Character Image Patterns as Big Data

Seiichi Uchida, Ryosuke Ishida, Akira Yoshida, Wenjie Cai, Yaokai Feng
Kyushu University, Fukuoka, Japan
uchida@ait.kyushu-u.ac.jp

Abstract

The ambitious goal of this research is to understand the real distribution of character patterns. Ideally, if we can collect all possible character patterns, we can totally understand how they are distributed in the image space. In addition, we also have the perfect character recognizer because we know the correct class for any character image. Of course, it is practically impossible to collect all those patterns — however, if we collect character patterns massively and analyze how the distribution changes according to the increase of patterns, we will be able to estimate the real distribution asymptotically. For this purpose, we use 822,714 manually ground-truthed 32×32 *handwritten digit patterns in this paper. The distribution of those patterns are observed by nearest neighbor analysis and network analysis, both of which do not make any approximation (such as low-dimensional representation) and thus do not corrupt the details of the distribution.*

***Keywords*-handwritten character patterns, distribution analysis, nearest neighbor, minimum spanning tree, big data**

I. Introduction

Needless to say, pattern distribution is the most important factor for pattern recognition. For example, if we know that patterns are characterized by Gaussian distributions, we can derive a quadratic discriminant function as the optimal classifier. If we know that patterns distribute in a low-dimensional manifold, we will introduce some dimensionality reduction techniques for feature extraction. If we know that patterns distribute as groups, we will introduce some clustering techniques for understanding representative patterns of individual groups.

Unfortunately, it is generally impossible to know the *true* distribution of patterns. For example, if we want to know the true distribution of 100×100 24-bit color images, we must collect ground-truthed $2^{240,000}$ images. However, if we make as much effort as possible to approach the true distribution asymptotically and then understand the trend of the distribution in various viewpoints, we can enhance the pattern recognition accuracy and/or develop new pattern recognition strategies.

This paper tries to understand the true pattern distribution as accurate as possible through various qualitative and quantitative analyses. For this ambitious purpose, 822,714 manually ground-truthed handwritten isolated digit patterns ("0",. . ., "9") are used. As emphasized in Section II, digit patterns are more suitable than general image patterns for this purpose in the several points.

As the tools for the distribution analyses, we will use *nearest neighbor (NN) analysis* and *network analysis*. NN analysis is simple but powerful in the sense that it can provide not only the 1-NN distance-based recognition accuracy but also the distribution of 1-NN distances, and so on. This distance distribution is useful to understand the global properties of the pattern distribution. As a unique application of NN analysis, we will also show results of image completion using 1-NN pattern.

Network analysis is also promising approach for analyzing the topological structure of pattern distribution. Nowadays, various large-scale network analyses are performed for complex network (e.g., scale-free network), graph mining, etc. In this paper, we will employ the minimum spanning tree (MST) as the network representation of massive character patterns and then analyze its structure in various ways.

Instead of popular lower-dimensional visualizations for distribution analysis, we adhere to use those approaches in this paper. This is because they realize simpler and lossless analyses. In other words, lower-dimensional visualization generally causes some approximation error, which may conceal important properties of "minorities" in the distribution.

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.190

II. Related Work

Recent computer hardware development allows us to process "big data". In fact, this trend is accelerating in image processing and pattern recognition. For example, we can access ground-truthed 11 million images of ImageNet [1]. (Now ImageNet contains 14 million images.) Torralba et al. [2] prepared a huge dataset with 80 million images gathered from the Internet and resized into 32×32 pixels. They showed that high recognition accuracy was achieved not by a complex and sophisticated recognition method but just by the simplest 1-NN rule with their massive image dataset. In addition, there are many fascinating trials using massive image datasets and rather simple techniques, such as [3], [4], [5], [6], [7], [8]. Although any of those datasets are far smaller than the collection of all possible images, those trials have often showed asymptotic analysis results on the effects of the size of datasets.

In this paper, we deal with "big data" of character image patterns. From the purpose of analyzing real pattern distribution, character image patterns possess the following merits over general image patterns. (i) Since there are only 10 classes for digits, it is possible to have an enough number of patterns per class for understanding the precise distribution of each class. (ii) Small and binary character images can form a compact feature space. (iii) The classes of character patterns can be defined with far less ambiguity than visual objects.

In past character recognition researches, rather smaller datasets have been used. The well-known MNIST dataset contains only 70,000 handwritten digit patterns. Smith et al. [9] have done one of the largest-scale researches with 223,000 (i.e., 1/4 of ours) handwritten isolated digit patterns, while mainly focusing on recognition accuracy rather than distribution analysis. The CASIA-HWDB2.0-2.2 [1] is a large handwritten Chinese character dataset which contains 3,895,135 patterns — around 550 patterns for each of 7,000 classes.

III. Experimental Setup

A. Dataset

Our handwritten digit image dataset is comprised of 822,714 patterns. Figure 1 shows several patterns from the dataset. All of the digit patterns were first isolated from their original scanned images. Then the ground-truth, i.e., correct class label ("0",. . ., "9"), was attached to each pattern carefully by manual inspections by several professional operators. Each pattern is a binary image (black and white) rescaled to 32 × 32 pixels.

NN analysis in Section IV was done by the leave-one-out manner, where every pattern of 822,714 patterns was treated as an input pattern and then its NN pattern were selected from the remaining 822,713 patterns. When we use a subset of the dataset, a certain number of patterns were randomly selected from the remaining 822,713 patterns.

[1] http://www.nlpr.ia.ac.cn/databases/handwriting/Home.html

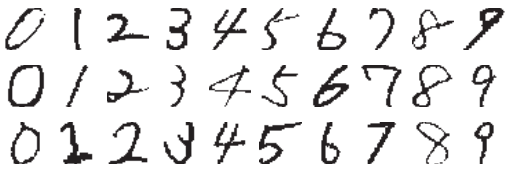

Figure 1. Digit images from our dataset.

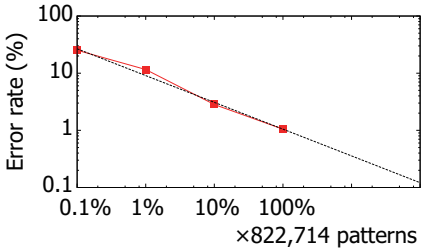

Figure 2. Error rates by 1-NN under different dataset sizes.

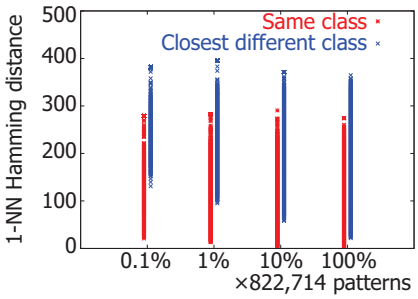

Figure 3. Distribution of 1-NN distance.

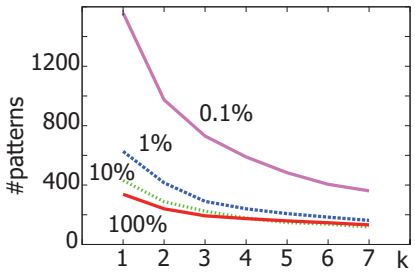

Figure 4. The number of patterns whose all k-NNs belong to other classes.

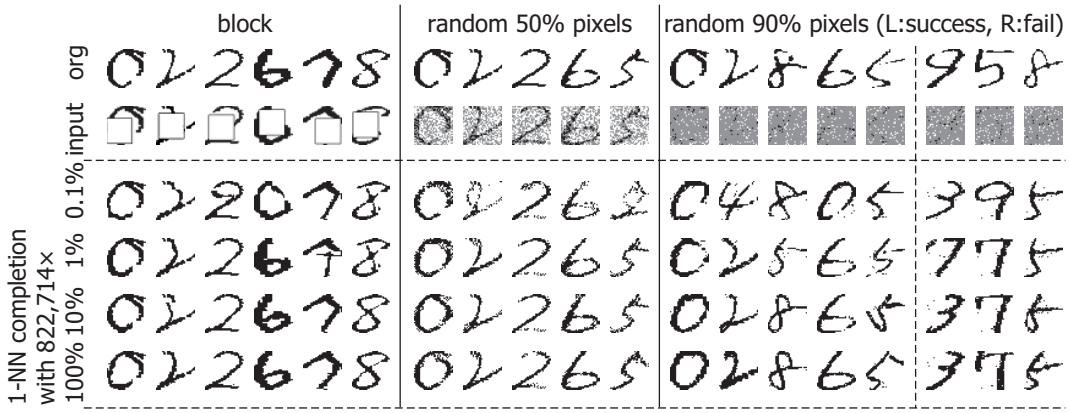

Figure 5. Image completion by 1-NN pattern.

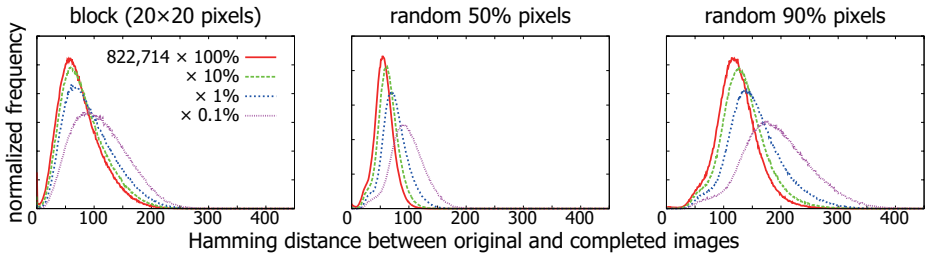

Figure 6. Image completion accuracy.

B. Feature and Distance Metric

Throughout this paper, the original simple feature, i.e., black and white pixel value is used in this paper, since the improvement of recognition accuracy is not our purpose. Consequently, each pattern is represented as a 1,024-dimensional binary vector and thus corresponds to a corner of the 1,024-dimensional hypercube. Note that the analyses on this discrete distribution is directly related to a continuous distribution by some feature extraction. This is because most feature extraction methods are based on some linear operations and thus properties in the original discrete distribution are mostly preserved even in the continuous distribution.

As the distance metric, we employ Hamming distance. The value of the Hamming distance can be interpreted intuitively. For example, if the Hamming distance is 100 between two 32×32 binary patterns, this indicates that the two patterns have different black/white value at 100 pixels (about 10% among 1,024 pixels).

IV. Nearest Neighbor Analysis

A. 1-NN Distance Analysis

Figure 2 plots the recognition accuracy by 1-NN discrimination under different dataset sizes. We can roughly find a parametric relation between the dataset size and the recognition accuracy. As shown in Fig. 2, we can estimate that, if the dataset size increases 10 times, the error rate decreases to 40%. For example, if we increase the dataset size to 100 times, i.e., if we have 82 million patterns, the error rate may become around 0.1%. Note that this logarithmic property coincides with the observation by Torralba [2].

Figure 3 plots the distributions of 1-NN distances to the same class and the closest different class under different dataset sizes. First, we observe the expected trend that both distances decrease by increasing dataset size. Second, we observe a more important fact that the distance to the closest different class is lower-bounded. In fact, even though 822 thousand patterns are densely distributed in their image space, any pair of images from *different* class have at least 20 pixels (about 2% of all 1,024 pixels) with different black/white color, whereas some pair of images from the same class are almost identical (that is, they have a distance around zero). This fact clearly indicates that there is a certain "gap" which differentiates one digit class from others.

Figure 4 shows the number of patterns whose all k-NNs belong to other class[2]. In other words, this graph shows the number of patterns surrounded by one or more different classes. We can consider those patterns as so-called outliers. The important fact is that

[2]Again, we used the leave-one-out manner. Accordingly, for each of all 822,714 patterns, its k-NNs from $p\%\times$ 822,713 are examined, where $p = 0.1, 1, 10$, and 100.

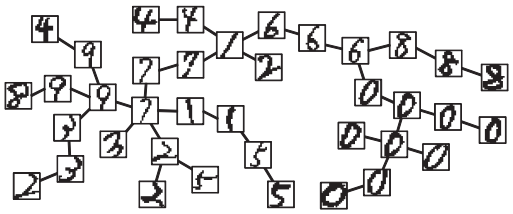

Figure 7. A tiny example of minimum spanning tree.

outliers *die hard*; even though we increase the dataset size, we had at least 200 outliers. Note that, as we observed from Fig. 3, those outliers should have 20 pixel difference from their neighbors. Consequently, these outliers are deviated from the distribution of its class with more than 20 pixel difference.

B. Image Completion by Nearest Neighbor

The power of a large dataset was observed through an image completion experiment. A set of pixels were removed from an original character image and then the values of those removed pixels are determined by referring the 1-NN pattern. The 1-NN pattern was selected by using the Hamming distance of the non-removed pixels.

Figure 5 shows the completion result on several patterns. More successful completion results are provided with larger dataset, regardless the type of pixel removal (i.e., block and random). Surprisingly, even if 90% pixels are removed, the completion by using the remaining 10% pixels are sometimes accurate.

Figure 6 shows the distribution of completion accuracy by Hamming distance under different number of removed pixels. When 90% pixels are removed, the average distance is around 100 with 822 thousand patterns. This indicates that, among 900 pixels (90% of 1,024 original pixels), 800 pixels are correctly completed. Note that this result indicates another possibility to make low resolution character image into higher resolution one by using a similar idea of face hallucination [3].

V. Network Analysis

A. Minimum Spanning Tree

In the following experiment, minimum spanning tree (MST) was employed for representing the structure of the character pattern distribution. The node and edge of MST are a single binary digit image and the distance between a pair of images. Note that in this

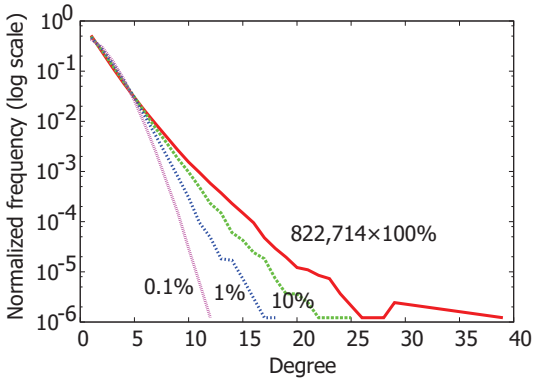

Figure 8. Distribution of node degree.

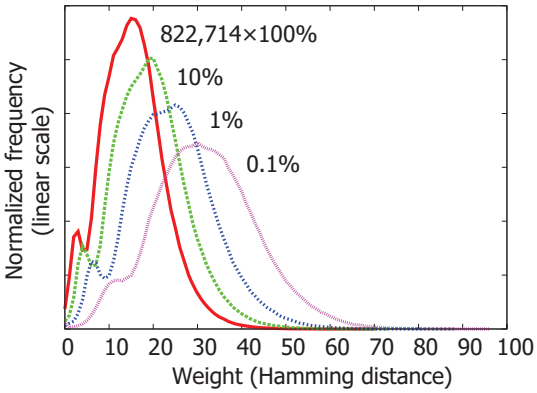

Figure 9. Distribution of edge weight.

experiment, 32×32 images were scaled to 16×16 to have a more dense distribution. Figure 7 shows the MST for a very tiny subset of our dataset.

MST has four characteristics suitable for large-scale distribution analysis. First, the proximity between similar patterns are preserved on MST and therefore global structure of the pattern distribution is also preserved. Second, similar patterns will form a cluster on MST. For example, even the small MST of Fig. 7, images of "0" form a cluster. Third, since MST is a tree, there is a unique and connected path between any pair of images (i.e, nodes). The path provides a "morphing" animation also a class transition between the pair. Fourth, there are efficient algorithms (e.g., Prim's algorithm) for constructing a large MST.

B. Statistics of MST

Figure 8 shows the distribution of node degree of the MST for 822 thousand digit images. We observe that the MST has more nodes with a large degree with a larger dataset. (In other words, we have more "hub" nodes as MST becomes larger.) This fact indicates that the distribution is non-uniform and there are clusters. The nodes around the center of a cluster have more

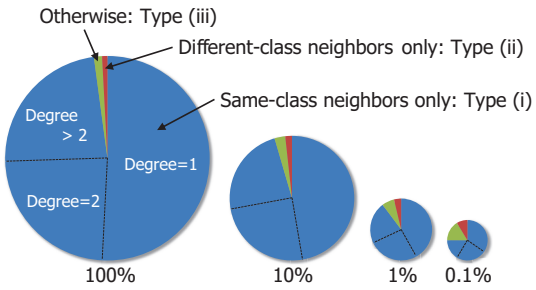

Figure 10. Classification of nodes by their class consistency.

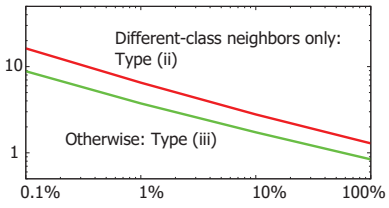

Figure 11. Percentage of nodes of Types (ii) and (iii).

neighbors as the dataset becomes larger.

Figure 9 shows the distribution of edge weight of the MST. Similarly to Fig. 9, it simply shows that the pattern distribution becomes more dense as the dataset size becomes larger.

1) Classification of Nodes by Class Consistency: Figure 10 shows the classification result of nodes by their *class consistency*. Class consistency of a node is one of three types, that is,

- Type (i): its all neighboring nodes (on MST) are from the same class,
- Type (ii): its all neighboring nodes are from other classes, and
- Type (iii): otherwise.

A node of Type (ii) can be considered as a kind of outliers. A node of Type (iii) is a "bridge" connecting two classes.

From Fig. 10, it is observed that most nodes are Type (i). In contrast, nodes of Types (ii) and (iii) are rare. Consequently, this result indicates each class forms a small number of large clusters on the MST. It coincides with the fact that support vectors, which exist around class boundaries, are often far less than the entire patterns.

Figure 11 plots the percentages of Types (ii) and (iii) as another graph. An important observation is that the percentages of both types are reduced to about 40%, if the dataset size increases 10 times. This observation leads two points. First, we can estimate the number of outliers, which are related to Type (ii). Second, this observation coincides with the result of Fig. 2 very well. In fact, the patterns of Types (ii) and (iii) have a large possibility to be misrecognized.

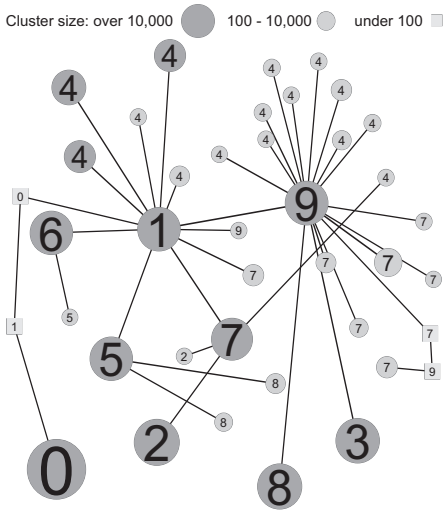

Figure 12. Cluster tree.

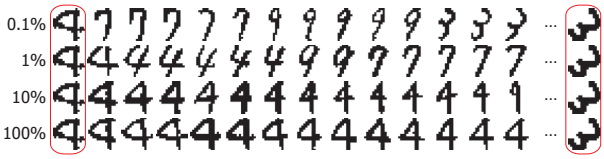

Figure 13. Image sequence along with the path between "4" and "3". Note that the first 14 patterns from the node "4" are plotted.

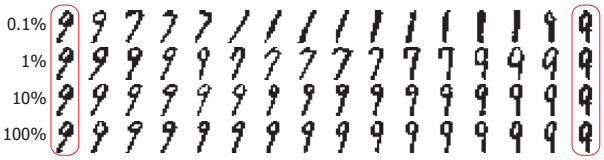

Figure 14. Image sequence along with the path between two different "9"s. Note that 16 patterns were selected along the path with the same interval.

C. Class Distribution on MST

As observed above, most nodes have neighbors from the same class and thus each class maybe forms a limited number of large clusters. For verifying this, we construct a *cluster tree*, where neighboring nodes from the same class are unified. For example, "0" nodes of Fig. 7 are unified into one node in the cluster tree. Two nodes spanning an edge of the cluster tree are Type (iii).

Figure 12 shows the cluster tree created from the

MST of all 822 thousand patterns. Each node, hereafter called supernode, is a cluster and its label is the class of the supernode and its size indicates the number of the unified nodes of the original MST. For better visibility, a supernode is omitted in this figure if it contains less than 100 nodes and its removal does not make the cluster tree disconnected.

Figure 12 shows that the cluster tree have several big supernodes. Specifically, every class (except "4") has a single big supernode and consequently it is shown that patterns of the class form one huge cluster in the original MST. Classes "4" and "7" have more supernodes and thus many clusters in the MST. We can say that those classes have several allographs and/or big overlaps with another class. For example, class "7" has typical two allographs, one of which is similar to "1" and the other "9".

Figure 12 also shows that class "1" is the main "hub" of the cluster tree. In fact, the path between any class pair (except for four class pairs, "2"-"7", "3"-"8", "3"-"9", and "8"- "9"), go through class "1". One reason is the fact that "1" has the most fundamental shape among digits. Specifically, most digits are vertically long and thus have similarity to "1".

Figure 13 shows the image sequence on the path between two patterns "4" and "3". The first 14 patterns from the node of "4" are plotted. It is observed that, as the dataset size increases, the neighboring patterns on the path becomes more similar to each other. This also represents how the pattern distribution becomes dense.

Figure 14 shows the image sequence on the path between two patterns from the same class "9". It is observed that, as the dataset size increases, the path is improved by using more similar patterns, which provides a better short-cut path. From another viewpoint, class "9" is scattered into small clusters with less patterns and then connected into a larger cluster with more patterns.

VI. Conclusion

The distribution of massive (about 822 thousand) handwritten digit patterns are analyzed by two methodologies, i.e., nearest neighbor analysis and network analysis. Although they are simple, they are free from any approximation in their representation and thus useful to observe the details of the distribution. Many observation have been conducted and the following facts were revealed:

- By increasing dataset size 10 times, the error rate decreases to 40%. This means that a further increase of the dataset size will improve the recognition accuracy, although its effect becomes smaller. This coincides with a network analysis result which shows that the percentage of patterns neighboring two different classes also decreases to 40%.
- By increasing dataset size, the NN patterns to the input pattern become closer. This fact was confirmed by observing not only the NN distance but also the edge weight distribution of the MST connecting all patterns.
- The existence of outliers, which were defined as patterns surrounded by patterns of different classes, was confirmed. It was difficult to change these outliers into inliers even with all of the 822,714 patterns.
- Nearest neighbor patterns are useful for image completion. In fact, even when 90% pixels are removed, it is possible to complete 89% (=800/900) of those pixels on average.
- The structure of MST connecting all 822 thousand patterns was observed quantitatively and qualitatively. An observation surely reveals that most classes form their own huge cluster in their distribution.

References

[1] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," Proc. CVPR, 2009.

[2] A. Torralba, R. Fergus, and W. T. Freeman, "80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition," IEEE Trans. PAMI, vol. 30, no. 11, pp. 1958-1970, 2008.

[3] C. Liu, H. -Y. Shum and W. T. Freeman, "Face Hallucination: Theory and Practice," Int. J. Comp. Vis., vol. 75, no. 1, pp. 115-134, 2007.

[4] J. Hays and A. A. Efros "Scene Completion Using Millions of Photographs", Proc. SIGGRAPH, 2007.

[5] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, A. Torralba, "SUN Database: Large-scale Scene Recognition from Abbey to Zoo," Proc. CVPR, 2010.

[6] F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting Image Databases from the Web," IEEE Trans. PAMI, vol. 33, no. 4, pp. 754-766, 2011.

[7] A. Karpenko and P. Aarabi, "Tiny Videos: A Large Data Set for Nonparametric Video Retrieval and Frame Classification," IEEE Trans. PAMI, vol. 33, no. 3, pp. 618-629, 2011.

[8] C. Liu, J. Yuen, and A. Torralba, "Nonparametric Scene Parsing via Label Transfer," IEEE Trans. PAMI, vol. 33, no. 12, pp. 2368–2382, 2011.

[9] S. J. Smith, M. 0. Bourgoin, K. Sims, and H. L. Voorhees, "Handwritten Character Classification Using Nearest Neighbor in Large Databases," IEEE Trans. PAMI, vol. 16, no 9, pp. 915-919, 1994.