

A New Feature Ranking Method in a HMM-Based Handwriting Recognition System

Sijun Kang and Venu Govindaraju
CEDAR, State University of New York at Buffalo

Abstract

In this paper we propose a new feature ranking method in a recognition system, by introducing the concept of the effectiveness of the distinguishing power of features and considering the correlation among features. To find the subset of most important features, first the best feature can be identified by its effective distinguishing power and put in an empty feature set. Then each of the remaining features is ranked based on their effective distinguishing capacity contribution and the highest-ranked feature is added to the selected subset. This process is repeated till the performance of the system reaches its peak or the effective distinguishing contribution falls below a certain value. The application of this method to an existing handwriting recognition system showed strong support for our methodology of feature ranking.

1. Introduction

Although many methods have been developed in the area of offline cursive word recognition, most of them contain a feature extraction module and a recognition module in their functionality blocks. No matter what preprocessing and segmentation methods are used, the features are the only input fed into the recognition system. The recognition result is decided by how much information the feature set contains and how well the recognition system can make use of the information provided by the feature set. Once the structure of the recognition module is fixed, extracting features out of the word image and selecting the best feature set becomes a critical task in building an efficient and accurate recognizer.

Hidden Markov Model (HMM) based recognition is the popularly used decoding method in offline handwriting word recognizers. Many feature

extraction techniques suitable for HMM algorithms have also been developed [1,2,3,5]. To decide which subset of features to use, one needs a methodology to rank them without actually evaluating them in the system because the number of subsets one can build grows exponentially with the number of total features.

This paper introduces a new measure for discrete feature ranking with the objective of choosing the best subset of features which can contribute the most in the final performance of a HMM-based recognition system. In section 2, we provide the formal definition of this new measurement. The description of its calculation with an existing recognition system and some experiments to support the new concept are presented in Section 3. Finally some conclusions and plans for future research are discussed.

2. A new feature ranking measure for feature set construction

We consider an existing HMM-based handwritten word recognizer [7] which uses high level structural features, such as loop, cross, cusp, etc. It ranks the words in a lexicon by matching the sequence of extracted features with the lexicon entries. We observe that while the discriminative power of a feature can be high, it might not contribute significantly to the final recognizer as compared to a feature with lesser discriminative power. One possible reason is the feature may not have been frequently observed in the word images and thus helps only in a very small number of cases. Another reason could be that highly correlated features are already in the feature set. In light of the first observation, we propose the concept of effective distinguishing power (ED) of a feature as a measure of feature ranking. Based on the second observation, the concept of effective distinguishing capacity contribution of a feature to a feature set (not containing this feature) is introduced.

2.1. Conditional perplexity, distinguishing power and effective distinguishing power

Conditional perplexity is an indicator based on the statistical notions of entropy and perplexity from information theory. It has been used in the speech recognition field by Bahl et al.[8] to evaluate the difficulty of a specific recognition task. It has also been used in the handwriting recognition domain by Grandidier et al.[4] to evaluate the discriminative power of a single feature in their strategy of improving feature sets in a discrete HMM-based handwriting recognition system.

Shannon defined the measure of entropy H as:

$$H = -\sum_i p_i \log p_i \quad (1)$$

where p_i is the probability of symbol i in a communication channel. H is used to determine the capacity of the channel required to transmit the signal. According to [4], the conditional entropy of a feature f_j is defined as:

$$H(f_j) = -\sum_{i=1}^{N_c} p(c_i | f_j) \cdot \log p(c_i | f_j) \quad (2)$$

where $\{c_i\}$ are the classes considered in the modeling and N_c is the number of classes. $H(f_j)$ reaches its maximum value of $\log N_c$ when:

$$p(c_i | f_j) = \frac{1}{N_c} \quad \forall c_i \quad (3)$$

In this case, f_j is considered as useless since there is no information embedded in feature f_j to discriminate between the N_c classes. $H(f_j)$ reaches its minimum value of 0 when there exists a class c_i such that:

$$p(c_i | f_j) = 1 \quad \text{and} \quad p(c_k | f_j) = 0 \quad \forall k \neq i \quad (4)$$

Because of the property mentioned above, $H(f_j)$ can be used to quantify the capability of feature f_j to discriminate between the classes $\{c_i\}$.

The conditional perplexity $PP(f_j)$ of a feature f_j is a measure derived from entropy $H(f_j)$ as follows:

$$PP(f_j) = 2^{H(f_j)} \quad (5)$$

This function varies between 1 and N_c ; thus it can be directly compared to the number of classes c_i involved. This is the advantage of using perplexity instead of the entropy.

2.2. Distinguishing power and effective distinguishing power

Definition: the distinguishing power of a feature f_j is defined as:

$$D(f_j) = EN_c - PP(f_j) \quad (6)$$

where EN_c is the number of classes possibly distinguished in the recognition process by f_j .

Definition: the effective distinguishing power of a feature f_j is defined as

$$ED(f_j) = D(f_j) \cdot \sum_{i=1}^{N_c} [p(f_j | c_i) \cdot p(c_i | f_j)] \quad (7)$$

where $[p(f_j | c_i) \cdot p(c_i | f_j)]$ represents the factor in which the distinguishing power of f_j can effectively help in recognizing class c_i .

2.3. Correlation angle between two features and effective distinguishing power contribution of a feature to a feature set

Definition: correlation angle θ between feature f_j and feature f_i with correlation value r ($-1 \leq r \leq 1$):

$$\theta(f_i, f_j) = \arccos(r) \quad (8)$$

Definition: the effective distinguishing capacity contribution $\Delta ED(f_j, E)$ of feature f_j to a feature set $E = \{f_i \mid i = n_1, n_2, \dots, n_k\}$, where $f_j \notin E$ and E is not empty:

$$\Delta D(f_j, E) = \min_{f_i \in E} ED(f_j) \cdot \sin \theta(f_j, f_i) \quad (9)$$

In the special case when E is empty, i.e., $E = \Phi$,

$$\Delta ED(f_j, \Phi) = ED(f_j) \quad (10)$$

2.4 Feature set construction based on the new measurements

With these new concepts, we can use the following new strategy to construct a high performance feature subset from a given feature set E .

- First, initialize the selected feature subset $S = \Phi$
- Then, repeatedly add the feature f in E which has the maximum value of effective distinguishing capacity contribution $\Delta ED(f_j, S)$ till this contribution reaches a low value threshold τ or the performance of the recognizer reaches its peak.

By using this strategy, we can construct a feature set with high collective distinguishing power and at the same time have having low correlation with each other.

3. Experiments

3.1. Experiment procedure

To show the effectiveness of the above new measurements, we design the following two experiments on a HMM-based handwriting word recognizer [7] with 16 structural features [6] in its feature space.

Experiment A: rank the features according to their conditional perplexity and construct a sequence of 9 feature sets by deleting the best feature from each feature set in each iteration. Then train and test the recognizer with each feature set separately.

Let us use $f_{j1}, f_{j2}, \dots, f_{j16}$ to denote the 16 features ranked by conditional perplexity and S_0, S_1, \dots, S_8 denote the 9 feature sets mentioned above, then we have

$$\begin{aligned} S_1 &= S_0 - \{f_{j1}\} \\ S_2 &= S_1 - \{f_{j2}\} \\ &\dots \\ S_8 &= S_7 - \{f_{j8}\} \end{aligned}$$

Experiment B: Rank the features according to their effective distinguishing power or their effective distinguishing capacity contribution and select the 8 best features one by one according to the strategy described above. Then construct a sequence of 9

feature sets with the same method used in experiment A. After that, the recognizer is trained and tested with each feature set separately. Let us denote the features ranked this way as $f_{k1}, f_{k2}, \dots, f_{k16}$ and the feature sets as T_1, \dots, T_8 .

Experiment C: Combine $f_{j1}, f_{j2}, \dots, f_{j8}$ to form a feature set S ; combine $f_{k1}, f_{k2}, \dots, f_{k8}$ to form a feature set T , then test the recognizer with S and T separately.

3.2. Experiment results

The experiments were done on a set of 988 word images extracted from CIIR historical document database [9] with alphabetic characters only. 494 of the images are used as the training set and the remaining 494 as the testing set. The lexicon is a set of 1000 words constructed with the image truth and randomly chosen words for all others. The classes used in the definition are the collection of all upper case and lower case letters. The 16 features and their symbolic representation are listed in following table:

Table 1 Feature symbols and description

UL	upward loop
ULC	upward long cusp
USC	upward short cusp
UA	upward arc
ULA	upward left-terminated arc
URA	upward right-terminated arc
C	circle
DL	downward loop
DLC	downward long cusp
DSC	downward short cusp
DA	downward arc
DLA	downward left-terminated arc
DRA	downward right-terminated arc
C	cross
B	bar
GAP	gap

All the feature measurements are computed from the training set. Since the truth of each training example is known, the backtracking procedure of the Viterbi algorithm in the HMM recognizer is used to automatically associate features with letters in a word. The probabilities are computed from the occurrences of all feature/class associations. The measurements of all features are shown in Table 2 in ranks according

to conditional perplexity. The first 8 features are selected for experiment A. Another 8 features are selected for experiment B according to their effective distinguishing power and are marked in the rightmost column.

The recognition performances for each of the feature sets constructed for all experiment are presented in Table 3. By comparing the performance of the feature sets created for experiment A and those created for experiment B, one can observe that the performance drops significantly faster in experiment B while more and more features with high effective distinguishing power are removed from the feature set. After cutting the 8 features with most effective distinguishing power, the top 1 accuracy drops to 7.0%, while the top 1 accuracy only degrades to 37.0% after cutting the 8 features with the best discriminative power according to the conditional perplexity. The results of experiment C show that the feature set S, consisting of the 8 features selected according to their conditional perplexity, achieves only 6.4% in top 1 accuracy while feature set T, consisting of the 8 features selected according to their effective distinguishing power, yields a higher accuracy of 42.9%.

4. Conclusion and future direction

We have introduced a new feature ranking method in a handwriting recognition system, by introducing the concept of the effectiveness of the distinguishing power of features and considering the correlation among the features. Our experiments show a strong support for our concept of feature ranking.

5. References

- [1] R. Plamondon and S. Srihari, *On-line and off-line handwriting recognition: A comprehensive survey*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, pp. 63–84, January 2000.
- [2] T. Steinherz, E. Rivlin, and N. Intrator, *Offline cursive script word recognition – a survey*, International Journal Document Analysis and Recognition, vol. 2, pp. 90–110, 1999.
- [3] A.L. Koerich, R. Sabourin, C.Y. Suen, *Large vocabulary off-line handwriting recognition: A survey*, Pattern Anal Applic (2003) 6, pp. 97-121.
- [4] F. Grandidier and R. Sabourin, C.Y. Suen, M. Gilloux, *A new strategy for improving feature sets in a discrete HMM-based handwriting recognition system*, Proceedings of the Seventh International Workshop on Frontiers in Handwriting Recognition, September 11-13 2000, Amsterdam, pp. 113-122.
- [5] A. El-Yacoubi, M. Gilloux, R. Sabourin, and C. Y. Suen, *An HMM-based approach for off-line unconstrained handwritten word modeling and recognition*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, pp. 752–760, August 1999.
- [6] H. Xue and V. Govindaraju, *Building skeletal graphs for structural feature extraction on handwriting images*, International Conference on Document Analysis and Recognition, (Seattle, Washington), pp. 96–100, September 2001.
- [7] H. Xue and V. Govindaraju, “*A Stochastic Model Combining Discrete Symbols and Continuous Attributes and Its Application to Handwriting Recognition*”, IEEE Transaction on Pattern Analysis and Machine Intelligence (submitted).
- [8] L.R. Bahl, F. Jelinek and R.L. Mercer, *A Maximum Likelihood Approach to Continuous Speech Recognition*. IEEE Trans. On PAMI 5 (1983) pp. 179-190
- [9] Data sets containing word images from the George Washington collection, Center for Intelligent Information Retrieval, University of Massachusetts Amherst.

Table 2: Feature measurements and subset selection in Exp A and B

f_i	Occurs	PP(f_i)	D(f_i)	ED(f_i)	Δ ED(f_i)	Exp A	Exp B
BAR	3	3.0000	42.0000	0.2212	0.2211	f_{j1}	
C	288	7.2399	37.7601	17.2161	17.2161	f_{j2}	f_{k1}
DL	52	9.2590	35.7410	4.5715	4.5219	f_{j3}	
UL	120	11.5933	33.4067	7.7493	7.6948	f_{j4}	f_{k8}
UA	390	14.2837	30.7163	11.9283	11.7610	f_{j5}	f_{k2}
URA	191	14.3838	30.6162	7.4578	7.3261	f_{j6}	
DRA	265	14.8282	30.1718	5.0067	4.9348	f_{j7}	
X	39	16.5765	28.4235	2.1131	2.1101	f_{j8}	
DLA	88	17.5674	27.4326	6.3318	6.3072		
ULA	78	17.9354	27.0646	2.0089	1.9930		
DA	655	19.5442	25.4558	9.8354	9.1669		f_{k7}
USC	770	21.0360	23.9640	10.2247	9.7336		f_{k5}
DSC	642	21.3375	23.6625	9.9886	9.6196		f_{k6}
ULC	813	21.5167	23.4833	11.2972	10.9030		f_{k4}
DLC	717	22.7323	22.2677	9.9225	7.2148		
GAP	1191	25.0921	19.9079	11.6606	11.6485		f_{k3}

Table 3: Experimental feature subsets and their recognition performances

Feature set	Features removed from its predecessor	Performance (top 1)	Performance (top 10)
S0		61.8%	88.1%
S1	BAR	61.6%	87.5%
S2	C	53.7%	83.5%
S3	DL	53.2%	81.8%
S4	UL	49.4%	82.1%
S5	UA	42.9%	78.6%
S6	URA	42.4%	75.9%
S7	DRA	37.8%	72.7%
S8	X	37.0%	68.9%
T1	C	54.0%	82.7%
T2	UA	49.4%	79.7%
T3	GAP	50.8%	80.2%
T4	ULC	37.5%	77.0%
T5	USC	39.4%	73.2%
T6	DSC	18.9%	53.5%
T7	DA	8.6%	40.8%
T8	UL	7.0%	35.4%
S		6.4%	23.5%
T		42.9%	78.6%