Open access • Proceedings Article • DOI:10.1109/ICDAR.2013.41

Part-Based Recognition of Arbitrary Fonts — Source link

Song Wang, Seiichi Uchida, Marcus Liwicki

Institutions: Kyushu University, German Research Centre for Artificial Intelligence

Published on: 25 Aug 2013 - International Conference on Document Analysis and Recognition

Topics: Intelligent character recognition, Intelligent word recognition, Three-dimensional face recognition, Document processing and 3D single-object recognition

Part-Based Recognition of Arbitrary Fonts

Wang Song, Seiichi Uchida
Kyushu University, Fukuoka, Japan
wangsong@human.ait.kyushu-u.ac.jp, uchida@ait.kyushu-u.ac.jp

Marcus Liwicki
DFKI, Kaiserslautern, Germany
Marcus.Liwicki@dfki.de

***Abstract*—In this paper, the part-based recognition method is introduced and applied to the arbitrary font recognition. The principle of the part-based method is to represent the character image as a set of parts and then recognize the image by finding the most possible parts set from the reference database. Since the part-based method does not rely on the global structure of a character, it is supposed to be robust against the variant appearances of the character. The experiment results indicate that it is possible to apply the part-based method to the font recognition, which is always considered as a difficult task by most of the researchers.**

Keywords*—part-based recognition, font recognition, local features

I. Introduction

Recently, with the development of the built-in camera in the mobile devices, now users can take photos anywhere and share them via Internet. With this trend, text recognition in images (scene character recognition [1]) begins to attract attention of the researchers who are working in the character recognition field. Many fascinating applications can be expected by using the scene character recognition technology, such as image search, mobile dictionary (translator) [2] and reading-life log. However, there are still many difficulties in scene character recognition; for example, the distortion and complex background for the text localization and the various character appearances for the text recognition. Up to now, many trials were made for the text localization and extraction [3]–[5], whereas the recognition was usually left to the document OCR engine.

Different from traditional document recognition, recognition of scene characters is a very difficult task. This is because general scene text often contains characters in specially-designed or decorated fonts, sometimes even the handwritten and 3D fonts. Figure 1 shows several examples of those characters. Clearly, the traditional OCR may have a very poor performance on those characters. Although there is some trial focused on the OCR which does not rely on the font information [6], within a single document image, a large quantity of text in the same font is necessary for this kind of method to recognize the characters. Usually, scene images contain only several words, which is extremely insufficient to the font-free OCR. Consequently, other methods should be considered for the recognition of arbitrary fonts.

Recently, part-based methods have been proposed for handwritten character recognition [7], [8] with promising results. In the part-based methods, a character image is recognized as a set of parts (local feature descriptors), and the global structure information is usually discarded. By only using the parts for the recognition, the part-based methods are very flexible to adapt to various character appearances and robust against the distortion of the characters.

The advantages of the part-based method are very important for the scene character recognition. This is because the scene characters are usually in various fonts (as the "S" shown in Fig. 1). Globally, those "S" look very different to each other. However, they all share some similar part, for example, the curve edge. In the part-based method, by using the similar parts, the different appearances of "S" can be correctly recognized.

In this paper, a general framework of the part-based method for an arbitrary font recognition[1] is introduced and three different part-based methods are proposed, especially one of them is first introduced in this paper. The three part-based methods are tested on the data of alphabets of multiple fonts as well as a commercial OCR. Based on the experiment results, an analysis is also made in order to find out the reason that why the different method has the different performance.

II. The Part-Based Methods - General framework and various implementations

In this section, the general framework of the part-based method is first introduced. Based on this framework, it is possible to generate different kinds of part-based methods. As examples, three different part-based methods are then introduced as well as the implementation methodology. A typical part-based method has the following properties.

- The query image is decomposed into several parts and each part is represented as a local feature descriptor. This process can be realized by some feature detection and description method, for example, the speeded-up robust features (SURF) [9]. All the local feature descriptors (hereafter called the keypoints) are used for representing this query image. The relative position of those keypoints is *not* considered in the recognition process.
- Reference keypoint database is created by extracting keypoints from the training images. Different combination of those reference keypoints can be seen as a "reference" (a set of reference keypoints) for an entire character. Please note that this reference may contain the keypoints from different training images or even different classes. We can also add restrictions to the combination of the reference keypoints to design different part-based methods.

[1]This paper is *not* related to "font recognition", which is the task to recognize the font type (e.g., bold, italic, Helvetica) of an input character. The purpose is to recognize its character class (e.g., "A", "b").

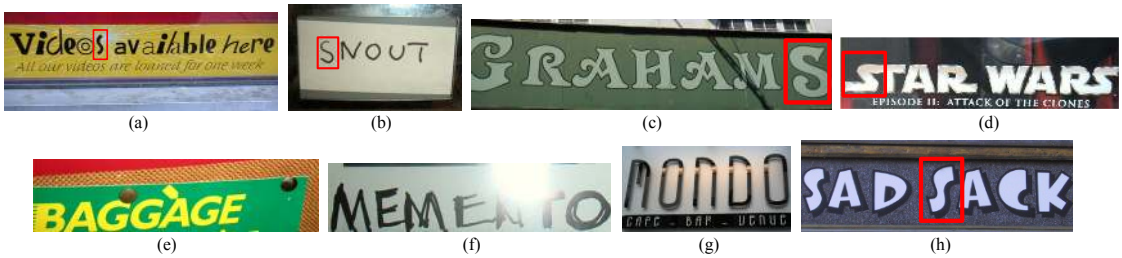

Fig. 1. Examples of font variation from ICDAR 2011 Robust Reading Competition dataset. The specially-designed font can be found in (a), (c), (g) and (h), decorated font in (e), handwritten font in (b) and (f), 3D font in (d). The red boxes are the same letter "S" in different font.

- The similarity of the query image and a certain class is measured by comparing the query keypoint set and the references. A distance is defined for this comparison and which is introduced in the following part.

A. The General Framework of the Part-Based Method

As noted above, the query image is represented as a set of query keypoints while the reference keypoint database is comprised of the reference keypoints. A set of reference keypoints, which has the same number of keypoints with the query keypoint set, is seen as a "reference". With a defined distance, the 1-nearest-neighbor (1NN) reference of the query keypoints set is searched. The class of this reference is evaluated as the final recognition result.

The process above can be represented as the following. Let Q denote the query image and the query keypoint set, which contains n query keypoints $k_1, \ldots, k_n$. A large reference keypoints database can be denoted as A and which contains all the reference keypoints $r_1, \ldots, r_L$, which are extracted from a large number of training patterns from all the classes. Let R denote a reference, which contains n reference keypoints $r_1, \ldots, r_n$ selected specially for Q from A. Here, based on the Euclidean distance in the feature space (the keypoint is actually an N-dimensional feature vector in the N-dimensional feature space), we define a distance between two keypoint sets—the query keypoint set Q and the reference R (QR distance):

$$D_{QR}(Q||R) = \sum_{i=1}^{n} ||k_i - r_{\tilde{R}(i)}||,$$

where the $r_{\tilde{R}(i)}$ is the Euclidean 1NN of k_i in R. The tilde " ˜ " indicates that this is optimized for k_i in R. With the QR distance, we have:

$$R_{\tilde{A}} = \underset{R}{\operatorname{argmin}}\, D_{QR}(Q||R),$$

where the $R_{\tilde{A}}$ is the 1NN R from A of Q by the minimum QR distance.

Assume that for any keypoint set, we can use a classification function f to evaluate its class C ($f : * \to C$, where $*$ can be Q, or R, or single keypoint k). In the part-based method, the $R_{\tilde{A}}$ is used for predicting the class of Q:

$$f(Q) = f(R_{\tilde{A}}). \quad (1)$$

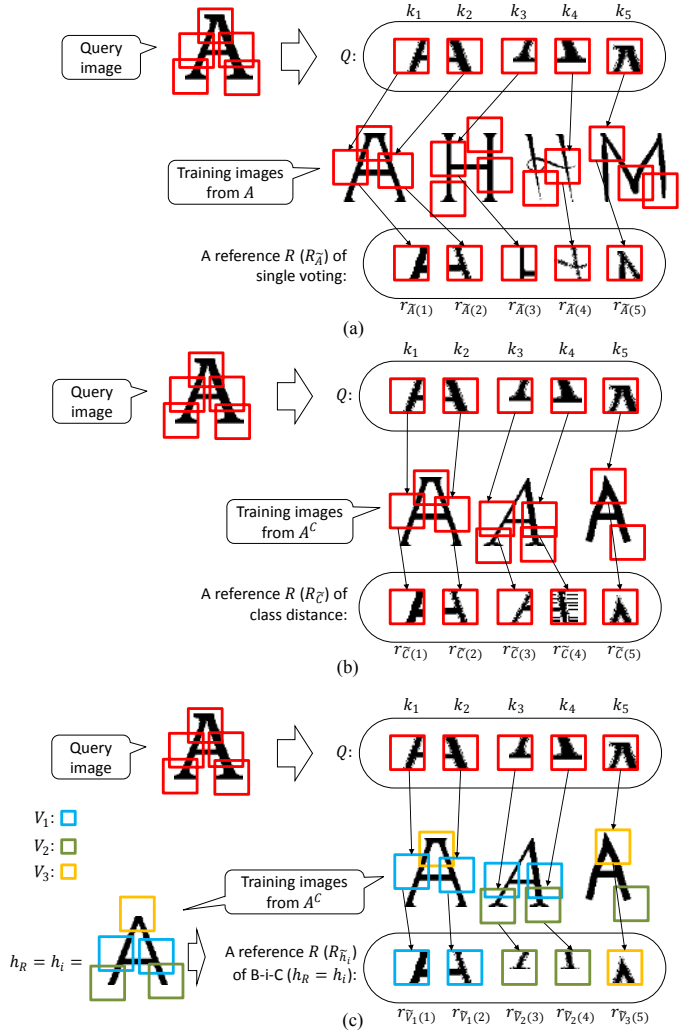

Fig. 2. The selecting of reference R in different part-based methods. In the figure, each color box stands for a keypoint and in the R selecting of the B-i-C method, the color of the box stands for the visual word.

B. Three Different Part-Based Methods

In the framework of the part-based method, there are still two questions which need to be discussed: first is how to select an R from A and the second is how to determine the evaluation function f of R. By using different strategies, different kinds of part-based methods can be generated. In the recognition of arbitrary font, those part-based methods may have different

performance and advantage. In order to find the optimal part-based method for arbitrary font recognition, an analysis of those methods should be conducted.

1) The Single Voting Method: In the part-based method of [7] (hereafter called the single voting method), a simplest strategy was used, that is, any combination of the reference keypoints can be seen as an R (as shown in Fig. 2 (a)):

$$R = \{r_1, \ldots, r_n | r_i \in A\}.$$

If all the r_i of R are from the same class C, naturally, we have $f(R) = C$. However, in the single voting method, since there is no restriction of selecting an R from A, thus r_i of R may be from different classes. Consequently, the function $f(R)$ provides the majority class in $f(r_1), \ldots, f(r_n)$. According to (1), the final result is $f(Q) = f(R_{\tilde{A}})$, where $R_{\tilde{A}}$ is the R optimized by keypoint-wise 1NN strategy, as noted before.

2) The Class Distance Method: In [8], a part-based method, called the class distance method, was proposed. Compared with the single voting method, the class distance method has a restriction of selecting R, that is, all the r_i of R should be extracted from the same class, as shown in Fig. 2 (b). By using this restriction, strange R (for example, the one contains r_i from multiple classes) in the single voting method can be removed. As a result, the class distance method is expected to have better performance than the single voting method.

If the R is from class C, let A^C denote all the reference keypoints $r_1, \ldots, r_M$ of class C, then:

$$R = \{r_1, \ldots, r_n | r_i \in A^C\}.$$

Since in the R, all the r_i are from the same class C, naturally, we have:

$$f(R) = C.$$

Similarly, according to (1), the final result is $f(Q) = f(R_{\tilde{A}})$.

3) The Bag-of-Features in a Class: Inspired by the bag-of-features method in the image classification, we have designed a new part-based method which used the visual words and the corresponding histograms as the restriction of selecting R. This method is called the "bag-of-features in class" method (the B-i-C method). In the B-i-C method, within the class C, all the reference keypoints of C are clustered into several visual words and then each training image can be represented as a histogram. Note that different from the traditional bag-of-features method, this clustering is done within a class but not on the whole training data. Assume in class C, there are J visual words $V_1, \ldots, V_J$. Let H denote all the histograms of the P training images in class C, which are $h_1, \ldots, h_P$.

In the B-i-C method, there are two restrictions of selecting R. The first restriction is the same with the class distance method, that is, all the r_i of R should be extracted from the same class. Assume that the R is from class C, then the same with the training images in class C, R should have a histogram h_R. The second restriction of B-i-C is $h_R \in H$ (as shown in Fig. 2 (c)). In other words, the h_R should be exactly the same with a certain h_i in H, or this R should not be selected. In summary, we have:

$$R = \{r_1, \ldots, r_n | r_i \in A^C\}, \text{and } h_R \in H.$$

This new restriction aims to make further reduction of the strange R. For example, an R from the class distance method may contain the r_i all from the left side of "H". Consequently, although the class of this "R" is "H", this R is more like "I". In the B-i-C method, those strange "R" can be removed by the new restriction. Then the same with the class distance method, if the r_i of R is from the class C, finally we have $f(R) = C$ and $f(Q) = f(R_{\tilde{A}})$.

C. Implementation Methodology

As noted above, the different combination of the reference keypoints can be seen as a reference R. Even if the number of the reference keypoints is small, the number of generated R will still become astronomical. For example, assume there are $1,000$ reference keypoints in A, and each query image Q contains 50 keypoints, then for the single voting method, the number of possible reference R is 1000^{50}. Therefore, it is impossible to compare the Q with every possible R. However, if we simply use the Euclidean 1NN of k_i in A or a certain class C, we can implement the three part-based methods easily.

Let $r_{\tilde{A}(i)}$ denote the Euclidean 1NN of k_i in A, clearly, as shown in Fig. 2 (a), in the single voting method, the $R_{\tilde{A}}$ is:

$$R_{\tilde{A}} = \{r_{\tilde{A}(1)}, r_{\tilde{A}(2)} \ldots, r_{\tilde{A}(n)}\}.$$

This equation means that if an R contains the Euclidean 1NN of all the k_i in A, then it is the 1NN of Q in A. Consequently, in the single voting method, we just need to search the $r_{\tilde{A}(i)}$ of every k_i and then the $R_{\tilde{A}}$ is found. As noted above, by counting the number of $r_{\tilde{A}(i)}$ of each class, the class with the maximum number of $r_{\tilde{A}(i)}$ is the final result $f(Q)$.

In the class distance method, let the $r_{\tilde{C}(i)}$ denote the Euclidean 1NN of k_i in A^C, $R_{\tilde{C}}$ denote the 1NN of Q in A^C, then similar with the single voting method, as shown in Fig. 2 (b), within each class C, we have:

$$R_{\tilde{C}} = \{r_{\tilde{C}(1)}, r_{\tilde{C}(2)} \ldots, r_{\tilde{C}(n)}\}.$$

By the equation above, we can find out the $R_{\tilde{C}}$ of Q in each class. The $R_{\tilde{C}}$ which has the minimum QR distance to Q is the $R_{\tilde{A}}$. The class of $R_{\tilde{C}}$ is the final result.

The implementation of B-i-C method is more complicated than the class distance method. This is because there is one more restriction in the B-i-C method: $h_R \in H$. For a given k_i, its Euclidean 1NN in V_j is noted as $r_{\tilde{V_j}(i)}$. The first step of B-i-C method is to find out the Euclidean 1NN of k_i in every visual word, which are $\{r_{\tilde{V_j}(1)}, \ldots, r_{\tilde{V_j}(n)}\}$.

Second, for a given h_i, under the condition $h_R = h_i$, we can also find the QR distance 1NN of Q, which is denoted as $R_{\tilde{h_i}}$. Clearly, all the keypoints of which should be the Euclidean 1NN of k_i in a certain V_j, which are $r_{\tilde{V_j}(i)}$ (as shown in Fig. 2 (c)). This is because all those keypoints are the Euclidean 1NN of k_i, thus the QR distance will increase if any other reference keypoint is used instead of the Euclidean 1NN. Of course, different kinds of combinations of the $r_{\tilde{V_j}(i)}$ keypoints may represent the same h_i, therefore we need to find out the combination which have the minimum QR distance with Q as the $R_{\tilde{h_i}}$. After find the $R_{\tilde{h_i}}$ of every h_i in H, the $R_{\tilde{C}}$ is the $R_{\tilde{h_i}}$ which has the minimum QR distance to Q. When the $R_{\tilde{C}}$ is found, we can find the $R_{\tilde{A}}$ just like the class distance method.

TABLE I. RECOGNITION RATES (%).

Methods	Recognition rate
Single voting	62.8
Class distance	73.5
B-i-C	71.5
OCR	56.7

III. EXPERIMENTS AND ANALYSIS

In this section, the three part-based methods mentioned above are tested on the data of arbitrary-font alphabets as well as a commercial OCR. Through the analysis of the experiment results, we may find out that what is important for the recognition of arbitrary fonts.

A. Experiment Setup

In the experiments, as shown in Fig. 3, we used the data which contains the alphabets in variety fonts. Those fonts were very different from the fonts that we usually use in the documents. We can find the specially-designed, decorated, handwritten and 3D fonts in the data and which usually can be found in the scene text.

For the experiments, we only considered 26 classes, which means the upper-case and the lower-case letters were seen as the same class. The SURF feature [9] was used as the keypoint detection and description method. From the data, for each class, 1,000 images were chosen randomly as the training data while 100 images were chosen as the test data. Each image was preprocessed to the same size of 200×200 and then the 120-pixel-wide border was added. This is to ensure that enough keypoints can be extracted from each image. Consequently, in the training data, there were 26,000 images and 8,056,673 keypoints were extracted; in the test data, there were 2,600 images and 724,561 keypoints were extracted.

B. Experiment Results

The results of the experiments are shown in Table I. Among all the methods, the OCR had the lowest recognition rate. This is because the OCR is designed for recognizing characters on ordinary documents. In the three part-based methods, the class distance method had the highest recognition rate, which was much higher than the recognition rate of the OCR. Since now most of the researchers are using the traditional OCR for the recognition of the scene characters, therefore if the part-based method can be used instead of the OCR, the total recognition rate of their system will be improved much. Moreover, since the part-based method does not rely on any global structure of the character, thus it is robust against the distortion and the noise of the character image, which usually can be found in the extracted scene character image.

The class distance method also had much better performance than the single voting method. This may be because the class distance method has more restriction on the selecting of R. However, although the B-i-C method has more restriction on selecting R than the class distance method, the recognition rate of the B-i-C method was not higher than it. This proves that too much restriction also has negative influence on the recognition. A more detailed analysis will be made in the following part.

Fig. 3. Results of the OCR. The upper part shows the correctly recognized images of the OCR and the lower part shows the misrecognized images (which were correctly recognized by the class distance method).

C. Analysis

First, the results of the OCR and the class distance method are compared. As shown in Fig. 3, most of the correctly recognized characters of the OCR are close to the regular font. However, there are still several decorated and 3D fonts which were correctly recognized by the OCR, like the <2>, <13>, <16>, <22>, <26> and <24>. This may be because, in those characters, the main stroke is still obvious and not distorted much. In the misrecognized characters, their font shapes become obviously different from the regular font, such as the handwritten fonts of <43> and <47>, the specially-designed fonts of <38>, <41>, <58>, <60> and <61>. The OCR was failed on those characters while the class distance could recognize them. Especially, as shown in Fig. 4, the class distance was able to recognize the character of <46>, which was heavily decorated. Those decorations can be seen as the noise. At the decoration part, there were many k_i which were closer to an $r_{\tilde{C}(i)}$ of wrong class. However, after the calculation of $D_{QR}(Q||R_{\tilde{C}})$, the correct class had the minimum QR distance to Q, thus the character was correctly recognized by the class distance method. There are also other similar examples like <37>, <44>, <51> and <55>.

Second, the results of the part-based methods are compared. Among all the three part-based methods, the single voting method has the simplest framework and which has no restriction on selecting R from A. Consequently, it had the

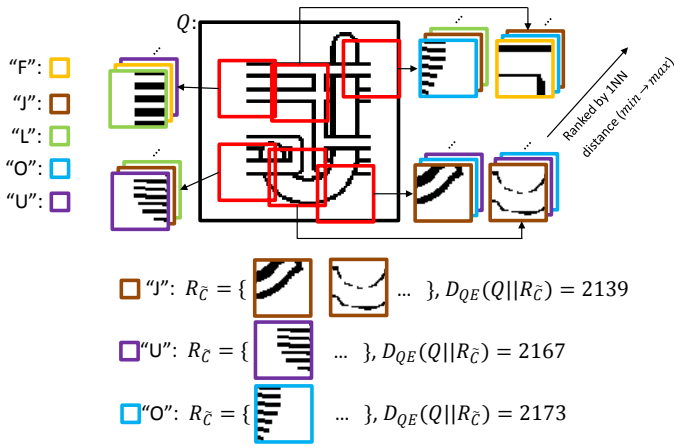

Fig. 4. The example of the recognition of the class distance method. The red boxes stand for the $\{k_i\}$. For each k_i, its corresponding $r_{\tilde{C}(i)}$ of different classes are shown by the boxes of other colors. Those $r_{\tilde{C}(i)}$ are also ranked by Euclidean distance to k_i. Note that only limited numbers of keypoints and classes are shown. The top three classes which have the minimum QR distance to Q are also shown.

lowest recognition rate among the three methods. However, although the B-i-C method has more complicated framework and more restrictions on selecting R than the class distance method, it did not outperform the class distance method in the experiments.

As shown in Fig. 5, the misrecognized characters of the B-i-C method are very ambiguous, such as <1> of "D" and <6> of "K", which are very similar to the letter "O" and "X". For those characters, in the B-i-C method, if there is no suitable h_i, the restriction of $h_R \in H$ may exert negative influence to the recognition. In contrast, in the class distance method, those characters may be correctly recognized by some discriminative keypoint.

In the misrecognized characters of the class distance method, the "parts" of the character are ambiguous. For example, in <8>, the character "I" is actually a part of the character "H". As the result, the k_i of "I" may have small distance to the r_i from the left or right side of "H". Thus its $R_{\tilde{C}}$ of "H", which only contains the r_i from the left or right side of "H", may be selected as the $R_{\tilde{A}}$. Consequently, this character of "I" was misrecognized as "H". Similar examples can also be found in <9> of "L", <11> of "P" and <13> of "C", which can be seen as a part of "E", "B", and "E". In the B-i-C methods, this ambiguity can be avoid by the restriction of $h_R \in H$. By using this restriction, for example, the $R_{\tilde{C}}$ of "H" must contain the r_i from the center of "H" to meet the condition $h_R \in H$. As the result, for the <8> of "I", its QR distance to $R_{\tilde{C}}$ of "H" will be obviously larger than the $R_{\tilde{C}}$ of "I" and thus the <8> can be correctly recognized by the B-i-C method.

From the above analysis, we can draw a conclusion that, the flexibility and the restriction of selecting R are both important for the recognition of arbitrary font. Although the B-i-C method may benefit from the restriction of $h_R \in H$, the flexibility of selecting R also decreases much. Consequently, on the arbitrary-font alphabets, the recognition rate of the B-i-C method is lower than the class distance method. We need

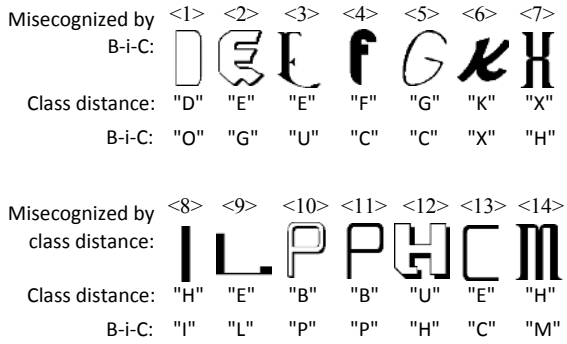

Fig. 5. Results of the class distance method and the B-i-C method. The upper characters were correctly recognized by the class distance method while misrecognized by the B-i-C method; the lower characters were correctly recognized by the B-i-C method while misrecognized by the class distance method.

to design proper restrictions of selecting R to make proper part-based method for the recognition of arbitrary font.

IV. Conclusion

In this paper, the three part-based methods and a commercial OCR system are evaluated on the data of alphabets of variety fonts. The result shows that the part-based method has the better performance than the OCR on the recognition of arbitrary fonts. More over, by the analysis of different part-based methods, it is possible to improve the performance of the part-based method on arbitrary font recognition by designing better restriction of R.

Acknowledgment

This work was supported in part by The Ministry of Education, Culture, Sports, Science and Technology in Japan under a Grant-in-Aid for Scientific Research No. 23300072.

References

[1] Keechul Jung, Kwang In Kim, and Anil K. Jain, "Text information Extraction in Images and Video: a Survey," Pattern Recognition, vol. 37, no. 5, pp. 977–997, 2003.

[2] Ismail Haritaoglu, "Scene Text Extraction and Translation for Handheld Devices," Proc. CVPR, pp. 408-413, 2001.

[3] B. Epshtein, E. Ofek and Y. Wexler, "Detecting Text in Natural Scenes with Stroke Width Transform," Proc. CVPR, 2010.

[4] K. Wang and S. Belongie, "Word Spotting in the Wild," Proc. ECCV, 2010.

[5] H. Goto, "Redefining the DCT-Based Feature for Scene Text Detection," IJDAR, vol. 11, no. 1, pp. 1–8, 2008.

[6] Andrew Kae, David A. Smith and Erik Learned-Miller, "Learning on the Fly: a Font-Free Approach Toward Multilingual OCR," IJDAR, vol. 14, no. 3, pp. 289–301, 2011.

[7] S. Uchida and M. Liwicki, "Part-Based Recognition of Handwritten Characters," Proc. ICFHR, pp. 545-550, 2010.

[8] S. Wang, S. Uchida and M. Liwicki, "Comparative Study of Part-Based Handwritten Character Recognition Methods," Proc. ICDAR, 2011.

[9] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robus Features," Proc. ECCV, 2006.