location occurrences shows that text regions are more in the center of the image, see Fig. 5. Text can also be placed off the center, however they rarely touch image borders. Image borders usually consist of background such as sky, road and grass [10].

Fig. 4. An example of saliency maps: (a) Original Images, (b) Color Boosting and (c) Curvature Shape Saliency. It is shown that colorful edge transitions are emphasized by color saliency while colorless edge transitions are emphasized by curvature saliency.

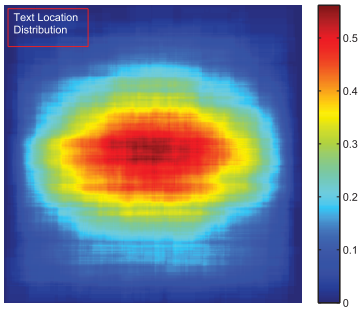


Fig. 5. Location occurrence probability of text in the ICDAR03 training dataset [1]. The probability distributions show that text rarely occurs at the image borders.

Fig. 6. Original images and text saliency maps obtained. Most of the background regions are filtered out while text regions are preserved by the proposed method. The proposed method is robust against photometric changes e.g., shadow and highlights, text size, style and orientation.

The ICDAR03 dataset mostly consists of images where text is in focus. However, the text location prior also holds for other datasets. Recently, a text detection dataset which consists of 67K images is collected by [37]. The images are from complex everyday scenes. Unlike ICDAR datasets, the images were not collected with text in mind and thus they contain a broad variety of text instances [37]. The authors compared the text distributions of the new dataset and the existing ICDAR datasets. They show that the Coco-Text dataset has a more uniform text distribution. However, image borders rarely contain text. Therefore, our observation about image borders that they usually contain non-text pixels also holds for the Coco-Text dataset (uncontrolled text detection dataset). To this end, pixels at the image borders are used as initial background seeds. Salient regions which are connected to image borders in color and curvature saliency maps are suppressed using algorithm 1. The original color and curvature saliency maps are used as intensity images.

B. Background Detection and Text Saliency

The refined color and curvature saliency maps are normalized to a fixed range $[0, 1]$ and linearly combined. Regions which do not have any response on this combined saliency map are considered as final background seeds. The background of the input image is constructed using these final background seeds using algorithm 1. Text saliency map is obtained by subtracting the background from the input image. The proposed method outputs a text saliency map which provides information about how likely a region contains text (See Fig. 6 for an illustration). This saliency map is further processed to extract textual cues.

IV. Textual Cue Encoding

Text saliency obtained in Section III is used to perform character recognition to extract textual cues. Two different approaches are used to perform character recognition. First, text saliency map is directly fed into leading commercial OCR engine (ABBYY). Second, text saliency map is further processed to obtain character candidates. Then, character candidates are fed into a state-of-the-art text recognition algorithm [54].

A. Character Recognizer - ABBYY

We first use ABBYY, leading commercial OCR engine, to perform character recognition on the text saliency. ABBYY receives an image as input and outputs recognized characters within that image. The gray level text saliency map is used as an input to ABBYY (no binarization is required). The recognized characters by ABBYY are directly used for textual cue encoding. The output of recognized characters are used to form bi- and trigrams without considering their spatial relations.

B. Character Recognizer - Flexible

Location information of the recognized characters cannot be obtained from ABBYY which restricts to perform spatially constrained textual cue encoding. Moreover, remarkable performance improvements on character recognition algorithms allows to perform more reliable textual cue encoding. Therefore, in addition to ABBYY, we use state-of-the-art character recognition algorithm [54] to perform textual cue encoding.

1) Character Localization: The proposed method generates a gray level text saliency map. To extract textual cues, we need to know where the text is. However, the generated saliency map does not provide bounding boxes to explicitly locate character areas. Therefore, a character localization method is applied on text saliency map.

A segmentation (binarization) based algorithm is used to locate text regions. We use the binarization algorithm from our earlier work [49]. The method is efficient and proven to work well for natural scene images. The method uses statistics of image intensities. It models the distribution by a generalized extreme value theory to determine the threshold for binarization. We refer [49] for the details of the algorithm. Each connected component generated after binarization process is considered as a character candidate.

The proposed method is not restricted to a segmentation algorithm. A sliding-window approach can also be used to

locate text with an exhaustive search in spatial and scale space. However, a sliding-window approach generates a large number of candidate regions which need to be processed by a recognizer. This increases the computational time. Therefore, a segmentation based approach is used to locate text.

2) Character Recognition: The character candidates generated by the approach outlined in Section IV-B.1 are used as input of a character recognition algorithm [54]. The method uses a four layer convolutional neural network. The network takes as input a gray-scale image, resizes it to 24×24 pixels, and normalizes the image by subtracting the patch mean divided by the standard deviation (i.e. whitening). The output is a probability $p(c|x)$ for the Latin alphabet, digits and a non-text class resulting in a total of 37 classes. These character probabilities are obtained by feeding last channels of the network into a soft-max manner.

3) Spatial Constraints: Recognized characters in Section IV-B2 are used to encode textual cues. However, the rich semantics of text cannot be fully conveyed by using only single characters. Character, 'P', 'E', 'A', 'T', can be combined into 'TAPE', 'PATE', or 'PEAT'. Therefore, rather than single characters, we propose to encode textual cues by forming bi- and trigrams. However, it is still not possible to distinguish between 'TAPE', 'PATE' and 'PEAT' by considering all possible bi- and trigram combinations of characters forming these words. Accordingly, we propose to encode textual cues with spatial constraints. For instance, recognized characters of the word 'TAPE' will form bigrams 'TA', 'AP', 'PE', 'TP' and 'AE' whereas the word 'PATE' will form 'PA', 'AT', 'TE', 'PT' and 'AE'.

To spatially constrain bi- and trigrams, we propose an approach similar to spatial pyramid [68] and Pyramidal Histogram Of Characters (PHOC) [81], [82]. Spatial pyramid encodes whether a particular visual feature appears in a specific spatial region of an image. PHOC aims the same to encode whether a particular character appears in a specific spatial region of a string. In contrast to these methods, we do not define exact image/string boundaries to localize information. We allow more general feature grouping by considering the following pairwise relations: *(i)* The ratio of the heights of two characters *(ii)* The angle between two character centers *(iii)* Euclidean distance between two consecutive character centers *(iv)* Shared area between two characters. Thresholds are set according to state-of-the-art text detection algorithms [3], [5], [53], [55] which form textlines between characters using the above constraints. The proposed method establishes connections between character candidates which satisfy the spatial constraints. These connections help to form bi- and trigrams. Then, textual representations are obtained by forming histograms of character combination occurrences. Each representation is independently normalized (L_1) and concatenated to obtain the final textual cue.

The proposed textual cue representation has certain benefits. First, it is not possible to distinguish anagrams (e.g. TAPE and PEAT) by considering all possible combinations of the characters forming the words. Therefore, the spatial constraints help to preserve the ordering of the characters. Second, the proposed method explicitly avoids word detection to extract

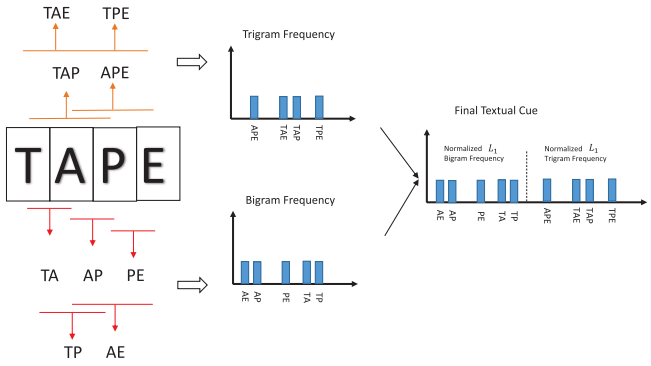


Fig. 7. An illustration on how the spatial constraints are used to encode textual cues. Each box represents a character candidate. These character candidates are used to form bi- and trigram representations. Bi- and trigrams are formed only when the spatial constraints are satisfied.

textual cues. The word formation step causes information loss due to introduced ad-hoc rules. If the connection between characters of the word is broken due to ad-hoc rules or missed characters, the proposed method can still extract textual cues. For instance, if the character "A" of "TAPE" is missed, word formation step would most likely not detect the word "TAPE". However, the proposed textual cue encoding would still encode "TP" and "TPE" (See Fig. 7 for an illustration).

V. Fusing Textual and Visual Information

Extracted textual and visual features are fused using three techniques, namely, early, late and kernel fusion. In this way, the relations between different modalities (i.e. text and visual features) are exploited at various levels of abstraction.

A. Early Fusion

Early fusion is performed at the feature level. After textual and visual features are extracted, the two modalities are concatenated in a single feature vector. Then, the Support Vector Machine (SVM) is used to classify this combined feature vector. Early fusion benefits from learning the regularities formed by the components independently based on different modalities. However, it restricts the choice of classifier and/or kernel to be the same for different modalities.

B. Late Fusion

Late fusion is performed at the decision level. After textual and visual features are extracted, SVM classifiers are trained on the unimodal features independently. A prediction is obtained for each modality. Then, these predictions are combined by averaging. In contrast to early fusion, various classifiers can be trained according to the modalities. On the other hand, late fusion does not exploit the feature level correlation among modalities. Moreover, the learning process for late fusion is more time consuming than early fusion since each modality requires training a different classifier.

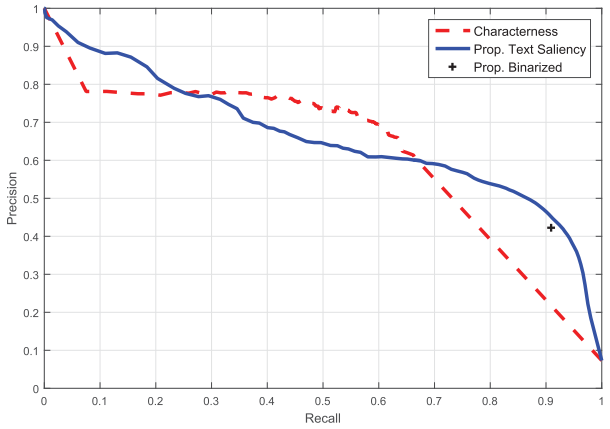


Fig. 8. Average precision-recall curves on ICDAR13 dataset. The proposed text saliency method (blue line) is fully unsupervised and achieves higher recall (after 0.7) with higher precision than Characterness [55] (red dashed line) which is a supervised text detection method. '+' represents the recall and precision values achieved after applying binarization [49] as described in Section IV-B.1.

C. Kernel Fusion

Kernel fusion is performed at the kernel level. After the textual and visual features are extracted, kernels are constructed on the unimodal features. Then, these kernels are combined by using the sum operation [56].

In this paper, we explore the above fusion techniques to exploit the relations between modality components for fine-grained image classification.

VI. Validation of Text Saliency

The proposed method provides a text saliency map. Reference [55] is similar to our approach because it generates a text saliency map too. Moreover, [55] is a supervised approach which outperforms other object saliency methods on text detection on ICDAR13 [58]. To demonstrate the effectiveness of the proposed method, we compare our results with [55]. The evaluation metric is used as in [55]. The pixel level annotations of the ICDAR13 are used as ground truth. The dataset consists of 229 images. For comparison reasons, the same set of randomly selected 100 images in [55] is used for testing.

A. Experiments and Results

We evaluate the effectiveness of our text saliency method based on the precision-recall (PR) curve as proposed by [55]. The generated text saliency map is normalized to the range of [0, 255]. Then, the PR curve for an image is computed by binarizing the saliency map using thresholds varying from 0 to 255. A full PR curve is generated by averaging PR curves over all test images.

The results are shown in Fig. 8. The proposed method (blue line) is fully unsupervised and achieves higher precision than [55] (red dashed line) at higher recall regions. We note that high recall character detection is crucial for textual cue extraction, since missed characters cannot be recovered. Consequently, missed characters also mean missed semantic information. We verify the importance of high recall by character recognition experiments in Section VII-B and fine-grained classification experiments in Section VIII-B (See Table II and Table V).

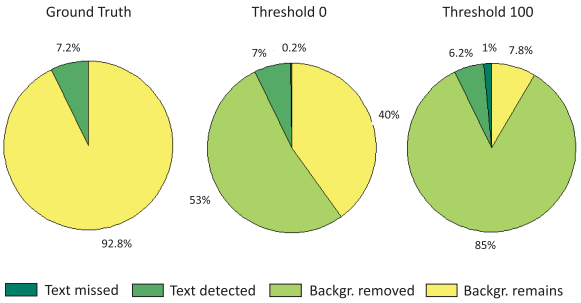


Fig. 9. The number of background and text pixels for the ICDAR13 dataset. The majority of the scenes consists of background pixels (93%). The text saliency map is normalized to the range of [0, 255]. Different thresholds indicate the number of text/background after the proposed method is applied. The results at threshold 0 shows that proposed method suppresses a large amount of background (53%) while preserving most of the text regions. The amount of suppressed background reaches (85%) whereas the amount of missed text detections is still reasonable at threshold 100.

To provide insights in the background removal process, we measure the total amount of background/text and the amount that remains after the proposed method has been applied. It is shown in Fig. 9 that 93% of the pixels of the overall images consists of background. We represented that the proposed method reaches higher precision values at higher recall regions compared to [55]. The reason is that the proposed method focuses on removing connected background pixels rather than the detection of text regions. The algorithm benefits from the background connectivity process to remove larger number of pixels. At the same time, the proposed method avoids to reach the text regions (due to contrast). Hence, the proposed method eliminates more background and misses a limited number of text pixels. Fig. 9 illustrates that a large number of background pixels are suppressed (53%) by the proposed method (at Threshold 0) while only a limited number of text pixels are missed (0.2%). The proposed method is suited as pre-processing step for text detection algorithms to reduce the search space (e.g. [70], [71]). The rest of the paper uses [49] to binarize the text saliency map. Systematically changing the threshold values for binarization is only used to obtain PR curve as in [55]. The proposed method achieves a recall of 90.95% and a precision of 42.35% when [49] is used for binarization.

VII. Character Recognition Evaluation

A. Dataset

End-to-end character recognition performance of our proposed method is evaluated on the publicly available ICDAR03 test dataset. The dataset consists of 5370 letters in 249 images. The dictionaries supplied by the dataset are not used to refine the recognition results.

B. Experiments and Results

We conduct two experiments. First, we evaluate the effect of background removal on the character recognition recall.

TABLE I

THE IMPACT OF BACKGROUND REMOVAL ON END-TO-END CHARACTER RECOGNITION ON ICDAR03. ABBYY IS A COMMERCIAL OCR ENGINE. *Orig.Image + ABBYY* USES ORIGINAL IMAGE WHEREAS *Proposed+ABBYY* USES TEXT SALIENCY MAP (BACKGROUND REMOVED INPUT IMAGES) AS INPUT FOR ABBYY CHARACTER RECOGNITION. THESE APPROACHES DIFFER FROM EACH OTHER DUE TO BACKGROUND REMOVAL. THIS SHOWS THAT ELIMINATING THE BACKGROUND FROM THE IMAGES INCREASES THE OCR ACCURACY

Method↓	Character Recognition Recall (%)
Orig. Image + ABBYY	37
Backg. Removed (Proposed) + ABBYY	**62**

TABLE II

END-TO-END CHARACTER RECOGNITION PERFORMANCE OF OUR PROPOSED METHODS AGAINST TEXT DETECTION METHODS [15], [22], [55], [78] ON ICDAR03. [54] IS USED FOR CHARACTER RECOGNITION FOR ALL THE METHODS. THE METHODS ONLY DIFFER IN CHARACTER DETECTION STEPS. THE RESULTS SHOW THAT PROPOSED METHOD SIGNIFICANTLY OUTPERFORMS [15], [22], [55], [78] ON END-TO-END CHARACTER RECOGNITION ON ICDAR03. *Proposed+ ABBYY* AND *Proposed+ [54]* DIFFER IN CHARACTER LOCALIZATION AND RECOGNITION STEPS. THE RESULTS SHOW THAT *Proposed+ [54]* OUTPERFORMS *Proposed+ABBYY*

Method↓	Character Recognition Recall (%)
Text Detection [15] + Text Recog. [54]	38
Text Detection [55] + Text Recog. [54]	53
Text Detection [22] + Text Recog. [54]	64
Text Detection [78] + Text Recog. [54]	64
Text Detection (Proposed) + Text Recog. (ABBYY)	62
Text Detection (Proposed) + Text Recog. [54]	**79**

Second, we compare the proposed method with state-of-the-art text detection methods [15], [22], [55]. The characters are detected using [15], [22], and [55]. Then, these characters are fed into character recognition algorithm [54]. We use character recognition recall as a performance measure to compare different methods. If a character has overlap greater than 0.5 with ground truth character and if it is recognized correctly, it is considered as true positive.

1) Experiment I: We compare *Proposed+ABBYY* against *Orig.Image+ABBYY*. For *Orig.Image+ABBYY* character recognition, the input image is directly fed into ABBYY without any processing. *Proposed+ABBYY* also uses ABBYY for character recognition. However, text saliency map is used as input. The results are shown in Table I. Although removing the background from the images may decrease the character detection recall in some cases, the experimental results show that background removal increases the overall character recognition recall of ABBYY by 25%. This is because traditional OCR systems are designed to work on documents with homogeneous backgrounds. Their performance deteriorates for natural scene images with cluttered and inhomogeneous background [72].

2) Experiment II: End-to-end character recognition results of our proposed methods are compared against state-of-the-art text detection methods [15], [22], [55]. Table II summarizes the results. We detect character candidates using state-of-the-art text detection algorithms [15], [22], [55], [78]. [54] is used to perform character recognition. End-to-end character recognition performance obtained using character candidates of the proposed method significantly outperforms (15%) other methods. References [15], [22], and [55] focus on balancing precision and recall. The authors introduce additional rules or learning-based steps to filter character candidates to increase the precision. Targeting a higher precision results in an increase of missed characters and negatively influences the subsequent text recognition process. This explains the reason why *([15], [22], [54])+ [54]* have lower end-to-end character recognition performance than *Proposed+ [54]*. The proposed method differs from traditional text detection approaches by having different tasks (textual cue extraction for fine-grained classification and logo retrieval) which requires different optimization methods. In this paper, we design a text detection method aiming at preserving complete textual information. The comparisons are made to validate our argument about not forming word candidates because the information loss during the word formation steps might be irreversible.

MSERs [78] is a widely used approach to obtain connected components for text detection task. Most of the referred algorithms are based on MSERs, however the additional word formation step makes direct comparison with MSERs not possible. Therefore, to directly compare the quality of the proposed text detection method with MSERs, we have performed an additional experiment. MSERs outputs on a gray-level image are directly used as input to character recognition algorithm [54] (no post processing or character filtering applied). The *MSER+ [54]* achieves 64% character recognition recall, whereas our method achieves 79%. The result shows that the proposed method outperforms MSER for character detection.

Proposed+ABBYY and *Proposed+[54]* differ in character candidate generation and character recognition steps. *Proposed+ [54]* (described in Section IV-B) uses the state-of-the-art character recognition method [54] rather than ABBYY as in *Proposed+ABBYY* (described in Section IV-A). A significant improvement (17%) over *Proposed+ABBYY* is obtained. This indicates that *Proposed+ABBYY* is restricted by the accuracy of ABBYY. *Proposed+ [54]* reaches state-of-the-art end-to-end character recognition accuracy up to 79% on this dataset.

3) Efficiency: The experiments are conducted on a laptop (Intel(R) Core(TM) i7-4810MQ Processor (2.80 GHz)) using Matlab. To run the method on a 480×640 resolution image takes 0.1s for each color and curvature saliency map extractions, 0.12s for recovering the background, 0.1s for binarization and 0.001s for each character candidate recognition.

VIII. FINE-GRAINED CLASSIFICATION

A. Dataset and Implementation Details

1) Dataset: We use sub-classes of the ImageNet[2] *building* and *place of business* sets to evaluate our fine-grained classification based on text and visual information. The dataset consists of 28 categories with 24,255 images, see Fig. 10 for the list of categories. In the experiments, we use all images from these categories. Note that many images may

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

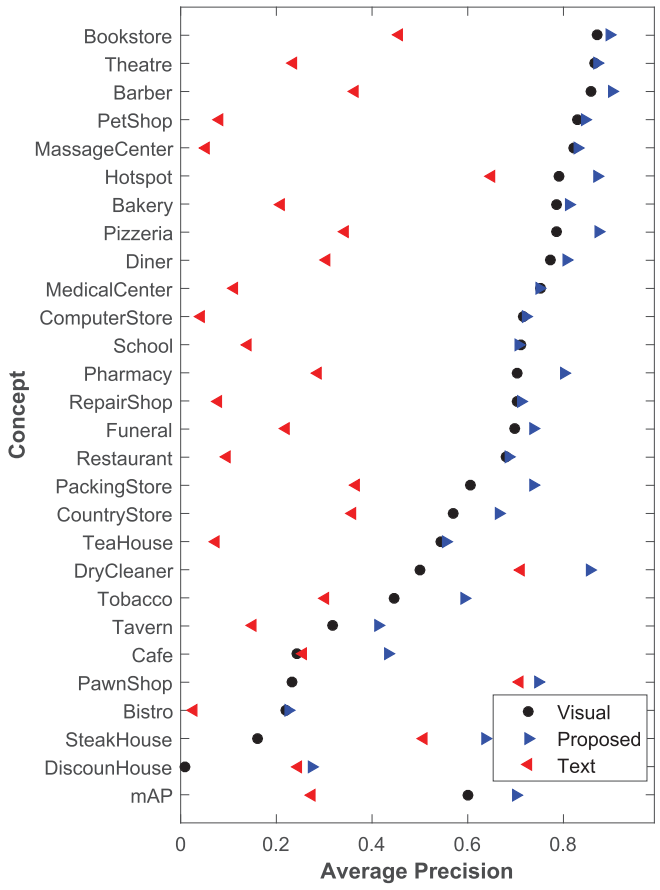


Fig. 10. Fine-grained classification results on each class for visual-only (DEEP-FT), textual-only and proposed method. The visual mAP is 60.3±0.2, text is 28.4±1.7 and proposed is 70.7±0.6. Adding textual cues significantly outperforms (10.4%) visual-only results.

not necessarily contain scene text fragments. The number of images that contain text varies e.g. *Bistro* (21%) and *Dry Cleaner* (89%).

2) Implementation Notes: We use average precision as the performance metric. We repeat all experiments three times to obtain standard deviation scores to validate the significance of the results. We use three types of features as visual baseline *1.* a standard bag of visual words (BOW) approach with SIFT using 4000 words with 1×3 and 2×2 spatial pyramid, *2.* we use a pre-trained model of GoogLeNet [23]. The GoogLeNet is trained on ILSVRC 2012, provided by the Caffe library.[3] We use the last average pooling layer as features (1024-dim), which are further normalized to unit lengths (DEEP), *3.* GoogLeNet is fine-tuned with a 28-way softmax classifier. The learning rate is 0.001, decreased by a factor of 10 every 5 epochs. The weight decay is set to 0.0005 and the momentum is 0.9. The network is fine-tuned for 20 epochs (DEEP-FT). We use the histogram intersection (for BOW) and linear (for DEEP and DEEP-FT) kernels in Libsvm and use its default value for the C parameter (=1) without any tuning. For text classification, we use the proposed method based on unconstrained and spatially constrained bi- and trigrams with the histogram intersection kernel with the same settings. The

[3]https://github.com/BVLC/caffe/tree/master/models/

TABLE III

THE INFLUENCE OF DIFFERENT TEXTUAL CUE ENCODINGS ON FINE-GRAINED CLASSIFICATION PERFORMANCE (mAP). ENCODING TEXTUAL CUES AS bi- OR TRIGRAMS PRODUCE SIMILAR RESULTS. ADDING PROPOSED SPATIAL CONSTRAINS ON TEXTUAL CUE ENCODINGS SIGNIFICANTLY OUTPERFORMS THE VERSION WITHOUT SPATIAL CONSTRAINS. COMBINING bi- AND TRIGRAM REPRESENTATIONS OUTPERFORMS EACH INDIVIDUAL REPRESENTATION

Textual Cue Encoding↓	Performance (mAP%)
unconstrained bigrams	13.1
unconstrained trigrams	12.6
Proposed spatially constrained bigrams	24.8
Proposed spatially constrained trigrams	24.0
Proposed spatially constrained [bi+tri]grams	**28.4**

visual and textual modalities are fused using *Kernel Fusion* described in section V.

B. Experiments and Results

We perform seven experiments. First, we evaluate the influence of different textual cue encodings on the final classification performance. Second, we quantify the impact of textual cues for fine-grained classification. Third, we compare our results against state-of-the-art text detection methods for textual cue extraction. Fourth, we assess the influence of three different fusion strategies. Fifth, we quantify the complementarity of the extracted features. Sixth, we quantify the impact of local and global visual cues and, their combinations with textual cue for fine-grained classification. Seventh, we discuss the influence of the performance change with respect to the amount of text in images.

1) Experiment I: We evaluate the following textual cue encodings: (1) unconstrained bigrams, (2) unconstrained trigrams (3) spatially constrained bigrams, (4) spatially constrained trigrams, and (5) spatially constrained bi- + trigrams. The results are summarized in Table III. Representing textual cues in terms of bi- or trigrams produce similar results. However, the combination outperforms each individual representation. Finally, the significant performance improvement between spatially constrained and unconstrained textual representations indicates that spatially constraining textual representations increases discriminative power of textual cues.

2) Experiment II: To quantify the influence of visual and textual cues, we compare the results of visual-only (BOW), visual-only (DEEP), visual-only (DEEP-FT), textual-only, textual+BOW, textual+DEEP and textual+DEEP-FT. The classification scores per category are shown in Fig. 10 and mAP is given in Table IV. The approach using textual cues, extracted by the proposed method, achieves an accuracy of 28.4%. The baselines using visual-only features achieve 34.9%, 54.5% and 60.3% for BOW, DEEP and DEEP-FT respectively. Using more discriminative visual features (DEEP-FT) significantly improves the visual baseline of BOW (26%). Using textual cues in combination with visual cues increases the mean average precision up to 47.9%, 66.2% and 70.7% for BOW, DEEP and DEEP-FT respectively. The performance gain due to combining visual and textual cues is

TABLE IV

THE IMPACT OF TEXTUAL AND VISUAL CUES FOR FINE-GRAINED CLASSIFICATION RESULTS (mAP). THE RESULTS SHOW THAT TEXTUAL CUES EXTRACTED BY THE PROPOSED METHOD ACHIEVE LIMITED ACCURACY, 28.4%. THE LOW PERFORMANCE OF VISUAL-ONLY SHOWS THAT VISUAL INFORMATION IS NOT SUFFICIENT. COMBINING TEXTUAL AND VISUAL CUES SIGNIFICANTLY OUTPERFORMS VISUAL-ONLY RESULTS WITH 13%, 12% AND 10.4% FOR BOW, DEEP AND DEEP-FT RESPECTIVELY. THIS SHOWS THAT TEXTUAL INFORMATION IS BENEFICIAL FOR FINE-GRAINED CLASSIFICATION AND IS COMPLEMENTARY TO THE VISUAL CUES. THE GAIN IN THE PERFORMANCE IS ALMOST THE SAME FOR THREE DIFFERENT VISUAL-ONLY BASELINES, EVEN THOUGH THEIR PERFORMANCE ARE AT DIFFERENT RANGES. THIS IS DUE TO THE FACT THAT TEXTUAL AND VISUAL CUES ARE FROM COMPLETELY DIFFERENT SOURCES

Source of Info. ↓	Performance (mAP%)
Textual-only	28.4
Visual-only (BOW)	34.9
Visual-only (DEEP)	53.3
Visual-only (DEEP-FT)	60.3
Textual + Visual (BOW)	**47.9**
Textual + Visual (DEEP)	**66.2**
Textual + Visual (DEEP-FT)	**70.7**

TABLE V

COMPARISON OF FINE-GRAINED CLASSIFICATION RESULTS OF TEXTUAL CUES USING THE PROPOSED TEXT DETECTION METHOD AND STATE-OF-THE-ART TEXT DETECTION METHODS [15], [22], [55]. *Textual* METHODS DIFFER IN TEXT DETECTION. TEXTUAL CUES ARE EXTRACTED FOR ALL THE METHODS AS PROPOSED IN THIS PAPER. REFERENCE [54] IS USED TO PERFORM CHARACTER RECOGNITION. Bi- AND TRIGRAMS ARE FORMED BETWEEN THE RECOGNIZED CHARACTERS BY CONSIDERING THE PROPOSED SPATIAL PAIRWISE CONSTRAINTS. THE PROPOSED METHOD SIGNIFICANTLY OUTPERFORMS [15], [22], [55] FOR TEXTUAL-ONLY CUES AND ALSO VISUAL+TEXTUAL CUES

Method ↓	Performance (mAP%)
Textual-only (Text Detection [15])	10.9
Textual-only (Text Detection [22])	17.8
Textual-only (Text Detection [55])	19.9
Textual-only (Text Detection Proposed)	**28.4**
Textual (Text Detection [15]) + Visual (DEEP-FT)	63.0
Textual (Text Detection [22]) + Visual (DEEP-FT)	66.4
Textual (Text Detection [55]) + Visual (DEEP-FT)	67.6
Textual (Text Detection Proposed) + Visual (BOW)	47.9
Textual (Text Detection Proposed) + Visual (DEEP)	66.2
Textual (Text Detection Proposed) + Visual (DEEP-FT)	**70.7**

TABLE VI

COMPARISON OF DIFFERENT STRATEGIES FOR MULTIMODAL INFORMATION FUSION

Fusion Strategy ↓	Performance (mAP%)
Early	**71.0**
Late - sum	70.0
Late - weighted sum	69.8
Late - product	65.0
Kernel	70.7

preserved, 13%, 12% and 10.4% for BOW, DEEP and DEEP-FT respectively (even though their individual performance are substantially different). Hence, textual information is beneficial for fine-grained classification and is complementary to visual cues.

Adding textual information to visual cues improves the accuracy of 26 out of the 28 classes. The low accuracy for textual cues compared to visual cues can be explained by the lack of scene text in many images, as is the case for the *Bistro* class. Nevertheless, text cues outperform visual cues for *Discount House*, *Steak House*, *PawnShop*, *Cafe* and *Dry Cleaner*. Intra-class variation for *Discount House* and *Steak House* are high. Therefore, the images within these classes are visually dissimilar and are difficult to group them together without text (even for humans).

3) Experiment III: The proposed method is compared against the state-of-the-art text detection methods [15], [22], [55] for textual cue extraction. The methods only differ in character detection. The same steps are followed for all the methods to extract textual cues as proposed in this paper. Table V illustrates that textual cues extracted by the proposed method outperforms [15], [22], and [55]. This is because [15], [22], and [55] focus on keeping the balance between precision and recall. These methods eliminate character candidates to increase precision which also increases the number of missed characters (lower recall). The missed characters are lost and cannot be recovered. This also means missed semantic information for fine-grained classification. Moreover, we compare our previous result in [2] for which we encode the textual cues using unconstrained bigrams, ABBYY for character recognition and BOW for visual features. Our previous result is 39%. *Textual(Proposed)+BOW* outperforms it significantly (9%). This shows the importance of *1.* spatially constraining bigrams, *2.* adding another layer of encoding (trigrams), *3.* high character recognition performance on fine-grained classification. Additionally, using more discriminative visual features [23] significantly improves the visual baseline. Consequently, the proposed method outperforms [2] by a large margin (32%).

4) Experiment IV: We assess the influence of the three different fusion strategies. These strategies are outlined in section V. Late fusion is performed in three different ways (*i*) predictions from each modality are combined equally using a sum operation (*sum*), (*ii*) another classifier is trained on top of predictions of the modalities to obtain weights to sum predictions from different modalities (*weighted sum*) and (*iii*) predictions from different modalities are added using a product operation (*product*). Table VI shows the results. *Early* performs slightly better than the rest. Early fusion benefits from learning the correlation between features of textual and visual modalities. *sum* and *weighted sum* performs in a similar way and better than *product*. Kernel fusion performs slightly lower than early fusion. However, *Kernel* is still beneficial since it does not restrict the choice of kernel to be the same for different modalities. Therefore, *Kernel* is used for comparisons.

5) Experiment V: In this experiment, we explore the complementarity of the extracted features. All combinations are performed using kernel fusion strategy. The results are summarized in Table VII. Textual features extracted by different methods only differ in detected text regions. Beside that,

TABLE VII

THE IMPACT OF COMBINING THE SAME AND DIFFERENT MODALITIES ON FINE-GRAINED CLASSIFICATION PERFORMANCE. TEXTUAL CUES ARE EXTRACTED IN THE SAME MANNER FOR *Textual(Proposed)* AND *Textual ([55])*. THEY DIFFER ONLY IN DETECTING TEXT REGIONS. COMBINING THEM DOES NOT INFLUENCE THE OVERALL ACCURACY. VISUAL CUES ARE EXTRACTED USING DIFFERENT FEATURES. THEREFORE, COMBINING BOW AND DEEP/DEEP-FT STILL IMPROVES DEEP/DEEP-FT ONLY. THE LARGEST GAINS ARE OBTAINED WHEN DIFFERENT MODALITIES ARE COMBINED

Source of Feat. ↓	Performance (mAP%)
Textual-only (Text Detection [15])	10.9
Textual-only (Text Detection [22])	17.8
Textual-only (Text Detection [55])	19.9
Textual-only (Text Detection Proposed)	28.4
Textual-only (Proposed+[15]+[22]+[55])	28.4
Visual-only (BOW)	34.9
Visual-only (DEEP)	53.3
Visual-only (DEEP-FT)	60.3
Visual-only (BOW + DEEP)	55.8
Visual-only (BOW + DEEP-FT)	62.9
Visual-only (DEEP + DEEP-FT)	62.1
Visual-only (BOW + DEEP + DEEP-FT)	63.5
Visual (BOW + DEEP + DEEP-FT) + Textual (Proposed)	**71.6**

textual cues are obtained in the same way. Therefore, combining textual cues together does not influence the overall accuracy. This also indicates that the proposed method covers all the information that [15], [22], [55] carries and more. BOW and GoogLeNet visual baselines are extracted using different features. Therefore, combining BOW and DEEP/DEEP-FT still improves DEEP/DEEP-FT only. This indicates that BOW and DEEP/DEEP-FT visual features are useful for fine-grained classification on this dataset. Combining all the extracted features still improves with 1% over the best result obtained in previous experiment (70.7%). Moreover, combining textual cues still significantly improves the performance even after combining all visual features. This indicates complementarity of visual and textual modalities.

6) Experiment VI: In this experiment, we quantify the impact of local and global visual cues and, their combinations with textual cue for fine-grained classification. To this end, we extract 100 regions from each image using EdgeBoxes [77] (object proposal algorithm). We represent each region by pre-trained (DEEP) and fine-tuned (DEEP-FT) GoogLeNet [23] features. To incorporate the local information we use the kernel function in [80] which is described as

$$k = \frac{1}{n_1}\frac{1}{n_2}\sum_{i=1}^{100}\sum_{j=1}^{100} d_i \cdot d_j, \tag{6}$$

$d_i \cdot d_j$ is the cosine similarity of the two regional features, and n_1 and n_2 are the normalization factors to ensure the self-similarity to be 1. The above kernel cross compares all the local regions. Table VIII summarizes the result. From Table VIII it can be derived that by including the local visual information, we improve the visual baseline performance from 53.3% to 58.4% when using pre-trained GoogLeNet (DEEP), and from 60.3% to 64.7% for fine-tuned GoogLeNet (DEEP-FT). Most importantly, even though the visual baselines are improved by including local regional visual information, adding the proposed textual cues can still dramatically boost the performance. We conclude that scene text in images provides highly complementary information to the visual cue, globally or locally.

7) Experiment VII: In this experiment, we discuss the performance change with respect to the amount of text in images. To this end, we have annotated text regions (as word bounding boxes) for the first 10 classes of the Con-Text dataset (in alphabetical order). We report per-class, *1.* the total number of images, *2.* the number of images with text, *3.* the percentage of images with text with respect to the total number of images, *4.* fine-grained classification rates (mAP) using only textual information (see Table IX).

"Dry Cleaner" obtains the maximum classification rate (71%). Hence, there is a strong correlation with the amount of text and the fine-grained classification rate since "Dry Cleaner" has the maximum number of images with text and also a high text percentage (89%). This holds also for the other classes which contain a high number of images with text, and a high text percentage (e.g. BookStore, Country Store). The proposed multimodal classification rate is low when the number of images with text and text percentage is limited (e.g. Bistro, ComputerStore). The text percentage alone is not a conclusive indicator for high classification rate (e.g. DiscountHouse). It is necessary that there is enough data (number of images with text) to learn from. Moreover, overlapping text is also important. For instance, "Country Store" has a high number of images and text percentage. However, the classification rate is relatively small compared to "Dry Cleaner". This is due to the diversity of text descriptions.

IX. LOGO RETRIEVAL

In logo retrieval, the aim is to find all images of a query logo in an image collection, e.g., *Starbucks*. Logos may consist of text alone or text is an important part of the logo itself such as *Starbucks*, *Ford*, *FedEX* and *Google*. However, in logo retrieval, recognized text in natural scene images has never been exploited before [63]–[65].

A. Dataset and Implementation Details

1) Dataset: Our approach is validated on *FlickrLogos-32* [67]. The dataset consists of 32 brand logos, e.g., *Texaco*, *Pepsi* and *Google* and 30 queries per logo resulting in a total of 960 queries. The search set contains 40 images per logo and 3000 non-logo images.

2) Implementation Notes: Again, we use average precision as the performance measure. To represent the visual cues, a standard bag of visual words (BOW) approach is used with a visual vocabulary of size 1 million. This visual representation reaches a retrieval accuracy of 54.8% mAP. This accuracy corresponds with what has been reported in [65]. To represent textual cues, we use spatially constrained bi- and trigrams and their combinations. The character candidates within the query bounding boxes are kept for the query images to represent textual cues. The images are ranked by the cosine similarity between the normalized textual representations. To combine the visual and textual cues, we use late fusion.

TABLE VIII

THE IMPACT OF LOCAL AND GLOBAL VISUAL CUES AND, THEIR COMBINATIONS WITH TEXTUAL CUE FOR FINE-GRAINED CLASSIFICATION. THE RESULTS ARE REPORTED IN MEAN AVERAGE PRECISION (%)

Source of Feat.↓	Pre-trained GoogLeNet (DEEP)	Fine-tuned GoogLeNet (DEEP-FT)
Global Visual Cue	53.3	60.3
Local Visual Cue	56.1	59.3
Local and Global Visual Cues	58.4	64.7
Local and Global Visual Cues + Textual Cues (Proposed)	69.3	73.6

TABLE IX

THE DISTRIBUTION OF IMAGES WITH TEXT AND WITHOUT TEXT FOR THE FIRST 10 CLASSES OF THE FINE-GRAINED DATASET

Class↓	Total Num. Images	Num. Images with Text	Text Percentage (%)	Performance (mAP%)
Bakery	1214	467	38	21.19
Barber	1573	635	40	35.81
Bistro	287	59	21	3.12
BookStore	1333	724	54	44.46
Cafe	839	438	52	25.80
DryCleaner	1195	1065	89	71.46
ComputerStore	287	142	49	5.38
CountryStore	1224	866	71	35.01
Diner	1136	781	69	30.95
DiscountHouse	43	42	98	25.63

TABLE X

THE INFLUENCE OF DIFFERENT TEXTUAL ENCODINGS E.G., bi- AND TRIGRAM AND bi- +TRIGRAM IN LOGO RETRIEVAL. Bi- AND TRIGRAMS PRODUCE SIMILAR RESULTS. COMBINING bi- AND TRIGRAM OUTPERFORMS EACH INDIVIDUAL REPRESENTATION

Textual Cue Encoding↓	Performance (mAP%)
bigrams	19.0
trigrams	18.9
[bi+tri]grams	**23.9**

TABLE XI

LOGO RETRIEVAL RESULTS (mAP) FOR TEXTUAL-ONLY, VISUAL-ONLY AND TEXTUAL+VISUAL. THE RESULTS SHOW THAT THE METHOD BASED ON TEXTUAL CUES EXTRACTED BY THE PROPOSED METHOD ACHIEVES AN ACCURACY OF 23.9%. COMBINING TEXTUAL AND VISUAL CUES INCREASES THE ACCURACY. THIS SHOWS THAT TEXTUAL INFORMATION IS BENEFICIAL FOR LOGO RETRIEVAL AND IS COMPLEMENTARY TO VISUAL CUES

Source of Info. ↓	Performance (mAP%)
textual-only	23.9
visual-only	54.8
textual+visual	**57.4**

B. Experiments and Results

We conduct two experiments. First, we quantify the influence of textual cues and the effect of different levels of textual cue encodings for logo retrieval. In the second experiment, we evaluate the complementarity of textual and visual cues for logo retrieval.

1) Experiment I: We evaluate different textual cue encodings: (1) bigrams, (2) trigrams and (3) bi- + trigrams. The results are summarized in Table X. Representing textual cues as bi- or trigrams produces similar results. However, combining both representations outperforms the individual encodings. This implies that bi- and trigrams have similar discriminative power, yet they capture information at different levels. Therefore, they are complementary to each other.

2) Experiment II: To assess the influence of textual cues, we compare the results of textual-only, visual-only and textual+visual. The results are summarized in Table XI. The results show that the method based on textual cues extracted by the proposed method achieves an accuracy of 23.9%. Using textual and visual cues increases mAP up to 57.4%. This implies that textual information is beneficial for logo retrieval. And that textual information is complementary to visual cues.

(a)

(b)

Fig. 11. Sample images from *FlickrLogos-32* to illustrate (a) where the proposed multimodal approach improves visual-only retrieval performance and (b) where it is not effective for visual-only retrieval performance.

Textual cues improve the retrieval accuracy of the logos which are formed by only text e.g., *esso*, *aldi* and *stella artois* (See Figure 11.a for sample images). However, textual cues are not effective, *(i)* when the characters of the logos are all connected e.g., *fedex* and *ford*. This is because the proposed character detection algorithm behaves the full word as a character. Therefore, proper recognition of characters

forming the word fails. *(ii)* when logos do not contain any text e.g., *apple* and *ferrari*. It is clear that the multimodal approach requires textual input. *(iii)* when the characters are not in a common text font e.g., *milka*. Character recognition algorithm has difficulties to correctly recognize characters when they are not in common text fonts. Figure 11.b illustrates some visual samples where the textual cues do not improve visual-only logo retrieval performance.

X. Conclusion

A method has been introduced to combine textual with visual cues for fine-grained classification and logo retrieval. While the state-of-the-art relies on visual cues only, this paper is the first work which proposes to combine recognized scene text and visual cues for fine-grained classification and logo retrieval. To extract text cues, we have proposed a generic, efficient and fully unsupervised algorithm for text detection. The proposed text detection method does not directly detect text regions but instead aims to detect background to infer text location. Remaining regions after eliminating background are considered as text regions. Then, text candidates have been processed by two methods to perform text recognition i.e. ABBYY commercial OCR machine and state-of-the-art character recognition algorithm [54]. Bi- and trigrams have been formed between the recognized characters by using proposed spatial encoding.

The proposed algorithm achieves state-of-the-art (end-to-end) character recognition accuracy on the ICDAR03. It is shown that bimodal information fusion of visual and textual cues increased the fine-grained classification accuracy by 10.4%. The proposed method outperforms state-of-the-art text detection methods [55] on text saliency and [15], [55], [55] on (end-to-end) character recognition and fine-grained classification. We improve earlier version of this paper [2] for fine-grained classification from 39.0% to 70.7% in *mAP*. Moreover, we applied our work also for logo retrieval. Textual cues proven to be complementary to visual cues for logo retrieval too. Combining textual and visual cues improves the logo retrieval performance over visual-only from 54.8% to 57.4%.

References

[1] S. Karaoglu, J. C. van Gemert, and T. Gevers, "Object reading: Text recognition for object recognition," in *Proc. ECCV Workshops*, 2012, pp. 456–465.

[2] S. Karaoglu, J. C. van Gemert, and T. Gevers, "Con-text: Text detection using background connectivity for fine-grained object classification," in *Proc. ACM-MM*, Oct. 2013, pp. 757–760.

[3] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *Proc. CVPR*, Jun. 2010, pp. 2963–2970.

[4] K. Wang, B. Babenko, and S. Belongie, "End-to-end scene text recognition," in *Proc. ICCV*, Nov. 2011, pp. 1457–1464.

[5] L. Neumann and J. Matas, "Text localization in real-world images using efficiently pruned exhaustive search," in *Proc. ICDAR*, Sep. 2011, pp. 687–691.

[6] M. Everingham, S. M. A. Eslami, L. van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[7] A. Mishra, K. Alahari, and C. V. Jawahar, "Image retrieval using textual cues," in *Proc. ICCV*, Dec. 2013, pp. 3040–3047.

[8] E. Gavves, B. Fernando, C. G. M. Snoek, A. W. M. Smeulders, and T. Tuytelaars, "Local alignments for fine-grained categorization," *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 191–212, 2015.

[9] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. ECCV*, 2014, pp. 834–849.

[10] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Comput. Vis. Image Understand.*, vol. 114, no. 6, pp. 712–722, 2010.

[11] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *Proc. ICCV*, Oct. 2007, pp. 1–8.

[12] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[13] T. Lu, S. Palaiahnakote, C. L. Tan, and W. Liu, "Text detection in multimodal video analysis," in *Video Text Detection*. London, U.K.: Springer, 2014, pp. 221–246.

[14] T. Lu, S. Palaiahnakote, C. L. Tan, and W. Liu, "Video text detection," in *Advances in Computer Vision and Pattern Recognition*. London, U.K.: Springer, 2014.

[15] L. Gómez and D. Karatzas, "Scene text recognition: No country for old men?" in *Proc. ACCV Workshops*, 2014, pp. 157–168.

[16] O. Augereau, N. Journet, A. Vialard, and J.-P. Domenger, "Improving classification of an industrial document image database by combining visual and textual features," in *Proc. Document Anal. Syst. (DAS)*, Apr. 2014, pp. 314–318.

[17] M. Rusiñol, V. Frinken, D. Karatzas, A. D. Bagdanov, and J. Lladós, "Multimodal page classification in administrative document image streams," *Int. J. Document Anal. Recognit.*, vol. 17, no. 4, pp. 331–341, 2014.

[18] B. Erol and J. J. Hull, "Semantic classification of business images," *Proc. SPIE*, vol. 6073, pp. 139–146, Jan. 2006.

[19] D. Doermann, E. Rivlin, and I. Weiss, "Applying algebraic and differential invariants for logo recognition," *Mach. Vis. Appl.*, vol. 9, no. 2, pp. 73–86, 1996.

[20] P. P. Roy, U. Pal, and J. Lladós, "Document seal detection using GHT and character proximity graphs," *Pattern Recognit.*, vol. 44, no. 6, pp. 1282–1295, Jun. 2011.

[21] S. S. Tsai *et al.*, "Combining image and text features: A hybrid approach to mobile book spine recognition," in *Proc. ACM-MM*, 2011, pp. 1029–1032.

[22] L. Neumann and J. Matas, "Real-time scene text localization and recognition," in *Proc. CVPR*, Jun. 2012, pp. 3538–3545.

[23] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. CVPR*, Jun. 2015, pp. 1–9.

[24] G. Wang, D. Hoiem, and D. Forsyth, "Building text features for object image classification," in *Proc. CVPR*, Jun. 2009, pp. 1367–1374.

[25] L. Elazary and L. Itti, "A Bayesian model for efficient visual search and recognition," *Vis. Res.*, vol. 50, no. 14, pp. 1338–1352, 2010.

[26] A. Torralba, "Contextual priming for object detection," *Int. J. Comput. Vis.*, vol. 53, no. 2, pp. 169–191, 2003.

[27] C. Yi and Y. Tian, "Scene text recognition in mobile applications by character descriptor and structure configuration," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2972–2982, Jul. 2014.

[28] M.-E. Nilsback and A. Zisserman, "A visual vocabulary for flower classification," in *Proc. CVPR*, Jun. 2006, pp. 1447–1454.

[29] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[30] E. Rahtu and J. Heikkilä, "A simple and efficient saliency detector for background subtraction," in *Proc. ICCV Workshops*, Sep./Oct. 2009, pp. 1137–1144.

[31] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 171–177, Jan. 2010.

[32] Q. Zhu, M. C. Yeh, and K. T. Cheng, "Multimodal fusion using learned text concepts for image categorization," in *Proc. ACM-MM*, 2006, pp. 211–220.

[33] A. Shahab, F. Shafait, A. Dengel, and S. Uchida, "How salient is scene text?" in *Proc. IAPR Int. Workshop Document Anal. Syst.*, Mar. 2012, pp. 317–321.

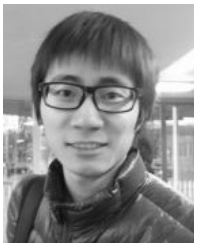
[34] A. Clavelli, D. Karatzas, J. Llados, M. Ferraro, and G. Boccignone, "Towards modelling an attention-based text localization process," in *Proc. Pattern Recognit. Image Anal.*, 2013, pp. 296–303.

[35] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. CVPR*, 2015, pp. 1072–1080.

[36] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.

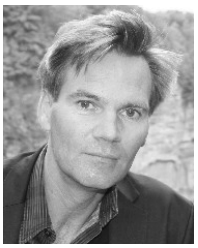
[37] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie. (2016). "COCO-text: Dataset and benchmark for text detection and recognition in natural images." [Online]. Available: https://arxiv.org/abs/1601.07140
[38] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. ECCV*, 2012, pp. 29–42.
[39] H. I. Koo and D. H. Kim, "Scene text detection via connected component clustering and nontext filtering," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2296–2305, Jun. 2013.
[40] Y.-P. Feng, H. Xinwen, and C.-L. Liu, "A hybrid approach to detect and localize texts in natural scene images," *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 800–813, Mar. 2011.
[41] L. Neumann and J. Matas, "A method for text localization and recognition in real-world images," in *Proc. ACCV*, 2010, pp. 770–783.
[42] S. Uchida, Y. Shigeyoshi, Y. Kunishige, and Y. Feng, "A keypoint-based approach toward scenery character detection," in *Proc. ICDAR*, Sep. 2011, pp. 819–823.
[43] T. Wang, D. J. Wu, A. Coates, and A. Y. Ng, "End-to-end text recognition with convolutional neural networks," in *Proc. ICPR*, Nov. 2012, pp. 3304–3308.
[44] Q. Sun, Y. Lu, and S. Sun, "A visual attention based approach to text extraction," in *Proc. ICPR*, Aug. 2010, pp. 3991–3995.
[45] H.-C. Wang and M. Pomplun, "The attraction of visual attention to texts in real-world scenes," in *Proc. Annu. Meeting Cognit. Sci. Soc.*, 2011, pp. 2733–2738.
[46] K. Wang, E. Rescorla, H. Shacham, and S. J. Belongie, "OpenScan: A fully transparent optical scan voting system," in *Proc. Electron. Voting Technol. Workshop*, Aug. 2010. [Online]. Availabla: https://www.usenix.org/conference/evtwote-10/openscan-fully-transparent-optical-scan-voting-system
[47] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.
[48] J. C. van Gemert, "Exploiting photographic style for category-level image classification by generalizing the spatial pyramid," in *Proc. ICMR*, 2011, pp. 1–14.
[49] B. Fernando, S. Karaoglu, and A. Trémeau, "Extreme value theory based text binarization in documents and natural scenes," in *Proc. ICMV*, 2010, pp. 144–151.
[50] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. ICCV*, Sep./Oct. 2009, pp. 2106–2113.
[51] J. van de Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 1, pp. 150–156, Jan. 2006.
[52] R. Valenti, N. Sebe, and T. Gevers, "Image saliency by isocentric curvedness and color," in *Proc. ICCV*, Sep./Oct. 2009, pp. 2185–2192.
[53] N. Ezaki, M. Bulacu, and L. Schomaker, "Text detection from natural scene images: Towards a system for visually impaired persons," in *Proc. ICPR*, Aug. 2004, pp. 683–686.
[54] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *Proc. ECCV*, 2014, pp. 512–528.
[55] Y. Li, W. Jia, C. Shen, and A. van den Hengel, "Characterness: An indicator of text in the wild," *IEEE Trans. Image Process.*, vol. 23, no. 4, pp. 1666–1677, Apr. 2014.
[56] S. Ayache, G. Quénot, and J. Gensel, "Classifier fusion for SVM-based multimedia semantic indexing," in *Proc. ECIR*, 2007, pp. 494–504.
[57] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions," in *Proc. IJDAR*, 2005, p. 682.
[58] D. Karatzas *et al.*, "ICDAR 2013 robust reading competition," in *Proc. ICDAR*, Aug. 2013, pp. 1484–1493.
[59] B. Yao, G. Bradski, and L. Fei-Fei, "A codebook-free and annotation-free approach for fine-grained image categorization," in *Proc. CVPR*, Jun. 2012, pp. 3466–3473.
[60] A. Joly and O. Buisson, "Logo retrieval with a contrario visual query expansion," in *Proc. ACM-MM*, 2009, pp. 581–584.
[61] C. Yao, X. Bai, and W. Liu, "A unified framework for multioriented text detection and recognition," *IEEE Trans. Image Process.*, vol. 23, no. 11, pp. 4737–4749, Nov. 2014.
[62] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. CVPR*, Jun. 2012, pp. 1083–1090.
[63] J. Revaud, M. Douze, and C. Schmid, "Correlation-based burstiness for logo retrieval," in *Proc. ACM-MM*, 2012, pp. 965–968.
[64] R. Tao, E. Gavves, C. G. M. Snoek, and A. W. M. Smeulders, "Locality in generic instance search from one example," in *Proc. CVPR*, Jun. 2014, pp. 2099–2106.
[65] S. Romberg and R. Lienhart, "Bundle min-hashing for logo recognition," in *Proc. ICMR*, 2013, pp. 113–120.
[66] C. Yi and Y. Tian, "Text string detection from natural scenes by structure-based partition and grouping," *IEEE Trans. Image Process.*, vol. 20, no. 9, pp. 2594–2605, Sep. 2011.
[67] S. Romberg, L. G. Pueyo, R. Lienhart, and R. van Zwol, "Scalable logo recognition in real-world images," in *Proc. ICMR*, 2011, p. 25.
[68] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, Jun. 2006, pp. 2169–2178.
[69] S. Lu, T. Chen, S. Tian, J.-H. Lim, and C.-L. Tan, "Scene text extraction based on edges and support vector regression," *Int. J. Document Anal. Recognit.*, vol. 18, no. 2, pp. 125–135, 2015.
[70] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Reading text in the wild with convolutional neural networks," *Int. J. Comput. Vis.*, vol. 116, no. 1, pp. 1–20, 2016.
[71] L. Gomez and D. Karatzas, "Object proposals for text extraction in the wild," in *Proc. ICDAR*, Aug. 2015, pp. 206–210.
[72] S. Milyaev, O. Barinova, T. Novikova, P. Kohli, and V. Lempitsky, "Fast and accurate scene text understanding with image binarization and off-the-shelf OCR," *Int. J. Document Anal. Recognit.*, vol. 18, no. 2, pp. 169–182, 2015.
[73] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarization," *Pattern Recognit.*, vol. 39, no. 3, pp. 317–327, 2006.
[74] T.-S. Chua *et al.*, "TRECVID 2004 search and feature extraction task by NUS PRIS," in *Proc. NIST TRECVID Workshop*, 2004.
[75] S. Lee *et al.*, "KU-ISPL TRECVID 2015 multimedia event detection system," in *Proc. TRECVID*, 2014.
[76] P. Over *et al.*, "TRECVID 2014—An overview of the goals, tasks, data, evaluation mechanisms and metrics," in *Proc. TRECVID*, 2014, p. 52.
[77] C. L. Zitnick and P. Dollär, "Edge boxes: Locating object proposals from edges," in *Proc. ECCV*, 2014, pp. 391–405.
[78] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comput.*, vol. 22, no. 10, pp. 761–767, 2004.
[79] Y. Movshovitz-Attias, Q. Yu, M. C. Stumpe, V. Shet, S. Arnoud, and L. Yatziv, "Ontological supervision for fine grained classification of street view storefronts," in *Proc. CVPR*, 2015, pp. 1693–1702.
[80] G. Tolias, R. Sicre, and H. Jégou. (2015). "Particular object retrieval with integral max-pooling of CNN activations." [Online]. Available: http://arxiv.org/abs/1511.05879
[81] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 12, pp. 2552–2566, Dec. 2014.
[82] S. K. Ghosh and E. Valveny, "Query by String word spotting based on character bi-Gram indexing," in *Proc. ICDAR*, Aug. 2015, pp. 881–885.

Sezer Karaoglu received the Ph.D. degree from the Computer Vision Group, Informatics Institute, University of Amsterdam (UvA), and the double master's degree. He was selected as Tuition Fee Scholar for a European master's degree from Color in Informatics and Media Technology Program. He is currently the CTO and a Co-founder of 3DUniversum, spin-offs with the Informatics Institute of the UvA. His research interests are 2D–3D computer vision and machine learning.



Ran Tao received the M.Sc. degree in computer science from Leiden University, The Netherlands, in 2011, and the Ph.D. degree in computer science from the University of Amsterdam, The Netherlands, in 2017. He is currently a Post-Doctoral Researcher with the QUVA Lab, the joint research lab of Qualcomm and the University of Amsterdam on deep learning and computer vision. His research interests include computer vision and machine learning, with a focus on instance search, object tracking, and deep learning.

Jan C. van Gemert received the Ph.D. degree from the University of Amsterdam in 2010. He was a Post-Doctoral Fellow with the University of Amsterdam and the École Normale Supérieure, Paris. He currently heads the Computer Vision Group, Delft University of Technology. He has authored over 50 peer-reviewed papers, is cited over 3,000 times. His research interests include learning visual encodings, image and video categorization, action and object recognition, and localization. He has an h-index of 20.



Theo Gevers is currently a Full Professor of Computer Vision, University of Amsterdam (UvA), Amsterdam, The Netherlands. He is also the Founder of Sightcorp and 3DUniversum, spin-offs of the Informatics Institute of the UvA. His main research interests are in the fundamentals of image understanding, 3-D object recognition, and color in computer vision.