A Database for Scene Text Detection and Recognition: USTB Street View 1000 Images (USTB-SV1K)

USTB Street View Text Detection 1000 Database (USTB-SV1K) [1] is used to evaluate text detection algorithms. It is collected and released publicly by Pattern Recognition and Information Retrieval Lab of USTB (http://prir.ustb.edu.cn/TexStar/MOMV-text-detection/). It contains 1000 street view images many of which have texts in multiple orientations and views. This makes the database more suitable for evaluate multi-orientation scene text detection methods.

USTB Street View Text Detection 1000 Database (USTB-SV1K) is collected from Google Street View. For each GPS location, the 360-angle-view image in Google Street View is actually composed of 91 patch images with 512 × 512 size. We decode the web link from Google, and automatically download all 91 patches, which are captured by video cameras with multi-orientation and multi-view texts. From all patches, we manually select images with texts which can be seen by person.

USTB Street View Text Detection 1000 Database (USTB-SV1K) contains 1000 street view (patch) images from 6 USA cities, i.e., New York, Boston, Los Angle, Washington DC, San Francisco, and Seattle. The set from each city includes about 160-180 images. There are three main challenges for detection and recognition on this dataset (see samples in Fig. 1). First, a plenty of texts are in multiple orientations and views. About 75%, 10% and 15% images are with (near) horizontal, multi-orientation, and multi-view (always with skewed distortions) texts respectively. Second, this dataset includes a lot of small or blurred texts (about 28%). Third, about one fourth of texts are specific street and business names, or parts of words, and can't be found in a common dictionary.

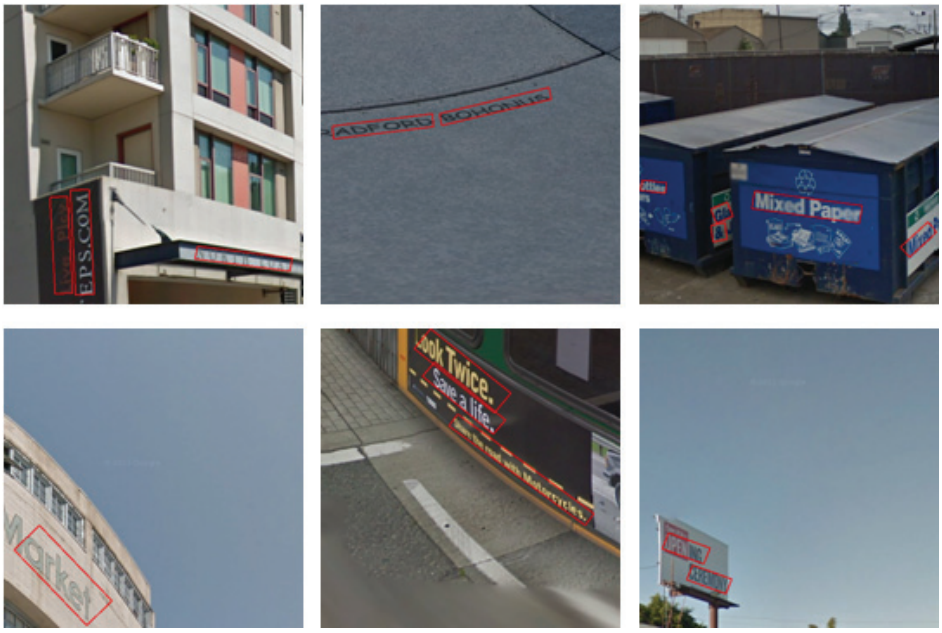

Fig. 1. Typical images from USTB-SV1K. Most multi-view scenes are also with skewed distortions. Detecting blurred, very small, or seriously distorted texts is challenging.

The dataset is divided into two parts: training set and the test set. The training set contains 500 images and the rest 500 images constitute the test set. All the 1000 images are fully annotated. The basic unit in this dataset is text line (see Figure 1) rather than word, which is used in the ICDAR datasets and MSRA-TD500 database.

The procedure of ground truth generation is shown in Fig. 1. While current evaluation methods for text detection are designed for horizontal texts only, we used a multi-orientation text detection evaluation protocol (see [2] for details). Minimum area rectangles are used in our protocol because they (rectangles in Fig. 1 (a)) are much tighter than axis-aligned rectangles (rectangles in Fig. 1 (b)).

Fig. 1. **(a) inclined rectangle.** **(b) axis-aligned rectangle.**

Format of ground truth files

Each image in the database corresponds to a ground truth file, in which each line records the information of one text. The format of the ground truth files is illustrated in Fig. 3.

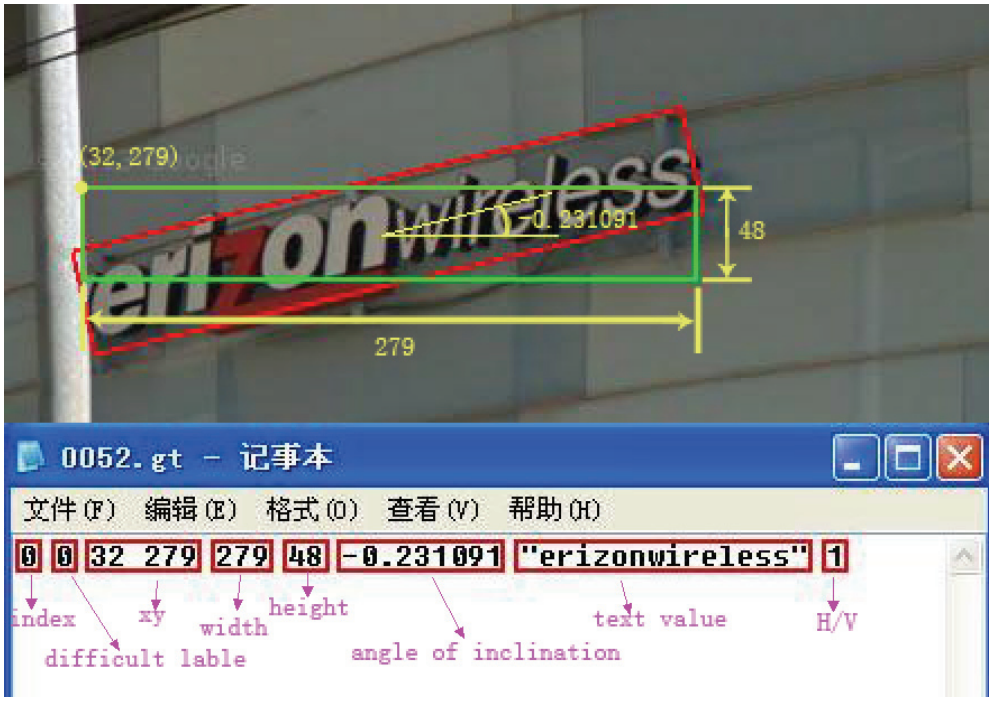

Fig. 3 Illustration of the ground truth file format. The index field can be ignored. The difficult label is "0" if the text content is easy to identify. If the text content is hard to recognize the text value and the difficult label will be "" and 1. The H/V is "1" if the rectangle is in the near horizontal direction and if the rectangle is in the near vertical direction it will be "2".

References

[1] X.-C. Yin, W.-Y. Pei, J. Zhang, and H.-W. Hao, "Multi-Orientation Scene Text Detection with Adaptive Clustering," submitted to IEEE TPAMI, 2014.

[2] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting Texts of Arbitrary Orientations in Natural Images," CVPR 2012.