A STUDY ON THE USE OF GABOR FEATURES FOR CHINESE OCR

Qiang Huo, Zhi-Dan Feng and Yong Ge

Department of Computer Science and Information Systems,
The University of Hong Kong, Pokfulam Road, Hong Kong, China
(E-mail: qhuo@csis.hku.hk)

ABSTRACT

In this paper, we revisit the topic of Gabor feature extraction for Chinese OCR. We adopt a very simple discriminant function to construct a maximum discriminant function based character recognizer. We experiment with a simple way of forming a feature vector for each character image by extracting Gabor features using one wavelength at locations uniformly sampled with one spatial resolution. Extensive experiments on large vocabulary Chinese OCR for both machine-printed and handwritten characters are performed by using large amount of training and testing data to demonstrate the effectiveness of the Gabor features for Chinese OCR. Using Gabor features as raw features, we have constructed several state-of-the-art Chinese OCR engines.

1. INTRODUCTION

After several decades of research, many advances have been achieved in the area of OCR (optical character recognition) for world's main languages. Statistical pattern recognition approach (including artificial neural network) remains one of the most popular approaches being adopted to construct character recognizers in real products. When we started our research in Chinese OCR several years ago, we decided to try first a maximum discriminant function based approach (e.g., [6]) to construct our Chinese character classifier. For this approach, suppose there are M character classes $\{C_i\}_{i=1}^{M}$, each being modeled by K_i prototypes, $\lambda_i = \{m_{ik}\}_{k=1}^{K_i}$, where each prototype m_{ik} is a D dimensional vector. We use $\Lambda = \{\lambda_i\}_{i=1}^{M}$ to denote the set of prototype parameters. The aim of our character recognizer is to classify an input binary character image $f(x,y)$ ($x, y = 0, 1, \cdots, N-1$; $f(x,y)$ takes the value of either 1 or 0) as one of the M classes. This is done in two steps: a feature analysis step and a pattern classification step. In the feature analysis step, the input image $f(x,y)$ is analyzed and D raw features are measured to form a feature vector X. In the pattern classification step, the feature vector X is compared with each of the M character models and a discriminant function is computed for each class C_i as follows:

$$g_i(X;\Lambda) = -\min_k ||X - m_{ik}||^2 = -\min_k \sum_{d=1}^{D} (X_d - m_{ikd})^2 \quad (1)$$

The class that gives the maximum discriminant function is considered to be the recognized class, i.e.,

$$X \in C_i \quad \text{if } i = \arg\max_j g_j(X;\Lambda) \quad (2)$$

and the value of $-g_i(X;\Lambda)$ serves as the dissimilarity score for the hypothesized character C_i. The set of parameters Λ can be trained from a set of training samples. It is well known that feature extraction is extremely important for this approach to achieve high yet robust recognition performance.

2. WHY GABOR FEATURES

After a literature survey, we realized that there are many feature extraction methods have been proposed in the OCR area, and none of them has become dominant. For the newcomers like us to this field, we were facing difficulties of choosing which method to use in our OCR system. Our experience in automatic speech recognition (ASR) told us that the currently dominant speech features used in ASR, namely mel-frequency cepstral coefficients (MFCC), mimic certain characteristics of human speech perception capability. This fact drives us to looking for the possible visual perception motivated features in image processing and computer vision areas that can also be used for character recognition. We found delightedly that there exist in literature such kind of features, namely Gabor features.

Marcelja [16] and Daugman [1] discovered that simple cells in the visual cortex can be modeled by Gabor functions [8]. The 2D Gabor functions proposed by Daugman are local spatial bandpass filters that achieve the theoretical limit for conjoint resolution of information in the 2D spatial and 2D Fourier domains [2]. New insights are also

This work was supported by a grant from the RGC of the Hong Kong SAR (Project No. HKU7020/98E) and two internal HKU CRCG grants.

gained in some recent studies by deriving 2D Gabor functions based on different criteria (e.g., [15, 17]). Families of self-similar 2D Gabor wavelets have been proposed and adopted for image analysis, representation, and compression (e.g., [3, 15]). Gabor filters have also been used extensively in various computer vision applications such as texture analysis, texture segmentation and classification, edge detection, etc. Furthermore, features extracted by using Gabor filters (we call them Gabor features) have been successfully applied to many pattern recognition applications such as face recognition (e.g., [14]), Iris pattern recognition (e.g., [4]), fingerprint recognition (e.g., [12]), and character recognition (e.g., [7, 18, 5, 20, 9, 10]). It seems that Gabor features are good candidates for visual perception related pattern recognition applications. Although the effectiveness of Gabor features has been demonstrated in several OCR studies such as the recognition of both machine-printed and handwritten alphanumeric characters (e.g., [7, 18, 10]), and small to medium vocabulary Chinese characters (e.g., [5, 20, 9], it is interesting to notice that in OCR area Gabor features have not become as popular as they have in face and Iris pattern recognition areas (e.g., [14, 4]). This situation is difficult for the new comers like us to understand, especially considering the following facts: 1) Gabor features are well-motivated and mathematically well-defined, 2) they are easy to understand, fine-tune and implement, 3) they have also been found less sensitive to noises, small range of translation, rotation, and scaling (e.g., [20, 9]). We thus decided to perform our own study on the use of Gabor features for the Chinese OCR.

The main purpose of this paper is to report in detail the experimental results we obtained for the *large vocabulary* recognition tasks of both machine-printed and handwritten Chinese characters with the above mentioned character classifier by using *large amount* of training and testing data. In the following, we first describe the specific Gabor feature extraction method we used in our experiments. Then, we describe in detail the character recognition tasks, the experimental setup and results.

3. USING GABOR FEATURES AS RAW FEATURES

Gabor features have been used in different ways to OCR. We adopt the following complex 2-D Gabor filter originally reported in [14] for face recognition:

$$G(x,y;\kappa,\vartheta_k) = G_1(x,y)[\cos(R) - \exp(-\frac{\sigma^2}{2})] + iG_1(x,y)\sin(R) \quad (3)$$

where $G_1(x,y) = \frac{\kappa^2}{\sigma^2}\exp[-\frac{\kappa^2(x^2+y^2)}{2\sigma^2}]$ with $\sigma = \pi$, $R = \kappa x\cos\vartheta_k + \kappa y\sin\vartheta_k$, $\kappa = \frac{2\pi}{\iota}$, and $\vartheta_k = \frac{\pi k}{\mathcal{M}}$ with $k = 0, 1, 2, ..., \mathcal{M}-1$. The parameters ι and ϑ_k are the wavelength and orientation of the above plane wave respectively.

Given a binary character image $f(x,y)$, at a sampling point (x_0, y_0), $\mathcal{M}$ Gabor features can be derived as the magnitudes of the $\mathcal{M}$ Gabor filter outputs as follows:

$$f_{\iota,k}(x_0,y_0) = |\sum_{x=-x_0}^{N-x_0-1}\sum_{y=-y_0}^{N-y_0-1} f(x_0+x, y_0+y)G(x,y;\kappa,\vartheta_k)|, \quad (4)$$

where $k = 0, 1, 2, ..., \mathcal{M}-1$. Consequently, for a given wavelength ι and by uniformly choosing $N_1 \times N_1$ spatial sampling points of (x_0, y_0)'s, we can derive $D = N_1 \times N_1 \times \mathcal{M}$ Gabor features from a given image $f(x,y)$ and use them to form the raw feature vector X (e.g. [10]). The optimal values of controlling parameters ι, N_1, and $\mathcal{M}$ are task- and data-dependent, thus can only be determined by experiments.

4. TASKS AND CHINESE CHARACTER CORPORA

We performed our experiments on two tasks. The first task is the recognition of machine-printed Chinese characters with a vocabulary of 6921 characters which include 6707 meaningful simplified Chinese characters derived by removing 56 meaningless single-radical characters from 6763 characters defined in GB2312-80 standard, 12 frequently used GBK Chinese characters, 62 alphanumeric characters, 140 punctuation marks and symbols. Over the years, a machine-printed Chinese character corpus has been constructed in our lab with in total $3,024,043$ character image samples from the above 6921 character classes. The original documents are from varied sources, such as newspapers, magazines, journals, books, printed lists of characters generated from many popular font libraries available on the market. The document quality and font sizes vary widely among the various sources. Many font styles including Song, Fang Song, Kai, He, Yuan, LiShu, WeiBei, XingKai, etc. are observed in our character corpus. Depending on the font size of the documents, the document pages have been digitized at a resolution ranging from 300 DPI to 500 DPI on several flatbed scanners. In our database, each character sample is stored as a normalized $N \times N$ ($N = 40$ here) binary image. We have chosen randomly about 20% of character samples for each character class to form a testing set and the remaining samples to form a training set. By this partition, there are 2,412,898 character samples in the training set and 611,145 character samples in the testing set. By applying the above feature extraction method to each character image in our training data set, we can derive a training set of feature vectors $\mathcal{X}^p = \{\mathcal{X}_i^p\}_{i=1}^M$ where $\mathcal{X}_i^p = \{X_{ij}^p\}$ represents the set of training samples for class C_i with X_{ij}^p being the j-th training sample.

The second task is the recognition of handwritten Chinese characters with a vocabulary of $4,616$ characters. This

vocabulary covers 4516 frequently used Chinese characters that have an internal code specified in GB 2312-80, 62 alphanumeric characters, 38 punctuation marks and symbols. More than five hundred writers were employed to write each of the characters neatly on grid-papers with 400 squares ($0.8cm \times 0.8cm$ for each) per page. These character pages were scanned and digitized at a resolution of 300 DPI on several flatbed scanners. Consequently, a handwritten Chinese character corpus was constructed with in total $2,410,335$ character image samples from the above $4,616$ character classes, each having more than 500 samples. We have chosen randomly 300 samples per class to form a training set and the remaining samples to form a testing set. By this partition, there are $1,384,800$ character samples in the training set and $1,025,535$ character samples in the testing set. In the following experiments, in both training and testing, each character sample is normalized into a $N \times N$ ($N = 40$ here) binary image by using a nonlinear shape normalization method based on line density by line interval as described in [19]. By applying the above feature extraction method to each normalized character image in our training data set, we can derive a training set of feature vectors $\mathcal{X}^h = \{\mathcal{X}_i^h\}_{i=1}^M$ where $\mathcal{X}_i^h = \{X_{ij}^h\}$ represents the set of training samples for class C_i with X_{ij}^h being the j-th training sample.

5. EXPERIMENTS AND RESULTS

In the current way of using Gabor features to form a feature vector for each character image, we only use one set of features derived by using one wavelength at one spatial sampling resolution. As we discussed above, the Gabor feature extraction procedure requires the setting of some control parameters which include the number of orientation $\mathcal{M}$, the wavelength ι, and the spatial sampling resolution parameter N_1. Several sets of experiments for the above two recognition tasks have been performed 1) to discern the effect of those control parameters on character recognition accuracy and 2) to check what the best set of control parameters should be. In all of the experiments, we use a K-means clustering method to estimate the set of prototype parameters $\{m_{ik}\}$ for each character class C_i which has 4 prototypes (i.e. $K_i = 4$). We also use 4 orientations (i.e. $\mathcal{M} = 4$) for Gabor feature extraction because we observed in a set of preliminary experiments that the result of using 8 orientations is only slightly better than that of 4 orientations.

Table 1 and Table 2 summarizes the open-test results in terms of recognition accuracies in % as a function of different wavelengths and spatial sampling resolutions for machine-printed and handwritten Chinese character recognition tasks respectively. For each fixed spatial sampling resolution, there exists an optimal wavelength which achieves the best performance. We observed that the wavelength $\iota = 8$ achieves consistently the best performance in machine-printed character recognition experiments, while the wavelength $\iota = 4\sqrt{2}$ achieves the best performance in handwritten character recognition experiments. As for the spatial sampling resolutions, it seems that 7×7 sampling is enough for machine-printed OCR while 8×8 is needed for handwritten OCR. Consequently, two good sets of control parameters are identified for Gabor feature extraction, i.e.

- $N_1 = 7, \mathcal{M} = 4, \iota = 8$ for machine-printed OCR,
- $N_1 = 8, \mathcal{M} = 4, \iota = 4\sqrt{2}$, for handwritten OCR.

They are used in constructing our Chinese OCR engines.

Table 1. A summary of open-test results (recognition accuracies in %) for machine-printed Chinese characters using different wavelengths and spatial sampling resolutions.

Spatial Sampling Resolution	Wavelength			
	$4\sqrt{2}$	8	$8\sqrt{2}$	16
5×5	96.85	99.08	95.26	88.11
6×6	97.09	99.17	95.57	84.42
7×7	97.37	99.24	96.18	81.94
8×8	97.35	99.23	96.48	81.68

Table 2. A summary of open-test results (recognition accuracies in %) for handwritten Chinese characters using different wavelengths and spatial sampling resolutions.

Spatial Sampling Resolution	Wavelength			
	4	$4\sqrt{2}$	8	$8\sqrt{2}$
6×6	73.28	85.28	84.36	68.35
7×7	79.65	87.73	85.45	68.72
8×8	83.19	88.90	85.07	66.55
10×10	86.53	89.40	85.01	66.63

For example, we have used the above setup to derive Gabor features to serve as the raw features extracted from a character image. By further 1) using the linear discriminant analysis (LDA) [6] for discriminative feature extraction and dimension reduction, and 2) using minimum classification error (MCE) as a criterion for model parameters training, we can construct a compact character recognizer which acieves a recognition accuracy of 99.64% on the testing set for machine-printed Chinese characters described above. More technical details are reported in [11].

In order to compare the effectiveness of the Gabor features with that of the other features used in OCR area, we choose to implement a currently popular feature extraction method in Chinese OCR, namely directional element features (DEF) as described in detail in [13]. The way of extracting DEFs and forming a feature vector for each character image is similar to that of Gabor features described above. At each spatial sampling point (x_0, y_0), 4 directional features are extracted. So, under the same spatial sampling

Table 3. A comparison of open-test results (recognition accuracies in %) for handwritten Chinese characters by using Gabor features and directional element features (DEF).

Spatial Sampling Resolution	Raw Features	
	Gabor	DEF
7×7	87.73	83.43
8×8	88.90	85.92

resolution, the Gabor feature vector and the DEF vector have the same dimension thus make them directly comparable. Table 3 compares the open-test results of using DEFs under two spatial sampling resolutions with that of using Gabor features for handwritten Chinese character recognition. Our results show that Gabor features perform consistently better than DEFs by using the simple maximum discriminant function based recognition strategy as described in the introduction section.

6. DISCUSSION AND CONCLUSION

In this paper, we have revisited the topic of Gabor feature extraction for Chinese OCR. As a first step, we adopted a very simple discriminant function to construct a maximum discriminant function based character recognizer. We experimented with a simple way to form a feature vector for each character image by extracting Gabor features using one wavelength at locations uniformly sampled with one spatial resolution. Extensive experiments on large vocabulary Chinese OCR for both machine-printed and handwritten characters are performed by using large amount of training and testing data. The effectiveness of the Gabor features are clearly demonstrated for Chinese OCR. Using Gabor features as raw features, we have constructed several state-of-the-art Chinese OCR engines. We've also been investigating other ways of using Gabor features in character modeling with some more complicated statistical modeling techniques such as hidden Markov model, 2D contextual stochastic model, etc. We will report those results elsewhere. To conclude, we want to join other researchers, who have tried Gabor features in OCR in the past, to recommend to the OCR research community the use of Gabor features in constructing OCR systems.

7. REFERENCES

[1] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vision Research*, Vol.20, pp.847-856, 1980.

[2] J. G. Daugman, "Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters," *J. Opt. Soc. Am. A*, Vol.2, No.7, pp.1160-1169, 1985.

[3] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," *IEEE Trans. on ASSP*, Vol.36, No.7, pp.1169-1179, 1988.

[4] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Trans. on PAMI*, Vol.15, No.11, pp.1148-1161, 1993.

[5] D. Deng et al., "Handwritten Chinese character recognition using spatial Gabor filters and self-organizing feature MAPs," *Proc. ICIP-1994*, 1994, pp.940-944.

[6] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, 1973.

[7] M. D. Garris et al., "Analysis of a biologically motivated neural network for character recognition," *Proc. Conf. Analysis of Neural Network Applications*, 1991, pp.160-175.

[8] D. Gabor, "Theory of communication," *J. IEE*, Vol.93, pp.429-459, 1946.

[9] Y. Hamamoto et al., "Recognition of handprinted Chinese characters using Gabor features," *Proc. ICDAR-95*, 1995, pp.II-819-823.

[10] Y. Hamamoto et al., "A Gabor filter-based method for recognizing handwritten numerals," *Pattern Recognition*, Vol.31, No.4, pp.395-400, 1998.

[11] Q. Huo, Y. Ge, Z.-D. Feng, "High performance Chinese OCR based on Gabor features, discriminative feature extraction and model training," *Proc. ICASSP-2001*, May 2001.

[12] A. K. Jain et al., "A multichannel approach to fingerprint classification," *IEEE Trans. on PAMI*, Vol.21, No.4, pp.348-359, 1999.

[13] N. Kato et al., "A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance," *IEEE Trans. PAMI*, Vol.21, No.3, pp.258-262, 1999.

[14] M. Lades et al., "Distortion invariant object recognition in the dynamic link architecture," *IEEE Trans. on Computer*, Vol. 42, No. 3, pp.300-311, 1993.

[15] T.-S. Lee, "Image representation using 2D Gabor wavelets," *IEEE Trans. on PAMI*, Vol.18, No.10, pp.959-971, 1996.

[16] S. Marcelja, "Mathematical description of the responses of simple cortical cells," *J. Opt. Soc. Am.* , Vol.70, pp.1297-1300, 1980.

[17] K. Okajima, "Two-dimensional Gabor-type receptive field as derived by mutual information maximization," *Neural Networks*, Vol.11, pp.441-447, 1998.

[18] A. Shustorovich, "A subspace projection approach to feature extraction: the two-dimensional Gabor transform for character recognition," *Neural Networks*, Vol.7, No.8, pp.1295-1301, 1994.

[19] J. Tsukumo88, H. Tanaka, "Classification of handprinted Chinese characters using nonlinear normalization methods, *Proc. ICPR*, Rome, Itly, 1988, pp.168-171.

[20] K. Yamada, "Optimal sampling intervals for Gabor features and printed Japanese character recognition," *Proc. ICDAR-95*, 1995, pp.I-150-153.