Front.Comput.Sci.
DOI

REVIEW ARTICLE

Part-Based Methods for Handwritten Digit Recognition

Song WANG (✉)[1], Seiichi UCHIDA[1], Marcus LIWICKI[2], Yaokai FENG[1]

1 Kyushu University, Fukuoka 819-0395, Japan
2 DFKI, Kaiserslautern D-67663, Germany

© Higher Education Press and Springer-Verlag Berlin Heidelberg 2012

Abstract In this paper, we intensively study the behavior of three part-based methods for handwritten digit recognition. The principle of the proposed methods is to represent a handwritten digit image as a set of parts and recognize the image by aggregating the recognition results of individual parts. Since part-based methods do not rely on the global structure of a character, they are expected to be more robust against various deformations which may damage the global structure. The proposed three methods are based on the same principle but different in their details, for example, the way of aggregating the individual results. Thus, those methods have different performances. Experimental results showed that even the simplest part-based method can achieve recognition rate as high as 98.42% while the improved one achieved 99.15%, which is comparable or even higher than some state-of-the-art method. This result is important because it reveals that characters can be recognized without their global structure. The results also show that the part-based method has robustness against deformations which usually appear in handwriting.

Keywords handwritten digit recognition, local features, part-based method

1 Introduction

Part-based methods have been proposed for visual object recognition with a simple framework but promising performance [1–3]. In those methods a query image is usually decomposed into parts. Each part is represented by a local feature descriptor. Some local recognition (or quantization) is conducted at each part and then the final recognition result is derived by aggregating all the local results.

Typical properties of general part-based methods are as follows.

- Part-based methods usually extract multiple (say, 100) local feature descriptors of image parts to represent a single image.
- Global features, especially, the position of a part in the image are often disregarded when evaluating the similarity of two images. This means that each image part is processed independently while not using its original location. Consequently, it is possible to improve the robustness against the variations in object appearance, the degradation of object, etc.
- The similarity depends on the comparison of two sets of local feature descriptors. Images with similar sets of local feature descriptors will be considered as images from the same class.
- Each class is sometimes represented by a large reference database of local feature descriptors extracted from multiple (i.e. different) images of the class in order to deal with more variations. It is also popular that those many local feature descriptors are quantized into a fewer number of representative ones.

Those properties of part-based methods seem to be beneficial for handwritten character recognition. First of all, the part-based method is flexible to adapt various appearances of handwritten characters (different fonts and writing styles, distortions, etc.). For example, a query character image may be recognized by the parts form different training images of the same class. Accordingly, even if there is no training character image whose appearance is globally similar to the query char-

Received month dd, yyyy; accepted month dd, yyyy

E-mail: wangsong@human.ait.kyushu-u.ac.jp

acter image, the query image will be recognized correctly by using its local similarity to a set of training images. Second, part-based methods are supposed to be robust on degraded character images (incomplete or deformed character images, partial occlusion, partial overlap and concatenation, broken or fragmented stroke, etc.). This is because the recognition result of a part-based method is determined by a group of image parts. A few degraded parts do not influence the final result severely.

For character recognition research, however, part-based methods have been rarely tried so far. This may be because most researchers have believed that global features are essential for representing characters. Another reason is that researchers have believed that the "part" of the character may be ambiguous and thus impossible to recognize the part with a reasonable accuracy. In fact, most of character recognition methods represent a whole character by a single global feature [4]. Hereafter, we call those conventional methods whole-based methods, in contrast to the part-based method.

Figure 1 shows the simplest way of realizing a part-based handwritten digit recognition. This method is comprised of a training step and two recognition steps, called feature-level recognition and image-level recognition. After the local feature detection, the character image is decomposed into local feature descriptors. In the feature-level recognition step, each local feature descriptor is recognized independently. In the image-level recognition step, all the feature-level recognition results are aggregated for the final recognition result. Various appearances of characters are not easy to be represented by some global structure or descriptor of a whole character. If we discard the global structure and only use the parts, we may gain benefits as mentioned above.

In this paper, detailed inspections of not only the simple part-based method of Fig. 1 but also its two improved versions are done, in order to find out how those part-based methods work in handwritten digit recognition. The main contributions of this paper are summarized as follows:

- The most important contribution is to prove the fact that handwritten characters can be recognized with a reasonably high accuracy by their parts. In other words, this paper proves that handwritten characters can be recognized without any global feature, which has been almost always used in the past trials.
- Comparison experiments were done in order to prove the advantages of the part-based method of Fig. 1 over the whole-based method.
- Two improved part-based methods versions have better performance. That is, these two methods indicate further potential of part-based recognition.

Table 1 Recognition rates (%) on MNIST database.

Methods	Recognition rate
Belongie et al. [17]	99.37
LeCun et al. [18]	98.90
Teow et al. [19]	99.41
Mayraz et al. [20]	98.30
Part-based method of single voting	98.42
Part-based method of class distance	99.15

2 Related Work

In image classification, because of the complexity of the image, many part-based classification methods and features were presented. In [5–7], several part-based features are designed for object recognition. In [8] and [9], the part-based features are used for view-invariant and rotation-invariant recognition. In [10–12], several features are presented for multi-resolution image recognition. In [13], a hierarchical part-based recognition method is presented. In [14] the part-based method is applied to image sequences segmentation. When using the parts instead of the whole image, the classification is more flexible and practical.

In character recognition research, various whole-based methods, which describe the whole character image as a single feature descriptor, have been proposed. The features of the whole-based method often use the contour or shape information of the characters to extract the feature vectors [15,16]. Like [17–20] many whole-based methods have been presented for handwritten digit recognition and in [21] different whole-based methods of handwritten digits were reviewed and compared experimentally. The best performance of those state-of-the-art methods on MNIST database are shown in Table 1 as while as two part-based methods (which will be introduced later). Clearly, the best recognition rate of part-based method is comparable with the state-of-the-art methods. Please note that to achieve the best performance of the part-based method, the whole training set of MNIST was used as while as the optimal size parameter.

Comparing to the whole-based method, a far smaller number of trials have been made for the part-based character recognition method. Suen et al. [22–25] have tried several part-based methods for character recognition; however, their trial still uses global features, that is, the global position of parts. There are some other trials which focused on the ro-

bustness against the global deformation of the characters. As mentioned in Section 1, the part-based method is expected to have such robustness. In [26–28], the part-based methods have been used to recognize the HIPs (human interaction proofs) which are often generated as the degraded digits or letters. In [29, 30], the part-based method has been conducted for character recognition in natural images. Characters in natural images are also known to be severely deformed and distorted. Some exceptional trials have been done quite recently in [31–34], where a part-based method is applied to an ancient manuscript recognition task. However, in above trials, the part-based method was employed to recognize uncommon characters (HIPs, characters in scene image, etc.), a more general handwritten character recognition task has not been analyzed so far.

Our trial in this paper is different from above trials in the following three points. First, in this paper, we first use the part-based method for general handwritten characters (digits) recognition. The recognition accuracy for those general character patterns from widely used database MNIST was observed in order to show the general performance of the part-based method in handwritten digit recognition task. Second, we have made deeper and more detailed analysis of the part-based method in order to find out that how it works in the recognition. Third, we have studied different part-based methods in order to find out the potential the part-based method.

3 The Simplest Part-Based Method—Single Voting Method

In this section, we will introduce the simplest part-based method called the *single voting method* of Fig. 1. (This method is named as "single voting" for a good contrast to its versions, which are introduced later.) As introduced in Section 1, the single voting method is comprised of a training step and two recognition steps, called feature-level recognition and image-level recognition. The single voting method is useful to observe the basic performance of the part-based method.

3.1 Local feature detection and description by SURF

As shown in Fig. 1, the character image is decomposed into local feature descriptors through a local feature detection and description tool. Speeded-up robust features [35] (SURF) is utilized throughout in this paper as such tool, whereas SIFT

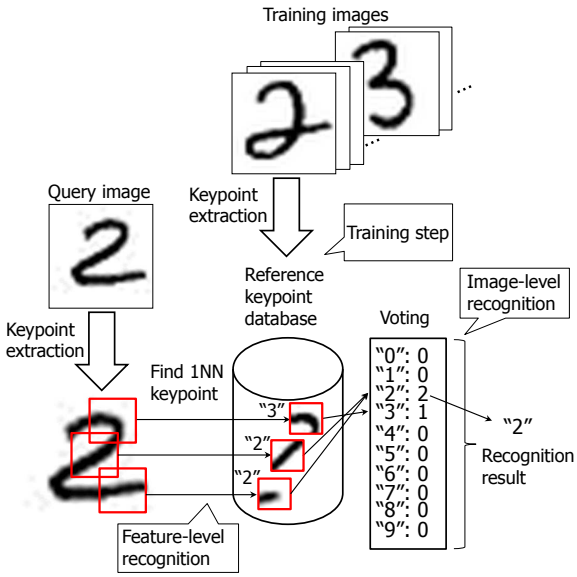

Fig. 1 The single voting method.

[36] and other recent local feature descriptors are also possible choice. SURF works as follows: SURF detects keypoints as the local maxima of Hessian values in a scale space and then describes the local area around each keypoint by a 128-dimensional vector as a feature descriptor. Hereafter, the feature descriptor is simply called the keypoint. The Euclidean distance of two keypoints in the feature space can be simply used to measure the dissimilarity between local parts.

It is important to note that in the following experiments, the SURF keypoint is modified to be scale-fixed and non-rotational although it is originally rotation and scale invariant. This modification has two meanings. First, it simplifies analysis of recognition results. Especially, it is possible to control the size of parts manually under this modification and thus to observe the effect of the size on the recognition performance. Second, the modification leads better recognition performance. In fact, if we use any rotation and scale invariant feature vector, the dissimilarity are often underestimated thus more ambiguous.

Since the SURF detector is based on the local maximum of approximate Hessian, keypoints are often located around corners and curves of character stroke. The number of keypoints from one training image varies and depends on the shape of the character. In the following experiment, about 59 keypoints were detected from a single digit image in average.

3.2 Training step

The training step of the single voting method is very simple; that is, SURF keypoints are extracted from the individual training images. All the extracted keypoints are called reference keypoints and stored into a database. Those keypoints extracted from the class C are labeled as C.

3.3 Recognition step

3.3.1 Feature-level recognition

In the feature-level recognition, query keypoints are firstly extracted from the query character image and then classified into one of the assumed classes (e.g., "0",. . ., "9" for handwritten digits). This classification is simply realized by finding the nearest-neighbor (1NN) of the query keypoint from the reference keypoint database (as shown in Fig. 1). Note that each query keypoint is classified independently, which means the query keypoints will not exert influence to each other in the feature-level recognition.

3.3.2 Image-level recognition

All the feature-level recognition results are aggregated in this step for determining the final recognition result of the query image. In the single voting method, this image-level recognition is conducted as a simple majority voting process. The votes are the classification results of the query keypoints. Since each query keypoint contributes one single vote in the majority voting process, it is called the single voting method. After the voting, the class with the maximum votes becomes the final recognition result.

4 Performance Evaluation of the Simplest Part-Based Method

Several experiments were done in order to evaluate the performance of the single voting method. As noted in Section 1, the part-based method is expected to have two advantages over the whole-based methods. The first advantage is higher recognition rate by combining parts from different images and the second the robustness against degradation of character images.

For an experimental comparison, a whole-based method was used as a typical character recognition framework. In the whole-based method, a single and large SURF keypoint is extracted from the whole image while considering the whole image as a large local part. Recognition was done by 1NN between these non-local SURF descriptors of the query and the training images. It is noteworthy that this SURF usage mimics the typical handwritten character recognition framework because a SURF feature vector is very similar to the directional feature [21], which has often been utilized in handwritten character recognition.

4.1 Experiments on handwritten digits

Handwritten digit images of MNIST were used as follows: for each class, the first 1,000 images of the "training" dataset of MNIST were used for creating the reference database; the "test" dataset (10,000 images in total) were used for recognition test. The reason to use only a part of the MNIST training dataset is because the part-based method of multiple voting, which will be introduced later, needs an extra training set of a large size. In order to make a comparison of the three part-based methods, here we saved the rest of the training set for the class distance method. As a result, the recognition rate was slightly different with the experiment of the whole training set. Each image (a 28×28 grayscale image) was preprocessed so that an enough number of keypoints were extracted by SURF. Specifically, it was magnified four times by the bicubic interpolation scaling after the addition of 10-pixel surrounding margin. Consequently, the image became 192×192 pixels.

For the single voting method, the database was comprised of about 59,000 reference keypoints. (That is, each of 1,000 training images provided 59 keypoints on average.) The average number of keypoints per training image was large on "0" and small on "1". Without any optimization algorithm, the time complexity for each 1NN search is $O(N)$ (N is the number of reference keypoints). In our experiment, the time consumption of each query image was about one minute. However, according to [37], if some optimization algorithm for nearest neighbor search was employed, this time consumption could be reduced to less than one second.

Figures 2 (a) and (b) show the keypoints of different fixed scales and the corresponding image-level recognition rates, respectively. A scale parameter s is used for describing the keypoint size. It can be seen that by increasing s, the size of local parts increases and thus each part contains more discriminative information. However, at the same time, the flexibility of the part-based method decreases. As the result, after reaching its highest recognition rate 97.8% at $s = 7$, the recognition rate of the single voting method begins to decrease when s increases. The whole-based method only had a recognition rate of 92.8% which was much lower than the

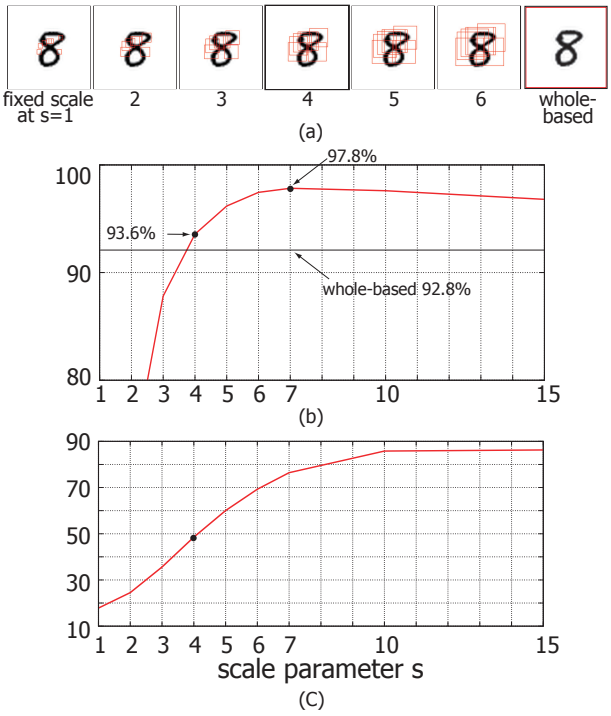

Fig. 2 (a) Size of local parts in of different scales. (b) Image-level recognition rate (%) and (c) feature-level recognition rate (%). In (a), only 7 keypoints (selected randomly) are shown.

highest recognition rate of the single voting method. This proves the first advantage of the part-based methods mentioned above.

Figure 2 (c) shows the feature-level recognition rates by different scales. Note that all of the feature-level recognition rates were much lower than the corresponding image-level recognition rate. Especially at $s = 4$, the feature-level recognition rate was only about 50%, while the corresponding image-level recognition rate was 93.8%. The confusion matrix of feature-level recognition ($s = 4$) are displayed in Table 2 (a). In each class, although only half of the results were correct, fortunately, the incorrect results are evenly distributed in the rest of the classes. Consequently, the correct class had a great chance to win in voting.

In Table 2 (a), we can see that the major misrecognition pairs (printed in boldface) were "1"↔"7" and "3"↔"5." This is simply because, for example, the lower parts of "3" and "5" are sometimes similar. It is interesting that there is no zero entry in the confusion matrix.

Table 2 (b) shows the confusion matrix of image-level recognition. The best recognition rate was 99.0% for "0" and the worst was 85.4% for "7", respectively. The pair "1"↔"7", which was a major misrecognition pair at feature-level, was also a major misrecognition pair at image-level. However, the other notorious pair "3"↔"5" was not a major misrecognition pair anymore; it is because their dissimilar upper parts allow a better discrimination. It is also noteworthy that classes with circular strokes (e.g., "2", "6", "9") were misrecognized as "0".

The difficulty of the feature-level recognition is illustrated in Fig. 3 (a), which shows the distribution of 1NN distances of each query keypoint to the correct class and that to the nearest incorrect class. There is a concentration along the diagonal line and thus the minimum distance to the incorrect class is often close to that to the correct class.

As shown in Fig. 3 (b), all test images are projected on a two-dimensional plane according to their v_{cor} and v_{incor} values. The former v_{cor} is the percentage of votes to the correct class at the image-level recognition. The latter v_{incor} is the percentage of votes to the top incorrect class. If $v_{\mathrm{cor}} > v_{\mathrm{incor}}$, a test image is correctly recognized at image-level recognition by majority voting. Since the feature-level recognition rate was around 50%, the peak of the distribution of v_{cor} is also around 50%. A more important thing is that the peak of v_{incor} is far lower and around 17%. This indicates that the misrecognitions at the feature-level recognition were not converged into a certain incorrect class but scattered into various incorrect classes. The above analysis is supported by this fact. Figure 3 (b) also indicates that most misrecognized images were located near the diagonal line. If a little more discriminative information is used in the voting process, those images may be correctly recognized.

As shown in Fig. 3 (c), the number of training images at each class affects the recognition rates drastically. In the extreme case, that is, if we use only a single training image for each class, feature-level recognition rate was degraded to 30%. This also proves that the keypoints are distributed with considerable overlaps and thus we need many reference keypoints for increasing the probability of finding a 1NN of the correct class. Note that the recognition rate was not saturated with 1,000 training images and thus will be improved if more training images are used.

According to [21], some state-of-the-art handwritten digit recognition methods can reach a recognition rate as high as 99.58% on MNIST database, which is much higher than the highest recognition rate of the single voting method. However, our recognition rate is enough to prove that handwritten digits can be recognized just by parts, that is, without any global structure. This fact leads further merits, for example, robustness to heavy deformations in handwritten characters. Moreover, the part-based method of character recognition is a very simple and flexible framework and thus has a potential

Table 2 Confusion matrix (%). ($s = 4$.)

(a) Feature-level recognition rate.

recognition result \ input	0	1	2	3	4	5	6	7	8	9
0	53.5	6.6	10.2	8.0	5.5	8.9	**12.3**	6.1	6.3	6.4
1	3.6	51.1	1.2	0.3	8.9	0.7	5.2	**12.9**	0.6	4.8
2	7.6	1.5	45.0	8.2	4.6	5.9	4.6	9.0	6.4	5.3
3	6.0	0.5	9.0	48.7	1.1	**16.5**	3.8	6.0	8.9	3.8
4	2.8	**9.8**	4.1	1.3	48.5	2.3	5.4	7.0	2.7	10.0
5	5.7	0.7	5.7	**15.1**	1.9	47.8	4.2	3.5	6.2	2.7
6	8.3	6.6	5.5	2.6	7.0	4.5	51.6	2.4	6.6	3.9
7	4.5	**15.9**	9.3	5.7	7.4	3.5	2.1	41.8	2.7	8.7
8	3.6	0.9	5.5	7.1	3.0	6.7	7.0	2.2	50.7	8.7
9	4.4	6.5	4.7	3.0	**12.1**	3.3	3.7	9.0	8.8	45.7

(b) Image-level recognition rate.

recognition result \ input	0	1	2	3	4	5	6	7	8	9
0	99.0	0.9	2.6	0.8	1.0	1.2	4.3	3.0	1.0	**3.7**
1	0.0	94.2	0.2	0.0	0.3	0.1	0.4	**5.8**	0.0	0.3
2	0.3	0.2	93.5	0.7	0.2	0.2	0.1	2.8	0.3	0.8
3	0.3	0.1	1.2	96.5	0.0	2.1	0.1	0.4	2.4	0.2
4	0.0	0.5	0.2	0.0	93.5	0.1	0.2	1.6	0.2	1.5
5	0.0	0.1	0.3	1.7	0.0	95.5	0.7	0.0	0.2	0.3
6	0.2	0.6	0.4	0.0	0.5	0.6	93.5	0.1	0.4	0.2
7	0.1	**3.4**	1.6	0.1	0.0	0.1	0.0	85.4	0.2	1.1
8	0.0	0.0	0.0	0.1	0.1	0.0	0.6	0.1	94.6	1.4
9	0.1	0.0	0.0	0.1	**4.4**	0.0	0.0	0.8	0.3	90.5

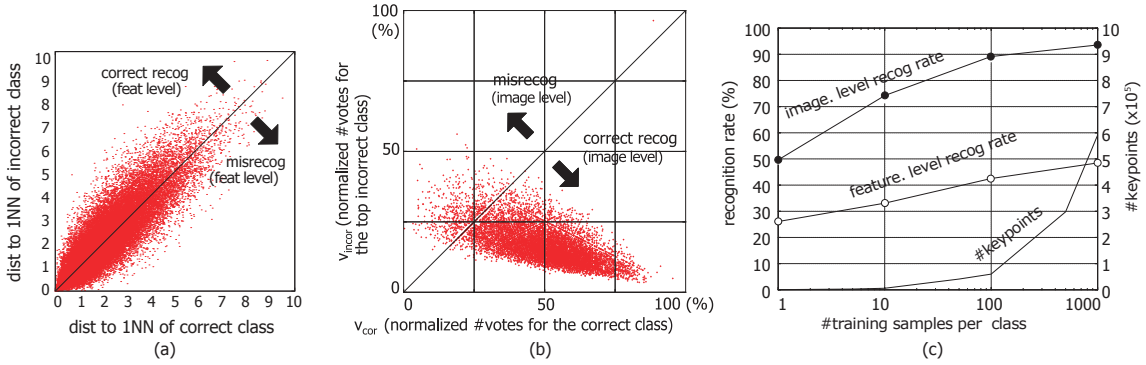

Fig. 3 Illustrations of keypoints and votes in the single voting method ($s = 4$). (a) shows the distribution of the keypoints by feature-level recognition result; (b) shows the distribution of the votes; (c) shows the influence of the size of the training set.

to be improved, which is shown in Section 5. Complementary combination with other state-of-the-art methods is a promising future research direction.

4.2 Robustness to Severe Deformation

An experiment was conducted on cropped handwritten digit images to show how the single voting method has more robustness to severe deformations than the whole-based method. The same training image set with the former experiments was used and each image was resized to 192×192 with additional margins. The same test image set was also used. However, one fourth of each original test image (7 pixels wide) was erased from the bottom, and then it was resized to normal as 28×28. Finally, those cropped images were resized to 192×192 with additional margins. Figure 4 (a) shows several normal images and cropped images. As mentioned above, the part-based method at $s = 4$ was tested.

Table 3 shows the recognition rates. The total recognition rate of the single voting method decreased from 93.6% to 81.3% while the whole-based method far more severely decreased from 92.8% to 44.9%. This proves the second advantage of the part-based method, that is, the part-based method is more robust against the deformation than the whole-based method. It is noteworthy that this result encourages us to use the part-based method for character string recognition where component characters are severely deformed by overlapping and/or inaccurate segmentation.

Table 3 Recognition rates (%) on cropped digits.

	class										
	total	0	1	2	3	4	5	6	7	8	9
single voting	81.3	97.6	98.6	74.9	62.9	90.4	75.8	86.8	78.1	80.1	66.3
whole-based	44.9	73.6	90.6	18.6	33.8	34.3	48.5	53.2	45.8	13.2	32.4

The recognition rate of each class is also displayed in Table 3. It can be seen that different classes have different decreases in recognition rate both in the single voting method and the whole-based method. As shown in Fig. 4 (a), only a part of the keypoints were influenced by the missed part of the image, thus in the single voting method the cropped image can still be correctly recognized. In (b) it shows all the correctly recognized query keypoints in each class in the single voting method. Through the recognized keypoints of

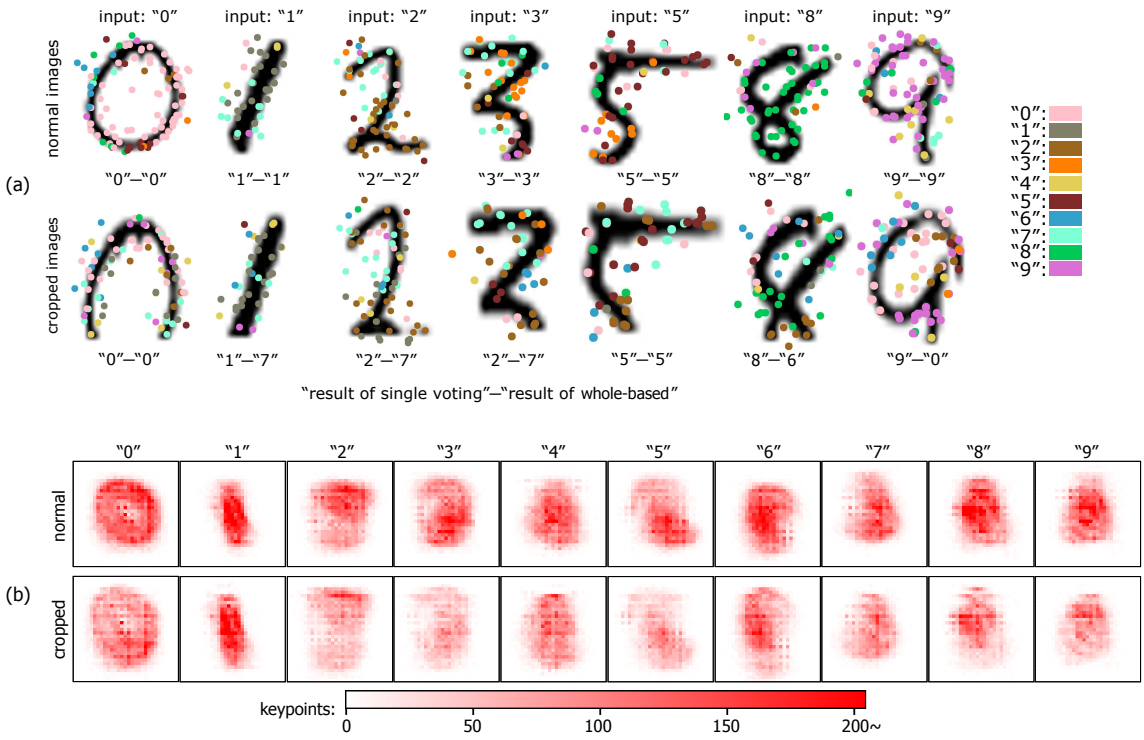

Fig. 4 Illustrations of cropped digit recognition. In (a), each circle in the image represents the center of a query keypoint in the single voting method, the color of the circle stands for the feature-level recognition result. The image-level recognition result is displayed below the image. In (b), the distribution of all correctly recognized query keypoints is illustrated. The color stands for the number of the keypoints of the same area.

normal images, we may find out which part of the character is important for the recognition in the single voting method.

5 Other Part-Based Methods

In this section, two part-based methods, called the multiple voting method and the class distance method, are newly introduced and compared to the single voting method. The main difference among three methods is their voting process. From the comparative study, we can find out the potential of the part-based methods for improving recognition rate and robustness.

5.1 The multiple voting method

As shown in Fig. 5 (a), in the multiple voting method each reference keypoint from the reference database stands for multiple classes by a distribution of class probabilities. For example, assuming a reference keypoint with 50% probability for class 1, 30% for class 2, 20% for class 3, and 0% for the other classes, the class distribution of this keypoint can be written as $(0.5, 0.3, 0.2, 0, 0, \cdots)$. Consequently, in the multiple voting method, the "votes" in voting process are actually the class distributions. It is natural to assume that a reference keypoint can stand for multiple classes. According to Table 2 (a), the reference keypoints were always referred as 1NN by query keypoints from other classes. This indicates that when a reference keypoint is referred by a query keypoint, the query keypoint can be from any class with a certain probability.

In the multiple voting method, since we have to obtain the class distribution of each reference keypoint, thus two training sets, 1 and 2, are used for this purpose. The training set 1 is used for extracting reference keypoints and the training set 2 is used for obtaining the class distribution of the reference keypoints. After keypoints are extracted from both of the training sets 1 and 2, 1NN of each keypoint from the set 2 is selected from the reference keypoints of the set 1. Then, if a reference keypoint is selected 3 times by class 1, 3 times by class 2, 4 times by class 3 and never by the other 7 classes, the class distribution of this reference keypoint can be obtained as $(0.3, 0.3, 0.4, 0, \cdots, 0)$. However, the total number of reference keypoints is different in each class. Therefore the distribution above needs to be normalized according to the total numbers of the reference keypoints of different classes.

As shown in Fig. 5 (a), the feature-level recognition and the image-level recognition of the multiple voting method are

the same with the single voting method, except that the multiple votes will be done according to the class distribution. The class with the maximum votes will be the final recognition result.

5.2 The Class Distance Method

As shown in Fig. 5 (b), the class distance method evaluate the probability of class C by assuming that all the keypoints of the query image belong to the same class C. According to [37], given a query image Q, let $k_1, \ldots, k_n$ denote all the keypoints of Q. If we have L reference keypoints from class C as $k_1^C, \ldots, k_L^C$, the class C of Q is determined by the following equation, where $k_{1NN}^C \in \{k_1^C, \ldots, k_L^C\}$ is the 1NN reference keypoint of k_i:

$$\hat{C} = \underset{C}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (k_i - k_{1\mathrm{NN}}^C)^2. \quad (1)$$

Although the theoretical detail is omitted here, the above equation is derived from a version of Kullback-Leibler divergence between keypoint feature distributions of the query image and the reference images of class C [37].

The three steps of the class distance method are shown in Fig. 5 (b). First, in the training step, a reference keypoint database is created for each class. Second, in feature-level recognition, for each query keypoint, a 1NN reference keypoint is searched for among the reference database of the class C. Third, in image-level recognition, the Euclidean distances between query keypoints and their 1NN keypoints (i.e., $k_{1\mathrm{NN}}^C$) are summarized as the distance between Q and class C, according to Eq.(1). The class with the minimum class distance will be seen as the final recognition result.

The class distance method superficially does not employ any voting process; however, Eq.(1) can be interpreted as a weighted voting scheme. In fact, we can see the 1NN distances to all classes from a query keypoint a class distribution around the query keypoint. Then the distributions of the query keypoints are summarized and the class with the minimum value becomes the final recognition result, which is just like the multiple voting method. However, there are two differences that (i) the multiple voting method uses the reference keypoint for obtaining the class distribution, whereas the class distance method uses the query keypoint, and (ii) the class distance method uses the 1NN distances to approximate the class distribution, whereas the multiple voting method uses the numbers of keypoints.

Table 4 Recognition rates (%) of three part-based methods.

size of training set	single voting	multiple voting	class distance
1000 images/class	93.6	95.4	97.9
50 images/class	86.1	93.6	92.8

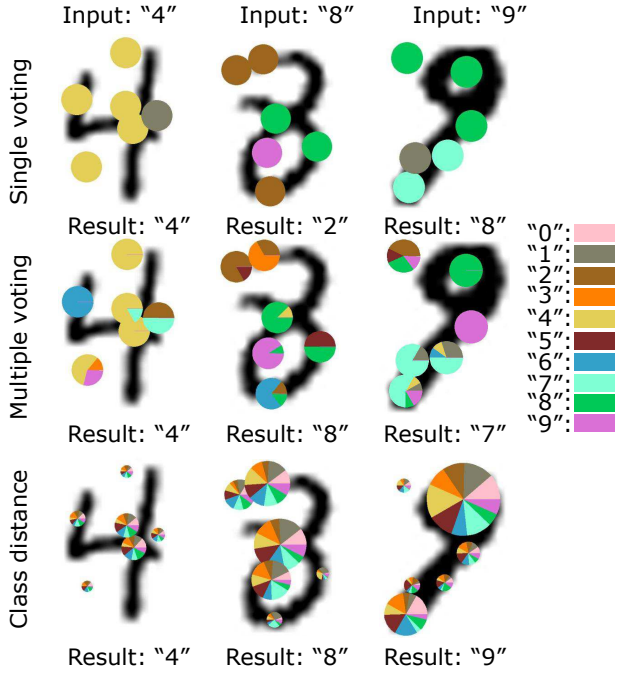

Fig. 6 Examples of the votes.

5.3 Comparative Experiments

The comparative experiments of the three part-based methods used the first 1,000 images of each class of the "training" dataset of MNIST for the training step. For the multiple voting method, the next 4,000 images of each class were used as training set 2. The whole MNIST "test" dataset was used for recognition test.

The recognition rate is displayed in the first row of Table 4. We can find out that the class distance method achieved the best accuracy among the three methods. This recognition rate is far higher than the recognition rate of the single voting method, which indicates that a well-designed framework can improve the recognition rate of the part-based method much.

Three examples from the experiments are shown in Fig. 6. We can observe how the different voting method affects the recognition result. In the figure each circle stands for one single vote in the voting process. The circle in the multiple voting method and the class distance method is a vote of class distribution. Note the opposite meaning of these circles; a larger portion of a class in a circle means a more class prob-

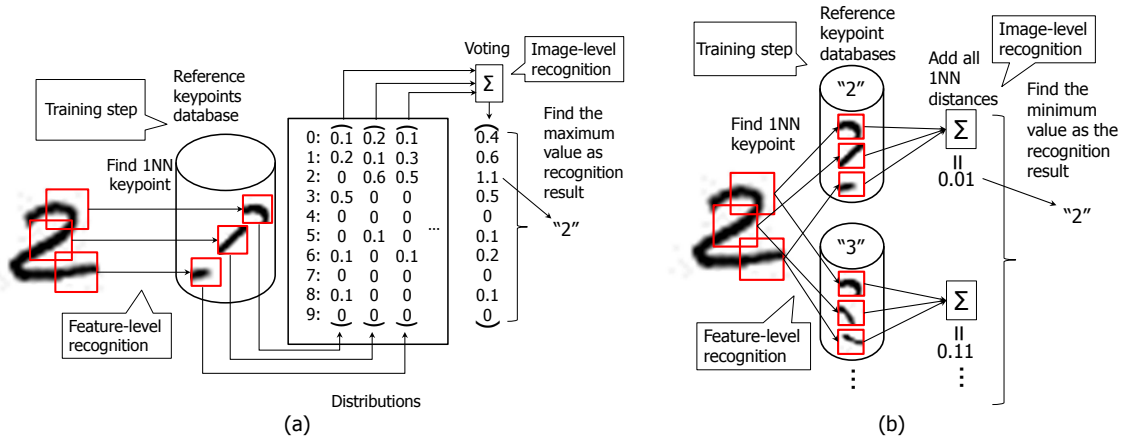

Fig. 5 Two versions of the part-based methods: (a) the multiple voting method and (b) the class distance method.

ability for the multiple voting method and a less class probability (i.e., a larger distance) for the class distance method.

For the input "4" all of the three methods were successful in recognition. For the input "8" the single voting method misrecognized the input image as "2". However, the votes change into the distributions in the multiple voting method. Since class "8" occupied the largest portion in total, the recognition result was "8" in the multiple voting method. For the input "9" both the single voting and multiple voting methods were failed while the class distance method correctly recognized the image as "9" by very small portion in the larger votes.

Another experiment was also done with an extremely small training set. The reference keypoints was extracted from 50 images per class. The average number of reference keypoints per class was only 2,968. The multiple voting method used the same training set 2 with the above experiment. The same test set was used with above experiments, which was the MNIST test set.

The second row of Table 4 shows the recognition rates of this experiment. We can find out that the multiple voting had a much better recognition rate than the single voting method. We can also find that the class distance method had a lower recognition rate than the multiple voting method. This is because the multiple voting method had a large training set 2, therefore the recognition rate of the multiple voting method didn't decrease as fast as the class distance method. It is important to note that this indicates that it is possible to reduce computational complexity by using the multiple voting method.

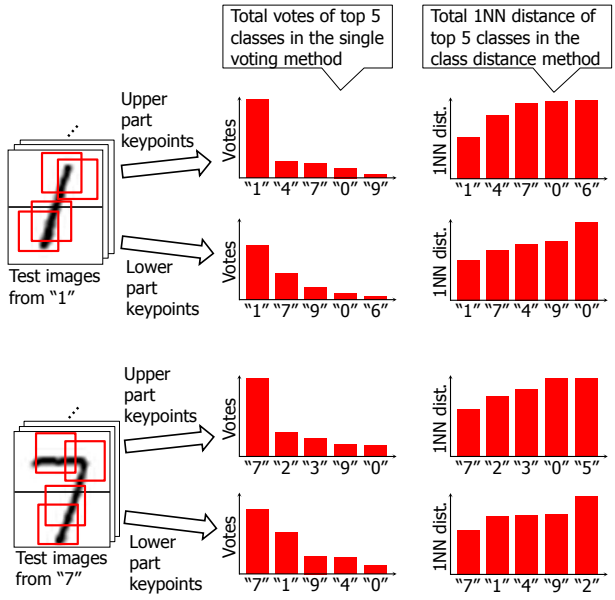

Fig. 7 Feature-level recognition on similar classes (with 1,000 images/class for training). The query keypoints extracted from the test images were separated into two parts by their location: the upper part keypoints and the lower part keypoints. The histograms show the results of those keypoints of the single voting method and the class distance method.

5.4 Discrimination of the Part-Based Method

In the part-based method, the global structure information is not considered, all the recognition is done based on the local parts. However, compared with the conventional methods, the lack of the global structure information does not degrade the discrimination of the part-based method. In Fig. 7, the feature-level recognition results of two similar classes ("1" and "7") are shown and for the analysis, the query keypoints were separated into the upper part keypoints and the lower part keypoints. As shown in Fig. 7, the query keypoints ex-

Table 5 Image-level recognition rate of the class distance method (%).

recognition result \ input	0	1	2	3	4	5	6	7	8	9
0	98.4	0.2	0.3	0.0	0.2	0.3	1.1	0.7	0.1	0.6
1	0.0	97.7	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1
2	0.4	0.2	99.0	0.8	0.0	0.4	0.1	1.9	0.1	1.5
3	0.1	0.1	0.3	98.4	0.0	0.4	0.0	0.0	0.1	0.2
4	0.0	0.4	0.1	0.0	99.1	0.0	0.1	0.6	0.0	0.5
5	0.3	0.1	0.2	0.7	0.0	97.9	0.4	0.0	0.2	0.6
6	0.2	0.2	0.0	0.0	0.1	0.3	97.7	0.0	0.2	0.0
7	0.1	1.0	0.1	0.1	0.1	0.2	0.0	96.0	0.0	0.2
8	0.5	0.0	0.0	0.0	0.2	0.2	0.5	0.1	99.0	0.5
9	0.0	0.1	0.0	0.0	0.3	0.0	0.0	0.6	0.3	95.8

tracted from the lower part of “1” and “7” are very similar. Consequently, from the histograms of the single voting method we can find that many lower part votes from test images of “1” went to the class “7” while in test images of “7” many of those lower part votes also went to the class “1”.

Globally, the upper part of “1” is also similar to the lower part of “7”, but in the histogram we can find out that only a few of the upper part votes from test images of “1” went to the class “7”, indeed most of the votes successfully went to the class “1”. This is because although the upper part of “1” and the lower part of “7” are very similar, the query keypoints from which are very different. As shown in Fig. 7, the upper part keypoint of “1” usually has a blank area at the top while in the lower part keypoint of “7” the top is always occupied by the vertical stroke. For one single keypoint, this is an obvious difference, and which leads to a vote distribution in upper part of “1” as shown in the histogram. Consequently, the ambiguity of the lower part can be overcome when we consider all the query keypoints of the test images from “1”. In conclusion, without the global structure information, the part-based method can still differentiate the similar shaped characters.

As mentioned in section 4, in the single voting method, there were several major misrecognition pairs. When we use the class distance method, as shown in Table 5, we can hardly find a major misrecognition pair. We may find a reason from the histograms of the class distance method in Fig. 7 (please note that in the class distance method, the smaller 1NN distance means more votes). As mentioned above, the vote in the class distance method contains more information than the single voting method, thus the histograms of the class distance is more closer to the true distribution. Since we can hardly find a main incorrect class (which is obviously smaller than other incorrect classes) from the histogram, there is no obvious misrecognition pair in Table 5.

From the above analysis, we can draw a conclusion that the discrimination of the part-based method is reliable. In feature-level recognition, one single query keypoint may be ambiguous; however, when all the query keypoints of the image are used in the image-level recognition, we can get a reliable result. In the future work, we can add more weight to those keypoints which are more distinguishable (like the upper part of “1”) to improve the recognition rate.

Table 6 Recognition rates (%) on CEDAR (uppercase letter).

	class distance	whole-based
recognition rate	78.9	70.1

5.5 Extension of the Part-Based Method on Handwritten Alphabets Recognition

As noted above, part-based methods have a simple framework and in which no special information of the characters is employed. For example, in the experiments mentioned above, the digit sample was only treated as an image. Consequently, it is very easy to extend the part-based method to different recognition tasks (alphabets, Chinese characters, etc.). In order to show the potential of part-based methods on different recognition tasks, an experiment based on the CEDAR dataset was conducted. In this experiment, the isolated handwritten letters (uppercase) of CEDAR were used. For each class, 100 images from the “training” dataset were used for training (except class “J”, “Q” and “Z”, whcih only contains 68, 3 and 24 images in total). All the images in test dataset were used for recognition test. The recognition rate of the class distance method and the whole-based method are shown in Table 6. Clearly, the class distance method had a much better performance than the whole-based method. This result proved that the part-based method still have the advantages when it applied to alphabets recognition. Note that here we only used 100 images per class for training. The recognition rate of the class distance method will be much better when more training images are used. In summary, the part-based method proposed in this paper can be applied to not only the handwritten digit recognition but also the character recognition of more classes.

6 Conclusion and Future Work

In this paper, the part-based method of digit recognition is discussed generally and deeply investigated. Without the usage of the global structure of characters, the part-based method can achieve promising recognition rates and have robustness against image deformation. The comparative study

of different part-based methods also shows the potential of the part-based methods. With different designed frameworks, the part-based method can have different advantages. In the class distance method, the recognition rate is improved much; in the multiple voting method, its robustness against the reduction of reference database size can be used to enhance the computation speed.

Future work will focus on the potential development of the part-based methods. A better framework of part-based method will be designed in order to obtain more accurate class distributions. With more accurate class distributions, a higher recognition rate can be expected. In addition, the part-based methods also have a potential to develop character string recognition methods without explicit segmentation into individual characters. This is similar to part-based object recognition, where a car is also recognized without explicit segmentation into tires, windows, body, etc. Furthermore, part-based methods with different local features will be studied.

References

1. Bart E. Ullman S. Class-based matching of object parts. In: Conference on Computer Vision and Pattern Recognition Workshop, 2004, 173–173
2. Zhang J. MarszaÅĆek M. Lazebnik S. et al. Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision, 2007, 73(2): 213–238
3. Mikolajczyk K. Schmid C. A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(10): 1615–1630
4. Plamondon R. Srihari S. Online and off-line handwriting recognition: a comprehensive survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(1): 63 –84
5. Carneiro G. The automatic design of feature spaces for local image descriptors using an ensemble of non-linear feature extractors. In: IEEE Conference on Computer Vision and Pattern Recognition, 2010, 3509 –3516
6. Lim K.L. Galoogahi H. Shape classification using local and global features. In: Fourth Pacific-Rim Symposium on Image and Video Technology, 2010, 115 –120
7. Keren D. Painter identification using local features and naive bayes. In: International Conference on Pattern Recognition, 2002, 474 – 477
8. Bart E. Byvatov E. Ullman S. View-invariant recognition using corresponding object fragments. In: Computer Vision, 2004,, 152–165
9. Song C. Yang F. Li P. Rotation invariant texture measured by local binary pattern for remote sensing image classification. In: International Workshop on Education Technology and Computer Science, 2010, 3–6
10. Liang P. Li S.F. Qin J.W. Multi-resolution local binary patterns for image classification. In: International Conference on Wavelet Analysis and Pattern Recognition, 2010, 164 –169
11. Suruliandi A. Srinivasan E. Ramar K. Image resolution dependency of local texture patterns in classification of color images. In: IEEE Annual India Conference, 2010, 1–6
12. Lazebnik S. Schmid C. Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Society Conference on Computer Vision and Pattern Recognition, 2006, 2169 – 2178
13. Ullman S. Epshtein B. Visual classification by a hierarchy of extended fragments. In: Toward Category-Level Object Recognition, 2006, 321–344
14. Ohm J.R. Ma P. Feature-based cluster segmentation of image sequences. In: International Conference on Image Processing, 1997, 178 –181
15. Wakabayashi T. Tsuruoka S. Kimura F. et al. On the size and variable transformation of feature vector for handwritten character recognition. IEICE Transactions Japan, 1993, J76-D-II(12): 2495–2503
16. Srikantan G. Lam S. Srihari S. Gradient-based contour encoding for character recognition. Pattern Recognition, 1996, 29(7): 1147–1160
17. Belongie S. Malik J. Puzicha J. Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(4): 509–522
18. Lecun Y. Bottou L. Bengio Y. et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278 –2324
19. Teow L.N. Loe K.F. Robust vision-based features and classification schemes for off-line handwritten digit recognition. Pattern Recognition, 2002, 35(11): 2355–2364
20. Mayraz G. Hinton G. Recognizing handwritten digits using hierarchical products of experts. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2002, 24(2): 189 –197
21. Liu C.L. Nakashima K. Sako H. et al. Handwritten digit recognition: benchmarking of state-of-the-art techniques. Pattern Recognition, 2003, 36(10): 2271 – 2285
22. Li Z.C. Suen C.Y. Crucial combinations of parts for handwritten alphanumeric characters. Mathematical and Computer Modelling, 2000, 31(8-9): 193 – 229
23. Li Z.C. Li H.J. Suen C.Y. et al. Recognition of handwritten characters by parts with multiple orientations. Mathematical and Computer Modelling, 2002, 35(3-4): 441 – 479
24. Suen C. Guo J. Li Z. Analysis and recognition of alphanumeric handprints by parts. IEEE Transactions on Systems, Man and Cybernetics, 1994, 24(4): 614–631
25. Li Z.C. Suen C.Y. The partition-combination method for recognition of handwritten characters. Pattern Recognition Letters, 2000, 21(8): 701 – 720
26. Chellapilla K. Simard P. Using machine learning to break visual human interaction proofs. In: Neural Information Processing Systems, 2004
27. Chellapilla K. Larson K. Simard P. et al. Computers beat humans at single character recognition in reading based human interaction proofs.

In: Conference on Email and Anti-Spam, 2005
28. Mori G. Malik J. Recognizing objects in adversarial clutter: Breaking a visual captcha. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, volume 1, Los Alamitos, CA, USA: IEEE Computer Society, 2003, 134–141
29. de Campos T.E. Babu B.R. Varma M. Character recognition in natural images. In: International Conference on Computer Vision Theory and Applications, 2009
30. Coates A. Carpenter B. Case C. et al. Text detection and character recognition in scene images with unsupervised feature learning. In: International Conference on Document Analysis and Recognition, 2011, 440–445
31. Diem M. Sablatnig R. Recognition of degraded handwritten characters using local features. In: International Conference on Document Analysis and Recognition, 2009, 221–225
32. Diem M. Sablatnig R. Are characters objects? In: International Conference on Frontiers in Handwriting Recognition, 2010, 565–570
33. Garz A. Diem M. Sablatnig R. Detecting text areas and decorative elements in ancient manuscripts. In: International Conference on Frontiers in Handwriting Recognition, Los Alamitos, CA, USA: IEEE Computer Society, 2010, 176–181
34. Sankar K. P. Jawahar C.V. Manmatha R. Nearest neighbor based collection ocr. In: International Workshop on Document Analysis Systems, 2010, 207–214
35. Bay H. Tuytelaars T. Van Gool L. SURF: Speeded up robust features. In: Computer Vision, 2006, 404–417
36. Lowe D.G. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004, 60(2): 91–110
37. Boiman O. Shechtman E. Irani M. In defense of nearest-neighbor based image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, 2008, 1–8

Song Wang received his B.Sc. degree in Physics from Hebei University, Baoding, China and M.E. degree in Computer Science from Huazhong University of Science and Technology, Wuhan, China. Since 2010, he has been a Ph.D student in the Department of Intelligent Systems of the Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan. His research interests include pattern recognition, off-line and on-line handwritten character recognition, image classification, scene character detection and document analysis.

Seiichi Uchida received B.E., M.E., and Dr. Eng. degrees from Kyushu University in 1990, 1992 and 1999, respectively. From 1992 to 1996, he joined SECOM Co., Ltd., Japan. Currently, he is a professor at Kyushu University. His research interests include pattern recognition and image processing. He received 2002 IEICE PRMU Research Encouraging Award, 2008 IEICE Best Paper Award, MIRU2006 Nagao Award (best paper award), MIRU2011 Excellent Paper Award, 2007 IAPR/ICDAR Best Paper Award, and 2010 ICFHR Best Paper Award. Dr. Uchida is a member of IEEE and IPSJ.

Marcus Liwicki received his M.S. degree in Computer Science from the Free University of Berlin, Germany, in 2004, and his PhD degree from the University of Bern, Switzerland, in 2007. Subsequently, he received the post-doctoral lecture qualification from the Technical University of Kaiserslautern, Germany, in 2011. Currently he is a senior researcher and private lecturer at the German Research Center for Artificial Intelligence (DFKI). His research interests include knowledge management, semantic desktop, electronic pen-input devices, on-line and off-line handwriting recognition and document analysis. From October 2009 to March 2010 he visited Kyushu University (Fukuoka, Japan) as a research fellow, supported by the Japanese Society for the Promotion of Science.

Yaokai Feng received his B.E. and M.E. degrees in Computer Science from Tianjin University, China, in 1986 and 1992, respectively. He received his Ph.D degree in Information Science from Kyushu University, Japan, in 2004. Now, he is an assistant professor at Kyushu University, Japan. His current research interests include database, pattern recognition, information retrieval and network security. In 2011, he received MIRU2011 Excellent Paper Award. He is a member of IPSJ and IEEE.