Quantifying the Unimportance of Prior Probabilities in a Computer Vision Problem[1]

David B. Sher and Jonathan J. Hull

Computer Science Dept.
SUNY at Buffalo
Buffalo NY 14260
sher@cs.buffalo.edu — sher%cs.buffalo.edu@ubvm.bitnet

Abstract

We present an empirical investigation of the importance of accurate assessment of prior probabilities in a typical visual classification problem, handwritten ZIP Code recognition. We investigated prior probabilities for individual digits and entire zipcodes; the results for priors of individual digits are summarized here and discussed in detail in [1]. In our studies of prior distributions over entire ZIP Code we found that qualitative information had a major effect on the efficacy of the algorithm while quantitative information is relatively unimportant.

1 Introduction

Applying Bayesian reasoning to a perception problem requires prior probabilities for the possible outcomes. For example, when classifying a handwritten digit as shown in figure 1 one must give prior probabilities for the digits from 0-9. These prior probabilities are difficult to determine and may change from time to time and place to place.

However, if the observed data results in large likelihood ratios then the same classification occurs for almost any prior probability distribution. For example: consider the statistical test of determining if a coin is fair or biased towards heads by a 2 to 1 ratio by 1000 independent flips. If 500 of the flips were heads then the posterior probability that the coin is fair is greater than .5 for all nonnegligible prior probabilities. Thus while a prior probability for the coin being fair is required, a prior of .5 results in the same answer as a prior of .01 or .99; hence, this method of classification is insensitive to the value of the prior in this case. Only if the number of heads is within the closed interval [579,591] will the prior being .01 or .99 matter; thus for only 13 unlikely outcomes do reasonable priors affect the classification. Note that 1000 trials corresponds to a 31 by 32 binary image.

In this research project we performed empirical testing to determine the sensitivity of a simple vision problem to several types of prior information. The problem we chose was ZIP Code recognition, because:

- Large data sets are available.
- A rich structure of prior information is available (for example only certain ZIP Codes are legal).
- The data fits into a small number of categories.
- Prior probabilities on the individual digits and prior probabilities on the context (the other digits in the ZIP Code) are available.

We studied two types of prior information:

1. Prior probabilities for the classification of individual digits. A prior probability distribution is given for the digits from 0 through 9. This research is summarized here and discussed in detail in [1].
2. Prior probabilities for entire ZIP Code. A prior probability distribution is given over 5 digit numbers. Such a distribution can be used to represent contextual information. For example the fact that the first two digits of ZIP Codes are often 11 can easily be represented in such a distribution.

To determine the sensitivity of a Bayesian classifier to prior information, we applied the classifier to two sets of prior information — a uniform prior and a correct prior. The difference in the accuracy of the classifier measures the sensitivity of the classifier to prior information. For prior distributions over single digits the difference from using the correct priors was insignificant. For prior distributions over the full set of ZIP Codes the improvement measured the presence of information rather than the degree of belief in it.

Our experiments indicate that visual classification problems are only sensitive to certain types of prior information. In particular qualitative information such as whether "10001" is a legal ZIP Code is much more important than quantitative information like "10320" receives twice as much mail as "10321".

Determining what kinds of prior information are important for correct classification eases the application of Bayesian techniques to the visual domain — especially since the most important information seems to be qualitative and thus easily determined.

2 Background

Statistical Pattern recognition [2] often takes a Bayesian approach to pattern classification. In the area of character recognition much similar work has been done on spelling correction. James Peterson studied the effect of dictionary size on word recognition [3], and Kashyap and Oomen [4] applied probabilistic methods to spelling correction. Transition probabilities also are often used with the Viterbi algorithm [5] for recognition of words [6, 7, 8]. George Nagy presents a general survey of statistical approaches to character recognition in [9].

Often for data restoration the maximum entropy principle is used to discover priors. Our results support the use of maximum entropy estimates of prior distributions, since we show that in most cases equal priors are nearly as effective as correct priors. Jaynes [10] is a strong proponent of applying maximum entropy to inverse problems such as image processing; Frieden [11] has thoroughly studied methods of applying maximum entropy to data reconstruction; Herman [12] applied maximum entropy to medical imaging; Andrews and Hunt [13] discuss using entropy to derive priors for Bayesian image processing; Gull and Skilling [14] studied which forms of entropy apply to images.

Using context for handwriting recognition has been studied by Duda and Hart [15]. A good survey of context work appears in [16]. Hull has studied the effect of context on character recognition extensively [17, 18, 19].

Much of computer vision is based on statistical pattern recognition[2] Other important work that takes this approach is Lowrance and Garvey [20], and Wesley and Hanson [21, 22] who use Dempster Shafer statistics; and Geman [23], Chellappa [24], Art Owen [25], and David Sher [26, 27, 1] who have taken a variety of Bayesian approaches to a

[1] We gratefully acknowledge the work of Carolyn Sher in editing this work and the financial support of the Office of Advanced Technology of the United States Postal Service and the Rome Air Development Center.

Int. Conf. on Pattern Recognition (ICPR), Atlantic City, N.J., June 17-21, 1990, 662-664.

variety of vision problems.

3 Experimental Design

The experiments measured accuracy of classification for a Bayesian classifier using a variety of priors. The data to be classified were handwritten ZIP Codes collected from mail pieces. The methodology for collecting this data is documented in [28]. The digits of the ZIP Codes were hand segmented (separated), normalized and binarized into 16 by 16 arrays of bits.

For each 16 by 16 array of bits, a, and for each digit d we computed the probability that a would be generated by someone trying to write a d, the *inverse probability distribution* for d from a. Given a prior distribution for the digits we can use Bayes' law to compute the probability that a is a representation of d.

To compute the probability that a is a representation of d, for example "0", we collected a training set of 0's, 1's, 2's, ..., a total of 8120 digit images were used for training. We used the training set as a basis for a nearest neighbor classifier described in [1].

Our first experiment was to discover the importance of prior probabilities for the individual digits of a ZIP Code. We compared using the inverse probabilities generated by our classifier with the true probabilities of digits in the test data to using the inverse probabilities with the uniform prior distribution and to a corrupted prior distribution. The correct prior distribution correctly classified 8.7% of the digits incorrectly classified by the corrupted distribution - not a very impressive improvement.

Our second study involved contextual information. How much can classification improve due to constraints from the other digits of the ZIP Code; for example, how useful it is to know that 14 is much more likely than 98 to begin a ZIP Code. This study measures the the decrease in error rate caused by increasing the accuracy of the prior distribution of ZIP Code probabilities in the detector.

The inverse probability that a set of 5 arrays was generated by a person attempting to generate 14260 is the product of the probabilities that she would write each array when she was trying to generate the corresponding digit of 14260 (the probability that the first array would be generated by a person trying to write a "1" multiplied by the probability that the second array would be generated by a person trying to write a "4" and so on). Hence the total inverse probability of 14260 is the product of the inverse probabilities of its digits, the probability that the digits were independently distorted, a common simplifying assumption. Given a prior distribution over the set of ZIP Codes and we applied Bayes' law to classify the digits and counted the correctly classified ZIP Codes.

We used two prior distributions over ZIP Codes. The first assigned equal probability to every valid ZIP Code (only about 50% of the 5 digit numbers are ZIP Codes). The second distribution assigned equal probability to the ZIP Codes in the test data. Using the second distribution, which is more informative, resulted in a more accurate classification of digits. We measured the accuracy yielded by entering a mixed prior into the detector; the *mixture* of two distributions, P_1 and P_2 is $aP_1 + (1-a)P_2$ with $a \in [0,1]$, this is the distribution of elements randomly selected with probability distribution P_1 with probability a and selected from distribution P_2 with probability $1-a$. Varying a between 0 and 1 yields a quantitative measure of the importance of prior information. Our experiments here demonstrated that qualitative rather than quantitative effects significantly improved our accuracy of classification.

4 Experimental Results

In our most significant exploration of the importance of prior information for classifying individual digits, we randomly selected from our test set digits according to a specified distribution, for example, we selected digits at random so that 64% would be 0's and the remaining 10% of the tests would be evenly distributed among the remaining digits. We then tested the improvement in accuracy from using the correct prior probabilities over using equal prior probabilities for all the digits as shown in table 1. In those experiments the percentage of 0's was adjusted and all other digits were given equal probability. The reduction in the number of errors was less than 7% even when 64% of the digits were 0, further indicating the unimportance of prior information at the level of digits.

Percentage 0's	Percent Correct		Percent Errors		Percent
	Correct Priors	Equal Priors	Correct Priors	Equal Priors	Reduction in Errors
64	95.22	94.89	4.78	5.11	6.5
55	94.22	94.00	5.78	6.00	3.7
46	93.56	93.33	6.44	6.67	3.4
37	93.00	92.89	7.00	7.11	1.5
28	92.44	92.33	7.56	7.67	1.4
19	92.22	92.33	7.78	7.67	-1.4

Table 1: Error rates using correct and equal priors on digits distributed according to a specified distribution

The experiment on prior information about complete ZIP Codes used the same template matcher as well as the same training and testing data. This experiment used the digits from each of the 312 ZIP Codes and determined the ZIP Code from the list of 41,595 valid USPS ZIP Codes with the maximum a-posteriori probability. The prior probability of a ZIP Code was computed as the product of the priors of its digits. This was multiplied by the probabilities of those digits as determined by the template matcher. A weighting factor f was also multiplied by each ZIP Code. The result was the posterior probability of the ZIP Code. The recognition decision was the ZIP Code with the maximum a-posteriori probability.

The weighting factor was varied to reflect different levels of contextual information. The minimum contextual information was represented by an equal probability for each ZIP Code (i.e., 1/41,595). The maximum contextual information was represented by an equal proba-

f	Correct		Errors		f	Correct		Errors	
	N	percent	N	percent		N	percent	N	percent
0.0	214	68.59%	98	31.41%	0.5	234	75.00%	78	25.00%
0.01	220	70.51%	92	29.49%	0.55	235	75.32%	77	24.68%
0.05	222	71.15%	90	28.85%	0.6	234	75.00%	78	25.00%
0.1	228	73.08%	84	26.92%	0.65	236	75.64%	76	24.36%
0.15	232	74.36%	80	25.64%	0.7	238	76.28%	74	23.72%
0.2	233	74.68%	79	25.32%	0.75	241	77.24%	71	22.76%
0.25	234	75.00%	78	25.00%	0.8	241	77.24%	71	22.76%
0.3	234	75.00%	78	25.00%	0.85	241	77.24%	71	22.76%
0.35	234	75.00%	78	25.00%	0.9	241	77.24%	71	22.76%
0.4	234	75.00%	78	25.00%	0.95	241	77.24%	71	22.76%
0.45	234	75.00%	78	25.00%	0.99	253	81.09%	59	18.91%
					1.0	297	95.19%	15	4.81%

Table 2: Results of ZIP Code recognition with varying levels of context

bility for each ZIP Code in the test set (i.e., 1/312) and a zero probability for every other ZIP Code. Intermediate levels of context were represented by mixtures of the two distributions that were computed as described in the previous section.

The results of the experiment that varied levels on ZIP Code context are shown in Table 2. With a minimum of contextual information (all ZIP Codes equally likely), 68.6 percent of the ZIP Codes in the test set are correctly recognized; performance improves as additional contextual support is provided up to 95.2 percent correct with the maximum contextual information. The increase in improvement is highly non-linear and that with $f = .95$, a correct rate of only 77 percent is achieved.

Figure 2 graphically demonstrates that the error rate is only significantly effected by the degree of mixture between the two prior distributions when the mixture is near 1 or 0. This means that there are three significantly different kinds of prior information:

1. Only the minimum context (legal ZIP codes)

Int. Conf. on Pattern Recognition (ICPR), Atlantic City, N.J., June 17-21, 1990, 662-664.

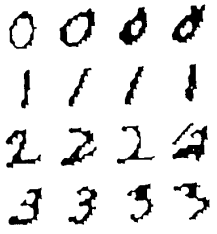

Figure 1: Handwritten Zipcode Digits taken from US Mail

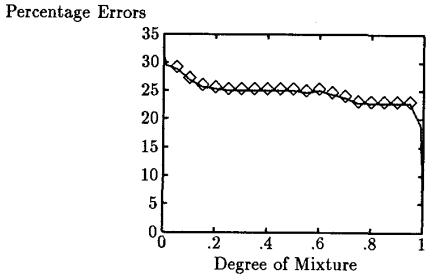

Figure 2: Graph of Error Rate for ZIP Code recognition vs. Degree of context

2. A mixture of the minimum context and the stronger context.

3. The stronger context (ZIP codes from the test set).

Because the error rate acts like a step function with regard to the degree of context, our experiments indicate the effect of contextual prior information is qualitative rather than quantitative.

5 Conclusion

We performed empirical experiments on the importance of prior information in a typical vision problem and discovered that prior probabilities for the individual digits have little effect on the accuracy of the resulting detector. We also studied the effect of contextual information and found that qualitative effects were much more important than quantitative effects. This study is evidence that precise estimation of prior probabilities is unnecessary in the domain of computer vision however accurate qualitative assessment of possibilities is important.

References

[1] David B. Sher and Jonathan J. Hull. Quantifying the unimportance of prior probabilities in a computer vision problem. *Pattern Recognition Letters*, 1990.

[2] Richard O. Duda and Peter E. Hart. *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., New York, London, Sydney, Toronto, 1973.

[3] James L. Peterson. A note on undetected typing errors. *Communications of the ACM*, 29(7):633–637, July 1986.

[4] R.L. Kashyap and B.J. Oommen. Spelling correction using probabilistic methods. *Pattern Recognition Letters*, 2(3):147–154, March 1984.

[5] G. D. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, march 1973.

[6] R. Shinghal and G.T. Toussaint. A bottom-up and top-down approach to using context in text recognition. *International Journal of Man-Machine Studies*, 11(2):201–212, 1979.

[7] D.L. Neuhoff. The viterbi algorithm as an aid in text recognition. *IEEE Transactions on Information Theory*, IT-21(2):222–226, 1975.

[8] J.J. Hull and S.N. Srihari. Experiments in text recognition with binary n-gram and viterbi algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-4(5):520–530, September 1982.

[9] George Nagy. *Handbook of :Statistics*, volume 2, chapter Optical character recognition: theory and practice, pages 621–649. 1982.

[10] E. T. Jaynes. *Maximum-Entropy and Bayesian Methods in Inverse Problems*, chapter Where do We Go from Here? D. Reidel Publishing Company, Dordrecht Boston Lancaster, 1985.

[11] B. Roy Frieden. *Maximum-Entropy and Bayesian Methods in Inverse Problems*, chapter Estimating Occurrence Laws with Maximum Probability, and the Transition to Entropic Estimators. D. Reidel Publishing Company, Dordrecht Boston Lancaster, 1985.

[12] Gabor T. Herman. *Maximum-Entropy and Bayesian Methods in Inverse Problems*, chapter Application of Maximum Entropy and Bayesian Optimization Methods to Image Reconstruction from Projections, pages 319–337. D. Reidel Publishing Company, Dordrecht Boston Lancaster, 1985.

[13] H. C. Andrews and B. R. Hunt. *Digital Image Restoration*, pages 187–211. Prentice-Hall, INC., Englewood Cliffs, New Jersey 07632, 1977.

[14] S. F. Gull and J. Skilling. *Maximum-Entropy and Bayesian Methods in Inverse Problems*, chapter The Entropy of an Image, pages 287–301. D. Reidel Publishing Company, Dordrecht Boston Lancaster, 1985.

[15] R. O. Duda and P. E. Hart. Experiments in the recognition of hand-printed text: Part ii context analysis. In *Afips Conference Proceedings*, volume 33, pages 1139–1149, 1968.

[16] E. Reuhkala. Recognition of strings of discrete symbols with special application to isolated word recognition. *Acta Polytechnica Scandanavia*, Ma 38:1–92, 1983.

[17] Jonathan J. Hull. Inter-word constraints in visual word recognition. In *Proceedings of the Conference of the Canadian Society for Computational Studies of Intelligence*, pages 134–138, May 1986.

[18] Jonathan J. Hull. The use of global context in text recognition. In *Proceedings of the 8th International Conference on Pattern Recognition*, October 1986.

[19] J.J. Hull. *A computational theory of visual word recognition*. PhD thesis, SUNY at Buffalo, Department of Computer Science, 1987.

[20] John D. Lowrance and Thomas D. Garvey. Evidential reasoning: An implementation for multisensor integration. Technical Report 307, SRI International Artificial Intelligence Center, Computer Science and Technology Division, December 1983.

[21] Leonard P. Wesley and Allen R. Hanson. The use of an evidential-based model for representing knowledge and reasoning about images in the visions system. *PAMI*, 4(5):14–25, Sept 1982.

[22] Leonard P. Wesley and Allen R. Hanson. The use of an evidential-based model for representing knowledge and reasoning about images in the visions system. *Proceedings of the Worksshop on Computer VIsion: Representation and Control*, pages 14–25, August 1982.

[23] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *PAMI*, 6(6):721–741, Novermber 1984.

[24] R. Chellappa. Fitting markov random field models to images. Technical Report 994, University of Maryland, COmputer Vision Laboratory, Computer Science Center, January 1981.

[25] Art Owen. A neighbourhood-based classifier for landsat data. *The Canadian Journal of Statistics*, 12(3):191–200, September 1984.

[26] D. Sher and S. Yen. The ergodic mondrian model — a source of markov random fields. Technical Report 89-3, S.U.N.Y. at Buffalo, May 1989.

[27] David Sher and E ren Chuang. Generating object location systems from complex object descriptions. In *Proceedings of the 6th Israeli Conference on Artificial Intelligence and Computer Vision*, pages 557–582, December 1989.

[28] S.N. Srihari, J.J. Hull, Chih-Chau L. Kuan, Ed Cohen, and Paul W. Palumbo. Handwritten address zip code recognition: A survey of techniques. Technical report, Department of Computer Science, SUNY at Buffalo, December 1987.

Int. Conf. on Pattern Recognition (ICPR), Atlantic City, N.J., June 17-21, 1990, 662-664.