NEOCR: A Configurable Dataset for Natural Image Text Recognition

Robert Nagy, Anders Dicker, Klaus Meyer-Wegener
Chair for Computer Science 6 (Data Management)
University of Erlangen-Nürnberg
Erlangen, Germany
Emails: {robert.nagy,anders.dicker,klaus.meyer-wegener}@cs.fau.de

***Abstract*—Recently growing attention has been paid to recognizing text in natural images. Natural image text OCR is far more complex than OCR in scanned documents. Text in real world environments appears in arbitrary colors, font sizes and font types, often affected by perspective distortion, lighting effects, textures or occlusion. Currently there are no datasets publicly available which cover all aspects of natural image OCR. We propose a comprehensive well-annotated configurable dataset for optical character recognition in natural images for the evaluation and comparison of approaches tackling with natural image text OCR. Based on the rich annotations of the proposed NEOCR dataset new and more precise evaluations are now possible, which give more detailed information on where improvements are most required in natural image text OCR.**

***Keywords*-ocr; dataset; scene text recognition; latin characters**

I. Introduction

Optical character recognition (OCR) for machine-printed documents and handwriting has a long history in computer science. For clean documents, current state-of-the-art methods achieve over 99% character recognition rates [1].

With the prevalence of digital cameras and mobile phones, an ever-growing amount of digital images are created. Many of these natural images contain text. The recognition of text in natural images opens a field of widespread applications, such as:

- help for visually impaired or blind [2] (e.g., reading text not transcribed in braille),
- mobile applications (e.g., translating photographed text for tourists and foreigners [3, 4, 5, 6]),
- object classification (e.g., multimodal fusion of text and visual information [7]),
- image annotation (e.g., for web search [8]),
- vision-based navigation and driving assistant systems [9].

Recently growing attention has been paid to recognizing text in real world images, also referred to as natural image text OCR [6] or scene text recognition [1]. Natural images are far more complex in contrast to machine-printed documents. Problems arise not only from background variations and surrounding objects in the image, but from the depicted text too, which usually takes on a great variety of appearances. In addition to the survey of [2], which compared the capturing devices, we summarized main characteristics of scanned document OCR and scene text recognition in table I.

For the evaluation and comparison of techniques developed specifically for natural image OCR, a publicly available well-annotated dataset is required. All current datasets (see section III) annotate only the words and bounding boxes in images. Also most text appears in horizontal arrangement, while in natural scenes humans are often confronted with text arranged vertically or circularly (text following a curved, wavy or circular line). Currently there is no well-annotated dataset publicly available that covers all aspects distinguishing scene text recognition from scanned document OCR.

We propose the NEOCR (Natural Environment OCR) dataset consisting of real world images extensively enriched with additional metadata. Based on this metadata several subdatasets can be created to identify and overcome weaknesses of OCR approaches on natural images. Main benefits of the proposed dataset compared to other related datasets are:

- annotation of *all* text visible in images,
- additional distortion quadrangles for a more precise ground truth representation of text regions,
- rich metadata for simple configuration of subdatasets with special characteristics for more detailed identification of shortcomings in OCR approaches.

The paper is organized as follows: In the next section we describe the construction of the new dataset and the annotation metadata in detail. In section III a short overview of currently available datasets for OCR in natural images is given and their characteristics are compared to the new NEOCR dataset. We describe new evaluation possibilities due to the rich annotation of the dataset and its future evolution in section IV.

II. Dataset

A comprehensive dataset with rich annotation for OCR in natural images is introduced. The images cover a broad range of characteristics that distinguish real world scenes

Criteria	Scanned documents	Natural image text
background	homogeneous, usually white or light paper	any color, even dark or textured
blurredness	sharp (depending on scanner)	possibly motion blur, blur because of depth of field
camera position	fixed, document lies on scanner's glass plate	variable, geometric and perspective distortions almost always present
character arrangement	clear horizontal lines	horizontal and vertical lines, rounded, wavy
colors	mostly black text on white background	high variability of colors, also light text on dark background (e.g. illuminated text) or only minor differences between tones
contrast	good (black/dark text on white/light background)	depends on colors, shadows, lighting, illumination, texture
font size	limited number of font sizes	high diversity in font sizes
font type (diversity in document)	usually 1-2 (limited) types of fonts	high diversity of fonts
font type (in general)	machine-print, handwriting	machine-print, handwriting, special (e.g. textured such as light bulbs)
noise	limited / negligible	shadows, lighting, texture, flash light, reflections, objects in the image
number of lines	usually several lines of text	often only one single line or word
occlusion	none	both horizontally, vertically or arbitrary possible
rotation (line arrangement)	horizontally aligned text lines or rotated by ± 90 degrees	arbitrary rotations
surface	text "attached" to plain paper	text freestanding (detached) or attached to objects with arbitrary nonplanar surfaces, high variability of distortions

Table I
Typical characteristics of OCR on scanned documents and natural image text recognition.

from scanned documents. The dataset contains a total of 659 images with 5238 bounding boxes (textfields). Images were captured by the authors and members of the lab using various digital cameras with diverse camera settings to achieve a natural variation of image characteristics. Afterwards images containing text were hand-selected with particular attention to achieving a high diversity in depicted text regions. This first release of the NEOCR dataset covers the following dimensions each by at least 100 textfields. Figure 1 shows examples from the NEOCR dataset for typical problems in natural image OCR.

Based on the rich annotation of optical, geometrical and typographical characteristics of bounding boxes, the NEOCR dataset can also be tailored into specific datasets to test new approaches for specialized scenarios. Additionally to bounding boxes, distortion quadrangles were added for a more accurate ground truth annotation of text regions and automatic derivation of rotation, scaling, translation and shearing values. These distortion quadrangles also enable a more precise representation of slanted text areas close to each other.

For image annotation, the web-based tool of [10] for the LabelMe dataset [11] was used. Due to the simple browser interface of LabelMe the NEOCR dataset can be extended continuously. Annotations are provided in XML for each image separately describing global image features, bounding boxes of text and its special characteristics. The XML-schema of LabelMe has been adapted and extended by tags for additional metadata. The annotation metadata is discussed in more detail in the following sections.

A. Global Image Metadata

General image metadata contains the filename, folder, source information and image properties. For each whole image its width, height, depth, brightness and contrast are annotated. Brightness values are obtained by extracting the luma channel (Y-channel) of the images and computing the mean value. The standard deviation of the luma channel is annotated as the contrast value. Both brightness and contrast values are obtained automatically using ImageMagick [12].

B. Textfield Metadata

All words and coherent text passages appearing in the images of the NEOCR dataset are marked by bounding boxes. Coherent text passages are several lines of text in same font size and type, color, texture and background (as they usually appear on memorial plaques or signs). All bounding boxes are rectangular and parallel to the axes. Additionally annotated distortion quadrangles inside the bounding boxes give a more accurate representation of text regions. The metadata is enriched by optical, geometrical and typographical characteristics.

(a) emboss, engrave (b) lens blur (c) perspective distortion

(d) crop, rotate, occlusion, circular (e) textured background (f) textured text

Figure 1. Example images from the NEOCR dataset depicting typical characteristics of natural image text recognition.

1) Optical Characteristics: Optical characteristics contain information about the blurredness, brightness, contrast, inversion (dark text on light or light text on dark background), noise and texture of a textfield.

Texture: Texture is very hard to measure automatically, because texture differences can form the text and text itself can be texture too. Following three categories have been defined:

- low: single color text with single color background,
- mid: multi-colored text or multi-colored background,
- high: multi-colored text and multi-colored background, or text without a continuous surface (e.g., luminous advertising built from light bulbs).

Brightness and contrast: Brightness and contrast values for bounding boxes are obtained the same way as for the whole image (see section II-A). As an attribute of the contrast characteristic we additionally annotate whether the dark text is represented on light background or vice versa (inverted).

Resolution: In contrast to 1000dpi and more in high resolution scanners, images taken by digital cameras achieve resolutions only up to 300dpi. The lower the focal length, the bigger the area captured by the lens. Depending on the pixel density and the size of the camera sensor small text can get unrecognizable. As a measure we define text resolution as the number of pixels in the bounding box divided by the number of characters.

Noise: Image noise can originate from the noise sensitivity of camera sensors or from image compression artifacts (e.g., in JPEG images). Usually, the higher the ISO values or the higher the compression rates, the bigger the noise in images. Because noise and texture are difficult to distinguish, we classify the bounding boxes into low, mid and high noise judged by eye.

Blurredness: Image blur can be divided into lens and motion blur. Lens blur can result from depth of field effects when using large aperture depending on the focal length and focus point. Similar blurring effects can also result from image compression. Motion blur can originate either from moving objects in the scene or camera shakes by the photographer. [13] gives an overview on different approaches for measuring image blur. As a measure for blurredness we annotated kurtosis to the bounding boxes. First edges are detected using a Laplacian-of-Gaussian filter. Afterwards the edge image is Fourier transformed and the steepness (kurtosis) of the spectral analysis is computed. The higher the kurtosis, the more blurred the image region.

2) Geometrical Characteristics: Character arrangement, distortion, occlusion and rotation are subsumed under geometrical characteristics.

Distortion: Because the camera sensor plane is almost never parallel to the photographed text's plane, text in natural images usually appears perspectively distorted. Several methods can be applied to represent distortion. In our annotations we used 8 floating point values as described in [14]. The 8 values can be represented as a matrix, where s_x and s_y describe scaling, r_x and r_y

rotation, t_x and t_y translation, and p_x and p_y shearing:

$$\begin{pmatrix} s_x & r_y & t_x \\ r_x & s_y & t_y \\ p_x & p_y & 1 \end{pmatrix} \quad (1)$$

The equations in [14] are defined for unit length bounding boxes. We adapted the equations for arbitrary sized bounding boxes. Based on the matrix and the original coordinates of the bounding box, the coordinates of the distorted quadrangle can be computed using the following two equations:

$$x' = \frac{s_x x + r_y y + t_x}{p_x x + p_y y + 1} \quad (2)$$

$$y' = \frac{r_x x + s_y y + t_y}{p_x x + p_y y + 1} \quad (3)$$

Rotation: Because of arbitrary camera directions and free positioning in the real world, text can appear diversely rotated in natural images. The rotation values are given in degrees as the offset measured from the horizontal axis given by the image itself. The text rotation angle is computed automatically based on the distortion parameters.

Arrangement: In natural images characters of a text can be arranged vertically too (e.g., some hotel signs). Also some text can follow curved baselines. In the annotations we distinguish between horizontally, vertically and circularly arranged text. Single characters were classified as horizontally arranged.

Occlusion: Depending on the chosen image detail by the photographer or objects present in the image, text can appear occluded in natural images. Because missing characters (vertical cover) and horizontal occlusion need to be treated separately, we distinguish between both in our annotations. Also the amount of cover is annotated as percentage value.

3) Typographical Characteristics: Typographical characteristics contain information about font type and language.

Typefaces: Typefaces of bounding boxes are classified into categories print, handwriting and special. The annotated text is case-sensitive, the font size can be derived from the resolution and bounding box size information. Font thickness is not annotated.

Language: Languages can be a very important information when using vocabularies for correcting errors in recognized text. Because the images were taken in several countries, 15 different languages are present in the NEOCR dataset, though the visible text is limited to latin characters. In some cases, text cannot be clearly assigned to any language. For these special cases we introduced categories for numbers, abbreviations and business names.

C. Summary

Figure 2 shows statistics on selected dimensions for the NEOCR dataset. The graphs prove the high diversity of the images in the dataset. The more accurate and rich annotation allows more detailed inspection and comparison of approaches for natural image text OCR. Further details for the annotations can be found on the NEOCR dataset website [15]. Some OCR algorithms rely on training data. For these approaches, a disjoint split of the NEOCR images in training and testing data is provided on the website.

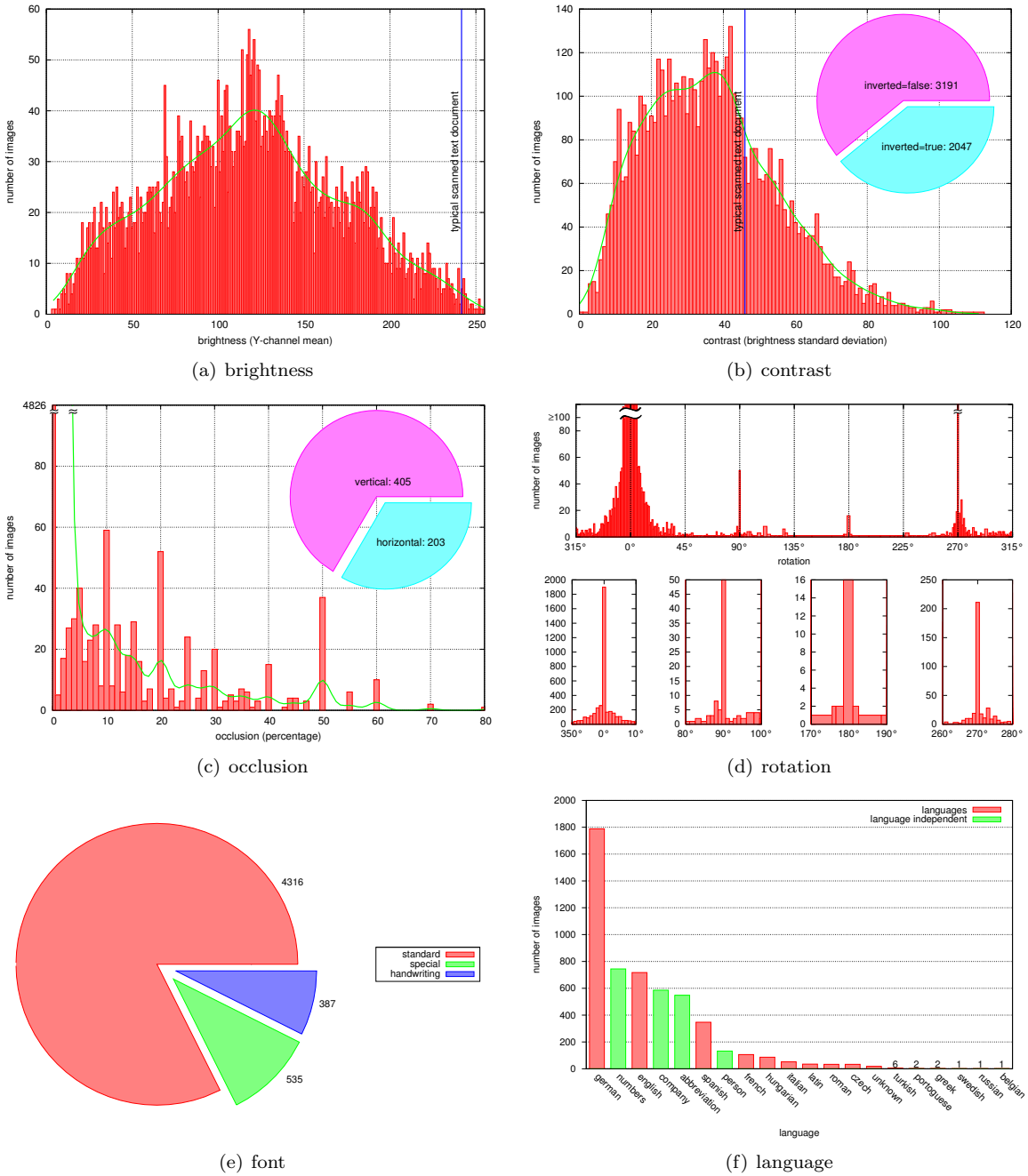
III. Related Work

Unfortunately, publicly available OCR datasets for scene text recognition are very scarce. The ICDAR 2003 Robust Reading dataset [16, 17, 18] is the most widely used in the community. The dataset contains 258 training and 251 test images annotated with a total of 2263 bounding boxes and text transcriptions. Bounding boxes are all parallel to the axes of the image, which is insufficient for marking text in natural scene images with their high variations of shapes and orientations. Although the images in the dataset show a considerable diversity in font types, the pictures are mostly focused on the depicted text and the dataset contains largely indoor scenes depicting book covers or closeups of device names. The dataset doesn't contain any vertically or circularly arranged text at all. The high diversity of natural images, such as shadows, light changes, illumination, character arrangement is not covered in the dataset.

The Chars74K dataset introduced by [19, 20] focuses on the recognition of Latin and Kannada characters in natural images. The dataset contains 1922 images mostly depicting sign boards, hoardings and advertisements from a frontal viewpoint. About 900 images have been annotated with bounding boxes for characters and words, of which only 312 images contain latin word annotations. Unfortunately, images with occlusion, low resolution or noise have been excluded and not all words visible in the images have been annotated.

[6] proposed the Street View Text dataset [21], which is based on images harvested from Google Street View [22]. The dataset contains 350 outdoor images depicting mostly business signs. At total of 904 rectangular textfields are annotated. Unfortunately, bounding boxes are parallel to the axes, which is insufficient for marking text variations in natural scenes. Another deficit is that not all words depicted in the image have been annotated.

In [23] a new stroke width based method was introduced for text recognition in natural scenes. The algorithm was evaluated using the ICDAR 2003 dataset and additionally on a newly proposed dataset (MS Text DB [24]). The 307 annotated images cover the characteristics of natural images more comprehensively as



(a) brightness (b) contrast

(c) occlusion (d) rotation

(e) font (f) language

Figure 2. Brightness, contrast, rotation, occlusion, font and language statistics proving the diversity of the proposed NEOCR dataset. Graphs 2(a) and 2(b) also show the usual value of a scanned text document taken from a computer science book. The number of images refers to the number of textfields marked by bounding boxes.

in the ICDAR dataset. Unfortunately, not all text visible in the images has been annotated and the bounding boxes are parallel to the axes.

Additionally, there also exist some special datasets of license plates, book covers or digits. Still sorely missed is a well-annotated dataset covering the aspects of natural images comprehensively, which could be applied for comparing different approaches and identifying gaps in natural image OCR.

Ground truth annotations in the related datasets presented above are limited to bounding box coordinates and text transcriptions. Therefore, our comparison of current datasets is limited to statistics on the number of annotated images, the number of annotated textfields

Dataset	#images	#boxes	avg. #char/box
ICDAR 2003	509	2263	6.15
Chars74K	312	2112	6.47
MS Text DB	307	1729	10.76
Street View Text	350	904	6.83
NEOCR	**659**	**5238**	**17.62**

Table II
Comparison of natural image text recognition datasets.

(bounding boxes) and the average number of characters per textfield. The Chars74K dataset is a special case, because it contains word annotations and redundantly its characters are also annotated. For this reason, only annotated words with a length bigger than 1 and consisting of latin characters or digits only were included in the statistics in table II.

Compared to other datasets dedicated to natural image OCR the NEOCR dataset contains much more annotated bounding boxes. Because not only words, but also phrases have been annotated in the NEOCR dataset, the average text length per bounding box is also higher. None of the related datasets has added metadata information to the annotated bounding boxes. NEOCR surpasses all other natural image OCR datasets with its rich additional metadata, that enables more detailed evaluations and more specific conclusions on weaknesses of OCR approaches.

IV. Conclusion

In this paper the NEOCR dataset has been presented for natural image text recognition. Besides the bounding box annotations, the dataset is enriched with additional metadata like rotation, occlusion or inversion. For a more accurate ground truth representation distortion quadrangles have been annotated too. Due to the rich annotation several subdatasets can be derived from the NEOCR dataset for testing new approaches in different situations. By the use of the dataset, differences among OCR approaches can be emphasized on a more detailed level and deficits can be identified more accurately. Scenarios like comparing the effect of vocabularies (due to the language metadata), the effect of distortion or rotation, character arrangement, contrast or the individual combination of these are now possible by using the NEOCR dataset. In future we plan to increase the number of annotated images by opening access to our adapted version of the LabelMe annotation tool.

References

[1] J. J. Weinman, E. Learned-Miller, and A. R. Hanson, "Scene text recognition using similarity and a lexicon with sparse belief propagation," *IEEE PAMI*, vol. 31, no. 10, pp. 1733–1746, October 2009.

[2] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: a survey," *IJDAR*, vol. 7, pp. 84–104, 2005.

[3] L. Z. Chang and S. Z. ZhiYing, "Robust pre-processing techniques for ocr applications on mobile devices," in *ACM Mobility*, September 2009.

[4] "Word Lens." [Online]. Available: http://questvisual.com/

[5] "knfbReader." [Online]. Available: http://www.knfbreader.com

[6] K. Wang and S. Belongie, "Word spotting in the wild," in *ECCV*, 2010, pp. 591–604.

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

[8] D. Lopresti and J. Zhou, "Locating and recognizing text in www images," *Information Retrieval*, vol. 2, no. 2-3, pp. 177–206, May 2000.

[9] W. Wu, X. Chen, and J. Yang, "Incremental detection of text on road signs from video with application to a driving assistant system," in *ACM MM*. New York, NY, USA: ACM, 2004, pp. 852–859.

[10] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *IJCV*, vol. 77, pp. 157–173, May 2008.

[11] "LabelMe Dataset." [Online]. Available: http://labelme.csail.mit.edu/

[12] "ImageMagick." [Online]. Available: http://www.imagemagick.org

[13] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (jnb)," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 717–728, 2009.

[14] G. Wolberg, *Digital Image Warping*. Los Alamitos, CA, USA: IEEE Computer Society Press, 1994.

[15] "NEOCR Dataset." [Online]. Available: http://www6.cs.fau.de/research/projects/pixtract/neocr

[16] "ICDAR Robust Reading Dataset," 2003. [Online]. Available: http://algoval.essex.ac.uk/icdar/Datasets.html

[17] S. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "Icdar 2003 robust reading competitions," in *ICDAR*, August 2003, pp. 682–687.

[18] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. M. Miyao, J. Zhu, W. Ou, C. Wolf, J.-M. Jolion, L. Todoran, M. Worring, and X. Lin, "Icdar 2003 robust reading competitions: Entries, results, and future directions," *IJDAR*, vol. 7, no. 2-3, pp. 105–122, 2005.

[19] "Chars74K Dataset," 2009. [Online]. Available: http://www.ee.surrey.ac.uk/CVSSP/demos/chars74k/

[20] T. E. de Campos, M. R. Babu, and M. Varma, "Character recognition in natural images," in *VISAPP*, 2009.

[21] "Street View Text Dataset," 2010. [Online]. Available: http://vision.ucsd.edu/~kai/svt/

[22] "Google Street View." [Online]. Available: http://maps.google.com

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

[24] "Microsoft Text Detection Database," 2010. [Online]. Available: http://research.microsoft.com/en-us/um/people/eyalofek/text_detection_database.zip