Analysis of Local Features for Handwritten Character Recognition

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

Marcus Liwicki
Kyushu University, Japan / DFKI, Germany
marcus.liwicki@dfki.de

***Abstract*—This paper investigates a part-based recognition method of handwritten digits. In the proposed method, the global structure of digit patterns is discarded by representing each pattern by just a set of local feature vectors. The method is then comprised of two steps. First, each of J local feature vectors of a target pattern is recognized into one of ten categories ("0"–"9") by the nearest neighbor discrimination with a large database of reference vectors. Second, the category of the target pattern is determined by the majority voting on the J local recognition results. Despite a pessimistic expectation, we have reached recognition rates much higher than 90% for the task of digit recognition.**

***Keywords*-local feature; handwritten character recognition; part-based recognition;**

I. Introduction

It seems that character recognition researchers have had a common sense that the global structure of each character is very essential for recognition. For example, many researchers have used stroke sequences, which directly represent the global structure of a character. Even a simple template matching method uses the global structure because it represents the structure on a two-dimensional pixel array.

In contrast, global structure has been easily discarded in visual object recognition. Specifically, in *part-based recognition* methods, each object is represented by a *set* of local features, each of which describes a local part around a keypoint. In general, many keypoints are detected automatically according to some criterion. It is important to note that part-based recognition methods usually discard the relative locations among local features. For example, the bag-of-keypoints approach, which recently became the most popular part-based recognition method, does not use any location information of the local features and thus totally discards the global structure.

The main motivation of this paper is to observe and analyze what happens if the global structure of handwritten characters is discarded from their recognition problem, i.e., what happens on part-based character recognition. We can easily have a pessimistic expectation that the recognition performance may be very poor — this is because it seems very difficult to recognize character patterns just by a set of some local parts of strokes. In fact, there are many similar circular (or crossing or straight) parts in English alphabets.

Against this pessimistic expectation, recognition results provided in this paper are very positive; for handwritten digits from MNIST database, we can have recognition rates much higher than 90%. While these rates do not exceed those by other state-of-the-art recognition methods (mainly based on global information), we still can be positive because we can expect strong advantages of the part-based recognition method on the tasks where conventional methods have difficulties. (These advantages will be listed in Section IV.)

II. Related Work

As noted in Section I, part-based recognition methods become more and more popular for visual object recognition. Those methods have evolved with keypoint detection and description techniques. Scale-invariant feature transform (SIFT) and its fast version speeded up robust features (SURF) [1] are two major choices. One of their important properties is that they provide local feature vectors invariant to rotation and scale change. Another property is that their feature vector describes a local part as a distribution of local edge directions.

In character recognition, however, only little attention has been paid to part-based recognition. In fact, features which represent the global structure of characters have been employed [2]. Very recently, a modified version of SIFT has been applied to the recognition of Chinese characters [3]. In [3], however, the keypoints are determined as dense regular grid points on the image and are therefore not truly part-based.

To the authors' best knowledge, the only work of part-based character recognition is Diem and Sablatnig [4], where degraded characters from historical manuscripts are to be recognized. The performance of their part-based recognizer achieved about 60 %. Unfortunately, their targets were somewhat peculiar and only few information about the testing environment was given. Furthermore, the recognition results were not analyzed carefully and thus it is still difficult to understand the characteristics of part-based recognition.

It is interesting to relate part-based character recognition to the process of human perception and human reading. When looking at an image, first the most conspicuous parts are perceived during the process of human vision. Especially for human reading, crossings, line endings, and curvatures play an important role [5]. Furthermore, ascenders and descenders are firstly processed [6]. This behavior is simulated

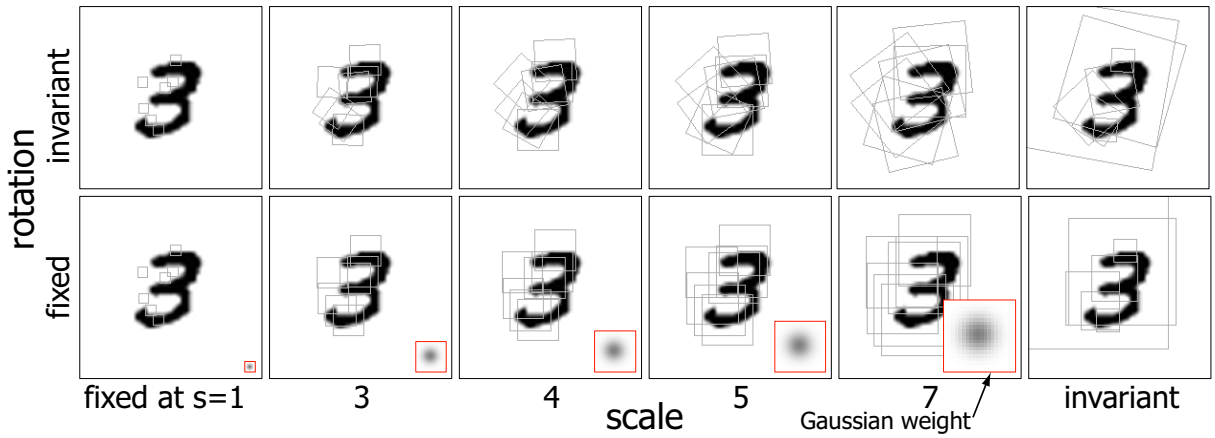

Figure 1. SURF keypoints under different conditions (in part).

by the part-based recognition method where the keypoints are determined around the most conspicuous parts.

III. Part-Based Character Recogntion

The part-based character recognition method used here is as follows. Note that although this method is not based on bag-of-keypoint approach, other state-of-the-art visual object recognition methods (e.g., [7]) are organized similarly.

Training: Let $\boldsymbol{T}_{c,i}$ denote the ith training pattern of the category c, where $c \in \{1,\ldots,C\}$ and $i \in \{1,\ldots,I_c\}$. The training step of the proposed method is done just by detecting $N_{c,i}$ keypoints from $\boldsymbol{T}_{c,i}$ and then describing the local part around every keypoint as a local feature vector $\boldsymbol{r}_{c,k}$. Hereafter, $\boldsymbol{r}_{c,k}$ is called *reference vector* and $\Omega = \{\boldsymbol{r}_{c,k} \mid \forall c,\ k = 1,\ldots,K_c\}$ is called *dictionary*. Note that $K_c = \sum_i N_{c,i}$. The technology to detect and describe the keypoints in our experiment will be discussed in Section V.

Feature-level recognition: Let $\boldsymbol{Q}$ denote an input pattern and $\{\boldsymbol{q}_j \mid j = 1,\ldots,J\}$ denote its local feature vectors. The recognition of $\boldsymbol{Q}$ is comprised of two steps. The first step is *feature-level recognition* where $\boldsymbol{q}_j$ is recognized by using the 1-NN rule on Ω. (Euclidean distance was used in the experiment.) Let $\tilde{c}_j$ denote the category of the 1-NN reference vector. Note that the original locations of $\boldsymbol{q}_j$ and $\boldsymbol{r}_{c,k}$ are discarded here; thus, for example, $\boldsymbol{r}_{c,k}$ from an upper area of a training pattern can be the 1-NN vector for $\boldsymbol{q}_j$ from the lower area of $\boldsymbol{Q}$.

Character-level recognition: The second step is *character-level recognition* where the category of $\boldsymbol{Q}$ is determined just by majority voting among $\{\tilde{c}_j\}$.

IV. Expected Advantages

The expected advantages of part-based character recognition methods are as follows:

- It will be robust to occlusion, decoration, and other partial degradations on character images.
- It can be considered as an unconstrained version of image distortion model [8], where the distance between two images evaluated under an optimal pixel-to-pixel correspondence, and thus be robust to severe deformations.
- It will relax the difficulty of segmentation. For example, it might be possible to recognize cursive scripts without segmentation just by having local recognition results at keypoints.
- Similarly, it can be applicable to text detection from scenery images.

V. Experiments

A. Dataset

As our dataset, 20,000 samples were extracted from the "training" dataset of MNIST handwritten digit database [9]. (Readers will find the recognition rates of MNIST achieved by various methods at [9].) For each category, the 2,000 samples were divided into 1,000 training patterns (thus $I_c = 1,000$) and 1,000 validation patterns. Most of the following results were evaluated by using these validation patterns as unknown input patterns. Note that although more training patterns provide a better recognition rate as shown later, we used only 1,000 training samples per category due to computational limitation.

Since the original image size (28×28) was too small to extract a considerable amount of keypoints, each pattern was magnified four times after the addition of 10-pixel surrounding margin. Consequently, the pattern size became 192×192.

B. Keypoint Detection and Description

As the keypoint detection and description technique, SURF [1] was employed. SURF is a fast version of SIFT

Table I
AVERAGE, STANDARD DEV, MAX, AND MIN OF $N_{c,i}$ FROM EACH TRAINING PATTERN.

	category										
	0	1	2	3	4	5	6	7	8	9	total
ave	78.9	38.5	61.5	60.1	57.1	60.1	61.8	54.8	61.9	56.3	59.1
stdev	12.3	11.1	10.9	12	10.3	12.2	12.7	10.8	11.6	10.7	14.8
max	123	83	99	101	102	107	112	95	104	94	123
min	37	12	32	30	30	31	29	24	32	25	12

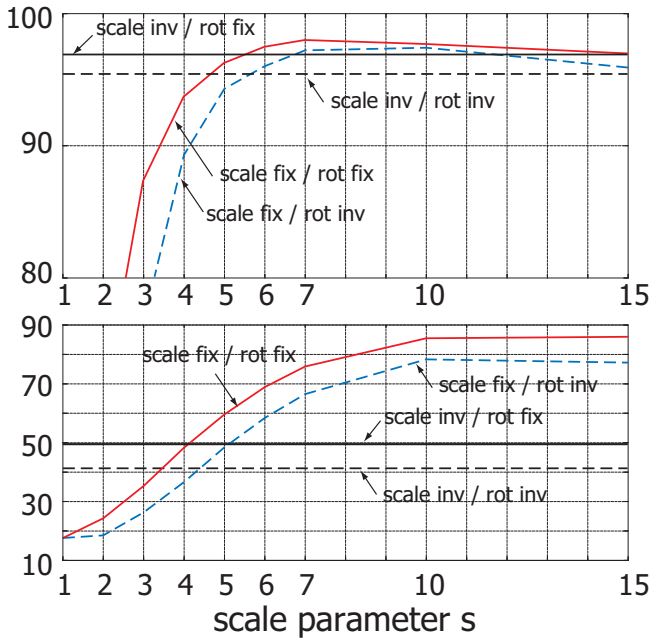

Figure 2. (Upper) character-level and (lower) feature-level recognition rates.

and its local feature vector is rotation and scale invariant; SURF automatically determines the size and orientation of its square area for describing the local feature around each keypoint.

We also conducted experiments after removing the rotation and scale invariance. In this case, the rotation (i.e., the orientation of the square area) was fixed at 0° and/or the scale (i.e., the size of the square area) was fixed by the parameter s.

C. Training Results

Table I shows the statistics of the dictionary, comprised of about 59,000 reference vectors. This means that the average number of keypoints per training pattern was about 59. This number increases on "0" and decreases on "1".

Figure 1 shows examples of keypoints detected. The original SURF feature (rotation and scale invariant) sometimes covers the entire character. Thus, this case is not "genuine" part-based recognition. In contrast, if we fix the scale by a small s, we can realize a part-based recognition. For example, when $s = 4$, the size of the square area is about 1/3 of the character size.

It is important to note that since SURF employs Gaussian weight on its feature vector, the truly effective area is narrower than the square area (illustrated underneath the digits in Fig. 1). If we consider the weight, the effective size is reduced to about 1/20 of the character size at $s = 4$.

Table II
CONFUSION MATRIX (%) ON THE CHARACTER-LEVEL. (ROT:FIX; SCALE:FIX($s = 4$))

input \ recognition result	0	1	2	3	4	5	6	7	8	9
0	98.0	0.0	0.4	0.5	0.1	0.2	0.5	0.0	0.2	0.1
1	1.0	91.8	0.4	0.0	0.5	0.1	0.9	**4.4**	0.0	0.9
2	1.7	0.1	95.9	0.6	0.0	0.3	0.5	0.8	0.1	0.0
3	1.2	0.1	0.9	95.0	0.0	1.9	0.0	0.5	0.3	0.1
4	0.8	0.6	0.0	0.0	94.1	0.0	0.4	0.3	0.1	**3.7**
5	1.3	0.1	0.1	2.5	0.1	94.7	0.8	0.0	0.2	0.2
6	2.9	0.3	0.1	0.0	0.0	1.0	95.3	0.0	0.4	0.0
7	2.2	**3.8**	2.8	0.3	2.8	0.1	0.0	87.5	0.0	0.5
8	0.7	0.0	0.1	2.1	0.3	0.1	0.3	0.0	96.0	0.4
9	**3.5**	0.4	0.4	0.5	2.3	0.7	0.2	1.0	1.3	89.7

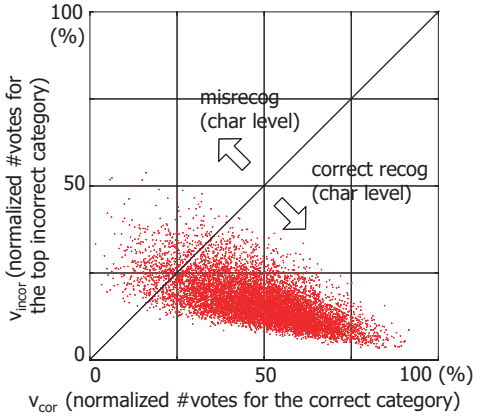

Figure 3. Distribution of $v_{\rm cor}$ and $v_{\rm incor}$. (Rot:fix; Scale:fix($s = 4$))

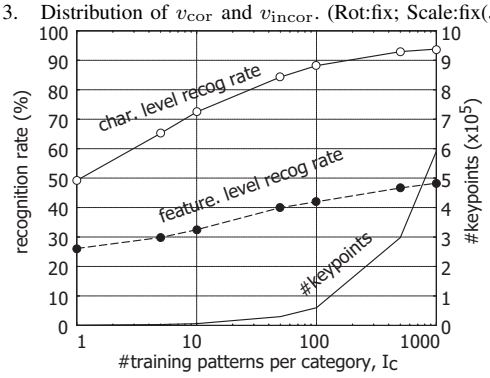

Figure 4. Recognition rate as a function of training set size. (Rot:fix; Scale:fix($s = 4$))

D. Character-Level Recognition

Figure 2 shows the character-level recognition rates. The original SURF feature (rotation and scale invariant) has achieved 95.4%. When the scale was fixed, a smaller s results in a lower recognition rate. However, it was beyond our expectation that we could still have 93.8% at $s = 4$ (and fixed rotation) [1] Again, the effective size of area for describing each local part is about 1/20 of the character size when $s = 4$; each local feature vector just describes a fragment of a stroke. Therefore, this result is very positive to the part-based character recognition method. The highest performance of 97% was achieved with a fixed rotation and $s = 7$. In this case the effective size is about 1/7 of the character size.

Table II(a) shows the confusion matrix in the case of fixed rotation and the fixed scale ($s = 4$). The best rate was 98.0% for "0" and the worst was 87.5% of "7", respectively. One of major misrecognition pairs was "1"↔"7". It is also remarkable that categories with circular strokes (e.g., "2", "6", "9") were misrecognized as "0".

E. Feature-Level Recognition

Figure 2 also shows feature-level recognition rates, which are the rates that $\boldsymbol{q}_j$ and its 1-NN reference vector have the same category. It is observed that the feature-level recognition rates are far worse than character-level recognition rates. In fact, only 50% of the feature vectors were correctly recognized in the case of the fixed rotation and the fixed scale ($s = 4$).

In other words, 50% of feature vectors got a 1-NN vector from a different category. This proves that we cannot assume that the feature vectors form clusters of categories in the feature space and thus we cannot employ recognition methods which assumes clusters. For example, the Bayes classifier based on Gaussian distributions will not work for feature-level recognition.

F. Effect of Majority Voting

In Fig. 3, all test patterns are projected on a two-dimensional plane according to their v_{cor} and v_{incor} values (%). The former is the normalized number of votes to the correct category at the character-level recognition. The latter is the normalized number of votes to the top incorrect category. If $v_{\mathrm{cor}} > v_{\mathrm{incor}}$, a test pattern is correctly recognized at character-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 implies that the misrecognitions at the feature-level recognition were not converged into a certain incorrect category but scattered into various incorrect categories. According to this fact, character-level recognition achieves recognition rates higher than 90% by majority voting.

[1] When the 10,000 test samples of the "test" dataset of MNIST were used for final evaluation, the recognition rate of this case was 93.6%. That is, there was no significant difference.

G. Effect of Training Set Size

As shown in Fig. 4, the number of training patterns (I_c) affects the recognition rates drastically. In the extreme case, that is, if we use only a single training pattern for each category, feature-level recognition rates are degraded to 30%. This also proves that the local features are distributed with considerable overlaps and thus we need many reference vectors for increasing the probability of finding a 1-NN of the correct category. Note that the recognition rate was not saturated with 1,000 training patterns and thus will be improved if we use more training patterns.

VI. Conclusion

A part-based character recognition experiment was conducted and, beyond one's expectation, positive results have been obtained. Although feature-level recognition is actually difficult, majority voting on character-level recognition helps to have the correct final recognition result. Even if we use a small part (about 1/20 of the character size) as the unit area of local feature description, we could achieve 93.8% recognition accuracy. Starting from this positive result, we will extend our application by considering the merits of the part-based recognition method.

Acknowledgment

The authors would like to thank Vincent Marsault (ENSEEIHT, France) for his contribution in launching this research.

References

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

[2] Ø. D. Trier, A. K. Jain, and T. Taxt, "Feature Extraction Methods for Character Recognition – A Survey," Pattern Recognit., vol. 29, no. 4, pp. 641–662, 1996.

[3] Z. Zhang, L. Jin, K. Ding, and X. Gao, "Character-SIFT: A Novel Feature for Offline Handwritten Chinese Character Recognition," Proc. ICDAR, pp. 763–767, 2009.

[4] M. Diem and R. Sablatnig, "Recognition of Degraded Handwritten Characters Using Local Features," Proc. ICDAR, pp. 221–225, 2009.

[5] L. Schomaker and E. Segers, "Finding Features Used in the Human Reading of Cursive Handwriting," IJDAR, vol. 2, no. 1, pp. 13–18, 1999.

[6] L. Avallone, C. De Stefano, C. Gambone, and A. Marcelli, "Visual Processes and Features in Human Reading of Cursive Handwriting," Proc. 12th Conf. Int. Graphonomics Soc., pp. 128–132, 2009.

[7] O. Boiman, E. Shechtman, M. Irani, "In Defense of Nearest-Neighbor Based Image Classification," Proc. CVPR, 2008.

[8] D. Keysers, T. Deselaers, C. Gollan, H. Ney, "Deformation Models for Image Recognition," IEEE Trans. PAMI, vol. 29, no. 8, pp. 1422-1435, 2007.

[9] `http://yann.lecun.com/exdb/mnist/`