Gabor filters for Document analysis in Indian Bilingual Documents

Peeta Basa Pati S Sabari Raju Nishikanta Pati A G Ramakrishnan
Department of Electrical Engineering
Indian Institute of Science
Bangalore 560 012, INDIA

Abstract

Reasonable success has been achieved at developing mono lingual OCR systems in Indian scripts. Scientists, optimistically, have started to look beyond. Development of bilingual OCR systems and OCR systems with capability to identify the text areas are some of the pointers to future activities in Indian scenario. The separation of text and non-text regions before considering the document image for OCR is an important task.

In this paper, we present a biologically inspired, multichannel filtering scheme for page layout analysis. The same scheme has been used for script recognition as well. Parameter tuning is mostly done heuristically. It has also been seen to be computationally viable for commercial OCR system development.

INTRODUCTION

Discriminating between the text and non-text regions of a document image is a complex and challenging task. It has kept the scientists, working in this field, puzzled for quite some time now [1, 2]. The non-text regions of a document image could consist of graphs, natural images and other kinds of sketches drawn with lines. In such a case, the task of isolating the text regions prior to going for character recognition is a must. In the case of bilingual documents, there is also a need for identifying the script of the text before starting character recognition. This simplifies segmentation of characters in the case of certain scripts and also reduces the search space in the database. Humans perform both the above mentioned tasks with amazing efficiency. There have been attempts to find out the way human brain achieves it. This process could be mimicked in system automation.

Biological visual systems are said to be involving a set of parallel quasi-independent filtering mechanisms at the cortical level [3, 4]. This supports the theory of multi-channel filtering of signals in systems that model biological vision [5]. Such a filter scheme employs a filter-bank with each filter modeling a single channel. This modeling job is best done by Gabor functions [6, 7]. Some researchers have already employed this scheme for texture segmentation [8, 9] while others have tried to use it for text page segmentation [10].

In this paper, we have taken a Gabor function based multichannel directional filtering approach for both text area separation and script identification at the word level.

MOTIVATION

Owing to the diversity of languages and scripts, English has proven to be the binding language in India. So bilingual documents (involving a regional language and English) are quite prevalent at present. For practical OCR systems to work in such a situation, bilingual OCR system is required.

Generally, there are two schools of approaches taken for such a task. One school believes in identification of the script of the text before taking the characters for recognition. This helps in reduced search in the database at the cost of the script recognition task. The other school has taken a combined database approach. In this case, the database of reference characters have the alphabets from the local script as well as the Roman script. Here the database is larger as it includes characters from both the scripts. Considering the case for a bilingual OCR system for Tamil script, there are 137 symbols used from Tamil script and 72 symbols come from Roman script. If a combined database approach is taken then for recognition of each symbol, the search space increases by about 53% for each Tamil character and about 190% for each Roman character. Besides, there is a danger of different characters of a given word getting classified to different scripts. Such an unwanted scene, however, is avoided by employing some post-processing scheme.

Some amount of work has already been reported for the recognition of the script. In the work reported by Chaudhuri et. al. [11], the authors have assumed uniscript text lines in multi-lingual documents. Though such an assumption is valid in some limited situations, in a general case we need to recognize the script at word level. Such an approach has been reported by Dhanya et. al. [12], which serves as the basis for the reported work.

Besides, most researchers have assumed that the documents are text only. But a general class of documents contain images, graphs, tables and sketches along with text. To the best of our knowledge, very little work has been reported on OCR system which tackle this issue, in Indian context. In the work presented by Chaudhuri et. al. [13], the authors assume the presence of only natural images in the text documents. They take a connected component analysis based method, in binarised document images, to separate out text regions from the rest. In the method reported by us, we assume that the text regions of a document image are predominantly high frequency regions. Hence we have taken a filter-bank approach to discriminate text areas from non-text areas.

0-7803-8243-9/04/$17.00 © 2004 IEEE

GABOR FILTER BANKS

A bank of Gabor filters are chosen for the tasks under consideration, namely, script recognition and text separation. This is because of the inherent advantages offered by Gabor function. They are: (i) it is the only function for which the lower bound of space bandwidth product is achieved, (ii) the shapes of Gabor filters resemble the receptive field profiles of the simple cells in the visual pathway, and (iii) they are direction specific band-pass filters. Gabor [14] introduced the function in 1946 in one-D case which was later extended to 2D case by Daugman [15].

A Gabor function is a Gaussian modulated by a complex sinusoid. In a 2D case, we put it mathematically,

$$h(x,y) \; = \; g(x',y') \; exp[j2\pi \, Ux] \tag{1}$$

where x' and y' are the x and y co-ordinates in the rotated rectangular co-ordinate system. U is the radial frequency in cycles/image width.

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{2}$$

$$g(x,y) = \frac{1}{(2\pi\sigma_x\sigma_y)} exp\left(-\frac{1}{2}\left[\left(\frac{x}{\sigma_x}\right)^2 + \left(\frac{y}{\sigma_y}\right)^2\right]\right) \tag{3}$$

σ_x and σ_y explain the spatial spread and the bandwidth of the filter function $h(x,y)$. If B_r is the radial frequency bandwidth in octaves and B_θ is the angular bandwidth in degrees, then,

$$\begin{aligned} \sigma_x &= \frac{\sqrt{2}}{2\pi U}\frac{2^{B_r}+1}{2^{B_r}-1} \\ \sigma_y &= \frac{\sqrt{2}}{2\pi U \tan\left(\frac{B_\theta}{2}\right)} \end{aligned} \tag{4}$$

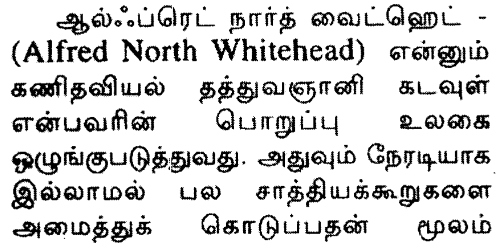
The power of discrimination of the different filters are dependent on the values of B_r and B_θ. The aspect ratio of $g(x,y)$, defined as:

$$\lambda \; = \; \frac{\sigma_y}{\sigma_x} \tag{5}$$

explains the symmetry measure of the filter.

Any combination of B_r, B_θ and U, involves two filters, one corresponding to the sine function and the other corresponding to the cosine function in the exponential term in Eqn.1. The cosine filter, also known as the real part of the filter function, is a even-symmetric filter and acts more like a low pass filter, while the sine part being odd-symmetric acts like a high pass filter.

Gabor filters of $B_r = 1$ octave and $B_\theta = 45^o$ at four different orientations ($\theta = 0^o$, 45^o, 90^o, 135^o) have been used for the reported work. The radial frequency U has been chosen, on a different-scale for text separation and script identification tasks, to suit the needs.



ஆல்ஃப்ரெட் நார்த் வைட்ஹெட் -
(Alfred North Whitehead) என்னும்
கணிதவியல் தத்துவஞானி கடவுள்
என்பவரின் பொறுப்பு உலகை
ஒழுங்குபடுத்துவது. அதுவும் நேரடியாக
இல்லாமல் பல சாத்தியக்கூறுகளை
அமைத்துக் கொடுப்பதன் மூலம்

Figure 1. An example bilingual document image.

DATA DESCRIPTION

Document images are scanned at 300 dpi by (a) Umax Astra 5400, and (b) HP Scanjet 2200c scanners. The pages are scanned in black and white photograph (gray level) format. The documents with only natural images are considered for page layout analysis. Here it has been assumed that the font sizes of the text portions vary between 10 to 18. The font size we have mentioned are as per definitions of Microsoft Inc. and used in MSWord.

All the documents considered, contained the local language as the major language (refer Fig. 1). English is the other language considered here. A set of about 1000 words each are collected from Tamil, Hindi, Odiya and Roman script from the above mentioned documents. 200 words in each script out of the collected 1000 words are used for the purpose of training, i.e., for the calculation of the linear discriminant function (LDF), X. Once the LDF is calculated, all the words available in the database, associated with the script, including the ones used for training, are used for testing the recognition accuracy. The words are available in binary image format and no normalization has been done on them.

TEXT AREA SEGMENTATION

The goal of text separation is to partition an image into text and non-text regions and to identify the region boundaries. Here we have assumed that natural images are present in the document along with text. We have taken a texture based segmentation algorithm motivated by the multi-channel filtering approach of human visual system (HVS). A space-frequency filter bank, using Gabor filters, has been designed and implemented for the purpose.

The filter response of a complex filter at any pixel (i,j) is:

$$E(u_l, \, \theta_k) \; = \; \sqrt{e(u_l,\theta_k)^2 + o(u_l,\theta_k)^2} \tag{6}$$

where $e(u_l,\theta_k)$ and $o(u_l,\theta_k)$ are the cosine(even) and sine(odd) filter outputs at that pixel location, respectively. The total local energy is the sum total of the outputs of all the complex filters at that given pixel.

$$E_T \; = \; \sum_{l=1}^{3}\sum_{k=1}^{4} E(u_l, \, \theta_k) \tag{7}$$

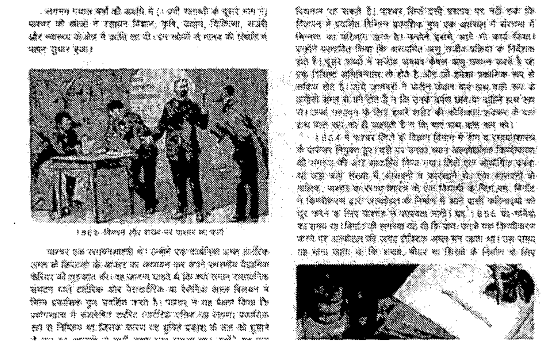


Figure 2. Document for Page Layout Analysis.

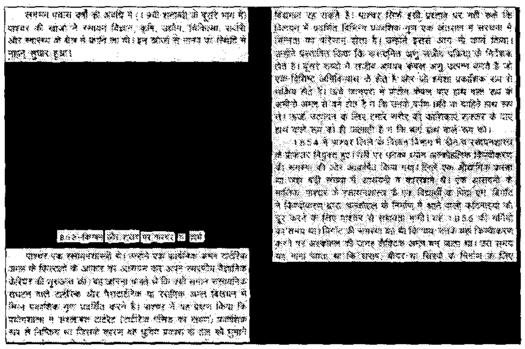


Figure 3. Result of Page Layout Analysis.

We low pass filter the magnitude response image of the Gabor filterbank (the E_T image) with a Gaussian filter. It is seen that this removes artifacts and irregularities present and, thereby, helps in text detection.

We evaluate block energy at each pixel by considering blocks of 15 × 15 size. An average of the block energies is evaluated across all the blocks in the image. This average energy is multiplied by a scale factor of 2 (decided heuristically) to set the threshold for text and non-text separation block wise, i.e., if the block energy is found to be above the threshold value, thus calculated, the block is considered to be belonging to text area.

Figure 2 shows a sample document image containing 2 images and Hindi text. The text separated regions, obtained by implementing the above mentioned algorithm, are shown in Fig. 3.

SCRIPT RECOGNITION

The recognition of the script at the word level has a significant effect on the process of OCR. We take a multi-channel filtering approach based on Gabor function for this task. We consider 4 different radial frequencies and 4 different angles of orientation ($\theta = 0^o, 45^o, 90^o\, and\, 135^o$). The four radial frequencies with four θ's give a combination of 16 odd and 16 even filters. A given word image is filtered with these 32 filters and the coefficients make a 32 dimensional feature vector v. Taking a set of sample words from each of the scripts (local script and Roman script), we decide a LDF, X, which best discriminates the scripts in the feature space.

The training feature space is represented by matrix A. If we have N samples each from the regional script and Roman, referred to as classes ω_1 and ω_2, respectively, then the size of A is $2N \times 32$. Each row of A is a feature vector v.

$$\begin{aligned} AX &= B \;\; or, \\ X &= A^{-1}B \end{aligned} \tag{8}$$

where B is a column vector of class values ($\omega_1 = 1 and \omega_2 = -1$). Since, A, in general, is a rectangular matrix, we compute the pseudo-inverse of the matrix A which is of the size $32 \times 2N$ and it combines with the class vector B of size $2N \times 1$ to give the vector X which is of size 32×1. This X vector represents a discriminating hyperplane in the 32-dimensional vector space which divides the space into class regions. All the patterns of a given class (words of any given script in our case) lie on one side of this plane. The class value of a new pattern (a test pattern) is decided based on which side of the plane it lies.

The above principle was implemented for script identification of the words taken from bilingual documents. Tamil, Hindi and Odiya were the languages in which the bilingual documents were considered. In all the above cases, English was the other language. The average time taken to recognize the script of a given word is 175ms on a Pentium-III based machine running at 1.4GHz. The results of such a recognition scheme has been presented in the table below:

Table 1. Script Recognition Performance of Gabor filter based technique on bilingual documents

Indian Script	Script Recognition Accuracy (both scripts)
Hindi	99.56
Tamil	96.02
Odiya	97.1

CONCLUSION AND DISCUSSION

In this paper, we have presented a technique which works for page layout analysis and for script recognition, at the word level, in bilingual text documents.

In a document, consisting of natural images and text portions, given a block of text area and another block of image area, it is natural that the text area contains more high frequency components. This is the property which we have exploited while taking this filter bank approach. The filters are designed to take care of this property. A clear discrimination is achieved between text and non-text areas based on the amount of energy present in the high frequency band.

In the work presented by Chaudhuri et. al. [13], the authors have taken a connected component analysis based approach. They also assume the presence of a top horizontal line called shirorekha or headline in the characters which join the characters in a word. Such an assumption, based on features of North Indian scripts, would prevent their method in being applied to many other Indian scripts due to the absence of this headline. However, our method being devoid of any such script dependent feature makes a more robust page analyser. In cases, where the natural image contains lot of high frequency components, it has been observed that the proposed scheme fails. This also fails if a hand drawn sketch is present in place of a natural image.

For script recognition, we have considered the total energy associated with the filter in the image, i.e., the amount of energy present in the image in the frequency band of the filter. Here we have exploited the property of the scripts considered. We have seen that symbols in Indian scripts are more curved than those in the Roman script. This led to the proposition of directional energy filtering approach and hence the Gabor filter bank. The results shown above seem to support our hypothesis. More importantly, since the Gabor function has been shown to be closely mimicking the cortical cells of HVS, it is probably the way humans also interpret images and texts. Since there is no work reported for script recognition at word level, to the best of our knowledge, the results of this work could not be compared.

Though we have assumed that the documents contain only natural images, in a broader scenario, we have to consider all kinds of clutter which includes graphs and sketches. Experimentation is underway to find the feasibility of the technique for separation of the areas which contain graphs, hand drawn sketches etc.

REFERENCES

[1] Y. Y. Tang, S. W. Lee, and C. Y. Suen, "Automatic document processing: A survey," Pattern Recognition, vol. 29, no. 12, pp. 1931–1952, 1996.

[2] L. O. Gorman and R. Kasturi, Document Image Analysis. Las Almitos, CA: IEEE Computer Society Press, 1994.

[3] F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," J. Physiol., vol. 197, pp. 551–556, 1968.

[4] R. L. De Valois, D. G. Albrecht, and L. G. Thorell, "Spatial-frequency selectivity of cells in macaque visual cortex," Vision Research, vol. 22, pp. 545–559, 1982.

[5] M. Porat and Y. Y. Zeevi, "The generalized gabor scheme of image representation in biological and machine vision," IEEE transaction on Pattern Analysis and Machine Intelligence, vol. 10, no. 4, pp. 452–467, 1988.

[6] S. Marcelja, "Mathematical description of the responses of simple cortical cells," J. Opt. Soc. Am., vol. 70, pp. 1297–1300, 1980.

[7] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," Vision Research, vol. 20, pp. 847–856, 198.

[8] A. K. Jain and F. Farrokhnia, "Unsupervised texture segmentation using Gabor filters," Pattern Recognition, vol. 24, no. 12, pp. 1167–1186, 1991.

[9] D. Dunn, W. E. Higgins, and J. Wakeley, "Texture segmentation using 2-D Gabor elementary functions," IEEE transaction on Pattern Analysis and Machine Intelligence, vol. 16, no. 2, pp. 130–149, 1994.

[10] W. Chan and G. Coghill, "Text analysis using local energy," Pattern Recognition, vol. 34, pp. 2523–2532, 2001.

[11] U. Pal and B. B. Chaudhuri, "Script line separation from Indian muli-script document," in Proceedings of the International Conference on Document Analysis and Recognition, pp. 406–409, 1999.

[12] D. Dhanya, A. G. Ramakrishnan, and P. B. Pati, "Script identification in printed bilingual docuements," Sadhana, vol. 27, pp. 73–82, 2002.

[13] A. R. Chaudhuri, A. K. Mandal, and B. B. Chaudhuri, "Page layout analyser for multilingual indian documents," in Proceedings of the Language Engineering Conference, pp. 24–32, 2002.

[14] D. Gabor, "Theory of communication," J. IEE (London), vol. 93, pp. 429–457, 1946.

[15] J. 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.