Handwritten Word Recognition using Conditional Random Fields

Shravya Shetty Harish Srinivasan
Sargur Srihari
Center of Excellence for Document Analysis and Recognition (CEDAR)
Department of Computer Science and Engineering, University at Buffalo
{sshetty, hs32, srihari}@cedar.buffalo.edu

Abstract

The paper describes a lexicon driven approach for word recognition on handwritten documents using Conditional Random Fields(CRFs). CRFs are discriminative models and do not make any assumptions about the underlying data and hence are known to be superior to Hidden Markov Models(HMMs) for sequence labeling problems. For word recognition, the document is first segmented into word images using an existing neural network based algorithm. Each word image is then over segmented into a number of small segments such that the combination of segments forms character images. Segment(s) is/are labeled as characters with probability evaluated from the CRF model. The total probability of a word image representing an entry from the lexicon is computed using a dynamic programming algorithm which evaluates the optimal combination of segments.

1 Introduction

Recognition of handwritten words from unconstrained scanned documents remains to be an open research problem. Complex structures in documents hinders segmentation into lines and words. Additionally the variability in handwriting imposes another level of challenge for word recognition. Figure 1 depicts the word recognition problem.

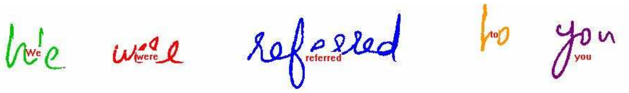

Figure 1. Word recognition problem: Handwritten previously segmented word images are to be recognized as shown

Word recognition can be classified under the following categories

1. Lexicon driven/Lexicon independent: Lexicon driven approach uses a fixed length lexicon, and associates each word image in the document with the top ranked matching lexicon. Lexicon independent methods rely solely on automatic character recognition.

2. Segmentation free/Segmentation based: Segmentation free methods try and recognize the entire word image using its global features. On the other hand segmentation based methods, rely on breaking the word image into smaller segments identifiable as characters and associate a character label to these segments. Here contextual dependencies between neighboring segments are exploited here using time series models such as Hidden Markov Models (HMMs).

Most success in this domain has been achieved using lexicon driven segmentation based approaches. Although in segmentation based methods, HMMs can be used to capture the spatial dependencies amongst neighboring segments, they are generative models and try to model the joint probability of the the data and the labels. Contrary to this Conditional Random Fields (CRFs) [1] model the conditional distribution and do not make any assumptions about the distribution of the data. Also, unlike HMMs, CRFs can capture a number of feature functions and each feature function can use the entire input data sequence. CRFs also avoid a fundamental limitation of maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed graphical models, which can be biased towards states with few successor states(label bias problem). HMMs can be used for segmentation free word recognition, the model captures the transition to the same character label or to the next character in the word. In [2] CRFs were used for recognition at the word level, but this approach leads to a large number of parameters as the model has to learn state parameters for every word in the lexicon and transition parameters for every possibe pair of neighboring words in the lexicon. Here, we use a lexicon driven approach based on character segmentation, we find HMMs to be unsuitable for this task because it is intuitively incorrect to use the transition of character labels when comparing a word image against a given word in the lexicon. CRFs on the other hand use transition features to relate the neighboring character labels with the features of the corresponding handwritten character segments. With regard to the above, we propose to use Conditional Random Fields for the recognition of handwritten words using a lexicon. The method is based on segmentation of word images into characters and identifying the lexicon with the highest probability of representing the word image under the CRF model. The character level segmentation is done by oversegmenting the word image and then computing the best possible grouping of these segments to form the characters of a given word in the lexicon. The rest of the paper is organized as follows. Section 2 describes the CRF model and its parameter estimation, followed by a detailed description of the methods for segmentation and feature extraction in section 3. Experiments and results are described in section 4 followed by conclusion in section 5.

2 Conditional Random Fields

The problem of word recognition as stated before is formulated as a sequence labeling task. The sequence to be labeled is a set of candidate character segments obtained by the segmentation of the word image. The probability that each segment/combination of segments represents a character in the lexicon is evaluated using the CRF model, to be described in section 2.1. The total probability of association between the word image and an entry in the lexicon is then computed using these individual character probabilities. Figure 2 shows the graphical model of CRF for the segmentation based word recognition problem. We first describe the CRF model in general and then its parameter estimation method.

2.1 CRF Model description

The probabilistic model of the Conditional Random Field used is given below.

$$P(\mathbf{y}|\mathbf{x},\theta) = \frac{e^{\psi(\mathbf{y},\mathbf{x};\theta)}}{\sum_{\mathbf{y}'} e^{\psi(\mathbf{y}',\mathbf{x};\theta)}} \qquad (1)$$

where $\mathbf{y_i}$ $\in$ {a-z,A-Z,0-9} and $\mathbf{x}$: Handwritten word image and θ : CRF model parameters. It is assumed that a word is segmented into m candidate character segments(details of the segmentation algorithm are described later in section 3.1).

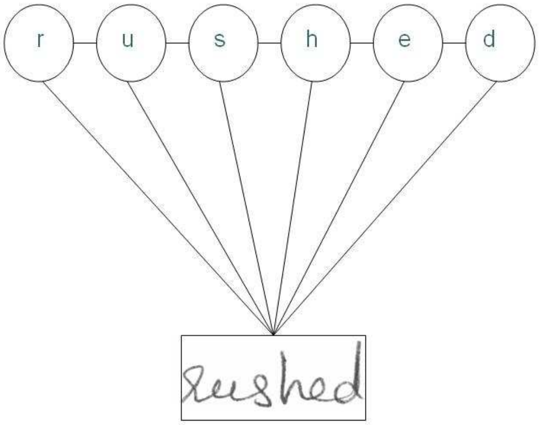

Figure 2. CRF model for segmentation based handwritten word recognition, for an example word“rushed”

$$\psi(\mathbf{y},x;\theta) = \sum_{j=1}^{m}\left(A(j,y_j,\mathbf{x};\theta^s) + \sum_{(j,k)\in E} I(j,k,y_j,y_k,\mathbf{x};\theta^t)\right) \qquad (2)$$

The first term in equation 2 is called the state term(sometimes called Association potential as mentioned in [3]) and it associates the characteristics of that character segment with its corresponding label. θ^s are called the state parameters for the CRF model. Analogous to it, the second term, captures the neighbor/contextual dependencies by associating pair wise interaction of the neighboring labels and the observed data(sometimes referred to as the interaction potential). θ^t are called the transition parameters of the CRF model. E is a set of edges that identify the neighbors of a character image. The specific set of neighbors used for this problem is described in detail in section 3.1. θ comprises of the state parameters,θ^s and the transition parameters,θ^t.
The association potential can be modeled as

$$A(j,y_j,\mathbf{x};\theta^s) = \sum_i (f_i^s(j,y_j,\mathbf{x}) \cdot \theta_{ij}^s)$$

where f_i^s is the i^{th} state feature extracted for that segment and θ_{ij}^s is the state parameter. The state features that are used for this problem are defined later in section 3.2 in table 3.2. The interaction potential $I(\cdot)$ is an inner product between the transition parameters θ^t and the transition features f_t.

$$I(j,k,y_j,y_k,\mathbf{x};\theta^t) = \sum_i (f_i^t(j,k,y_j,y_k,\mathbf{x}) \cdot \theta_{ijk}^t)$$

2.2 Parameter estimation

There are numerous ways to estimate the parameters of this CRF model [4]. In this paper we use conjugate gradient ascent to maximize the likelihood of the data. We assume that given the data, the labels of characters follow first order Markov property. For the problem of word recognition, we assume that the label of the current character depends only on the preceding character label. The maximum likelihood estimate of the parameters, θ are given by equation 3.

$$\theta_{ML} = \arg\max_{\theta} \prod_{i=1}^{M} P(y_i|y_{\mathcal{N}_i}, \mathbf{x}, \theta) \tag{3}$$

where $P(y_i|y_{\mathcal{N}_i}, \mathbf{x}, \theta)$ (Probability of the label y_i for a particular patch i given the labels of its neighbors, $y_{\mathcal{N}_i}$), is given below.

$$P(y_i|y_{\mathcal{N}_i}, \mathbf{x}, \theta) = \frac{e^{\psi(y_i, \mathbf{x}; \theta)}}{\sum_a e^{\psi(y_i = a, \mathbf{x}; \theta)}} \tag{4}$$

where $\psi(y_i, x; \theta)$ is defined in equation 2.
The joint probability of recognizing the word image as a word in the lexicon can be factorized into potential terms, each containing just the label and the features of a character in the word, and that of its preceding character in the word. Equation 5 shows the factorization of the probability for the word 'the' as given by the CRF model.

$$P(\mathbf{y} =' the'|\mathbf{x}, \theta) = \frac{e^{\psi(y='t', x; \theta)}}{\sum_{y'='a'}^{'z'} e^{\psi(y', x; \theta)}} \frac{e^{\psi(y='h', y_n='t', x; \theta)}}{\sum_{y'='a'}^{'z'} e^{\psi(y', y_n='t', x; \theta)}} \frac{e^{\psi(y='e', y_n='h', x; \theta)}}{\sum_{y'='a'}^{'z'} e^{\psi(y', y_n='h', x; \theta)}} \tag{5}$$

This above factorization enables the use of dynamic programming as will be discussed later in section 3.3.

3 Segmentation and feature extraction

The scanned handwritten document is first binarized using Otsu's thresholding algorithm or by other adaptive thresholding schemes. The binarized image is represented using chain code for faster processing. Line segmentation is first performed using bi-variate Gaussian densities to model each line [5]. Each line is then split into words using a neural network(NN) based algorithm. The NN classifies the gap between two connected components as a word gap or not.

3.1 Segmentation

Each of these word images is further segmented into a number of finer segments considering the following rules which are described in [6]:

1. The number of segments per character must be between 1 and 4.

2. All touching characters should be separated.

Hence, the result is a an over segmented word image. We define *candidate character image* as one or more(maximum 4) of these fine segments. This is the input to the CRF word recognizer. Ligatures extracted from the contour of the word image are strong candidates for segmentation points, but they alone are not sufficient. Hence, concavity features in the upper contour and convexity features in the lower contour are used in addition to ligatures[6].

3.2 Features

State and transition features are extracted for each *candidate character image*. State features include physical features like height, width, aspect ratio and position in the text of the lexicon. In addition we also use as features the *similarity* between the *candidate character image* and the prototype character image for that particular character in the lexicon. To obtain prototype character images, a large training set of characters were clustered using global and local image level features(WMR features)that were extracted. The cluster centers of these images, were converted to a form of codebook by indexing them using their features. This codebook was then used to compute the similarity with the *candidate character images* by comparing their feature values using a variety of metrics(listed in the state feature summary table 3.2). Transition features include the vertical overlap between the two neighboring segments, the difference in the height, difference in width, difference in aspect ratio, the total width of the bigram. Table 3.2 summarizes these. Note that the transition features are for every possible pair of character labels.

3.3 Dynamic programming

The objective of word recognition is to find the word in the lexicon and a grouping of segments for the word which maximizes $P(y, s|x)$ where y is the word, s is the grouping of segments and x is the observed word image.

$$\arg \max_{y \in Lexicon} \left[\arg\max_{s} P(y|x, s) P(s|x) \right] \tag{6}$$

$P(y|x, s)$ is estimated using equation 1 and it is normalized by the length of the word to account for $P(s|x)$. The character segments need to be grouped to find the optimal combination for a particular word in the lexicon. The optimal grouping is the one that maximizes the joint probability as in example equation 5. Rather than evaluate all possible groupings, a dynamic programming based approach can

Feature	Description
Position	Position of character in the lexicon normalized by the length.
Place	Whether the character appears in the beginning, middle or at the end.
Height	Height(pixels) of the *candidate character image*
Width	Width(pixels) of the *candidate character image*
Aspect ratio	Ratio of the height of the character to its width
Euclidean Distance	Euclidean Distance of the character to its prototype cluster center
Manhattan Distance	Manhattan Distance of the character to its prototype cluster center
Tanimoto Distance	Tanimoto Distance of the character to its prototype cluster center
Inner Product	Inner Product of the character WMR features and its prototype cluster center features
KNN Distance	Distance of the character from its 5 nearest prototype images
Height Deviation	Deviation of the height of the character from its expected height
Top Deviation	Deviation of the position of the top of the character from its expected top position
Bottom Deviation	Deviation of the position of the top of the character from its expected top position

Table 1. State features for the CRF model

Feature	Description
Label	Label of the character pair eg. a,b or q,u etc.
Vertical overlap	Vertical overlap(pixels) between the two *candidate character images*
Height difference	Difference in Height(pixels) between the *candidate character images*
Width difference	Difference in Width(pixels) between the *candidate character images*
Aspect ratio difference	Difference in aspect ratio between the *candidate character images*
Bigram width	Sum of individual widths(pixels) of the *candidate character images*.

Table 2. Transition features for the CRF model

be utilized to speed up this process. The approach computes the probability of the *candidate character image* to end at the e_{th} segment. The log joint probability of characters $1 \dots j$ ending at the e^{th} segment in the image is given by the dynamic programming equation as in equation 7.

$$L_j(e) = max_{1 \leq b \leq e} \left[L(b, e) + L_{j-1}(b-1) \right] \qquad (7)$$

where $L(b, e)$ is the log probability of the current character(j^{th} to begin at the b^{th}, and end at the e^{th} segment in the image. $L_{j-1}(b-1)$ similarly represents the log joint probability of the characters $1 \dots j-1$ ending at the $b-1^{th}$ segment. If there are M segments in the image and n characters in the word, then using the above approach, $L_n(M)$ will represent the maximum log joint probability for the particular word. The match between the lexicon and the image, is then computed by normalizing this maximum log joint probability by the length(number of characters) of the lexicon. The normalization is essential since, longer lexicons will tend to have lower probabilities prior to normalization.

4 Experiments

The CEDAR letter database used for the experiments is a collection of unconstrained full page handwritten documents written by a number of writers representative of the US population. Each full page consists of about 120 unique words. The evaluation was done only for lower case characters. The experiments were done in two phases. In the first phase, a total of 1000 words were picked at random and word recognition was performed. The performance of a Segmentation based Dynamic Programming approach[6](SDP) was compared against the CRF method on this data set. The SDP method, is a distance based approach that decides on the optimal grouping of the segments which minimized the total distance between the characters and their prototype cluster centers. Figure 4 shows a comparison of the two methods in terms of the number of times the correct lexicon was identified in the top k ranks. This was performed on a lexicon of 120 words. The CRF method as seen consistently outperforms the SDP method. Figure 4 compares the performance of word recognition using the CRF model with varying lexicon sizes of $10, 120$ and 300. Note that, there is only a small decrease in performance on increasing the lexicon size from 120 to 300. In the second phase of the experiments, 12 full page documents written by different authors were randomly picked and all the words in them were subjected to recognition using the CRF and the SDP[6] method. A comparison of the number of times the correct lexicon(out of 120 words) was *not* identified as the top choice for the two methods is shown in table 4. Again, the CRF based method is seen to give a lower number of recognition errors.

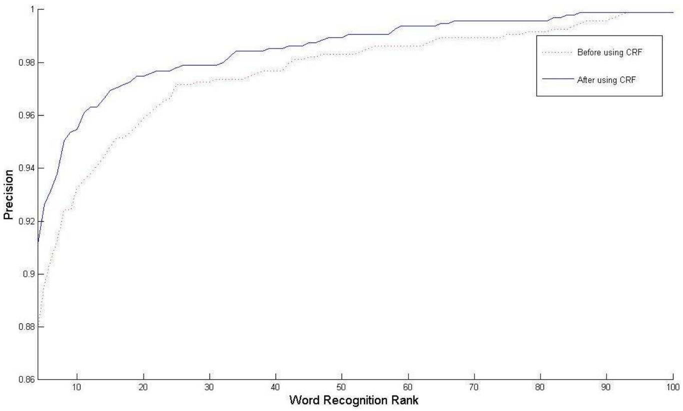

Figure 3. Comparison of the CRF method to Segmentation based Dynamic Programming method[6](SDP in legend). The y axis is the percentage of times the correct lexicon came up in the top k ranks($k \equiv x - axis$)

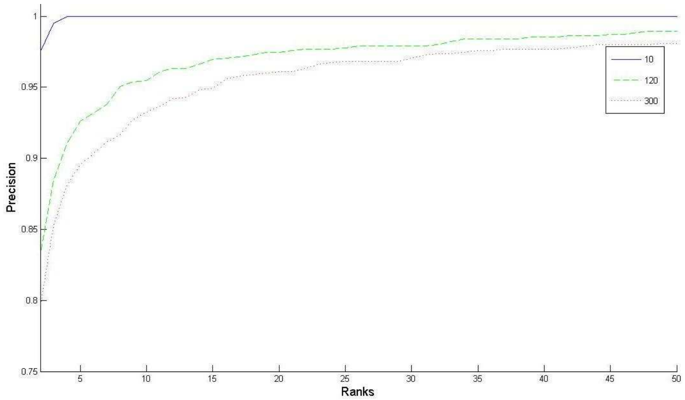

Figure 4. Word recognition precision for different lexicon sizes for the CRF model. The y axis is the percentage of times the correct lexicon came up in the top k ranks($k \equiv x - axis$)

Total number of words	1139
No of incorrectly recognized words by *CRF*	216
No of incorrectly recognized words by *SDP* method	240

Table 3. Comparison of word recognition in full page documents between the CRF and the SDP[6] method.

5 Summary and Conclusion

A novel use of Conditional Random Fields to recognize pre-segmented handwritten words was proposed and discussed. The model utilizes features exploiting the inter-dependencies between neighboring character segments along with features specific to a character image. Language models can be used to correct the word recognition results and further increases the recognition accuracy.

References

[1] J. Lafferty, A. Macullum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequential data,” *Eighteenth international conference on Machine Learning(ICML-2001)*, pp. 282–289, 2001.

[2] S. Feng, R. Manmatha, and A. McCallum, “Exploring the use of conditional random field models and hmms for historical handwritten document recognition,” *Second International Conference on Document Image Analysis for Libraries(DIAL)*, 2006.

[3] S. Kumar and M. Hebert, “Discriminative fields for modeling spatial dependencies in natural images,” *Advances in Neural information processing systems(NIPS-2003)*, 2003.

[4] H. Wallach, “Efficient training of conditional random fields,” *Proc. 6th Annual CLUK Research Colloquium*, 2002.

[5] M. Arivazhagan, H. Srinivasan, and S. Srihari, “A statistical approach to line segmentation in handwritten documents,” *To apper in proceedings of Document Recognition and Retrieval XIV SPIE*, 2007.

[6] G. Kim and V. Govindaraju, “A lexicon driven approach to handwritten word recognition for real-time applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. Vol. 19(4), pp. pp. 366–379, April 1997.