A methodology for deriving probabilistic correctness measures from recognizers

Bouchaffra, D.; Govindaraju, V.; Srihari, S.;
Dept. of Comput. Sci., State Univ. of New York, Buffalo, NY

This paper appears in: Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on
Issue Date : 23-25 Jun 1998
On page(s): 930 - 935
Meeting Date : 23 Jun 1998-25 Jun 1998
ISSN : 1063-6919
Print ISBN: 0-8186-8497-6

INSPEC Accession Number: 5985950
Digital Object Identifier : 10.1109/CVPR.1998.698716
Date of Current Version : 06 August 2002

ABSTRACT

This paper describes the derivation of probability of correctness from scores assigned by most recognizers. Motivation for this research is three-fold: (i) probability values can be used to rerank the output of any recognizer by using a new set of training data; if the training data is sufficiently large and representative of the test data, the recognition rates are seen to improve significantly, (ii) derivation of probability values puts the output of different recognizers on the same scale; this makes comparison across recognizers trivial, and (iii) word recognition can be readily extended to phrase and sentence recognition because the integration of language models becomes straightforward. We have conducted an extensive set of experiments. The results show a reranking of recognition choices based on the derived probability values leading to an enhancement in performance

A Methodology for Deriving Probabilistic Correctness Measures from Recognizers

Djamel Bouchaffra, Venu Govindaraju and Sargur N. Srihari

{bouchaff,govind,srihari}@cedar.buffalo.edu
Center of Excellence for Document Analysis and Recognition (CEDAR)
Department of Computer Science (CEDAR)
State University of New York at Buffalo
Amherst, New York 14228–2567

Abstract

This paper describes the derivation of probability of correctness from scores assigned by most recognizers. Motivation for this research is three-fold: (i) probability values can be used to rerank the output of any recognizer by using a new set of training data; if the training data is sufficiently large and representative of the test data, the recognition rates are seen to improve significantly, (ii) derivation of probability values puts the output of different recognizers on the same scale; this makes comparison across recognizers trivial, and (iii) word recognition can be readily extended to phrase and sentence recognition because the integration of language models becomes straightforward.

We have conducted an extensive set of experiments. The results show a reranking of recognition choices based on the derived probability values leading to an enhancement in performance. The performance of many different digit recognizers improved by 1-4% points on a blind set of images.

1 Introduction

We present in this paper the groundwork for the use of Bayesian methodology in integration of recognizers with any subsequent processing by deriving meaningful probabilistic measures from recognizers. We also address the important notion of scalability of scores [9] and show how scores from different recognizers can be compared. Such normalization of scores under a common scale promotes effective combination of recognizers. Finally, it is our conjecture that the probability values themselves are more precise in what they convey than the typically output scores of recognizers.

Previously, researchers have assumed in majority of the work [10, 15], that the recognizer merely provides a ranked list of classes for each input pattern and is associated with distance measures which are largely ignored by subsequent stages. Our interest is in deriving probabilistic correctess measures from word recognizers, such that their output will be suitable for integration with subsequent stages such as linguistic processing [1, 2, 3, 14] in sentence recognition and classifier combination [9, 13]. Further, we expect that the recognition rate to improve because of the additional re-training required by our methodology.

1.1 Word Recognition Background

In this paper we will draw examples from handwritten word recognition to illustrate our point. However, the methodology described is equally applicable to all pattern classification tasks. The practical implementation of word recognizers use a lexicon of limited size (Figure 1). Given the image of a handwritten word and a lexicon of possible words, the task is one of ranking the lexicon based on the "goodness" of match between each lexicon entry and the word image. Typically, the word recognizer computes a measure of "similarity" between each lexicon entry and the word image and uses this measure to sort the lexicon in descending order of the similarity measure [6, 12]. The lexicon entry with the highest similarity is the top choice of the recognizer. The top m choices are often referred to as the confusion set ($\mathcal{A}$). They constitute the lexicon entry that is the true match to the input image and the neighboring lexicon entries that are similar in some feature space.

The similarity measures returned by recognizers are also referred to in the literature as "confidence scores" and "distance measures": $C(\omega_i|X)$, the confidence of the recognize on class ω_i by analyzing pattern X. Given the same pattern, two recognizers

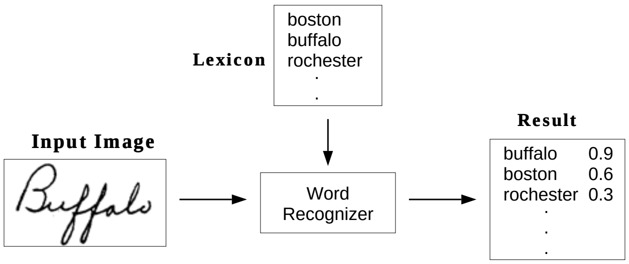


Figure 1: I/O behavior of a word recognizer. Input is the word image and a lexicon of possible choices. Output is the lexicon sorted by some confidence measure. The top m choices form the neighborhood or the confusion set.

may return the exact same ranking of the lexicon with different associated scores. Further, even the scores of the recognizers can be identical, but the information conveyed is still different. In order to interpret what the significance of a particular score returned by a recognizer is, one needs to study the behavior of the recognizer over several input patterns.

1.2 Intutive Problem Description

Following is the analogy that should clarify the point being made. Teacher A and Teacher B want to evaluate the proficiency of students in Math. Both teachers give an exam to n students ($S_1, \ldots S_n$) and grade the students' responses. Teacher A examines the student's response and gives a score of 80% to the student (S_j) with the best answers (other students get lower scores). Teacher B gives a score of 90% to the same student (S_j) which is also the highest score in his grading. Observing the above, the following questions become pertinent.

1. Is student S_j the most proficient in Math among the students examined?
2. Is the opinion of Teacher B about S_j's proficiency stronger than that of Teacher A, given that Teacher B gave the student a higher score?
3. When Teacher A (B) gives the best student a score of 80% (90%), is he correct in his selection of the best student?

An intutive answer to the first question would be that indeed student S_j is the most proficient among the students $S_1 \ldots S_n$, if he scores the highest consistently over many exams. This eliminates the possibility of chance occurrence. However, it must be ensured that all the exams administered are of equal difficulty.

An intutive answer to the second question is that the information given is insufficient to make the determination. The grading policies of both teachers have to be studied over a large number of tests to quantify their grading behavior. In one teacher's mind, 80% can mean more than what 90% means to another teacher. The reason that comparison across teachers is difficult is because each teacher has a different notion of what a particular score means, *i.e.*, the scores are on a different scale.

It is the third question that we address in this paper- the students are the lexicon entries, the teacher is the word recognizer, and the input image is one exam. Our objective is to assess to what degree the recognizer is correct in its ability to label word images. In particular, we are interested in the ability of the recognizer to choose the best class. The degree of correctness over a large number of trials should provide the probability of correctness. It is fair to assume that the probability of correctness increases as scores increase. However, the probability of correctness does not necessarily become 1 when the score of the top choice is 100%.

1.3 Motivation

No matter what the particular algorithm may be, all word recognizers invariably compute the "goodness" of match between the image and the symbolic representation of the word. While the distance measures returned by recognizers are adequate for most applications where recognition is the final stage of the application, we believe that there is a need for true probabilistic measures. The need for deriving probability values for the purpose of expressing signal and language information in a single framework has recently been underscored by Hull [8].

This research also makes possible the notion of having a common scale. Irrespective of how different recognizers, using different paradigms, arrive at their distance scores (or confidence values), when they are converted to probabilities as described above, they are all at a common scale because they have all been derived by the same process.

Finally, the most important implication of this research from a practical viewpoint is that it describes a methodology of re-ranking the output of a recognizer based on probability values (derived from the confidence scores) which opens the possibility of improving the recognition rate of the recognizer. In fact, our expriments with digit recognizers (reported in section 4) show that the improvement is very pronounced in inherently poor recognizers. This is quite remarkable given the fact that we have no access to the features and classification process of the recognizer. The improvements are due to additional training brought to bear upon the problem which reveals insights into the behavior of the recognizer.

1.4 Paper Organization

Section 2 outlines the precise statement of the problem in general mathematical terms and describes the difficulty in finding a solution. We elaborate the methodology developed for deriving probability values from confidence scores returned by recognizers. Section 3 presents the application of sentence recognition to support our claim that probability values from word recognizers allow natural integration with other stages of a recognition engine. Section 4 describes the experiments supporting our methodology for both words as well as digits.

2 DPS: Deriving Probability given Score

Our task is to take a recognizer given as a blackbox, observe its behavior on an input pattern, and derive probability of correctness of its output. The general problem can be mathematically described as follows.

Let $\Omega = \{\omega_1, \omega_2, ..., \omega_c\}$ be the set of c classes and X be the input pattern. We define P_Ω^X as the probability distribution on Ω generated by the pattern X. Traditionally, a recognizer returns $C(\omega_i|X)$, where C is the "confidence" measure. Our objective is to derive $P(\omega_i|X)$, the probability for a class to be the true class (top choice) given the input pattern X. Using Bayes' rule

$$P(\omega_i|X) = \frac{P(X|\omega_i) \times P(\omega_i)}{P(X)} \tag{1}$$

Without any additional information, we can reasonably assume that all classes ω_i, have the same chance of being the truth. In other words, the "truthing" distribution of classes is uniform and, $P(\omega_i)$ can be approximated by $\frac{1}{c}$. However, it is difficult to compute the *state conditional* probability $P(X|\omega_i)$. It corresponds to the probability of an input pattern given a particular class as a truth. $P(X)$ is difficult to compute because there are infinite number of possible shapes in which X can appear.

We circumvent the difficulty in computing $P(\omega_i|X)$, by falling back to the classic definition of probability. This definition records the frequency of an event of interest over many trials. However, given a particular pattern, a recognizer provides the same ranking of classes, no matter how many times the trials are repeated. In other words, there is no notion of *randomness* inherent in this process. One way of introducing randomness is to generate repeated trials with a large number of patterns that are "similar" to the input pattern. The notion of "similarity" used will be defined further. A set of similar patterns, $\overline{X}$ are generated to constitute the trials of the process. $\overline{X}$ is built during a re-training phase (given that the recognizer in the blackbox was trained before) and contains the input pattern X and its neighbors. In our analogy, this is the process of constructing more exams for the students where all the exams are of the same difficulty level as the first. It is this "re-training" that brings additional "knowledge" to bear upon the problem leading to potentially improved recognition rate.

Using the basic definition of probability we can estimate $P(\omega_i|X)$ by counting the number of times ω_i is the top choice in the ranked lexicon during $|\overline{X}|$ trials where $|\overline{X}|$ is the number of input patterns contained in the set $\overline{X}$.

$$\hat{P}(\omega_i|X) \simeq \frac{\sum_{u \in \overline{X}} \zeta_\Omega^u(\omega_i)}{|\overline{X}|} \tag{2}$$

where:

$$\zeta_\Omega^u(\omega_i) = \begin{cases} 1 & if\ \omega_i \in \Omega\ is\ the\ top\ choice\ given\ u \in \overline{X} \\ 0 & otherwise \end{cases}, \tag{3}$$

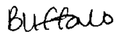
Since the identity of X is not known, it is not easy to find the neighbors of X. The task becomes tractable if we make the assumption that the true class is always present in the confusion set $\mathcal{A}$ (top m choices output by the recognizer). Based on this assumption, the neighborhood set $\overline{X}$ can be written as

$$\overline{X} = \bigcup_{i=1}^{i=m} \overline{X}_j^i, \tag{4}$$

where $\overline{X}_j^i$ represents the subset of images that belong to class ω_i and $C(\omega_i|X \equiv\ image\ of\ class\ \omega_i) \in [j-\delta, j+\delta]$ and δ is a small number described in the following subsection.

2.1 Re-training

$C(\omega_i|X \equiv\ image\ of\ class\ \omega_i)$ is computed for every image of every class and serves as the image's tag. Thus, all the images belonging to a particular class can be quantized into clusters based on the range of score received. We can choose the δ interval so that the images within the cluster of a particular class are all "similar" in the feature space of the recognizer used. One cluster from the images of each class in $\mathcal{A}$ will form the set of trial images $\overline{X}$ for retraining.



Buffalo	5.32
Beauty	4.10
Bounty	2.80
Niagara	2.33
North	1.25

Figure 2: ASCII words that form set $\mathcal{A}$ and their scores, given the word image ``Buffalo''.

For the purpose of illustrating our methodology, we have adopted the following clustering procedure. The difference between the maximum and minimum scores received is used to find the range of scores for images of one class. This range was equally divided into equal intervals of size 2δ. Images with score j belong to cluster X^i_j, where X belongs to the class ω_i.

533 samples of images of the word "BUFFALO" were taken. A word recognizer [12] was invoked with the following 2 inputs: (i) the sample of image of "BUFFALO" and (ii) a single lexicon entry: BUFFALO. The idea is to tag each image with the score $C(\omega_{BUFFALO}|X\ belongs\ to\ \omega_{BUFFALO})$, the distance between the input sample and the word recognizer's notion of a prototype of "BUFFALO". Table 1 shows the clusters obtained by choosing $2\times\delta = 10$.

Other clustering techniques [5, 7, 11] can be used as well. Some interesting observations become noteworthy. First, we notice that the word recognizer does reflect our intutive sense of similarity. Writing styles seem to get sloppy as the cluster distance increases. Second, noise in the image contributes to a lower score even if the writing style is quite good.

If the classes in the confusion set $\mathcal{A}$ are as seen in Figure 2 with their respective scores, then $\overline{X}$ is

$$\overline{X}^{Buffalo}_{5.32} \cup \overline{X}^{Beauty}_{4.10} \cup \overline{X}^{Bounty}_{2.80} \cup \overline{X}^{Niagara}_{2.33} \cup \overline{X}^{North}_{1.25}. \quad (5)$$

We can evaluate probability values as:

$$P(\omega_{BUFFALO}|X\ unknown\ image) \simeq \frac{Nb^{|\overline{X}|}_{BUFFALO}}{|\overline{X}|}, \quad (6)$$

where $Nb^{|\overline{X}|}_{BUFFALO}$ is the number of times "BUFFALO" is present as the top choice among $|\overline{X}|$ trials.

3 Sentence Recognition Application

Sentence recognition applications deal with: (i) an input sentence image, (ii) word recognition results for each word image in the sentence, and (iii) a language source expressed by the word/part-of-speech distributions [1, 2, 3, 4].

A sentence image $I = i_1 i_2 ... i_n$ is assigned to the sequence of clusters $\overline{I} = \overline{I}_1 \overline{I}_2 ... \overline{I}_n$ where each $\overline{I}_k$, $\{k \in [1..n]\}$ is a cluster containing the k-th input word image and n is the number of handwritten word images. This provides a mapping from the space of word images to the space of clusters of word images.

The problem consists of determining an optimal word/part-of-speech path $(W,T)^*$ given the input sentence image $I = i_1 i_2 ... i_n$. The optimal path $(W,T)^*$ can be written as:

$$(W,T)^* = arg \max_{(W,T)} P((W,T)|I) \quad (7)$$

Equation 8 illustrates the use of a Bayesian framework to integrate signal and language can be written as a product of two terms.

The language part of the model is given by Equation 9.

4 Experimental Results

We have conducted experiments with handwritten word recognizers and handwritten digit recognizers. In case of handwritten word recognizers, we cannot report actual improvement of recognition rates because of the sparseness of data. It is difficult to find many samples of the same word. On the other hand, samples of handwritten digits are abundantly available.

4.1 Word Recognition

There were 1621 images of 27 different words collected for the experiment. Following is the procedure adopted.

1. Clusters are created by submitting each image in the re-training data set (several samples of the same ASCII exist) to the word recognizer with only the exact truth in the lexicon.

2. The samples are put in 5 different clusters by simply dividing the range of scores into 5 zones and allowing each sample image to fall into a zone.

3. The word recognizer runs on each image in the test set.

4. For each of the top m choices returned, all samples of images corresponding to the same class and the score range are put together into a single group ($\overline{X}$).

5. Each of the sample images in this group are sent to the word recognizer.

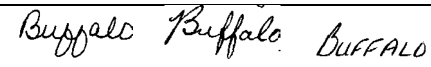
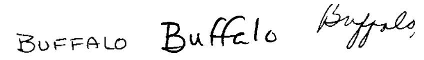
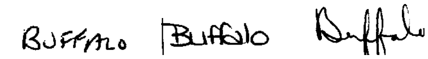
Cluster$_j$	Range of Confidence	Cluster Size	Samples
$\overline{X}_1^{Buffalo}$	0.00 ... 1.99	8	
$\overline{X}_2^{Buffalo}$	2.00 ... 3.99	326	
$\overline{X}_3^{Buffalo}$	4.00 ... 5.99	173	
$\overline{X}_4^{Buffalo}$	6.00 ... 7.99	22	
$\overline{X}_5^{Buffalo}$	8.00 ... 9.99	4	

Table 1: 533 images of Buffalo divided into 5 clusters based on the confidence scores received during the re-training phase; δ used is 1

$$(W,T)^* = arg \max_{(W,T)} \left[\prod_{j=1}^{j=n} L((w_j,t_j),(w_{j-1},t_{j-1}),(w_{j-2},t_{j-2})) \times P(w_j|i_j) \right] \tag{8}$$

6. Probability values are computed based on the presence of the word in the top choice.

Following tables show the output of the word recognizer on 3 examples. The truth (identity) is unknown to the testing program. We display the probability value for each ASCII word and rerank the confusion set with respect to these probability values.

TRUTH: "EVERYTHING"

TOP1	ASCII	SCORES	PROBABILITIES
1:	insurance	6.022	0.364
2:	everything	6.087	0.273
5:	deductible	7.036	0.273
4:	valuables	6.968	0.091
3:	completely	6.954	0.000

TRUTH: "ESTIMATES"

TOP1	ASCII	SCORES	PROBABILITIES
1 :	insurance	6.340	0.357
4 :	estimates	6.736	0.214
3 :	valuables	6.693	0.071
2 :	family	6.533	0.071
5 :	expensive	6.967	0.000

TRUTH: "RUINED"

TOP1	ASCII	SCORES	PROBABILITIES
1 :	ruined	5.387	0.353
3 :	insurance	6.033	0.235
4 :	valuables	6.101	0.235
2 :	between	5.995	0.059
5 :	caused	6.271	0.000

4.2 Digit Recognition

Five digit recognizers [13] (GSC, GRAD, KP, HISTO, POLY) were tested for recognition rate improvement. DPS (Derivation of Probability given Score) uses the methodology described in this paper to improve upon the output provided by the digit recognizers. The performance improvement is shown in Table 3 for different blind image sets.

5 Summary

We have presented a methodology for deriving probability values given recognition scores assigned to classes. We have argued that these probability values are more informative and useful in a variety of applications including, but not limited to, multiple classifier methodologies and statistical language modeling. We have shown how the task of integrating language information and word recognition can be done very simply by treating them as probability scores.

We have also noted that in case of poor handwriting styles, the probability measures actually improve the recognition process. One clear advantage of the method described here is that the method works with the recognizers as blackboxes. Improving the performance of the recognizer simply involves re-training with a new set of images. The additional knowledge acquired during the re-training contributes to improved recognition rates as seen in the case of five digit recognizers tested.

The work presented in this paper opens several new research areas.

1. In the future, we would like to study the impact of the size of the clusters on the recognition rates. The clustering scheme must ensure that the images in a cluster are similar. For some clustering schemes this can cause certain clusters to have very few images and some to be overly dense.

2. Another area of promise is the method of find-

$$L((w_j,t_j),(w_{j-1},t_{j-1}),(w_{j-2},t_{j-2})) = \frac{P((w_j,t_j)|(w_{j-1},t_{j-1}),(w_{j-2},t_{j-2}))}{P(w_j,t_j)} \quad (9)$$

	GSC		GRAD		KP		HISTO		POLY	
Images	%	%	%	%	%	%	%	%	%	%
100	96	97	97	97	98	98	50	51	85	88
500	98	98	96.2	96.2	97.8	98	51.6	51.6	88	88
4000			97.8	97.8	98.1		49.4	49.9		

Figure 3: Improvement in recognition rate of digit recognizers by using the DPS method

ing $\overline{X}$. What we have proposed in this paper is just one of many possible approaches. We have focused on the confusion set $\mathcal{A}$ to create the set of trials. Each cluster $\overline{X}_j^i$ is "pure" in that all its images are of the same class (ω_i). Given that the purpose of creating $\overline{X}$ is to find all the images that can confuse with the input image we should also include in the clusters those images that are incorrectly labeled as ω_i. This creates a more realistic set of trials. We will experiment with this method and expect to see an even bigger improvement in performance.

3. The idea of deriving probabilities from other types of scores can be extended to many applications. Recognition systems with several components (*e.g.*, bank check recognition system, address recognition system *etc.*,) can use this methodology to put the final scores in a Bayesian framework. We would like to see the impact of this work in reading handwritten addresses in the future.

References

[1] ARPA, "Human Language Technology", Proceedings of a Workshop Held at Plainsboro, New Jersey, Sponsored by Advanced Research Projects Agency, Information Science and Technology Office, San Francisco, California: Morgan Kaufmann Publishers, Inc., 1994.

[2] D. Bouchaffra, E. Koontz, V. Kripasundar and R. K. Srihari, "Integrating Signal and Language Context to Improve Handwritten Phrase Recognition: Alternative Approaches", in *Preliminary Papers of the Sixth International Workshop on Artificial Intelligence and Statistics*, Jan., 4-7, 1997.

[3] D. Bouchaffra, E. Koontz, V. Kripasundar and R. K. Srihari, "Incorporating diverse information sources in handwriting recognition postprocessing", in *International Journal of Imaging Systems and Technology*, Vol. 7, pp 320-329, published by John Wiley & Sons Inc., 1996.

[4] S. J. DeRose, "Grammatical Category Disambiguation by Statistical Optimization", in *Computational Linguistics* 14(1), pp. 31-39, 1988.

[5] R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", Wiley, New York, 1973.

[6] J. Favata and S.N. Srihari, "Off-line sentence level recognition", in *International Workshop and Frontier of Handwriting Recognition* (IWFHR V), 1996.

[7] J. Hartigan, "Clustering Algorithms", New York, Wiley 1975.

[8] J. J. Hull, "Incorporating Language Syntax in Visual Text Recognition with a Statistical Model", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, NO. 12, Dec. 1996.

[9] T. K. Ho, J. J. Hull and S.N. Srihari, "Decision Combination in Multiple Classifier Systems", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, NO. 1, Jan 1994.

[10] S. Impedovo, "Frontiers in handwriting recognition", in *S. Impedovo editor, Fundamentals in Handwriting Recognition*, pages 7-39, Springer-Verlag Germany, 1994.

[11] A. K. Jain and R.C. Dubes, "Algorithms for Clustering Data", Englewood Cliffs, N.J. Printice-Hall, 1988

[12] G. Kim and V. Govindaraju, "A Lexicon Driven Approach to Handwritten Word Recognition for Real-time Applications", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4), pp. 366-379, 1997.

[13] D. S. Lee, "A theory of Classifier Combination: the Neural Network Approach", PhD Thesis Dissertation, Department of Computer Science, State University of New York at Buffal, 1995.

[14] R. Rosenfeld, "Adaptive Statistical Language Modeling: A Maximum Entropy Approach", PhD thesis, School of Computer Science, Carnegie Mellon University, 1994.

[15] C. Y. Suen et al., "Computer recognition of unconstrained handwritten numerals", in *IEEE Proc. 80, 1162-1180*, 1992.