Discrimination between Printed and Handwritten Text in Documents

M.S. Shirdhonkar
B. L. D. E. A's, College of Engineering,
Bijapur, India

Manesh B. Kokare
S.G.G.S, Institute of Engineering and Technology,
Nanded, India

ABSTRACT
Recognition techniques for printed and handwritten text in scanned documents are significantly different. In this paper, we propose method to automatically identify the signature in the scanned document images. This helps to retrieve the document images based on the signature. A simple region growing algorithm is used to segment the document into a number of patches. A patch is composed of many closely located components. A component is a one piece of connected foreground pixels (say 8 connectivity). We extracted the state features of all the patches to identify the signature in the document images. A label for each such segmented patch is inferred using neural network model (NN) and support vector machine (SVM). These models are flexible enough to include signature as a type of handwriting and isolate it from machine-print. From experimental results we found that classification rate for SVM is superior over NN.

General Terms
Pattern Recognition, data mining, document image retrieval.

Keywords
Document analysis, text identification, machine vision, signature detection and retrieval

1. INTRODUCTION
Great number of applications uses documents presenting printed text and handwriting. Old documents, petitions, requests, applications for college admission, letters, requirements, memorandums, envelopes and bank checks are some examples. As the most pervasive method of individual identification and document authentication, signature present convincing evidence and provide an important form of indexing for effective document image processing and retrieval in a broad range of applications. Complex documents present a great challenge to the field of document recognition and retrieval. The combined presence of noise, handwriting, signature, logos, machine-print with different fonts, and rule lines impose a lot of restrictions to algorithms that work relatively well on simple documents. The primary task of processing these complex documents is that of isolating the different contents present in the document. Once the contents such as handwriting, machine-print, signature and noise are separated out, they can now be called as indexed documents which are ready to be used by a context-based image retrieval system.

The main contributions of this paper are summarized as follows. In this paper we have presented labeling of signature in the scanned document images. Features of all the database images are extracted using nine state features. The significance of these features is able to distinguish handwritten signature and printed text efficiently. The process of identification of handwritten signature in document image is useful to retrieve the document image based on signature.

The rest of the paper is organized into five sections. Section 2, focus is made on proposed system. In section 3, focus is made on classification model. In section 4, focus is made on experimental results. In section 5, we conclude this paper..

2. PROPOSED SYSTEM
The developed system considers application letters. Blank regions, lines, printed and handwritten words can be found all over these documents. However, they do not present logos, figures, tables, graphs or another type of element. The problem is formulated as follows, given a document, segment the document into a number of patches and label each of the segments as one of machine-print or handwriting. The class of handwriting includes those of signature and class of machine-print includes printed text of different fonts. A block diagram shown in Figure 1, describes the steps involved in the labeling scanned documents.

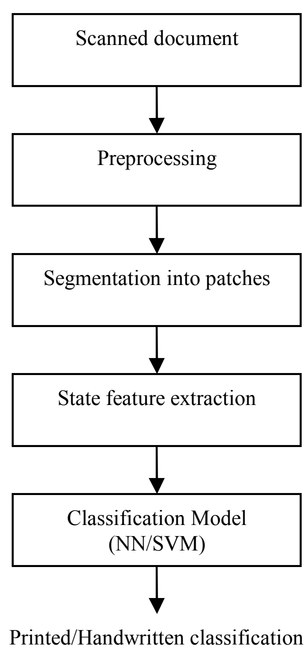

Fig. 1 Block diagram describing the steps involved in system

It has three main steps: preprocessing, segmentation and feature extraction

2.1 Preprocessing steps

The following operations are accomplished in this step. Firstly, median filter is applied to decrease the noise in the image. This noise can appear on document acquisition or digitalization. Secondly, the text is separated from background by automatic thresholding. The Otsu bi-level approach [1], is used to define the threshold. Finally, the document image is submitted to morphologic opening (that is erosion followed by dilation) by a 3x1 symmetrical structuring element. This is designed with two purposes: eliminating reminiscent noises of the previous phase and soften vertical contours of the text. These are essential for proper computation of some features to be extracted.

2.2 Segmentation

In computer vision segmentation refers to the process of partitioning a digital image into multiple regions (sets of pixels). Image segmentation here is typically used to locate objects especially words printed as well as hand-written that form regions of interest [2], [3].

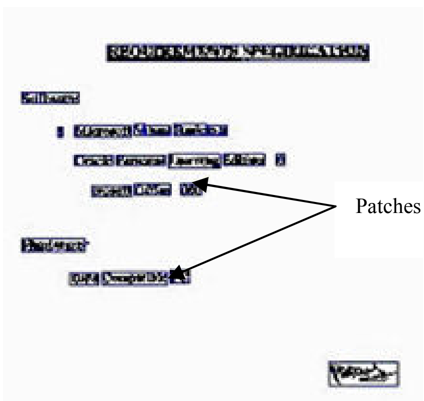

Fig. 2 Segmentation on a sample document

A patch is defined to be a region in a document such that, if a rectangular window (size determined dynamically for each document) is drawn with each foreground pixel within the patch at its center, then the window shall not contain any foreground pixel from another patch. Figure 2 shows a sample document showing the result of the segmentation process and each patch is marked with a box around it.

The algorithm for generating these patches is a region growing algorithm and a brief description is given below.

Step 1: Initialize every pixel to be a separate patch.
Step 2: Start with a foreground pixel that is not already marked.
Step 3: With this pixel as the center, draw a rectangular window around it. The size of the window is optimized to get the desired size of patches on a validation set of documents.
Step 4: All foreground pixels of connected components with pixel enclosed within these rectangular windows are marked as belonging to the same patch as that of the center pixel.
Step 5: Repeat steps 2 through 4 until all pixels are marked.
Step 6: Patches with fewer pixels than a predetermined threshold is ignored as noise and are not attempted to be labeled as one of machine-print, has signature, noise.

2.3 Feature extraction

Feature extraction involves extracting the meaningful information from the images. So that it reduces the storage required and hence the system becomes faster and effective in image retrieval. Once the features are extracted, they are stored in the database for future use. The degree to which a computer can extract meaningful information from the image is the most powerful key to the advancement of intelligent image interpreting systems. One of the biggest advantages of feature extraction is that, it significantly reduces the information (compared to the original image) to represent an image for understanding the content of that image. State features try to associate each patch to a label using the characteristics of that patch alone [3], [4]. Eight state features, are extracted for each patch and these are listed in Table 1, and described as follows

1) Height: This is determined by computing the difference between the maximum y co-ordinate value and minimum y co-ordinate value of a patch.

2) Aspect Ratio: This is determined by computing the ratio of the width and height of the patch, the width of the patch is determined by taking the difference between the maximum and minimum x co-ordinate values of a patch.

3) Density: This is determined by counting the number of foreground pixels (white). The density is then computed as the ratio of this count by total number of pixels within the patch.

4) Percentage of text above: The y co-ordinate values of the patches that are lesser than the y co-ordinate value of the patch in question will form the patches that lie above the current patch. The percentage is derived by the difference of the number of patches above the patch and the total number of patches in the document image multiplied by 100.

5) Maximum run length: Runlength is the number of continuous pixels in a particular direction, this is determined by first computing the number of continuous pixels in each row within a patch and is stored in an array then the maximum value within this array forms the maximum runlength in the horizontal direction.

6) Average runlength: The array with the horizontal runlengths within the patch as described above are then divided by the number of runlengths, thus giving the average runlength.

7) Horizontal transition: The horizontal transitions are determined by counting the number of times the pixel value changes from black to white or vice versa within a patch from left to right within the patch.

8) Vertical transitions: The vertical transitions are determined by counting the number of times the pixel value changes from black to white or vice versa within a patch from top to bottom of the patch.

Table 1. State features

State features	Description
Height	Maximum height of the patch
Aspect ratio	Width/Height of the patch
Density	Density of foreground pixels within the patch
Percentage of text above	Relative location of the patch with respect to the entire document
Maximum run Length	The maximum horizontal run length within a Patch
Average run Length	The average horizontal run length within a Patch
Horizontal Transitions	A count of the number of times the pixel value transitions from White to black horizontally.
Vertical Transitions	A count of the number of times the pixel value transitions from white to black vertically

3. CLASSIFICATION

Many classical machine learning algorithms may be applied to classification, which includes Bayesian learning, Neural Network (NN), Support Vector Machine (SVM), Conditional Random Field (CRF) and so on. In our system we used NN and SVM to classify the handwritten/printed text of scanned document. Experimental result on the database indicates that a classification performance using SVM is superior over NN.

A neural network model uses the back-propagation algorithm for training. The features form the input to the model during the training. The hidden neurons in the network are computed using the following equation is

$$\frac{No.of\ input\ features + No.of\ output\ labels}{2} + \sqrt{X} \quad (1)$$

Where X is the Number of training samples. The neural network had 2 output neurons and 24 hidden neurons. The parameter learning was done using back propagation algorithm.

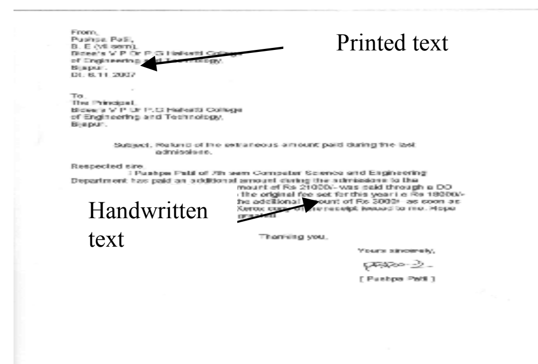

Fig. A sample document with labeled as machine Printed/handwritten

The SVM provide a new approach to the problem of pattern recognition with clear connections to the underlying statistical learning theory. SVM links the problems they are designed for with a large body of existing work on kernel based methods [6], [7]. Here we briefly introduce the basic concepts of two classes SVM. On pattern classification problems, SVMs provide very good generalization performance in empirical applications. We begin our discussion of support vector machines by returning to the two-class classification problem using linear models of the form

$$y(X) = W^T \phi(X) + b \quad (2)$$

Where $\phi(X)$ denotes a fixed feature-space transformation, and we have made the bias parameter b explicit. The training data set comprises N input vectors $X_1, \dots, X_N$, with corresponding target values $t_1, \cdots t_N$, and new data points are classified according to the sign of $y(X)$. The features are written to an input text file that consists of the training set and the testing set involves inputting the features of a particular patch in query image to the network which then determines its label/class.

4. EXPERIMENTAL RESULTS

In our experiment we have used ten different documents which contains total of 1050 patches. Six documents are used for training purpose, which contain 655 patches. The remaining 396 patches from document are used as the testing set. This consists of signature and printed text. A sample document with labeled as machine printed/handwritten refer to Figure 3.The comparison of the labeling accuracy (recall) and precision values on these patches are shown in Table 2. The resulting labeled documents can be effectively used in content based image retrieval [8]. For performance evaluation of the system, it is significant to define a suitable metric. Two metrics are employed in our experiments as follows.

Recall of label 'a' is

$$= \frac{Amount\ of\ correctly\ classified\ data\ of\ label\ 'a'}{Total\ amount\ of\ data\ of\ label\ 'a'} \qquad (3)$$

Precision of label 'a' is

$$= \frac{Amount\ of\ correctly\ classified\ data\ of\ label\ 'a'}{Total\ amount\ of\ text\ classified\ to\ be\ of\ label\ 'a'} \qquad (4)$$

Table 2: The results of labeling the documents

Type Of Label	SVM		NN	
	Recall	**Precision**	**Recall**	**Precision**
Machine Printed text	92	96	86	96
Handwritten signature	100	75	100	50

5. CONCLUSION

The paper presents the labeling of printed and handwritten signature in document image using nine different local state features. These features capture information efficiently for each patch in order to distinguish the handwritten signature and printed text. The neural network model and support vector machines are used for classification/labeling problem. The labeled documents can be effectively used in document image retrieval based on signature as query. The performance of SVM model is superior over NN.

6. REFERENCES

[1] N.Otsu, A. 1979,Threshold Selection Method from Gray – Level Histograms. In IEEE Transactions on Systems, Man and Cybernetics, v.9, n 1, pp. 62-66.

[2] Shravya Shetty, Harish Srinivasan, Matthew Beal and Sargur Srihari. 2007. Segmentation and labeling of documents using conditional random Fields. Center of Excellence for document analysis and recognition (CEDAR), University of Buffalo, and State University of New York.

[3] J. Laffery, A. Macullum and F. Perira.2001.Conditional random Fields: Probabilistic Model for segmenting and labeling sequential data. Eighteenth International Conference on Machine Learning, pp.282-289.

[4] Shravya Shetty, Harish Srinivas and Sargur Srihari.2007. Use of Conitional Random Fields for signature based retrieval of scanned documents. Center of Excellence for Document analysis and recognition (CEDAR), University of Buffalo, State University of New York, pp. 1-15.

[5] Rafael C. Gonzales's, Richard E.Words and Steven L, Digital Image using MATLAB, Eddins, Low Price Edition.

[6] Christopher M. Bishop Pattern Recognition and Machine Learning

[7] Christopher J.C.Burges.1998.A Tutorial on support vector Machines for Pattern recognition. Bell Lab. Lucent Technologies,pp. 121-167.

[8] Guangyu Zhu, Yefeng Zheng, and David Doermann.2008. Signature-based Document image retrieval. ECCV, Part III, LNCS 5304, pp.752-765..