Markov Random Field Based Text Identification from Annotated Machine Printed Documents

Xujun Peng[1], Srirangaraj Setlur[1], Venu Govindaraju[1], Ramachandrula Sitaram[2] and Kiran Bhuvanagiri[2]

[1] Center for Unified Biometrics and Sensors
Dept of Computer Science and Engineering
SUNY at Buffalo, Amherst, NY 14228, USA
{xpeng, setlur, govind}@buffalo.edu

[2] HP Labs India
Hosur Main Road, Adugodi
Bangalore 560030, India
{sitaram, kiran.kumar2}@hp.com

Abstract

In this paper, we describe an approach to segment handwritten text, machine printed text and noise from annotated machine printed documents. Three categories of word level features are extracted. We use a modified K-Means clustering algorithm for classification followed by a relabeling procedure using Markov Random Field(MRF) based on a concept of neighboring patches and Belief Propagation(BP) rules. Experimental results on an imbalanced data set show that our approach achieves an overall recall of 96.33% .

1. Introduction

Unlike the retrieval of machine printed documents, where high OCR accuracy can be expected, the retrieval of noisy annotated documents which contain both handwritten text and machine printed text is still a challenge because document retrieval in the context of handwriting has not been widely explored.

The pre-processing of mixed documents to isolate handwritten and machine printed text and to remove noise is an important step in the design of systems for OCR, author identification, signature verification and document retrieval.

In [3], Jose et al. suggest two types of features (content related features and shape related features) to characterize handwritten text on bank check images. Jang et al. [9] propose a type of geometric features to classify machine printed and handwritten addresses on mail-pieces. Considering a text word as a sequence signal, Guo and Ma [7] separate handwritten material from documents using a Hidden Markov Model. In order to identify Arabic handwritten text in mixed documents, Farooq et al. [4] use an EM based probabilistic NN model. In [15], Zheng et al. propose a Gibbs network which is optimized by Highest Confidence First algorithm for text classification and extend their work to signature detection [16]. A similar approach but using Conditional Random Field is proposed by Shetty et al. [13]. Recently, considerable work has been done on image binarization [2] using Markov Random Field(MRF) which is inspired by MRF's success in the area of image restoration [5]. MRF has also been used to label text[15].

In this paper, we present a MRF based approach to separate handwritten text, machine printed text and noise from annotated documents. Text segmentation and feature extraction are covered in section 2. In section 3, we initially classify 3 different kinds of patches using G-Means, which are then refined using Markov Random Fields based on the concept of a system of neighbors and Belief Propagation(BP) update rules as proposed in section 4. Experimental results and our conclusions are presented in section 5 and section 6.

2. Preprocessing

Our pre-processing consists of two steps: (i) text segmentation and (ii) feature extraction.

2.1. Text segmentation

Prior to classification, each binarized document is segmented into patches which are small snippets of the image. In our MRF based framework, we model each document as a random field which consists of a number of patches. A $m \times n$ sized window is used for dilation of the original binarized image and the bounding box of each connected component after dilation is defined as a patch. The size of the window is empirically chosen such that the resultant patch typically represents a handwritten or machine-printed word. Patches whose size is smaller than a threshold t_l or larger than a threshold t_h are eliminated as noise.

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.237

2.2. Feature extraction

Three different categories of features are considered for classification of a given patch into one of three classes viz., handwritten text, machine-printed text and noise.

- **Patch Level Features:** 12 patch level statistic features are extracted for each patch of size $w \times h$ as shown in Table 1.
- **Connected Component Features:** Connected components in each patch are extracted from the original (non-dilated) binarized image and 9 connected component based features are considered as described in Table 1.
- **Gabor Features:** Gabor filters can serve as directional band-pass filters which are modulations of a complex sinusoidal and Gaussian function. The 2-D Gabor filter is defined as Eq.1 in the space domain (details of parameters can be found in [11, 12, 6]):

$$g_{\lambda,\theta,\varphi,\delta,\gamma}(x,y) = K \exp\left(-\frac{x'^2+\gamma^2 y'^2}{2\delta^2}\right)\cos\left(2\pi\frac{x'}{\lambda}+\varphi\right) \quad (1)$$

where

$x' = x\cos\theta + y\sin\theta, \quad y' = -x\sin\theta + y\cos\theta$

and λ is wavelength of the cosine factor of the Gabor filter kernel, θ is orientation, φ is the phase offset, γ is the aspect ratio and δ is the squared deviation of the Gaussian function. A set of different θ and λ lead to the 8 gabor filters used in our experiments.

3 Initial Classification

The initial labeling of the three different kinds of patches during training is carried out using a modified K-Means clustering method known as G-Means [8].

Unlike normal K-Means clustering where one has to estimate k, G-Means determines the number of clusters based on the distribution of the training data. The essence of G-Means is to split the data set using normal K-Means until each cluster is Gaussian-like in its distribution. The procedure of G-Means is described in following table.

In our experiments, G-Means is used to further cluster the three different classes individually to obtain a total of N centers. Then each feature point in the training data set is assigned to a label $i(0 \leqslant i < N)$ which is the index of the closest center $\overline{c}_i$ and the co-variance of each cluster is calculated as:

$$\Sigma_i = \frac{1}{M}\sum_{k=1}^{M}(x_k - \overline{c}_i)(x_k - \overline{c}_i)^T \quad (0 \leqslant i < N) \quad (2)$$

G-Means Procedure	
1:	Initialize data set as a cluster $C_i = \{x \mid x \in class(i)\}$.
2:	Project samples within cluster onto an optimal projection direction v to get corresponding one-dimensional data set $\hat{C}_i = \{y \mid y = \langle x, v\rangle / \|v\|^2\}$.
3:	Estimate the confidence of statistic A to determine cluster's distribution using Anderson-Darling test [1]: $A^2 = -n - \sum_{k=1}^{n}\frac{2k-1}{n}[\ln G(y_k) + \ln(1 - G(y_{n+1-k}))]$ where n is the size of the data set, y_k is the sorted sample from $\hat{C}_i$ and $G(x)$ is the normal distribution function.
4:	If $A < \alpha$, where α is a pre-defined threshold, C_i is regarded as a Gaussian-like distribution and is stored as a qualified cluster with its center $\overline{c}_i = \sum x_k / n$. Otherwise, data set is split into 2 clusters C_{i+1} and C_{i+2} using normal K-Means clustering.
5:	For each new cluster, go to step 2, until every cluster has a Gaussian-like distribution.

where M is the number of feature points within the cluster, x_k is a sample in this cluster and $\overline{c}_i$ is the center of the cluster.

In the classification phase, the Mahalanobis distance from a feature point to each cluster is calculated and this feature point is assigned to the closest center(label):

$$L(x) = \arg\min_i D_m(x, \overline{c}_i) = \arg\min_i \sqrt{(x-\overline{c}_i)^T \Sigma_i^{-1}(x-\overline{c}_i)} \quad (3)$$

Since the class to which each center belongs (handwritten text, machine printed text or noise) is already known during the training phase, it is easy to map the label of the test feature points to one of the three classes.

For convenience, we use the terminology *index* i of center $\overline{c}_i$ and *label* L_i interchangeably in the following sections.

4 Relabeling

Due to overlaps in feature space, misclassification cannot be avoided using a single classifier. Therefore, post-processing or relabeling is needed. The intuition for relabeling is that a patch surrounded by patches from a single different class has a high probability of belonging to that class. Markov Random Field is a kind of network which describes the statistical dependency between observed and hidden states in the net and is a suitable model to relabel patches in our scenario.

4.1 MRF Topology

We use the topology shown in Fig.1 for our MRF. Each grey node x_i in the hidden layer exclusively corresponds to

Patch Feature	Description
Location	Relative location of patch with respect to entire document: x, y
Relative width and height	Relative width and height of patch with respect to its nearest neighbor.
Foreground density	The number of foreground pixels divided by the size of the patch: $d = \sum_{x,y} I(x,y)/(w \times h)$
Average stroke width	The number of foreground pixels divided by the length of the contour: $s = \sum_{x,y} I(x,y)/l$
Crossing number	The number of pixels whose intensity differs from its neighbor vertically and horizontally: $c_x = \sum_{x,y} I(x,y) \oplus I(x+1,y)/h, \quad c_y = \sum_{x,y} I(x,y) \oplus I(x,y+1)/w$
Variance of projection	Variance of horizontal and vertical projection.
Maximum run length	Maximum runlength within the patch in the horizontal and vertical directions.
Connected Component Feature	**Description**
Components number	The number of connected components within a patch: n
Maximum width and height	Maximum width and height of a connected component within the patch: max_w, max_h
Mean of width and height	Average width and height of components within the patch: $ave_w = \frac{1}{n}\sum_i w(i), \quad ave_h = \frac{1}{n}\sum_i h(i)$
Variance of width and height	Width and height variation of components within the patch: $var_w = \sqrt{\frac{1}{n}\sum_i (w(i) - ave_w)^2}, \quad var_h = \sqrt{\frac{1}{n}\sum_i (h(i) - ave_h)^2}$
Hole ratio	Total hole area within patch divided by patch's size.
Overlap ratio	Overlap area between connected components divided by the size of the patch.

Table 1. Patch level and connected component level features

a document patch and will be assigned to a label $L_i (0 \leqslant L_i < N)$ after initial classification. Hidden nodes are connected to their four spatially closest neighbors which are defined in the next subsection. Each white node y_i in the observation layer is a feature point for that patch and connects to its hidden node. Edges between nodes carry messages which indicate the similarity (hidden layer) or dependency(observation layer) between the nodes.

Messages in a network propagate in two opposite directions ([10, 14, 5]): 1) a node receives incoming messages from its neighbors which can be used for maximum a posteriori (MAP) estimation of belief at a certain node i

$$\hat{x_i} = \arg\max_{x_i} \psi(x_i, y_i) \prod_{j \in N(i)} m_{ij}(x_i) \tag{4}$$

2) outgoing messages from a node to its neighbors e.g. from node j to i which leads to the message update rule

$$m_{ij}(x_i) = \max_{x_j} \psi(x_i, x_j)\psi(x_j, y_j) \prod_{k \in N(j) \setminus i} m_{jk}(x_j) \tag{5}$$

where $\psi(x, y)$ is the compatibility function representing the similarity or dependency between two nodes, $m_{ij}(x_i)$ represents a message from node j to i, $j \in N(i)$ are the neighbors of node i and $k \in N(j) \setminus i$ are the neighbors of node j except node i.

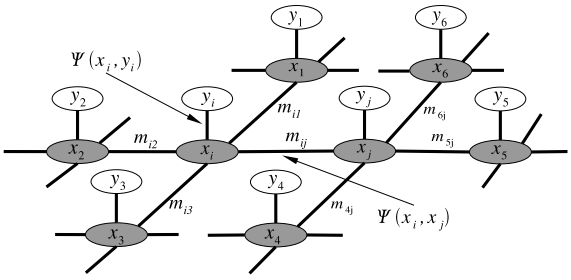

Figure 1. The topology of the MRF. Each grey node x_i which is a label for a text patch connects to its four nearest neighbors. Each observation node y_i which is a feature point in the feature space connects to its hidden label x_i. Edges between any pair of nodes indicate their similarity.

4.2 System of Neighbors

The spatial distance between each pair of patches in a document is defined as:

$$D_n(i,j) = \frac{(dx_{i,j} - \hat{x})^2}{2\hat{x}^2} + \frac{(dy_{i,j} - \hat{y})^2}{2\hat{y}^2} \tag{6}$$

In this Gaussian-like function, $[dx_{i,j}, dy_{i,j}]$ represent the convex-hull distances between patches i and j in the hor-

izontal and vertical directions, $\hat{x}$ is the dominant gap between words and $\hat{y}$ is the dominant gap between text lines over the entire document. Dominant gaps $\hat{x}$ and $\hat{y}$ can be estimated using histograms. Based on the spatial distance, the four closest neighbors are considered for each patch. In Fig.2, the four nearest neighbors of the patch contained in the red rectangle are represented by the four black rectangles.

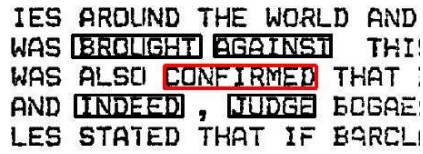

Figure 2. A patch and its nearest neighbors.

4.3 Belief Propagation(BP) Update Rules

Unlike other methods [15] that use patch clique occurrence frequency to relabel patches, we consider similarities between two hidden nodes in both the space as well as the feature domains. The compatibility function between hidden nodes x_i and x_j is defined as

$$\psi(x_i, x_j) = 1 + \alpha e^{-D_n(i,j)} + \beta e^{-D_e(L_i, L_j)} \quad (7)$$

where $D_n(i, j)$ is the spatial distance calculated from Eq.(6) between two neighboring patches and $D_e(L_i, L_j)$ is the Euclidean distance between the two hidden nodes that represent the assigned centers in the feature space. α and β are two parameters that control the influence between neighbors.

The compatibility between a hidden node and its corresponding observed node is:

$$\psi(x_i, y_i) = e^{1/(\lambda D_m(c_i, o_i))} \quad (8)$$

where $D_m(c_i, o_i)$ is the Mahalanobis distance, c_i is the center of the hidden node in feature space and o_i is the feature point representing the patch. Parameter λ controls the influence of the observed node on the hidden node.

All initial messages m_{ij} in Eqs.(4) and (5) are set to 1 and updated using Eqs.(7) and (8) until there is no label flip during message propagation.

5 Experiments

94 documents from the Tobacco industrial litigation archives were used in our experiments. The original binarized documents were dilated using a 5×5 window and segmented using the method described in Section 2.1.

(a) (b)

Figure 3. Mixed hand/machine text and low resolution patches

A total of 29685 patches(19842 machine printed patches, 832 handwritten patches and 9011 noise patches) were extracted. 15409 patches from 48 documents were used for training. 117 centers for machine printed patches, 5 centers for handwritten patches and 53 centers for noise patches were obtained using G-Means based clustering on the training data. The remaining 14276 patches from 46 documents were used for testing.

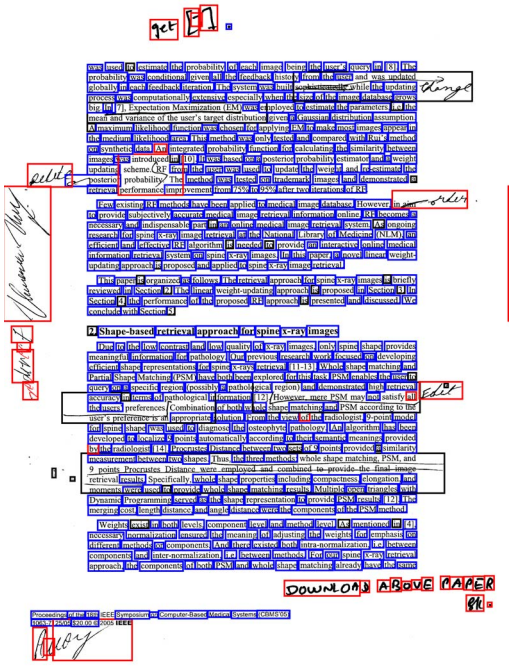

Figure 4. An example document with the labeled results from the system. Blue boxes represent machine printed text, red boxes represent handwritten text and black boxes represent noise.

All test patches were initially assigned to the closest centers in the feature space as described in Section 3. The MRF based BP rules from Section 4 were then used to relabel all

	G-Means		MRF	
	Precision	Recall	Precision	Recall
Machine-printed	98.91%	94.11%	99.20%	96.59%
Handwritten	47.97%	90.91%	64.99%	96.01%
Noise	92.40%	95.19%	94.84%	96.40%
Overall	N/A	93.40%	N/A	96.33%

Table 2. Result of classification & MRF relabeling

the patches. *Precision* and *recall* measures were computed to estimate the performance of our approach.

$$precision(i) = \frac{\#\ of\ patches\ correctly\ classified\ as\ class\ i}{\#\ of\ patches\ classified\ as\ class\ i} \quad (9)$$

$$recall(i) = \frac{\#\ of\ patches\ correctly\ classified\ as\ class\ i}{\#\ of\ patches\ belonging\ to\ class\ i} \quad (10)$$

Table 2 shows that the precision and recall for each class increased after using MRF relabeling, especially for handwritten text. Overall recall increased from 93.4% to 96.33%. In our test set, there were about 20 patches which contained both handwritten and machine printed text and over 300 patches whose resolution was very low as shown in Fig.3. The mixed handwritten/text patches should probably be treated as a separate class since it is not clear how these should be labeled and hence were excluded from our evaluation. The low resolution patches were labeled as machine print but were typically classified by the system as handwritten patches and resulted in the precision metric for the handwritten text being low. The problem of the low resolution patches can be overcome by adding more such samples into the training set.

Testing was also done on a separate HP image set of 66 documents that were very different from the Tobacco litigation data set. Fig.4 shows one example result from the HP data set. Most patches are correctly classified even though the system was only trained on the Tobacco litigation set.

6 Conclusions

In this paper, we present a Markov Random Field based method to classify three different kinds of text(machine printed, handwritten and noise). To integrate MRF into the initial classifier, we use G-Means to cluster each class individually and use those centers as our hidden nodes in the MRF. A novel Gaussian-like function is used to compute distance between patches to locate neighbors. Experiments show that our method has better classification performance than a single classifier. Future work includes the use of smaller patches and use of other classification techniques.

References

[1] T. W. Anderson and D. A. Darling. Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Annals of Mathematical Statistics 1952*, pages 193–212.

[2] H. Cao and V. Govindaraju. Handwritten carbon form preprocessing based on markov random field. *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR) '07*, pages 1–7, 2007.

[3] J. B. D. S. Eduardo, B. Dubuisson, and F. Bortolozzi. Characterizing and distinguishing text in bank cheque images. *Proc. XV Brazilian Symposium on Computer Graphics and Image Processing*, pages 203–209, 2002.

[4] F. Farooq, K. Sridharan, and V. Govindaraju. Identifying handwritten text in mixed documents. *Proc. 18th International Conference on Pattern Recognition(ICPR) 2006*, 2:1142–1145, 2006.

[5] W. Freeman, O. Carmichael, and E. Pasztor. Learning low-level vision. *International Journal of Computer Vision*, 40(1):25–47, 2000.

[6] S. E. Grigorescu, N. Petkov, and P. Kruizinga. Comparison of texture features based on gabor filters. *IEEE Trans. on Image Processing*, 11(10):1160–1167, 2002.

[7] J. Guo and M. Ma. Separating handwritten material from machine printed text using hidden markov models. *Proc. Sixth International Conference on Document Analysis and Recognition*, pages 439–443, 2001.

[8] G. Hamerly and C. Elkan. Learning the k in k-means. *In Proc. 17th NIPS*, 2003.

[9] S. I. Jang, S. H. Jeong, and Y.-S. Nam. Classification of machine-printed and handwritten addresses on korean mail piece images using geometric features. *Proc. 17th International Conference on Pattern Recognition(ICPR) 2004*, 2:383–386, 2004.

[10] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 1988.

[11] N. Petkov. Biologically motivated computationally intensive approaches to image pattern recognition. *Future Generation Computer Systems*, 11(4-5):451–465, 1995.

[12] N. Petkov and P. Kruizinga. Computational models of visual neurons specialised in the detection of periodic and aperiodic oriented visual stimuli: bar and grating cells. *Biological Cybernetics*, 76(2):83–96, 1997.

[13] S. Shetty, H. Srinivasan, M. Beal, and S. Srihari. Segmentation and labeling of documents using conditional random fields. *Proc. Document Recognition and Retrieval IV, Proceedings of SPIE*, pages 6500U–1–11, 2007.

[14] L. Xiong, F. Wang, and C. Zhang. Multilevel belief propagation for fast inference on markov random fields. *Proc. Seventh IEEE International Conference on Data Mining(ICDM) 2007*, pages 371–380, 28-31 Oct. 2007.

[15] Y. Zheng, H. Li, and D. Doermann. Machine printed text and handwriting identification in noisy document images. *IEEE J PAMI*, 26(3):337–353, 2004.

[16] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger. Multi-scale structural saliency for signature detection. *Proc. IEEE Conference on Computer Vision and Pattern Recognition(CVPR) '07*, pages 1–8, 2007.