UDK621.3:(53+54+621+66), ISSN0352-9045 Informacije MIDEM 41(2011)2, Ljubljana

CVL OCR DB, AN ANNOTATED IMAGE DATABASE OF TEXT IN NATURAL SCENES, AND ITS USABILITY

Andrej Ikica, Peter Peer

Computer Vision Laboratory (CVL), Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

Key words: computer vision, text detection, optical character recognition, natural scenes

Abstract: Text detection and optical character recognition (OCR) in images of natural scenes is a fairly new computer vision area but yet very useful in numerous applicative areas. Although many implementations gain promising results, they are evaluated mostly on the private image collections that are very hard or even impossible to get. Therefore, it is very difficult to compare them objectively. Since our aim is to help the research community in standardizing the evaluation of the text detection and OCR methods, we present CVL OCR DB, a public database of annotated images of text in diverse natural scenes, captured at varying weather and lighting conditions. All the images in the database are annotated with the text region and single character location information, making CVL OCR DB suitable for testing and evaluating both text detection and OCR methods. Moreover, all the single characters are also cropped from the original images and stored individually, turning our database into a huge collection of characters suitable for training and testing OCR classifiers.

Anotirana podatkovna baza slik teksta v naravnih scenah CVL OCR DB in njena uporaba

Kjučne besede: prenos radiofrekvenčnega signala prek optičnega vlakna, oddaljena antenska enota, celična arhitektura, vlakenski dostop, fiksno in mobilno zlivanje

Izvleček: Detekcija teksta in optična razpoznava simbolov (OCR) na slikah naravnih scen je razmeroma novo področje računalniškega vida, pa vendar zelo uporabna na številnih aplikativnih področjih. Mnoge implementacije dosegajo spodbudne rezultate, vendar njihova evalvacija večinoma poteka na privatnih zbirkah slik, ki so težko dostopne ali celo nedostopne, zato je metode med seboj zelo težko objektivno primerjati. Naš namen je pomagati raziskovalni skupnosti pri standardizaciji evalvacije omenjenih metod. Zato predstavljamo CVL OCR DB, javno bazo anotiranih slik teksta v naravnih scenah, ki so zajete pod različnimi vremenskimi in svetlobnimi pogoji. Vse slike v bazi vsebujejo informacijo o lokacijah prisotnih tekstovnih regij in posameznih črk, kar omogoča testiranje in evalvacijo tako metod detekcije teksta, kot tudi metod razpoznave simbolov. Vsi posamezni znaki so dodatno izrezani iz originalnih slik ter individualno shranjeni, kar naredi našo podatkovno bazo ogromno zbirko znakov, primerno za učenje in testiranje klasifikatorjev OCR.

1. Introduction

Due to a broad range of applicative areas, text detection and OCR in natural scene images have gained a lot of research interest in the last decade. Since the research area is just beginning to evolve, there are obviously no widely accepted standards for both evaluation methodology as well as the common evaluation database, as it is typical in other computer vision areas such as face detection and recognition /1,2/. Indeed, there already exist image databases of text in natural scenes such as the ICDAR database /3/. These collections, however, include only a couple of hundreds of images and are more or less intended for private use or for competitions, whereas our idea is to establish a public collection of thousands of images that would organically grow over time.

Our approach has several advantages. First, all the existing methods can be validated and objectively compared to each other on the same dataset and against the same ground-truth data. Second, with CVL OCR DB it is easy to compare the new text detection and OCR approaches with the state-of-the-art methods. Many authors claim to achieve better results compared to other (older) methods, although they often do not explicitly reveal which datasets have been used for evaluation or they simply evaluate methods on their private datasets. Finally, CVL OCR DB is a public database, which hopefully will be expanded by a diverse number of researchers and volunteers, so all the image data will actually represent a vast variety of possible scenarios. Many researchers, unfortunately, establish their private image datasets with the drawbacks of their methods in mind. Thus, they (not being aware of) avoid integrating "worst-case scenario" images into their datasets. We already tried to avoid this issue in the process of collecting and annotating image data by intentionally including the volunteers who had no previous experience in text detection and OCR whatsoever. Therefore, the collected data were actually the images seen by an average human observer, not a computer vision expert. The results were often contrary to our expectations and included complicated fonts and styles, very complex backgrounds, characters substituted by other symbols, text skews, rotations etc.

The article is organized as follows. First, the CVL OCR DB construction life-cycle is described in section 2, followed by the description of the capturing methodology in Section 3. The conceptual model of annotation and the model implementation are presented in Section 4 and Section 5, respectively. In Section 6 the annotation process is described and in Section 7 we present the practical aspects of our database and evaluation results of several text detection

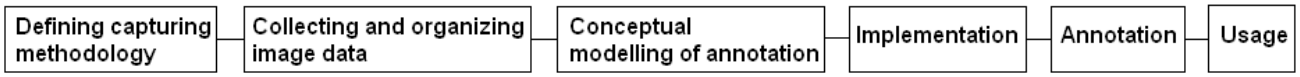

Fig. 1. The CVL OCR DB construction life-cycle.

methods on the CVL OCR DB subset. Finally, the article is concluded with the discussion and the terms for using CVL OCR DB in Section 8.

2. Construction life-cycle

The CVL OCR DB construction life-cycle is divided into 6 stages as shown in Fig. 1. After defining the coarse capturing methodology, the image data was collected and organized. Through the process of collecting and organizing the image data the methodology was constantly refined in order to capture as broader set of different scenarios as possible in the best possible way. The conceptual model of annotation developed afterwards was used for the actual implementation and the ground-truth data annotation. Finally, the last stage is dissemination and usage of the database for evaluation of different computer vision algorithms.

3. Capturing methodology

In the image data collecting stage both the authors and the volunteers were involved. We captured a great number of images of text in natural scenes, including images of signboards, shop names, traffic signs and jumbo posters (Fig. 2). All the images were captured at varying weather and lighting conditions with either a compact camera (up to 7 mega pixels) or a mobile phone (up to 5 mega pixels), but were eventually resized to fit the computer screen. There are three main reasons for resizing the images. First, the resized (smaller) images contain enough visual information, therefore, the original (bigger) images would unnecessarily occupy the disk space. Second, smaller images can be processed much faster. The third reason, however, lies in the fact that precise annotation (selecting bounding polygons and cropping characters) can be performed much easier on the images that do not stretch beyond the computer screen dimensions.

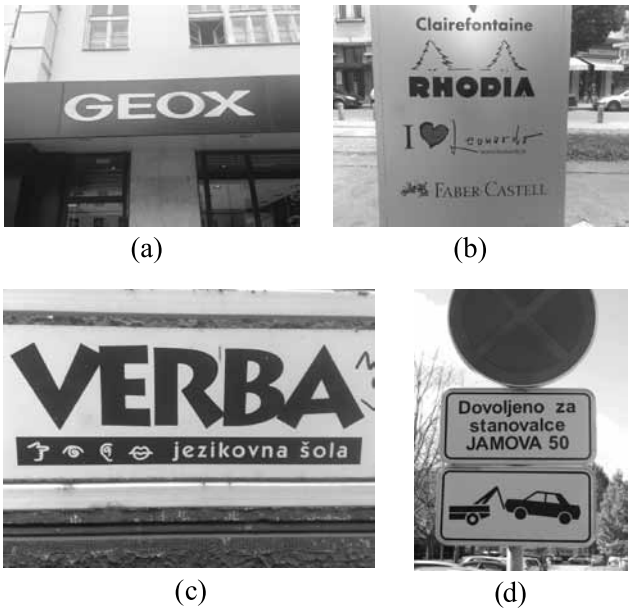

Fig. 2. Examples of captured text images.

All the collected images are divided into three main categories: "day", "night" and "artificial" (tree structure in Fig. 3). The "day" category corresponds to the images captured at daylight, the "night" category corresponds to the images captured at night time and, finally, the "artificial" category corresponds to the images captured on locations where only artificial light is present (eg. shopping centers). Further on, the "day" category is divided into 4 subcategories: "normal" (normal sunlight), "sun" (very intense sunlight), "fog" and "rain". Similarly the "night" category is divided into 2 categories: "dusk" and "night" (the actual night time). This top level structure is maintained in a directory structure (Fig. 3). Every leaf node of the diagram in Fig. 3 corresponds to the actual image repository for a given category/subcategory. In other words, all the images of the particular category/subcategory are physically stored in these repository directories. We will refer to these directories as data folders.

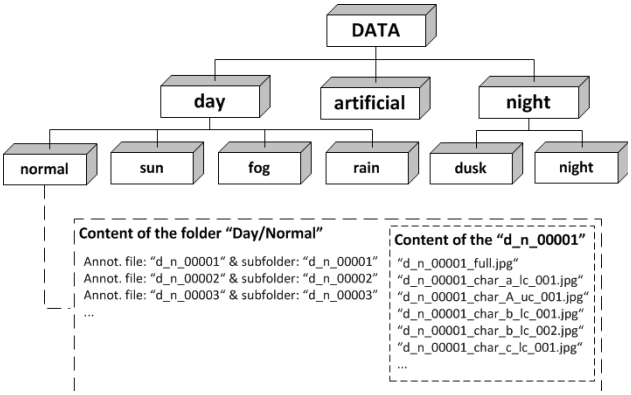

Fig. 3. File organization of CVL OCR DB.

Each data folder contains a list of annotation files. The annotation file naming convention is the following: The filename starts with one of the following prefixes: "d_n" (day/normal), "d_s" (day/sun), "d_f" (day/fog), "d_r" (day/rain), "a" (artificial), "n_d" (night/dusk) and "n_n" (night, night), depending on the category/subcategory of the data folder. The prefix is followed by the underscore and a 5-digit image index (including the leading zeros). For example: an annotation file for a 325th image in the day/fog category is named as "d_f_00325". Similarly a 12345th annotation file in the artificial category is named as "a_12345". The 5-digit index has an upper bound of maximum 99,999 images in a certain data folder, which is more than enough for the purpose.

Every annotation file is linked to an actual image and a list of character images in it (Section 6 describes the process of getting the character images in the annotation stage). Both are stored in a subdirectory of the particular data folder. We will refer to these subdirectories as the image folders. The image folder naming convention is the same as the annotation file naming convention.

The actual images in the image folders are stored as JPEG files. Their naming convention is again simple and very similar to the naming convention used so far. The full image is named after the image folder it resides in and is concluded with the "_full.jpg" suffix. For example, a full text scene image in a "d_f_00325" subdirectory is named "d_f_00325_full.jpg". Single character naming convention is a little more complicated, since the single characters can occur more than once. For instance, the text scene image in Fig. 2b contains 2 letters "C", 4 letters "a" and 3 letters "A". Every character image filename starts with a subdirectory prefix followed by the "_char_<char_symbol>_<lc|uc>_<3_digit_index>.jpg" (see dashed rectangle in Fig. 3 for details), where "lc" corresponds to the lowercase and "uc" to the uppercase. For example, if our imaginary image "d_f_00325_full.jpg" consisted of 2 letters "C", 1 letter "a" and 1 letter "A", the corresponding character image filenames would be:

"d_f_00325_char_C_uc_001.jpg",
"d_f_00325_char_C_uc_002.jpg",
"d_f_00325_char_a_lc_001.jpg",
"d_f_00325_char_A_uc_001.jpg".

The "lc/uc" labels are used to avoid treating the upper and lowercase character filenames as the same filename. Without these labels some operating systems would treat "d_f_00325_char_a_001.jpg" and "d_f_00325_char_A_001.jpg" as the same filename. Moreover, the "lc/uc" labels are a great advantage when, for example, researcher wants to find all the uppercase letters in the database. The given naming convention allows a maximum of 999 of same characters per single full image, which is more than enough for the purpose.

It might seem that the proposed naming convention introduces redundant data into our methodology – the image prefixes and folder names, for example, contain the same attributes. There are two main reasons why we proposed such a naming convention. First, the category is included in the filename to exploit the operating system functionality for simple queries – all the images of a certain category/subcategory can be obtained via simple filename operating system search. Second, we want our database to be consistent even in the case of human errors – for example, when some image files are accidentally moved to other folders.

4. Conceptual model of annotation

Another added value of CVL OCR DB is the annotated ground-truth data, ie. the annotation files which correspond to the text in images. For every image in the database there exists a corresponding annotation file. As shown on the left side in Fig. 4, every annotation file ("AnnotationFile" class) is linked to the actual image ("TextSceneImg" class) and multiple single character images ("CharImage" class) cropped from the image. Single characters can be used for OCR classifier training and testing. Fig. 5 shows an example of an image with three text regions. For clarity, cropped characters of the first region are displayed as well on the right side of the figure.

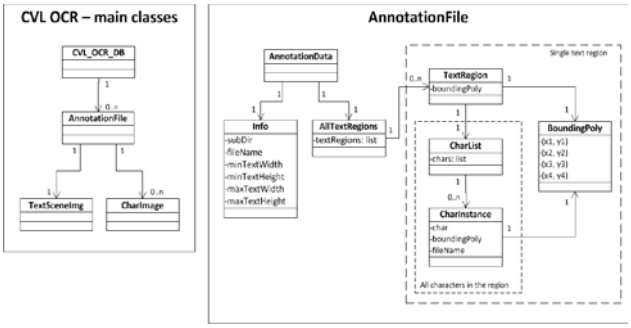

Fig. 4. The CVL OCR DB conceptual model.

Fig. 5. Text regions (left) and single characters of the first text region (right).

The annotation file (see the right side in Fig. 4) contains two main classes: the "Info" class and the "AllTextRegions" class. The "Info" classincludes the actual image filename ("filename"), its location ("subdir") and the size constraints (min. and max. dimensions) of the text regions present in the particular image. In the database usage stage, all the text regions detected by the text detection methods violating the size constraints are simply ignored, making the evaluation process equally fair for all the methods. "AllTextRegions" class is a list of all text regions present in the image. Each text region, namely "TextRegion" class, corresponds to the physical text region in the image and is described with a bounding polygon ("boundingPoly") and a list of characters it contains. For example, the first text region in Fig. 5 contains characters "N", "O", "K", "I" and "A". Each character "CharInstance" in the "CharList" is described with the UTF-8 character code, its filename and again with the bounding polygon.

5. Model implementation

Since the XML files represent a very flexible, easy-to-read data format, they are a reasonable choice for describing the annotation files. Therefore, the conceptual model was implemented in the XML Schema Definition (XSD) Language /4/. The schema defines the grammar and the XML documents that follow its rules are nothing more than words in language defined by the schema. An implementation decision was taken as well, to follow a recommendation that all tags be elements with no attributes /5/. Fig. 6 shows an XSD schema for the XML annotation files derived from the conceptual model of annotation.

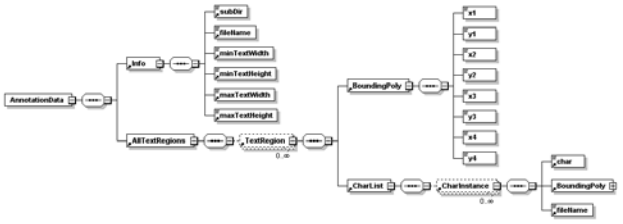

Fig. 6. XML implementation of the conceptual model - XSD schema for XML annotation files.

6. Annotation

For the annotation purposes a program for fast and effective annotation, namely TextAnnotator, was implemented. An average user can select multiple images, mark the text regions (with bounding polygons) and crop the individual characters. The annotation program automatically stores the image data in the appropriate data/image folders and creates new XML annotation files or updates the existing. In order to preserve database consistency, all XML annotation files are validated against the XSD schema (Fig. 6). Since the images are often affected by perspective and projective transformations, the text can appear rotated and skewed. To solve these issues, a 4-point bounding polygon is chosen (instead of a rectangular bounding box) to represent the text regions. Moreover, the rotated single characters pose a serious problem to OCR classifiers, so the cropped characters can be rotated by a human annotator by simply dragging the mouse in the desired direction.

Currently there are much over 1,000 annotated images of text in natural scenes in the CVL OCR DB. We encourage other researchers to participate in expanding the database by annotating and uploading new text scene images (see Discussion and URL sections of the article for participation details).

7. Using cvl ocr db in practice

CVL OCR DB has an enormous practical potential, since it is a step towards standardizing the evaluation approach for text detection and OCR methods on images of text in natural scenes. It is of vital importance to be able to objectively compare methods on the same input images and the same ground-truth data.

CVL OCR DB has already been used to compare several text detection methods /6/. We compared our proposed text detection method based on the projection profiles /7,8/ against two other text detection methods, namely Ezaki /9/ and Shivakumara /10/ methods. The evaluation results in terms of precision *p*, recall *r* and quality measure *f* /6/ are shown in Table 1.

Table 1. Practical example of CVL OCR DB usage: evaluation results of different text detection methods (bigger number means better result).

Text detection method	p	r	f
Shivakumara /8/	58,7%	51,0%	54,6%
Ezaki /7/	58,0%	53,1%	55,4%
Proposed method /4/	70,9%	55,2%	62,1%

8. Discussion

In this article we presented CVL OCR DB, a public annotated image database of text in natural scenes. CVL OCR DB has several advantages. First of all, the database is public and can therefore be expanded with new images. Second, CVL OCR DB comes with the TextAnnotator program which is easy to use and enables fast database expansion without affecting the structure and internal relationships. And finally, CVL OCR DB was constructed through several iterations and revisions, so the exceptions, special cases and boundary conditions are taken into account. It is important to understand that CVL OCR DB differs from the raw image databases that contain only images and no ground-truth data whatsoever. Our database is ground-truth annotated – not only in terms of region and character locations, but also in terms of capture and lighting conditions. The important aspect is definitely its multi-functionality, since it can serve for both text detection and OCR purposes.

Our aim is to try to contribute to the research community by proposing a functional and easy-to-use framework for evaluating and comparing text detection and OCR methods. We would like to encourage the research community to participate in CVL OCR DB usage – not only for the evaluation purposes but also as active contributors who help expanding it by adding and annotating new images. All the necessary material, including text annotation program, documentation and usage agreement can be downloaded from the CVL OCR DB web portal:
URL
CVL OCR DB web portal: http://www.lrv.fri.uni-lj.si/~peterp/CVLOCRDB/.

References

/1./ Resources for Face Detection (2010). Retrieved December 20 2010 from http://vision.ai.uiuc.edu/mhyang/face-detection-survey.html#face-database

/2./ Face Databases (2010). Retrieved December 20 2010 from http://web.mit.edu/emeyers/www/face_databases.html

/3./ S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "ICDAR 2003 robust reading competitions", Proc. of 7th Int. Conf. on Document Analysis and Recognition (ICDAR 2003), Vol. II, 3-6 August 2003, pp. 682–687

/4./ XML Schema Definition (XSD) Language (2010). Retrieved December 20 2010 from http://www.w3.org/TR/xmlschema-0/

/5./ XML tags and attributes (2010). Retrieved December 20 2010 from http://www.oasis-open.org/cover/elementsAndAttrs.html

/6./ A. Ikica, P. Peer, "An improved edge profile based method for text detection in natural images", Eurocon 2011

/7./ J. Park, G. Lee, E. Kim, J. Lim, S. Kim, H. Yang, M. Lee, S. Hwang, "Automatic detection and recognition of Korean text in outdoor signboard images", Pattern Recognition Letters, 2010.

/8./ T. N. Dinh, J. Park, G. Lee, "Korean Text Detection and Binarization in Color Signboards", Proc. of The Seventh Int. Conf. on Advanced Language Processing and Web Information Technology (ALPIT 2008), pp. 235-240

/9./ N. Ezaki, M. Bulacu, L. Schomaker, "Text Detection from Natural Scene Images: Towards a System for Visually Impaired Persons", Int. Conf. on Pattern Recognition (ICPR 2004), vol. II, pp. 683-686

/10./ P. Shivakumara, T. Q. Phan, C. L. Tan, "Video text detection based on filters and edge features", Int. Conf. on Multimedia & Expo (ICME 2009), pp. 514-517

Andrej Ikica, Peter Peer

Computer Vision Laboratory (CVL)
Faculty of Computer and Information Science,
University of Ljubljana
Tržaška 25, Ljubljana, Slovenia
andrej.ikica, peter.peer}@fri.uni-lj.si

Prispelo: 27.05.2010 *Sprejeto: 24.06.2011*