Appendix A

Chinese Handwriting Recognition

As a side experiment, we used the feature selection techniques discussed in Chapter 7 to recognize Chinese handwriting. Our goals are: (1) to demonstrate feature selection is important because it is the crucial part for the recognition job. (2) to compare the feature set found by the feature selection algorithms with a human expert's selection.

1.1 Feature selection for Chinese handwriting recognition

Although most of the research in handwriting recognition is for on-line systems [Singer et al, 94], there is no doubt that off-line systems are also very important especially in domains such as automatic tax form processing.

To date, research for Chinese and Japanese character recognition is still preliminary[1]. Because the number of Kanji, i.e. Chinese characters, is over fifty thousand, it is hard to rely on any general-purpose global model to recognize all Chinese characters. Alternatively, a promising approach is to separate the Chinese characters into several groups. For each group, a local model is developed to distinguish the different characters.

---

1. There are some Chinese and Japanese recognition products on the market. The product introductions claim that their accuracy is over 90%. However, we do not know what kind of principles they apply. And we notice some of those products can only recognize rigidly written characters.

Although it may be possible to build the local models off-line, manually, it is better if we have an on-line automatic configuration mechanism. Not only does this automatic system save software developers from tedious and time-consuming work, but also it is adaptive and can learn different personal handwriting styles.

In this section, we propose an idea to recognize Chinese and Japanese handwriting off-line, with automatically configured adaptive local models. We also give a prototype of this system.

Chinese characters are constructed by ten fundamental strokes.

$$\text{、 一 丨 丿 乀 ⇁ 亅 乛 乚 〈}$$
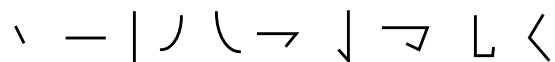
The different combinations with different relative positioning determine different characters. For example, there are eight different Chinese characters plus "F" and the Japanese character "ki" containing two horizontal lines and one vertical line, illustrated in Figure A-1.

In this prototype system, some features are useful for recognition, while others may not be so significant, or, can be substituted, referring to Figure A-2. Notice: (1) The human expert's selection, as shown in Figure A-2(a), is not the only functional set, there exist multiple options. (2) Among the multiple functional feature sets, some of them may lead to more accurate recognition than the others.

To find the features including those not-so-significant, we can follow these three steps:

- Figure out the horizontal lines, vertical lines, and other strokes, respectively.
- Sort the lines from top to bottom, or from left to right.
- Calculate all the possible features according to prior knowledge. In the case of Figure A-1, each stroke has two ends. The features can be the distances from the ends of each stroke to
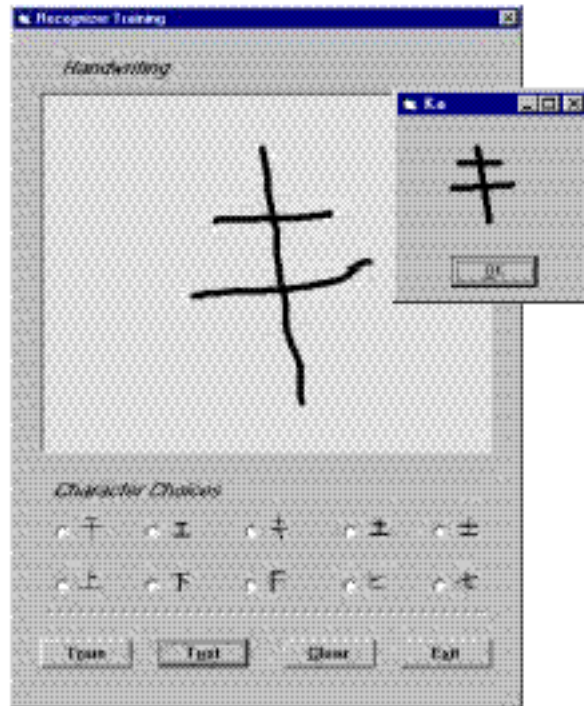
Figure A-1: A prototype of Chinese handwriting recognition system.

(a) Features selected by human expert

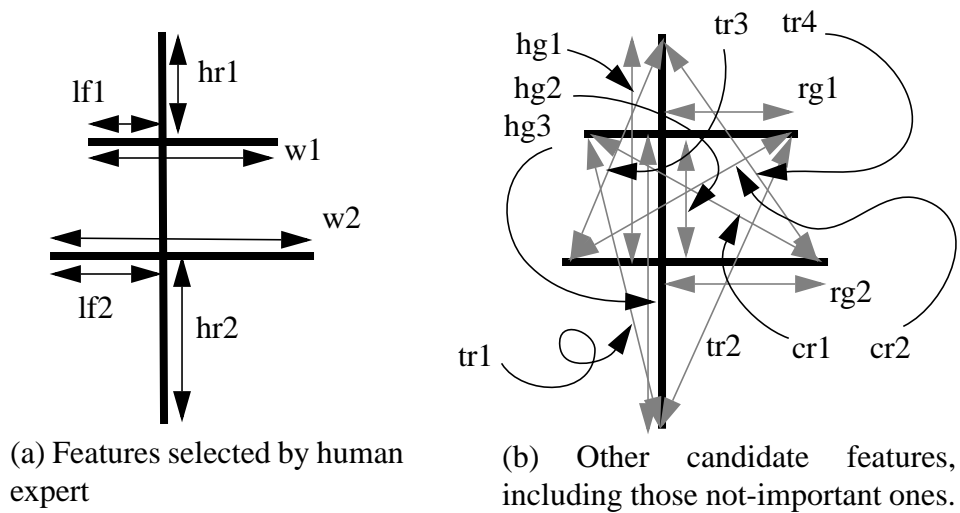(b) Other candidate features, including those not-important ones.

Figure A-2: The features used for the Chinese handwriting recognition prototype system.

those of others, as well as the distances to all the intersections, illustrated in Figure A-2.

After we have found the candidate features, we can apply the various feature selection algorithms to select the proper features for the recognition job. In the experiment, we try four feature selection algorithms: Super-greedy (Super), Greedy (Greedy), Restricted Forward Selection (RFS) and conventional Forward Selection (FS). We request that any selected feature sets contain no more than eight components. To evaluate the goodness of the selected feature sets, we calculate their 20-fold scores. Since our procedure is carefully designed to avoid overfitting, the smaller a feature set's score is, the more accurately this feature set is able to recognize any one out of the ten characters. We also count the numbers of seconds consumed by the four algorithms so as to compare their computational costs.

Table A-1: Kanji feature selection

Selection Methods	*m = 8 = Max Number of features*			*m = 12*	
	Selected feature set	*20fold score*	*Cost*	*20fold score*	*Cost*
Super	w2, lf1, lf2, hg1, tr1, tr2, tr3, tr4	0.038	532	0.018	529
Greedy	w2, lf1, lf2, hg1, tr2, tr3, tr4	0.041	767	0.022	916
RFS	hr1, w1, lf1, lf2, hg1, hg3, cr2, tr1	0.018	1414	0.016	1570
FS	hr1, hr2, w1, lf1, lf2, hg1, tr3	0.016	3586	0.018	4829
Human	hr1, hr2, w1, w2, lf1, lf2	0.016	--	0.016	--

In Table A-1, we observe that different selection algorithms may find different sets of features. When we carefully study these various sets with respect to Figure A-2, we find all of them are functional. Second, we find that the feature sets selected by RFS and FS are very similar to the human expert's preference, but different from the sets found by Super and Greedy. Third, although all of these feature sets have satisfactory accuracy, those found by the greedier algorithms lead to less accurate recognition performance. However, if we allow more components

to enter the feature sets, even the greedier algorithms' selections become more powerful. Finally, the greedier algorithms are cheaper than the others.

1.2 Future work

The prototype system is sufficient to demonstrate the importance and capability of the feature selection algorithms. But to pursue a good Chinese handwriting recognition system, some further work has to be done. Since this topic is a digression from the discussion of feature selection, we only give a brief introduction.

For more complicated kanji, for example 藏 which means "hide" and "Tibet", the number of possible features will explode. Fortunately, every Chinese character can be split into some standard particles, and the number of these standard particles is no more than one hundred. Indexed by these particles and their relative positioning, any Chinese character can be represented by no more than five digits. One example is illustrated in Figure A-3. This technique is called Wang-coding or Five-stroke coding, which has become one of the national standard typing methods in China.

叭　　23 (□)　24 (八)　1 (left to right)

兄　　23 (□)　24 (八)　2 (up and down)

Figure A-3: An illustration of Wang-coding of a Chinese character.

Now the remaining difficulty is how to find those standard particles from any Chinese characters. One promising approach is *A** search.