NEOCR Dataset Metadata Documentation

Robert Nagy, Anders Dicker, Klaus Meyer-Wegener
Chair for Computer Science 6 (Data Management)
University of Erlangen-Nürnberg
Martensstr. 3., 91058 Erlangen, Germany

March 14, 2011

Contents

1 NEOCR Dataset

NEOCR is a comprehensive configurable dataset with rich annotation for OCR in natural images. The images cover a broad range of characteristics that distinguish real world scenes from scanned documents. Example images are shown in figure 1. The dataset contains a total of 659 images with 5238 textfields. Images were captured by the authors and other members of the chair using various digital cameras with diverse camera settings. Because the images were taken in several countries, 15 different languages are present in the dataset, though the visible text is limited to latin characters.

For image annotation, the web-based annotation tool provided by [3] for the LabelMe[1] dataset was used. 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 dataset and the XSD file can be downloaded from the NEOCR Dataset website[2].

[1] http://labelme.csail.mit.edu/
[2] http://www6.cs.fau.de/research/projects/pixtract/neocr/

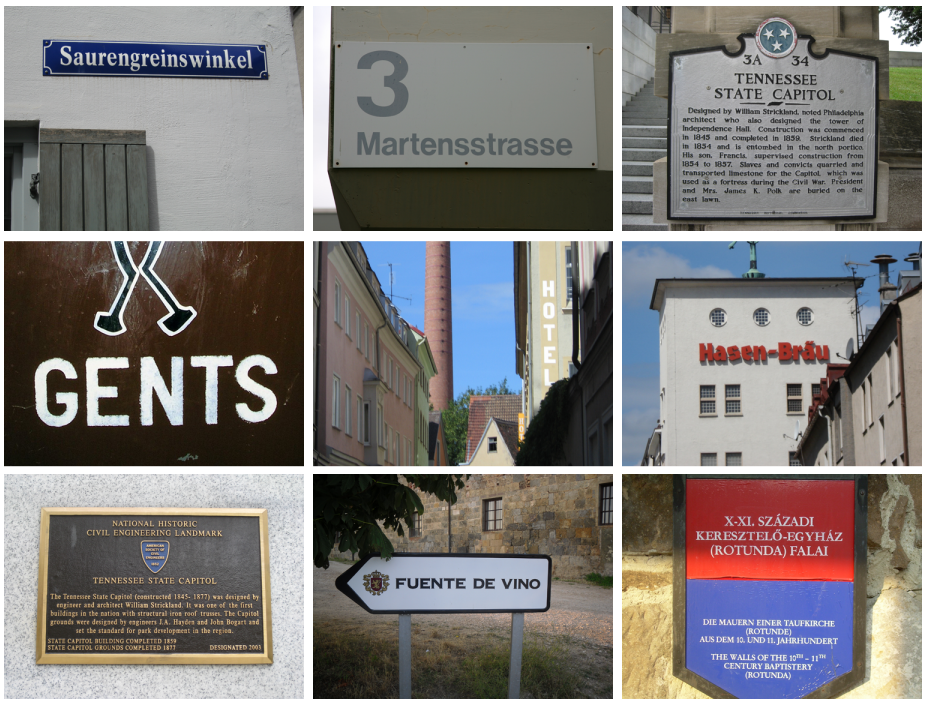

Figure 1: Example images from the NEOCR dataset. Note that the dataset also includes images with text in different languages, text with vertical character arrangement, light text on dark and dark text on light background, occlusion, good and bad contrast.

2 Metadata

One of the main advantages of the NEOCR Dataset is its rich annotation. Next to bounding boxes and its contained text, optical, geometrical and typographical characteristics have been annotated too. Therefore, the NEOCR dataset can also be separated as required based on selected dimensions and specific datasets can be derived to test new approaches for different scenarios.

2.1 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 using ImageMagick[3]. For each image all visible text is annotated using bounding boxes.

2.2 Bounding Box 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. The following optical, geometrical and typographical metadata are annotated for each bounding box separately.

2.2.1 Optical Characteristics

Texture. Texture is very hard to measure automatically, because texture differences can form the text and text itself can be texture too. We defined three categories to which we assigned the bounding boxes:

- low: single color text and single color background,
- mid: multi-colored text or multi-color 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 2.1). 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 only achieve resolutions up to 300dpi. The lower focal length is used, the bigger area is 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.

[3] http://www.imagemagick.org

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 ISO values or the higher compression rates are used, 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. [1] and [2] give overviews 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 (LoG). 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.

2.2.2 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 [4]. The 8 values can be represented as a matrix, where s_x and s_y describes scaling, r_x and r_y rotation, t_x and t_y translation, and p_x and p_y shearing.

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

The equations in [4] 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:

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

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

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 offset measured from the horizontal axis given by the image itself. The text rotation is computed based on the distortion parameters.

Arrangement. In natural images characters of a text can be arranged also vertically (e.g., hotel signs). Also some text can follow curved baselines. In the annotations we distinguish between horizontally, vertically and circularly arranged text. Some bounding boxes contain only one single character, which we 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.

2.2.3 Typographical Characteristics

Typefaces. Typefaces of bounding boxes are classified into categories standard, handwriting and special. Font thickness, size and upper or lower case chraracteristics are not annotated.

Language. Languages can be a very important information when using vocabularies for correcting errors in recognized text. Because we took images in several countries, the NEOCR dataset contains text in different languages, though they contain latin characters only. Altogether the dataset contains text in following languages (in alphabetical order): Belgian Czech, English, French, German, Greek, Hungarian, Italian, Latin, Portuguese, Russian, Spanish, Swedish, Turkish. In some cases, text cannot be clearly assigned to any language. For these special cases we introduced categories for numbers, roman numbers, person names, abbreviations and business names.

2.3 Example Annotation

In table 1 an example annotation for figure 2 is shown.

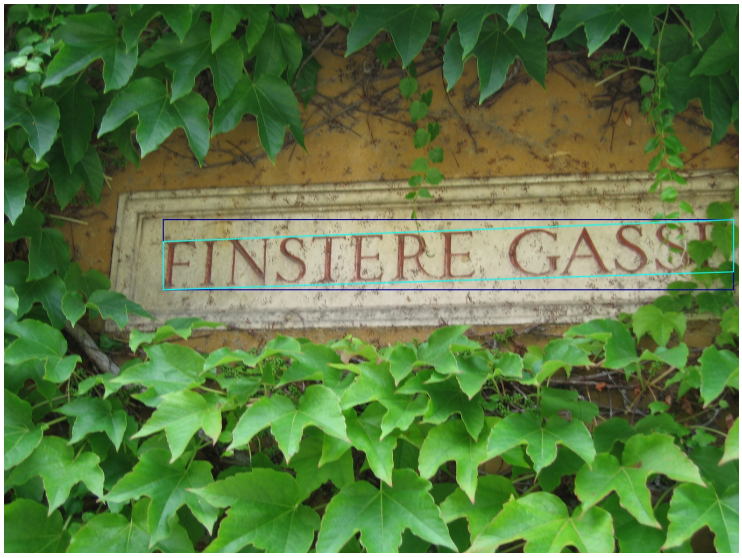

Figure 2: Example image from the NEOCR dataset. The annotated metadata is shown in table 1.

References

[1] C. F. Batten. Autofocusing and astigmatism correction in the scanning electron microscope. Master's thesis, University of Cambridge, August 2000.

[2] 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*, 18(4):717–728, 2009.

[3] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, May 2008.

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

CATEGORY	DATATYPE	VALUES RANGE	EXAMPLE VALUE
texture	string	low, mid, high	mid
brightness	float	[0;255]	164.493
contrast	float	[0;123]	36.6992
inversion	boolean	true, false	false
resolution	float	[1;1000000]	49810
noise	string	low, mid, high	low
blurredness	float	[1;100000]	231.787
distortion	8 float values	[-1;1]	sx: 0.92, sy: 0.67, rx: -0.04, ry: 0, tx: 0, ty: 92, px: -3.28-05, py: 0
rotation	float	[0;360]	2.00934289847729
occlusion	integer	[0;100]	5
occlusion direction	string	horizontal, vertical	vertical
typeface	string	standard, special, handwriting	standard
language	string	german, english, spanish, hungarian, italian, latin, french, belgian, russian, turkish, greek, swedish, czech, portoguese, numbers, roman date, abbreviation, company, person, unknown	german
difficult	boolean	true, false	false

Table 1: Example annotations for figure 2.