Word Image Matching as a Techique for Degraded Text Recognition

Jonathan J. Hull, Siamak Khoubyari, and Tin Kam Ho
Center of Excellence for Document Analysis and Recognition
Department of Computer Science
State University of New York at Buffalo
Buffalo, New York 14260 USA
hull@cs.buffalo.edu

Abstract

A technique is presented that determines equivalences between word images in a passage of text. A clustering procedure is applied to group visually similar words. Initial hypotheses for the identities of words are then generated by matching the word groups to language statistics that predict the frequency at which certain words will occur. This is followed by a recognition step that assigns identifications to the images in the clusters. This paper concentrates on the clustering algorithm. A clustering technique is presented and its performance on a running text of 1062 word images is determined. It is shown that the clustering algorithm can correctly locate groups of short function words with better than a 95 percent correct rate.

1. Introduction

The recognition of digital images of text is typically performed by segmentation and recognition of isolated characters. This is often followed by a contextual postprocessing step in which word-level context is used to correct for errors in individual character recognition by matching character decisions to words in a dictionary. An alternative approach to word recognition bypasses the character recognition phase and matches the features of whole words to entries in a dictionary [4]. A recent version of this method uses multiple classifiers and outputs a ranking of the dictionary [3]. This technique has proven to be tolerant to a wide range of image noise and is thus ideally suited to the recognition of degraded word images.

The use of global context above the level of individual words to improve text recognition performance has also been considered. For example, the semantics of a constrained domain (chess games) has been used to correct for character recognition errors [1]. Language-level syntax has also been employed to improve word recognition by reducing the number of alternatives for a word's identity based on the hypothesized syntactic categories for two adjacent words [5].

This paper proposes to use an additional global contextual knowledge source that can be derived from passages of running text, namely, the repetitions of words. It is known from the analysis of language statistics that certain words occur more frequently than others. For example, as shown in Table 1, in a typical English language document seven percent of the words are "the" and three percent are "of". Furthermore, the ten most frequent words in English make up more than 23 percent of the word tokens and the top twenty words comprise about 29 percent of the sample. Thus, nearly one third of the words in a passage of text could be recognized by a method that could distinguish only twenty word images.

We propose to utilize such knowledge by matching word images to one another with a clustering algorithm. The clusters are then matched to language statistics to derive word identifications. The advantages of this approach include its tolerance to image degradation. By matching whole word images, the internal featural context of words is used to compensate for features that are missing or distorted. This is an improvement on a previous technique that matched only isolated character images to a-priori statistics [2]. Word-level constraints have also been used in a substitution cipher formulation [8].

Matching of isolated words also integrates naturally with other global contextual knowledge sources. These include probabilities that pairs of words will follow one another. For example, the probability of the word *the* occurring given that the word *of* has occurred is

0-8186-2915-0/92 $3.00

word	freq	word	freq	word	freq	word	freq
THE	0.071	IN	0.022	ON	0.008	AS	0.006
OF	0.032	FOR	0.010	HE	0.007	BY	0.006
AND	0.024	THAT	0.009	AT	0.007	IT	0.005
A	0.024	IS	0.008	WITH	0.006	HIS	0.005
TO	0.023	WAS	0.008	BE	0.006	SAID	0.004

Table 1. The top twenty most frequent words in a sample of over 90,000 words of newspaper reportage.

0.283. Thus, about 28 percent of the images in a cluster of *'of*'s should be followed by words in a cluster of *'the*'s. The transition probabilities for the ten most frequent words in English are shown in Table 2.

The remainder of this paper presents a method for word image matching that is based on an initial clustering of word-level feature descriptions and assignment of possible word identities. Experimental results are presented that show the ability of the algorithm to locate clusters of function words.

2. Overall Algorithm Description

The algorithmic framework in which the word matching process is intended to operate is illustrated in Figure 1. In this approach, an image of a passage of text is segmented into words and the clustering process is applied to find equivalence classes among the word images. Various constraints are used to limit the images included in the clusters. This is followed by a recognition step in which word transition probabilities and other knowledge sources are used to assign identities to the words in the clusters.

The rest of this paper explores the clustering process in detail. An experimental investigation of various aspects of the clustering are discussed.

	the	of	and	a	to	in	for	that	is	was
the										
of	283			40				4		
and	79	5		27	9	10	3	9	6	5
a										
to	130			26				2		
in	173		1	39			1	2		
for	223			88		2		3		
that	147	18		25	3	13	2	2	27	6
is	80	4		102	27	17	3	23		
was	68	1		67	17	24	3	6		

Table 2. Transition probabilities (multipled by 1000) for the ten most frequent words in the entire Brown Corpus

3. Word Image Matching

The word matching process first calculates a feature description for every word in a passage of text. The feature descriptions are then compared and a distance is calculated between them. Images with small distances should be equivalent to one another.

Because this approach should be tolerant to noise such as broken, smeared, touching, or degraded characters, a robust method for matching was chosen that is based on the analysis of the whole shape of a word. A set of features that is referred to as the *stroke direction distribution* is used to describe the shape of a word. It captures the spatial distribution of black pixels belonging to strokes of various directions.

The features are extracted using the local direction contribution method suggested for use with isolated Chinese characters [7]. At each black pixel in the image, the longest continuous run of black pixels in each of the four directions east-west, northeast-southwest, north-south, and northwest-southeast is computed. The pixel is labeled with the direction in which the run length is a maximum. That is, each black pixel is labeled as part of a stroke of one of the four directions.

The word image is first divided into a four-by-ten grid and the number of labeled black pixels of each type in each grid cell are counted. The counts are then normalized by the total number of black pixels in the image. The stroke direction distribution is represented by a 160-dimensional feature vector, which stores the normalized counts of black pixels of each of the four types in the 40 cells. A city-block distance metric is used to compare the feature vectors of two word images.

The resultant feature vectors are input to a hierarchical, bottom-up clustering algorithm. The distances between the feature vectors of all the words are computed and the pair of words with the minimum distance is located. These two words are merged into a cluster if there are no *constraints* that prohibit the merge. A cluster is then represented by its centroid vector. The evaluation and merging process is repeated until no further merges are possible. At this point each cluster should contain identical words.

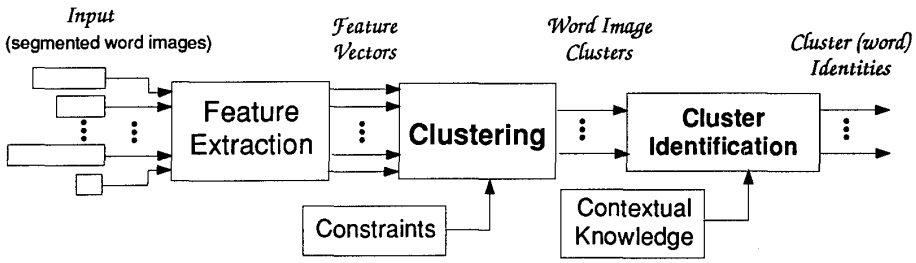

Figure 1. Overall algorithm design.

The constaints used in the clustering procedure are:

1. *Word non-adjacency:* Two clusters are not merged if they contain words that are physically adjacent in the original text sample. This is valid because of the low probability of two adjacent words having the same identity.
2. *Word image length:* Two clusters are merged only if the average lengths of the images they contain are similar.
3. *Word image density:* Two clusters are merged only if the densities (percentage of black pixels) of the images they contain are similar.
4. *Cluster distance:* A merge of two clusters is not performed if the absolute value of the distance between them is above a threshold.
5. *Inter-level distance increase:* A merge of two clusters is not performed if the rate of change in the distance between subclusters and the hypothesized cluster is large. Even though the distance between two clusters is less than a threshold, a large rate of change can still indicate that the words in the two clusters are different.

4. Experimental Environment

A series of experiments were conducted with the word image matching algorithm on images that were generated to represent a portion of the Brown Corpus [6]. This is a sample of over one million words of running text that was constructed to represent modern, edited American English. The corpus is divided into 500 samples of about 2000 words each. The samples were chosen from 15 different subject categories that span a range from Newspaper Reportage to Military Science. The corpus contains graphic codes that allow images to be generated that reasonably approximate the appearance of the original text.

One sample was chosen from the corpus for experimentation because it was a continuous article written by a single author. Many other samples had been split into several disjoint pieces. The subject category was "Belles Lettres" and the specific sample was G02.

The first 1062 words of the sample were printed in an 11 point Times Roman font on plain white paper by a laser printer. The resultant pages were then scanned at 200 pixels per inch in 8 bit grayscale on a desktop digitizer and binarized.

5. Performance Analysis

The ability of the word matching procedure to reliably locate groups of equivalent images within the same passage of text was analyzed. Of primary interest was the ability to locate clusters of short function words. These can be a difficult problem for isolated character recognition algorithms and postprocessing techniques which have little context to rely on.

The results of an experiment on the 1062 word images described earlier are shown in Table 3. Two error rates illustrate performance. The internal error rate is the percentage of words in a cluster that are different from the most frequent word. The external error rate is the percentage of a given word class that is not contained in the primary cluster for that word.

The results show that all of the ten most frequent words are grouped into separate clusters. In some cases a few extra words are included in the clusters. However, none of the top ten words are split into other clusters,

	all constraints					w/o non-adjacency			
word	no. in sample	cluster size	% error internal	no. in cluster	% error external	cluster size	% error internal	no. in cluster	% error external
the	84	84	0.0	84	0.0	98	14.3	84	0.0
of	62	62	0.0	62	0.0	62	0.0	62	0.0
and	18	23	21.7	18	0.0	27	33.3	18	0.0
a	20	20	0.0	20	0.0	20	0.0	20	0.0
to	26	26	0.0	26	0.0	29	10.3	26	0.0
in	27	30	10.0	27	0.0	27	0.0	27	0.0
for	11	11	0.0	11	0.0	98	88.8	11	0.0
that	9	10	10.0	9	0.0	19	52.6	9	0.0
is	13	13	0.0	13	0.0	13	0.0	13	0.0
was	19	20	5.0	19	0.0	27	29.6	19	0.0
avg.			4.7		0.0		22.9		0.0

Table 3. Clustering results for sample G02 with and without the word non-adjacency constraint

i.e., there is a zero percent external error rate.

The effect of removing the first constraint (word non-adjacency) from the clustering procedure is also shown in Table 3. The internal error rate rises to an average of 22.9 percent over all the clusters. However, the zero percent external error rate is maintained. It should be noted that if the correct word identities were assigned to the clusters, only 10 images out of the 1062 would have been incorrectly recognized when all the constraints were used. However, 131 words would have been incorrectly recognized when the non-adjacency constraint was removed.

6. Discussion and Conclusions

A word matching procedure that locates equivalences between groups of word images in a passage of text was presented. This method is more tolerant to image noise than a traditional word recognition approach since determining whether two words are the same can be much easier than determining their classification.

An clustering algorithm was presented that locates all the groups of equivalent word images in a passage of text. A feature description based on an analysis of the shape of a whole word was used as well as a set of constraints about both the language and the images themselves. Future work on this technique will include investigation of multiple feature sets and the use of other knowledge sources, such as language syntax and semantics to improve performance.

References

1. H. S. Baird and K. Thompson, "Reading Chess," *IEEE Transactions on Pattern Analysis and Machine Intelligence 12* (1990), 552-559.
2. R. G. Casey and G. Nagy, "An autonomous reading machine," *IEEE Transactions on Computers C-17*, 4 (May, 1968).
3. T. K. Ho, J. J. Hull and S. N. Srihari, "Word recognition with multi-level contextual knowledge," *International Conference on Document Analysis and Recognition*, Saint Malo, France, September 30-October 2, 1991, 905-915.
4. J. J. Hull, "Hypothesis generation in a computational model for visual word recognition," *IEEE Expert 1*, 3 (Fall, 1986), 63-70.
5. J. J. Hull, "Feature selection and language syntax in text recognition," in *From Pixels to Features*, J. C. Simon (editor), North Holland, 1989, 249-260.
6. H. Kucera and W. N. Francis, *Computational analysis of present-day American English*, Brown University Press, Providence, Rhode Island, 1967.
7. S. Mori, K. Yamamoto and M. Yasuda, "Research on machine recognition of handprinted characters," *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-6*, 4 (July 1984), 386-405.
8. G. Nagy, S. Seth and K. Einspahr, "Decoding substitution ciphers by means of word matching with application to OCR," *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-9*, 5 (September, 1987), 710-715.