A Neural Network Based Color Document Segmentation

Hsiao-Yu Han
Industrial Technology Research Institute
Hsinchu, Taiwan

Abstract

Document segmentation is defined as distinguishing different parts of the document image based on contents. In this paper, the document image is segmented into texts, pictures, and background. The algorithm we proposed includes background removal, block segmentation, feature extraction, and recognition. In background removal, we use local thresholds to extract foreground of the image. In block segmentation, run-length smoothing algorithm and connected component analysis are applied to divide the document image into a set of regions. And then, the features including image features and geometry features from the regions are extracted. Finally, these features are fed into the classifier which is a three-layer back-propagation neural network. The output of the neural network is the result of the recognition: texts or pictures. Through the experiments, we know that most document images with simple backgrounds can be segmented well by the method we proposed. Therefore, there are several advantages in our document segmentation system. 1. Localized thresholds to distinguish foreground from background based on color concepts. 2. Able to discriminate texts from pictures by extraction of good features. 3. Use a trainable neural network as the classifier where the structure can be adjusted flexibly. 4. Precise segmentation since the classifier is trained by mass of document images.

Introduction

Document image analysis (DIA) is the theory and practice of recovering the symbol structure of digital images scanned from paper or produced by computers. DIA has been developed since optical character recognition was used in business about 40 years ago. Nowadays, document imaging is a billion-dollar business.

In this paper, an algorithm is developed for document segmentation which is an important issue in DIA and is the first step for many image applications. Generally, document segmentation is to segment a document into several classes and extract the regions of interested. A critical issue in document segmentation is the extraction of good features. These features must be capable of distinguishing different classes in document images. Relations between neighboring pixels are one of the mostly used features. In [1] and [2], vertical cross-correlation between neighboring pixels is applied to segment images into texts and pictures. In [2] and [3], run-length statistics are important for distinguishing texts and pictures. Transients that are defined as variations between neighboring pixels are the other common used features.[3-5] Geometry features, including the size of a block,[6] ratio of width to height of a block,[3,5] are the other types of useful traits. In [7], features consist of wavelet coefficients that are obtained by wavelet transforms.

The other important issue in document segmentation is the classifier. Most researchers use rule-based classifier where classification rules are predetermined and applied for the selected classes.[2,3,5,7-9] The extracted features are compared with the rules to determine what kinds of classes the blocks belong to. In [4], the features are operated in the fuzzy classifier where different fuzzy rules derived from feature analysis and membership management are used, and the result of the defuzzification is the threshold values for each pixel. In [1], a neural network is used to be the classifier. Features are extracted for the input of the trained neural network, and output of the network is the classification result: text and picture.

Texture analysis is applied for document segmentation in recent years,[10,11] where different classes are turned into different kinds of textures. Document images are decomposed into feature images using a bank of filters, including Gabor filters, Laplacians of Gaussians, IIR filters and etc. Different filters are assumed to capture some specific local characteristics of the input texture, such as spatial frequency, directionality, edgeness, etc.

A document segmentation method is proposed in this paper. Color information is considered to remove the background of the document images. In addition to the geometry feature and the image information, a unique feature related to histogram analysis is computed. A neural network based classifier is used to segment the document images into text and picture.

Algorithm

The flow chart of our proposed algorithm for document segmentation is presented in figure 1. There are four processes in this method: background removal, block segmentation, feature extraction, and recognition. In background removal, adaptive threshold of local region is

computed to separate foreground and background. A binary image in which foreground pixels are denoted in black and background pixels are denoted in white is obtained after background removal. Run length smoothing algorithm and connected component analysis are used in the procedure of block segmentation to divide the binary image into a set of regions. Feature vectors are extracted from each region in the third step of the procedure for document segmentation that we called feature extraction. At the last step for recognition, feature vectors of each region are inputted to a classifier based on a back-propagation neural network, output of which is the class of each region.

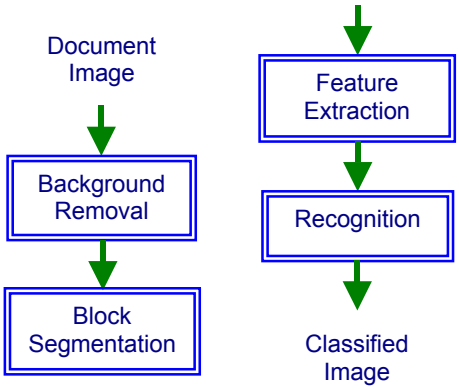


Figure 1. Algorithm for document segmentation

Background Removal

Background removal is the first and quite important step for document segmentation. The purpose of background removal obviously is to remove the background of the document image that is usually defined as smooth color or blank region. After removing the background of the image, the foreground that contains texts and pictures is obtained. The result of the segmentation will be poor with the uncompleted foreground which means not all the background is removed or some foreground is removed in background removal.

Adaptive local thresholding is used to remove the background of the image in our method. Thresholds that computed by averaging pixel values of the local region is a variable depending on the region. If the pixel value is greater than the threshold, the pixel value is set to 1 that means foreground pixel. If the pixel value is less than the threshold, the pixel value is set to 0 that means background pixel.

Color information of the document image is neglected since only the gray thresholds of the image are computed. In order to achieve better results, the color information must be considered. The adaptive local thresholding is also used on R, G and B color planes. The process of the pixels in the image for background removal is showed as equation (1).

$$
\begin{aligned}
P1 &= \begin{cases} 1, \text{if } P_{Gray} > \text{Threshold}_{Gray} \\ 0, \text{otherwise} \end{cases} \\
P2 &= \begin{cases} 1, \text{if } P_{R} > \text{Threshold}_{Red} \\ 0, \text{otherwise} \end{cases} \\
P3 &= \begin{cases} 1, \text{if } P_{G} > \text{Threshold}_{Green} \\ 0, \text{otherwise} \end{cases} \\
P4 &= \begin{cases} 1, \text{if } P_{B} > \text{Threshold}_{Blue} \\ 0, \text{otherwise} \end{cases} \\
P &= P1 \text{ OR } P2 \text{ OR } P3 \text{ OR } P4
\end{aligned}
\tag{1}
$$

Block Segmentation

Block segmentation decomposes a binary document image into a set of regions that are rectangular blocks. First, related foreground pixels are combined into the same region by the run-length smoothing algorithm (RLSA). Let the foreground pixels are denoted by 1, and the background pixels are denoted by 0 in the binary image. A sequence of 0 or 1 is called a run. Two runs of 1 will be merged if and only if there is not a sufficient space between the runs. The run-length smoothing algorithm consists of the following steps:

1. Set a constant threshold C.
2. If the run length of 0 in horizontal direction of the binary image is greater than C, these 0s remain unchanged.
3. If the run length of 0 in horizontal direction of the binary image is less than C, these 0s are all changed to 1s.

For example, consider the following string

10010000011111

Constant threshold C is set to 4. The first run of 0 is 2 so the two 0s are converted to 1; on the other hand, the second run of 0 is 5 so the five 0s remain unchanged. As a result, the original string become

11110000011111

The second step is to locate each foreground region in a smoothing binary image obtained from the run-length smoothing algorithm. The problem of locating each region can be treated as finding connected components. In the image, each component is denoted by a unique label, and all the pixels belonged to the same component are labeled with the same label.

Two passes in which the image is scanned in raster order are required in traditional connected component analysis. In the first pass, the pixels are labeled with temporary labels and equivalent table is built. The current pixel under scan is labeled by considering the labels of neighboring pixels that are assigned previously. If the neighboring pixels have the same label, the current pixel is labeled with the same label. If the neighboring pixels have different labels, the current pixel is labeled with the oldest

label. If all the neighboring pixels have never been assigned with labels, a new label is assigned to the current pixel. In the second pass, the temporary labels are replaced with the representative labels of the equivalent class.

Feature Extraction

After block segmentation, the pixels belonged to the same object are clustered together, and the objects in the foreground are distinguished from each other. In feature extraction, the suitable features of each block in the image are obtained.

The suitable features means the features extracted from texts block are obviously different from those extracted from picture block.

Three kinds of features are adopted in our method. First, the geometry feature, the aspect ratio of the block, is computed by equation (2).

$$\text{Aspect Ratio} = \frac{\Delta x}{\Delta y} \tag{2}$$

where Δx is the width of the block, Δy is the height of the block.

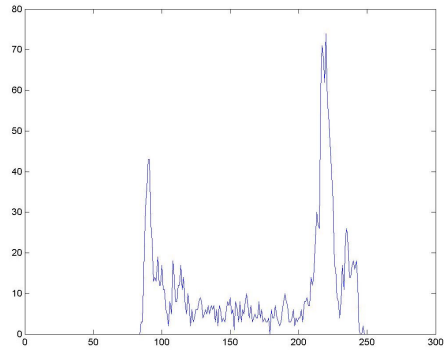
The ratio of the text block is usually extremely large or small. On the other hand, the picture blocks are fewer formed in thin rectangular. The second feature we used is the density of foreground pixels in a block. Higher ratio usually means higher probability that the block belonged to the picture region.

The last feature is obtained by histogram analysis. Histogram of the block is first derived. Observing from the histograms of text blocks and picture blocks, we find that there are two dominant peaks in the histogram of text blocks, one of which is the text part, and the other is the background. The ratio of the two dominant peaks is computed; nevertheless, the variation of the histogram is always so strong that the peaks are not able to be obtained directly from the original histogram. Hence, the histogram is passed through the median filter twice before the ratio is computed.

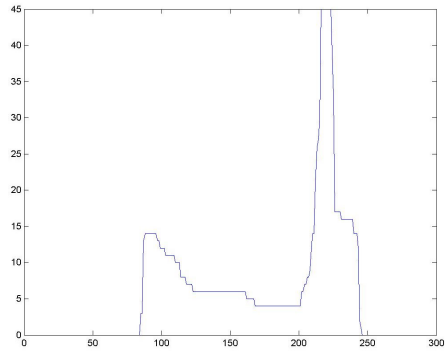
Assume the histogram of the block is depicted in figure 2(a), and after passing through the median filter twice, the result is shown in figure 2(b).

Recognition

A more sophisticated classifier is used to instead of the simple rule-based classifier. Since the classifier can learn more refined relationships than the simple rule-based methods, the classifier can provide more accurate recognition. We use a three-layer back-propagation neural network as the classifier. This classifier learns the relationship between the coefficients from given sets of samples. The back-propagation neural network can decide a high-dimensionally decision boundary. It also can distinguish texts and pictures and is capable of learning with relatively few training samples.



(a)



(b)

Figure 2. Histogram analysis: (a) original histogram, (b) histogram after passed through median filter twice.

There are one input layer, one hidden layer, and one output layer in our neural network. The number of input node is the same as the dimension of feature vectors. Nine hidden nodes are in hidden layer and only one output node is in output layer. We use images of texts and images of pictures to be training samples. After extracting feature points from an image, the feature vectors are fed into the neural network. The desired output of the picture is 1 and the desired output of the texts is 0.

The back-propagation learning algorithm is one of the most important historical developments in neural networks. The learning algorithm is applied to multilayer feedforward networks consisting of processing elements with continuous differentiable activation functions. Such networks associated with the back-propagation learning algorithm are also called back-propagation neural networks. A typical three-layer back-propagation neural network is shown in Figure 3. Given a training set of input-output pairs $\{(\mathbf{x}^{(k)}, \mathbf{d}^{(k)})\}$, $k = 1, 2, \ldots, p$, the algorithm provides a procedure to update the weights in a back-propagation network to classify the given input patterns correctly. The basis of this weight-updating algorithm is simply the gradient-decent method as used for simple perceptrons with differentiable units.

For a given input-output pair ($\mathbf{x}^{(k)}$, $\mathbf{d}^{(k)}$), the back-propagation algorithm performs two phases of data flow. First, the input pattern $\mathbf{x}^{(k)}$ is propagated from the input layer to the output layer and, as a result of this forward flow of data, it produces an actual output $\mathbf{y}^{(k)}$. Then the error signals resulting from the difference between $\mathbf{d}^{(k)}$ and $\mathbf{y}^{(k)}$ are back-propagated from the output layer to the previous layers for them to update their weights.

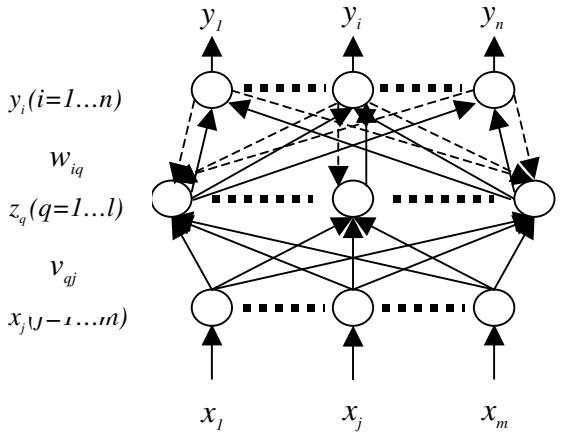


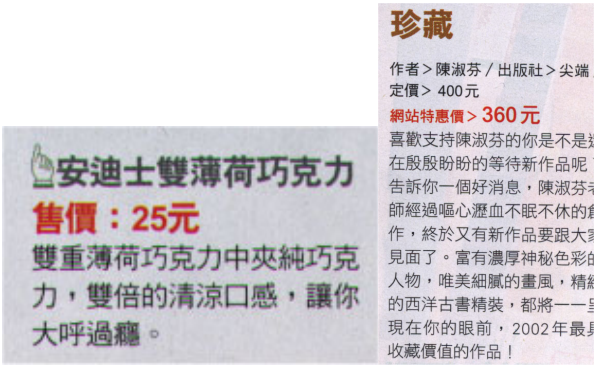
Figure 3. A Typical three-layer back-propagation neural network.

Experiment Results

The proposed algorithm for document segmentation is implemented on a PentiumⅢ 1GHz PC with 512MB RAM and programmed on Borland C++ Builder 5.0. The experiments are mainly partitioned into two parts: off-line training and on-line segmentation. In off-line training, 150 24-bit texts images and 150 24-bit picture images that are scanned from books or photoed by us are in the training image database. Some examples of training images are shown in figure 4. After training, a weighted matrix of the hidden layer, a biased matrix of the hidden layer, a weighted matrix of the output layer, and a biased matrix of the output layer are obtained.

In on-line segmentation, a document image mixture of texts and pictures is first acquired. A typical document image of size 277×347 is shown in figure 5. In background removal of the procedure for document segmentation, the background of the document is removed, and a binary image is obtained as illustrated in figure 6(a) in which foreground pixels are denoted by black. After block segmentation, each object block is located as shown in figure 6(b). Feature vectors combined of three features of each block are extracted and fed into the trained neural network in horizontal raster order. Figure 6(c) is the segmentation result of figure 5 in which the background is filled in white, the texts blocks are filled in black and the picture block is filled in black diagonal lines. Compare to the original document image, texts, picture and background are segmented correctly in this case.

(a)



(b)

Figure 4. Examples of training images. (a) pictures, (b) texts.

Figure 5. Original document image

(a)

(b)

(c)

Figure 6. Color Document Segmentation: (a) background removal, (b) block segmentation, (c) document segmentation result.

Conclusion

In this paper, the approach we proposed is for color document segmentation. We segment the document image into texts, pictures, and background. There are four mainly procedures for document segmentation in the algorithm: background removal, block segmentation, feature extraction and recognition. A binary image with black foreground and white background is obtained by adaptive local thresholding which is relative to color information. After the block segmentation procedure, which includes run-length smoothing algorithm and connected component analysis, the binary image is divided into a set of blocks. Finally, extract features from each block and feed them into the classifier that is the trained three-layer back-propagation neural network.

Through the experiments, the system we develop is able to segment document images with simple background correctly. Our system features the followings:

1. Localized thresholds to distinguish foreground from background based on color concepts.
2. Able to discriminate texts from pictures by extraction of good features.
3. Use a trainable neural network as the classifier where the structure can be adjusted flexibly.
4. Precise segmentation since the classifier is trained by mass of document images.

Although the document segmentation we proposed yields good segmentation results, the system is constrained under some situations. First, the backgrounds in the images must be simple that means the backgrounds must be lighter color. Second, the texts and pictures must be apart from each other. If texts are on the pictures, texts and pictures will be treated as the combined objects in the system. Third, mass of document images must be collected and sent to the neural network for training.

Document segmentation is generally applied to data storage, data transmission, data retrieval, optical characters recognition, image enhancement and printing technology. In data storage and data transmission, pictures and texts are compressed in different methods respectively. The size of compressed document image is much smaller than that of the original image. In data retrieval, partial information in a document image is able to be retrieved automatically.

The texts part of the document image are extracted and then processed by OCR directly in a completed OCR system. Due to the perceptions from people to texts and pictures are so different, the enhancement methods applied to texts and pictures are also different. The enhancement on texts is to sharpen the edges in the image to make the texts clearer while the enhancement on pictures is to enhance the color, contrast, illumination and etc. Document segmentation also can be applied as preprocessing in printing technology. The printing processes including halftone process, image enhancement, printing resolution, and color conversion can change based on document segmentation of the contents of the original images.

References

1. S.K. Yip, Z. Chi, "Page segmentation and content classification for automatic document image processing," 2001 Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2001.
2. J. Sauvola, M. Pietikainen, "Page segmentation and classification using fast feature extraction and connectivity analysis," 1995 Proceedings of the Third International Conference on Document Analysis and Recognition, Volume: 2, 1995.
3. F.Y. Shih, Shy-Shyan Chen, "Adaptive document block segmentation and classification," IEEE Transactions on Systems, Man and Cybernetics, Part B, Volume: 26, Issue: 5, Oct. 1996.
4. J. Sauvola, T. Seppanen, S. Haapakoski, M. Pietikainen, "Adaptive document binarization," 1997 Proceedings of the Fourth International Conference on Document Analysis and Recognition, Volume: 1, 1997.
5. F.Y. Shih, S. -S. Chen, D.C.D. Hung, P.A. Ng, "A document segmentation, classification and recognition system," ICSI '92 Proceedings of the Second International Conference on Systems Integration, 1992.
6. Hong-Ming Suen, Jhing-Fa Wang, "Color document image segmentation for automated document entry systems," TENCON '96. Proceedings of Digital Signal Processing Applications, Volume: 1, 1996.
7. Jia Li, R. M. Gray, "Context-based multiscale classification of document images using wavelet coefficient distributions," IEEE Transactions on Image Processing, Volume: 9, Issue: 9, Sept. 2000.
8. J. L. Fisher, S. C. Hinds, D. P. D'Amato, "A rule-based system for document image segmentation," 1990 Proceedings of 10th International Conference on Pattern Recognition, Volume: i, 1990.
9. Han Wang, S.Z. Li, S. Ragupathi, "Document segmentation and classification with top-down approach," KES '97. Proceedings of 1997 First International Conference on Knowledge-Based Intelligent Electronic Systems, Volume: 1, 1997.
10. A. K. Jain, K. Karu, "Learning texture discrimination masks," IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 18 Issue: 2, Feb. 1996.
11. J. S. Payne, T. J. Stonham, D. Patel, "Document segmentation using texture analysis," 1994 Proceedings of the 12th IAPR International. Conference on Pattern Recognition, Vol. 2 - Conference B: Computer Vision & Image Processing, Volume: 2, 1994.

Biography

Hsiao-Yu Han received the B.S. degree in electrical engineering from National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C., in 1999, and the M.S. degree in electrical and control engineering from National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., in 2001.

Since 2001, he has been an engineer at Opto-Electronics & Systems Laboratories in Industrial Technology Research Institute, Taiwan. His current researches are in the areas of image processing, printing system, and machine vision.