Pattern Recognition Letters 11 (1990) 379-383
North-Holland

June 1990

Quantifying the unimportance of prior probabilities in a computer vision problem

David B. SHER and Jonathan J. HULL
Computer Science Dept., SUNY at Buffalo, Buffalo, NY 14260, USA

Received 20 October 1989
Revised 19 December 1989

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 found that little accuracy was gained by accurate assessment of the prior distribution over the individual digits.

Key words: Bayesian inference, computer vision, image analysis, decision theory, character recognition, pattern classification.

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.

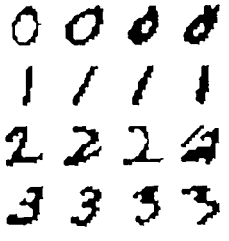

Figure 1. Handwritten zip code digits taken from US mail.

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 0.5 for all non-negligible prior probabilities. Thus while a prior probability for the coin being fair is required, a prior of 0.5 results in the same answer as a prior of 0.01 or 0.99; 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 0.01 or 0.99 matter; thus for only 13 unlikely outcomes do reasonable priors affect the classification. Note that 1000 trials correspond 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.

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.

0167-8655/90/$03.50

- 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.

To determine the sensitivity of Bayesian estimation to prior information when applied to recognition of handwritten zipcodes, we applied a Bayesian 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 improvement from using the correct priors was insignificant.

Our experiments indicate that certain visual classification problems are only sensitive to certain types of prior information. Low level vision problems usually involve making decisions on the basis of large amounts of data which often yield large likelihood ratios. Large likelihood ratios cause insensitivity to priors in Bayesian classification.

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 [6] 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 [26], and Kashyap and Oomen [21] applied probabilistic methods to spelling correction. Transition probabilities also are often used with the Viterbi algorithm [7] for recognition of words [31, 24, 14]. George Nagy presents a general survey of statistical approaches to character recognition in [23].

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 [20] is a strong proponent of applying maximum entropy to inverse problems such as image processing; Frieden [8, 10, 9] has thoroughly studied methods of applying maximum entropy to data reconstruction; Herman [13] applied maximum entropy to medical imaging; Andrews and Hunt [1] discuss using entropy to derive priors for Bayesian image processing; Gull and Skilling [12] studied which forms of entropy apply to images.

Using context for handwriting recognition has been studied by Duda and Hart [5]. A good survey of context work appears in [27]. Hull has studied the effect of context on character recognition extensively [15, 17, 18, 19].

Much of computer vision is based on statistical pattern recognition [6]. Other important work that takes his approach is Lowrance and Garvey [22], and Wesley and Hanson [32, 33] who use Dempster Shafer statistics; and Geman [4, 11], Witkin [34], Chellappa [2], Art Owen [25], and David Sher [30, 28, 29] who have taken a variety of Bayesian approaches to a variety of vision problems.

3. Experiments

The experiments measured the accuracy of a Bayesian classifier for handwritten ZIP Codes collected from mail pieces, using a variety of priors. The methodology for collecting this data is discussed in [16]. 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 assumed that digits

were generated by randomly, independently changing the values of the bits in the array with probability p. Thus the probability that a is generated by a single element of the training set is a function of the hamming distance, h, between the observed data and the training set element: $p^h(1-p)^{256-h}$. The probability that any of the training set's 0's would generate a was computed as the average of the probabilities that each of the 0's would generate a. Thus we can compute the inverse probability for each digit given a. Visual perception algorithms commonly assume independence of errors; using a more sophisticated error model results in a slower more complex algorithm without necessarily yielding a better result.

We compared using the inverse probabilities generated by our classifier with the true probabilities of digits in the test data (shown in Figure 1) 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 experiment used 1560 digits as test data and 8120 digit images as training data that were extracted from handwritten ZIP Code images. The test data's digits were not contained in the training set. Each digit was input to the template matcher and it returned a ranked list of how well the digits zero through nine matched the input along with a likelihood score. This score was then multiplied by the *a priori* probability for that digit occurring in a certain position in the ZIP Code to yield an *a posteriori* probability. The class with the maximum *a posteriori* probability was output.

We applied this scheme with three *a priori* distributions. The 'Correct' prior information was represented by probabilities that were determined by the occurrence of digits at positions one to five in the 312 ZIP Codes of the test set. This distribution is shown in Table 1. 'Equal' prior information was represented by a distribution in which the prior probability of each of the ten digits was 0.1. The 'Incorrect' prior information was represented by a corruption of the best distribution in which the ranking within each position was reversed. For example, in position one, if the digit with the maximum a priori probability was the zero and the minimum was the five, the zero was assigned the probability of the five and the five the probability of the zero. The same process was applied to the other digits in the other positions. The worst distribution is shown in Table 2.

Table 1
Probabilities for digits by position in the ZIP code

	position in ZIP Code				
digit	1	2	3	4	5
0	0.141	0.096	0.135	0.202	0.176
1	0.058	0.067	0.090	0.099	0.135
2	0.096	0.125	0.138	0.099	0.083
3	0.167	0.160	0.093	0.109	0.090
4	0.157	0.125	0.125	0.093	0.099
5	0.128	0.141	0.096	0.055	0.087
6	0.064	0.061	0.080	0.122	0.131
7	0.096	0.055	0.099	0.080	0.074
8	0.035	0.106	0.067	0.055	0.055
9	0.058	0.064	0.077	0.087	0.071

The results of the digit recognition experiment are shown in Table 3. It is seen that the correct recognition rate with the worst priors is 89.7 percent. With equal priors the correct rate increased to 90.1

Table 2
Incorrect probabilities for digits by position in the ZIP code

	position in ZIP Code				
digit	1	2	3	4	5
0	0.058	0.106	0.077	0.055	0.055
1	0.141	0.125	0.099	0.087	0.071
2	0.096	0.064	0.067	0.093	0.099
3	0.035	0.055	0.096	0.080	0.087
4	0.058	0.067	0.080	0.099	0.083
5	0.064	0.061	0.093	0.202	0.090
6	0.128	0.141	0.125	0.055	0.074
7	0.096	0.160	0.090	0.109	0.131
8	0.167	0.096	0.138	0.122	0.176
9	0.157	0.125	0.135	0.099	0.135

Table 3
Results of digit recognition experiment

	Prior probabilities		
	Incorrect	Equal	Correct
% error	10.3	9.9	9.4
% improvement over equal priors	−4.0	0.0	5.0

Table 4
Error rates using correct and equal priors on digits distributed according to a specified distribution

Percentage 0's	Percent errors		Percent reduction in errors
	Correct priors	Equal priors	
64	4.78	5.11	6.5
55	5.78	6.00	3.7
46	6.44	6.67	3.4
37	7.00	7.11	1.5
28	7.56	7.67	1.4
19	7.78	7.67	−1.4

percent and with the best priors 90.6 percent of the digits in the test set were correctly recognized. Thus there is an insignificant difference in performance.

To further investigate the importance of prior information about 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 4. 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.

4. 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. This study is evidence that precise estimation of prior probabilities may be unnecessary in certain domains of computer vision.

References

[1] Andrews, H.C. and B.R. Hunt (1977). *Digital Image Restoration.* Prentice-Hall, Englewood Cliffs, NJ, 187–211.

[2] Chellappa, R. (1981). Fitting Markov random field models to images. Technical Report 994, University of Maryland, Computer Vision Laboratory, Computer Science Center, January 1981.

[3] Dalkey, N.C. (1985). *Maximum-Entropy and Bayesian Methods in Inverse Problems.* Chapter: Inductive Inference and the Maximum Entropy Principle. Reidel, Dordrecht, 351–364.

[4] Derin, H., H. Elliott, R. Cristi and D. Geman (1984). Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields. *IEEE Trans. Pattern Anal. Machine Intell.* 6 (6), 707–720.

[5] Duda, R.O. and P.E. Hart (1968). Experiments in the recognition of handprinted text: Part II context analysis. *AFIPS Conf. Proc.* 33, 1139–1149.

[6] Duda, R.O. and P.E. Hart (1973). *Pattern Classification and Scene Analysis.* Wiley, New York.

[7] Forney, G.D. (1973). The Viterbi algorithm. *Proc. IEEE* 61 (3), 268–278.

[8] Frieden, B.R. (1972). Restoring with maximum entropy. *J. Opt. Soc. Amer.* 62 (4), 511–518.

[9] Frieden, B.R. (1985). *Maximum-Entropy and Bayesian Methods in Inverse Problems.* Chapter: Estimating Occurrence Laws with Maximum Probability, and the Transition to Entropic Estimators. Reidel, Dordrecht.

[10] Frieden, B.R. and Zoltani (1985). Maximum bounded entropy. *Applied Optics* 24, 201.

[11] Geman, S. and D. Geman (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* 6 (6), 721–741.

[12] Gull, S.F. and J. Skilling (1985). *Maximum-Entropy and Bayesian Methods in Inverse Problems.* Chapter: The Entropy of an Image. Reidel, Dordrecht, 287–301.

[13] Herman, G.T. (1985). *Maximum-Entropy and Bayesian Methods in Inverse Problems.* Chapter: Application of Maximum Entropy and Bayesian Optimization Methods to Image Reconstruction from Projections. Reidel, Dordrecht, 319–337.

[14] Hull, J.J. and S.N. Srihari (1982). Experiments in text recognition with binary *n*-gram and Viterbi algorithms. *IEEE Trans. Pattern Anal. Machine Intell.* 4 (5), 520–530.

[15] Hull, J.J., S.N. Srihari and R. Choudhari (1983). An integrated algorithm for text recognition: comparison with a cascaded algorithm. *IEEE Trans. Pattern Anal. Machine Intell.* 5 (4), 384–394.

[16] Hull, J.J., S.N. Srihari, E. Cohen, C.-C.L. Kuan, P. Cullen and P.W. Palumbo (1988). A blackboard-based approach to handwritten zip code recognition. *Internat. Conf. on Pattern Recognition*, November 1988, 111–113.

[17] Hull, J.J. (1986). Inter-word constraints in visual word recognition. *Proc. Conf. Canadian Society for Computational Studies of Intelligence*, May 1986, 134–138.

[18] Hull, J.J. (1986). The use of global context in text recognition. *Proc. 8th Internat. Conf. on Pattern Recognition*, October 1986.

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

[20] Jaynes, E.T. (1985). *Maximim-Entropy and Bayesian Methods in Inverse Problems*. Chapter Where do We Go from Here? Reidel, Dordrecht.

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

[22] Lowrance, J.D. and T.D. Garvey (1983). Evidential reasoning: An implementation for multisensor integration. Technical Report 307, SRI International Artificial Intelligence Center, Computer Science and Technology Division, December 1983.

[23] Nagy, G. (1982). *Handbook of Statistics*, Vol. 2. Chapter Optical character recognition: theory and practice, 621–649.

[24] Neuhoff, D.L. (1975). The Viterbi algorithm as an aid in text recognition. *IEEE Trans. Information Theory* 21 (2), 222–226.

[25] Owen, A. (1984). A neighbourhood-based classifier for landsat data. *Canad. J. Statistics* 12 (3), 191–200.

[26] Peterson, J.L. (1986). A note on undetected typing errors. *Comm. ACM* 29 (7), 633–637.

[27] Reuhkala, E. (1983). Recognition of strings of discrete symbols with special application to isolated word recognition. *Acta Polytechnica Scandinavia* 38, 1–92.

[28] Sher, D.B. (1987). Generating robust operators from specialized ones. *IEEE Computer Society Workshop on Computer Vision*, Miami, FL, November 1987. IEEE Press, New York.

[29] Sher, D.B. (1987). A probabilistic approach to low-level vision. PhD Thesis, Computer Science Department, University of Rochester, August 1987.

[30] Sher, D.B. (1987). Tunable facet model likelihood generators for boundary pixel detection. *IEEE Computer Society Workshop on Computer Vision*, Miami, FL, November 1987. IEEE Press, New York.

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

[32] Wesley, L.P. and A.R. Hanson (1982). The use of an evidential-based model for representing knowledge and reasoning about images in the visions system. *IEEE Trans. Pattern Anal. Machine Intell.* 4 (5), 14–25.

[33] Wesley, L.P. and A.R. Hanson (1982). The use of an evidential-based model for representing knowledge and reasoning about images in the visions system. *Proc. Workshop on Computer Vision: Representation and Control*, August 1982, 14–25.

[34] Witkin, A.P. (1981). Recovering surface shape and orientation from texture. *Artificial Intelligence* 17 (1–3), 17–45.