OPTIMAL DESIGN OF TRANSFORM CODERS AND QUANTIZERS FOR IMAGE CLASSIFICATION

Soumya Jana and Pierre Moulin

University of Illinois at Urbana-Champaign
Beckman Institute, Coordinate Science Lab, and ECE Department
405 N. Mathews Ave., Urbana, IL 61801
Email: {*jana,moulin*}*@ifp.uiuc.edu*

ABSTRACT

In a variety of applications (including automatic target recognition) image classification algorithms operate on compressed image data. This paper explores the design of optimal transform coders and scalar quantizers using Chernoff bounds on probability of misclassification as the measure of classification accuracy. This design improves classification performance but the mean square error (as well as the visual quality) of the coded image degrades. However, by appropriately combining classification accuracy and mean square error in the cost function, one can achieve good classification with low (visual) distortion, which is desirable in classification systems requiring visual authentication.

1. INTRODUCTION

There exists a variety of applications in which image classification algorithms operate on compressed image data. Examples include Automatic Target Recognition [1, 2, 3, 4], detection of abnormalities in compressed medical images, and classification of compressed aerial images [5]. Recently there have been attempts to design the compression algorithm to optimize measures of classification performance. In [5], vector quantization (VQ) schemes are designed so as to optimize a weighted sum of Bayes risk for classification and mean–squared reconstruction error (MSE), subject to bit rate constraints. In [3], information–theoretic bounds on detection performance are used to evaluate the performance of transform coders.

This paper extends our recent work [6] in which a technique was proposed for optimal design of transform coders that minimize Chernoff bounds on probability of misclassification (P_e) under an average bit rate constraint. The choice of Chernoff bounds on P_e is attractive because of its tractability and asymptotic tightness [7]. The design of optimal quantizers and the problem of bit allocation, are explored analytically using a high–rate quantization assumption. Results are furnished to illustrate potential improvement of classification accuracy over conventional MSE-based designs. A tradeoff between classification accuracy and MSE is investigated which is useful in scenarios where images are used both for classification and visual evaluation purposes. Results are presented for a binary classification problem but can be extended to M–ary classification problems.

THIS RESEARCH WAS SUPPORTED BY ARO UNDER CONTRACT NUMBERS ARO DAAH-04-95-1-0494 AND ARMY WUHT-011398-S1, AND BY DARPA UNDER CONTRACT F49620-98-1-0498, ADMINISTERED BY AFOSR.

2. PROBLEM STATEMENT

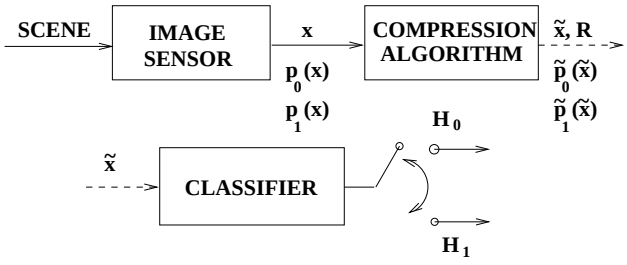

Figure 1: Binary Classification Using Compressed Data

Fig. 1 presents the generic binary classification problem considered in this paper. N–pixel image data $\mathbf{x}$ are acquired by image sensors. The classification algorithm does not have access to $\mathbf{x}$ but operates on a compressed version $\tilde{\mathbf{x}}$ of these original data which is coded at an average rate of R bits per pixel. The classification problem is posed as a statistical hypothesis testing problem. Specifically, there are two hypotheses H_0 and H_1, with respective *a priori* probabilities μ_0 and μ_1. The distribution of $\mathbf{x}$ under hypothesis H_i is denoted by $p_i(\mathbf{x})$, $i = 0, 1$. The resulting distributions on $\tilde{\mathbf{x}}$ are denoted by $\tilde{p}_i(\tilde{\mathbf{x}})$, $i = 0, 1$. An unconstrained compression algorithm would ideally solve the classification problem and only transmit the binary decision. For computational reasons, it may not be possible to solve this problem prior to transmission; and structural constraints may be imposed on the compression scheme. In this paper, we assume a linear transform coder followed by a bank of uniform quantizers and independent coding of each quantized transform coefficient, as shown in Fig. 2. The bank of uniform quantizers with step sizes $\{\Delta_k\}_{k=1}^N$, can be viewed as one with fixed equal step sizes Δ, following a diagonal transform with diagonal entries $\left\{\frac{\Delta_k}{\Delta}\right\}_{k=1}^N$. Since the diagonal transform can be absorbed in $\mathbf{T}$, the step sizes of the quantizers can be assumed equal with no

loss of generality.

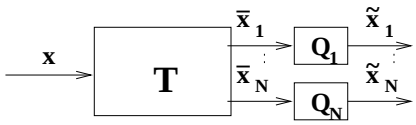

Figure 2: What is the optimal linear transform $\mathbf{T}$?

We seek the optimal transform $\mathbf{T}$, so that, ideally, P_e based on compressed data $\tilde{\mathbf{x}}$ is minimized. But

$$P_e = \sum_{\tilde{\mathbf{x}}} \min\left[\mu_0 \tilde{p_0}(\tilde{\mathbf{x}}), \mu_1 \tilde{p_1}(\tilde{\mathbf{x}})\right]$$

is an intractable function of $\mathbf{T}$, so we use information-theoretic bounds on P_e as the cost function. Specifically, consider Chernoff bounds which are upper bounds on P_e [7]:

$$P_e \leq \mu_0^{1-s} \mu_1^s e^{-D_s(\tilde{p}_0, \tilde{p}_1)}, \forall\, 0 < s < 1,$$

where,

$$D_s(\tilde{p}_0, \tilde{p}_1) = -\ln \sum_{\tilde{\mathbf{x}}} \tilde{p}_0(\tilde{\mathbf{x}}) \left(\frac{\tilde{p}_1(\tilde{\mathbf{x}})}{\tilde{p}_0(\tilde{\mathbf{x}})}\right)^s \qquad (1)$$

is the Chernoff distance between the distributions $\tilde{p}_0$ and $\tilde{p}_1$. We view (1) as our performance index. Large Chernoff distance between the distributions indicates that good discrimination between them is possible. Assuming the coder is designed for the mixture distribution $\mu_0 p_0 + \mu_1 p_1$, we measure bit rate by the first–order entropy of the mixture distribution, $R = \sum_k H(\mu_0 \tilde{p}_{0k} + \mu_1 \tilde{p}_{1k})$, where $H(\cdot)$ denotes entropy [8], and $\tilde{p}_{ik}$ denotes the distribution of $\tilde{x}_k$, kth quantized coefficient, under hypothesis H_i. The optimization problem with rate constraints can be formulated as a Lagrangian optimization problem,

$$\min_{\mathbf{T}} \left[-D_s(\tilde{p}_0, \tilde{p}_1) + \nu \sum_k H(\mu_0 \tilde{p}_{0k} + \mu_1 \tilde{p}_{1k})\right] \qquad (2)$$

where ν is the Lagrange multiplier which is selected so as to meet the bit rate constraint. At high bit rates, $\nu \to 0$; at low bit rates, $\nu \to \infty$.

3. OPTIMAL TRANSFORM

We would like to solve problem (2) for images. Assume that $\mathbf{x}$ under both H_0 and H_1 is a stationary, periodic process with symmetric extensions. In that case, the discrete cosine transform (DCT) produces decorrelated components of $\overline{\mathbf{x}}$ under both H_0 and H_1. Then we shall show that the optimal transform $\mathbf{T}$ to use, in the sense (2), is DCT followed by a diagonal transform (refer Fig. 2). The optimal transform under an MSE design criterion also has a similar structure, and so does the one designed to optimize a linear combination of classification accuracy and MSE.

Let us define $\mathbf{T} \triangleq \mathbf{UGVC}$, where $\mathbf{U}$, $\mathbf{V}$ are unitary matrices, $\mathbf{G}$ is a diagonal matrix with positive entries and $\mathbf{C}$ is the DCT matrix. Then (2) can be posed as a nested minimization problem.

$$\min_{\mathbf{G}} \left[\min_{\mathbf{U},\mathbf{V}} \left\{-D_s(\tilde{p}_0, \tilde{p}_1) + \nu \sum_k H(\mu_0 \tilde{p}_{0k} + \mu_1 \tilde{p}_{1k})\right\}\right] \qquad (3)$$

We would like to show that the optimal $\mathbf{U}$ and $\mathbf{V}$ are identity matrices. At this point, we state the following results for fine quantization without proof. In addition, $\mathbf{x}$ is assumed zero-mean[1] Gaussian under both rival hypotheses.
***Result* 1:** If $\mathbf{G}$ and $\mathbf{V}$ are set to identity, then choosing $\mathbf{U}$ to be any permutation matrix gives a local minimum of (3). Moreover, this $\mathbf{U}$ is also a global minimizer when $N = 2$. Global optimality for all N is strongly suspected.
***Result* 2:** If $\mathbf{G}$ is fixed, it can be shown using *Result* 1 that, the choice of $\mathbf{U}$ as a permutation matrix, such that diagonal elements of $\mathbf{UGU^T}$ are arranged in a certain order depending on p_0 and p_1, and $\mathbf{V}$ as any permutation matrix is a stationary point of the inner minimization problem in (3). We suspect stronger optimality properties (at least local optimality) of the above assignment of $\mathbf{U}$ and $\mathbf{V}$.
***Result* 3:** Setting $\mathbf{U}$ and $\mathbf{V}$ to arbitrary permutation matrices and performing the outer minimization in (3) lead to a stationary point that has the required arrangement of diagonal elements of $\mathbf{UGU^T}$.

In view of *Result* 3, we set both $\mathbf{U}$ and $\mathbf{V}$ to identity matrix without any loss of generality. Let $\mathbf{G}^*$ be the minimizer of the outer minimization problem in (3). Thus the overall optimal transform is $\mathbf{T}^* = \mathbf{G}^*\mathbf{C}$, which is DCT followed by a diagonal transform. Referring to Fig. 2, we can absorb $\mathbf{G}^*$ in the quantizers, and view DCT as the optimal transform. In this perspective, the quantizers have different step sizes, Δ_k, $k = 1, 2, ..., N$, and finding their optimal values is equivalent to finding $\mathbf{G}^*$. We shall assume this structure for the optimal transform and quantizers from now on.

4. OPTIMAL QUANTIZERS

The DCT coefficients being independent under each hypothesis, the Chernoff distance (1) is additive over these coefficients,

$$D_s(p_0, p_1) = \sum_{k=1}^{N} D_s(\overline{p}_{0k}, \overline{p}_{1k}), \qquad (4)$$

where, $\overline{p}_{ik}$ denotes the distribution of $\overline{x}_k$, kth transformed coefficient, under hypothesis H_i. Subsequently, the outer minimization problem in (3) can be decoupled to N scalar minimization problems:

$$\min_{\Delta_k} \left[-D_s(\tilde{p}_{0k}, \tilde{p}_{1k}) + \nu H(\mu_0 \tilde{p}_{0k} + \mu_1 \tilde{p}_{1k})\right],\ k = 1, 2, ..., N, \qquad (5)$$

subject to the rate constraint $R = \sum_k H(\mu_0 \tilde{p}_{0k} + \mu_1 \tilde{p}_{1k})$. Although these optimizations have to be performed numerically in general, further analytical exploration is possible in fine quantization regime (as is the case with conventional MSE designs).

Fine Quantization : In view of (4), the first term in (5) can be replaced by the loss in Chernoff distance,

[1]More correctly, the discrimination information provided by mean is ignored.

$D_s(\overline{p}_{0k}, \overline{p}_{1k}) - D_s(\tilde{p}_{0k}, \tilde{p}_{1k})$, which is quadratic in Δ_k as $\Delta_k \to 0$ [9], with the proportionality constant,

$$\beta_k = \frac{s(1-s)}{24} \frac{\left(\overline{\sigma}_{0k}^2 - \overline{\sigma}_{1k}^2\right)^2}{\overline{\sigma}_{0k}^{(-s)} \overline{\sigma}_{1k}^{(s-1)}} \times \begin{cases} \frac{1}{\overline{\sigma}_{0k}^4((3-s)\overline{\sigma}_{1k}^2 - (2-s)\overline{\sigma}_{0k}^2)^{3/2}}, & \forall \overline{\sigma}_{0k} \leq \overline{\sigma}_{1k} \\ \frac{1}{\overline{\sigma}_{1k}^4((2+s)\overline{\sigma}_{0k}^2 - (1+s)\overline{\sigma}_{1k}^2)^{3/2}}, & \forall \overline{\sigma}_{0k} > \overline{\sigma}_{1k} \end{cases}$$

where, $\overline{p}_{ik}$ is zero–mean Gaussian distribution with variance $\overline{\sigma}_{ik}$, $i = 0, 1$. The second term in (5), entropy, satisfies the asymptotic relation,

$$H(\mu_0 \tilde{p}_{0k} + \mu_1 \tilde{p}_{1k}) \sim h(\mu_0 \overline{p}_{0k} + \mu_1 \overline{p}_{1k}) - \log(\Delta_k) + \alpha_k \Delta_k^2,$$

as $\Delta_k \to 0$, where $\alpha_k = \frac{1}{24} I(\mu_0 \overline{p}_{0k} + \mu_1 \overline{p}_{1k})$, and $I(\cdot)$ and $h(\cdot)$ denote the Fisher Information and the differential entropy of the argument respectively [8]. Thus problem (5) is asymptotically equivalent to:

$$\min_{\Delta_k} \left[\beta_k \Delta_k^2 + \nu\{h(\mu_0 \overline{p}_{0k} + \mu_1 \overline{p}_{1k}) - \log(\Delta_k) + \alpha_k \Delta_k^2\}\right].$$

It is easy to verify that the optimal step sizes are given by,

$$\Delta_k = \sqrt{\frac{\nu}{2(\beta_k + \nu\alpha_k)}}, \quad k = 1, 2, ..., N.$$

At high bit rates, $\nu \to 0$ and we have $\Delta_k \sim \sqrt{\frac{\nu}{2\beta_k}}$, which equalizes the loss in Chernoff distance among all k's.

Until now, we concentrated on Chernoff distance as our design criterion. More generally, MSE or a weighted combination of MSE and Chernoff distance can be optimized using a similar technique. There the first term of the cost function in (5) is replaced by the linear combination $(1-\phi)\beta_k \Delta_k^2 + \phi \frac{\Delta_k^2}{12}$, where $\phi \in [0, 1]$ is a weighting factor. Note that $\phi = 1$ leads to the conventional MSE-based design, whereas $\phi = 0$ reduces to (5).

5. PERFORMANCE

We have outlined the optimal design of a transform coder that minimizes the loss in classification accuracy. We have also presented how a similar procedure can be used to design a transform coder that optimizes a linear combination of MSE and classification accuracy. Now we include an example of texture classification to illustrate the merit of the proposed design over classical design based on only MSE.

In our example, we seek to discriminate between two image classes, formed on the basis of 'Water' and 'Sand' images, taken from *MIT VisTex database* [10]. Under the assumptions of Sec. 3, the probability distributions p_0 and p_1 are estimated block-wise (16×16) on the basis of the actual image values in the neighboring blocks in both image classes. In the first part of our example, we use fine quantization. As Fig. 3 shows, the optimized coder ($\phi = 0$) achieves better classification performance than does a conventional coder that minimizes MSE ($\phi = 1$), using standard bit allocation techniques. Also from Fig. 4, we note that the MSE performance of the coder that optimizes classification performance ($\phi = 0$) is not good. However, we have observed that, by optimizing a combination of these criteria, one can achieve high classification accuracy as well as low MSE. For instance, in this example a linear combination with $\phi = 0.001$, performs very well in terms of classification accuracy as well as mean square error.

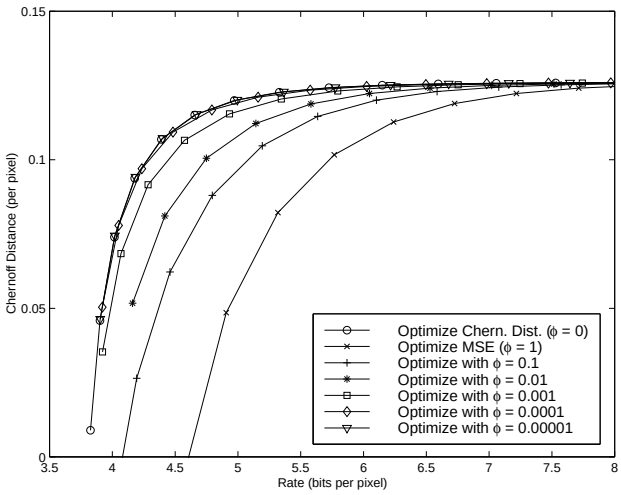

Figure 3: Fine Quantization: Classification performance of optimized coders for the rival image classes 'Water' and 'Sand'.

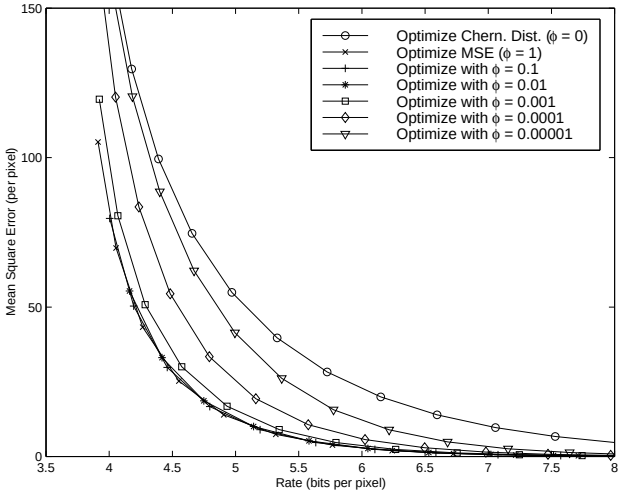

Figure 4: Fine Quantization: Mean square error of optimized coders for the rival image classes 'Water' and 'Sand'.

Fig. 3 and 4 assume high bit rates. However in most practical cases of interest, the rate constraints are more severe, and one needs to use coarse quantization. In this regime asymptotic approximations may be inaccurate. So the design process is computationally intensive. However, design (5) can still be used, and the main conclusion about optimized coders remains valid. In Fig. 5, we illustrate our classification problem at rates ≈ 0.6 bits/pel. As in higher rate, we find the optimized coder ($\phi = 0$) achieves

a higher classification accuracy but leads to poor MSE and visual quality. On the other hand the MSE-optimized coder ($\phi = 1$) leads to good visual quality but poor classification. However we observe that a coder, optimizing a linear combination of these criteria with $\phi = 0.0001$, retains good visual quality (and low MSE) while maintaining high classification accuracy. In real life scenarios (like military applications), where a classification error could lead to disaster, human supervision may be necessary. In such applications, good visual quality and high classification performance are simultaneously desirable, and the proposed design could be very attractive.

6. REFERENCES

[1] J.-W. Nahm and M.J.T. Smith, "Very Low Bit Rate Data Compression Using a Quality Measure Based on Target Detection Performance," *Proc. SPIE*, 1995.

[2] S.R.F. Sims, "Data Compression Issues in Automatic Target Recognition and the Measurement of Distortion," *Opt. Eng.*, Vol. 36, No. 10, pp. 2671—2674, Oct. 1997.

[3] A. Jain, P. Moulin, M. I. Miller and K. Ramchandran, "Performance Bounds for ATR Based on Compressed Data," *Proc. Army Workshop on Data Compression Techniques for Missile Guidance Data Links*, Huntsville, AL, Dec. 1998. Available from `http://www.ifp.uiuc.edu/~moulin/acw98.ps`.

[4] A. Jain, P. Moulin, M. I. Miller and K. Ramchandran, "Information-Theoretic Bounds on Target Recognition Performance," *SPIE's 14th International Symposium on AeroSense 2000*, Orlando, FL, April 2000.

[5] K.O. Perlmutter, S.M. Perlmutter, R.M. Gray, R. A. Olshen, and K. L. Oehler, "Bayes Risk Weighted Vector Quantization with Posterior Estimation for Image Compression and Classification," *IEEE Trans. Im. Proc.*, Vol. 5, No. 2, pp. 347—360, Feb. 1996.

[6] S. Jana and P. Moulin, "Optimal Design of Transform Coders for Image Classification," Proc. Conf. on Information Sciences and Systems, Baltimore, MD, March 1999. Available from `http://www.ifp.uiuc.edu/~moulin/Papers/ciss99.ps.Z`.

[7] H. L. Van Trees, *Detection, Estimation and Modulation Theory, Part I*, John Wiley & Sons, 1968.

[8] T. M. Cover, and J. A. Thomas, *Elements of information Theory*, John Wiley & Sons, 1991.

[9] H. V. Poor, "Fine Quantization in Signal Detection and Estimation," *IEEE Trans. Info. Theory*, Vol. 34, No. 5, pp. 960—972, Sept. 1988.

[10] `http://vismod.www.media.mit.edu/vismod/imagery/VisionTexture/`

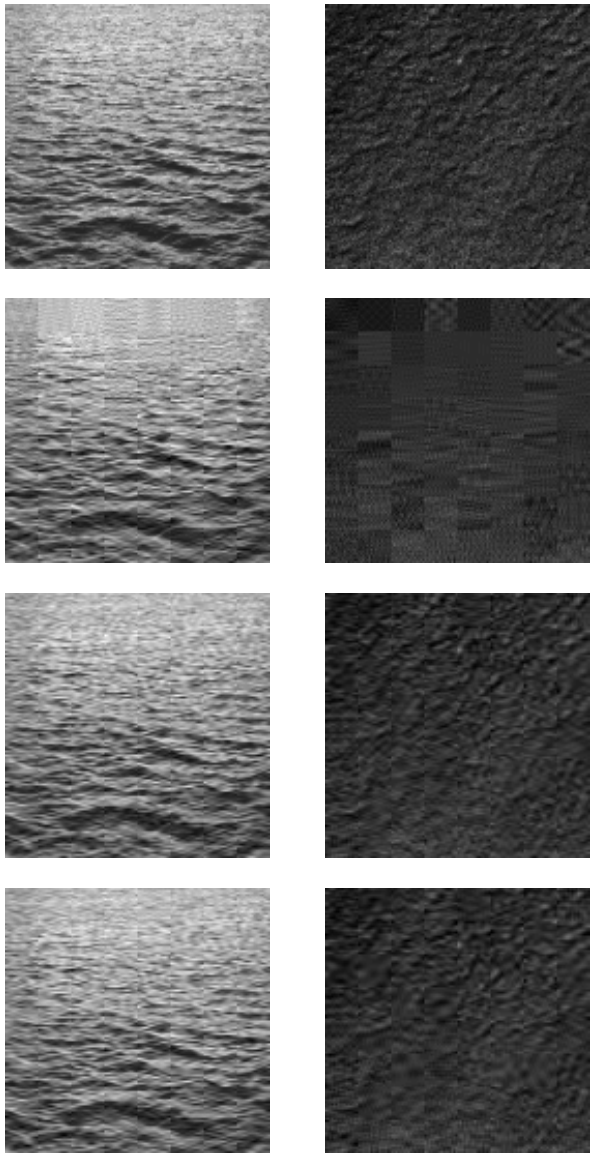

Figure 5: Texture classification example: Coarse quantization. Row 1: Original $\mathbf{1}28 \times 128$ Water and Sand images [10]; Row 2: Water and Sand images reproduced by the coder optimized for Chernoff distance ($\phi = 0$), Rate = 0.59 bits/pel, Chernoff dist. = 0.0904 /pel, MSE = 280 /pel; Row 3: Water and Sand images reproduced by the coder optimized for MSE ($\phi = 1$), Rate = 0.62 bits/pel, Chernoff dist. = 0.0659 /pel, MSE = 126 /pel; Row 4: Water and Sand images reproduced by the coder optimized for a linear combination ($\phi = 0.0001$), Rate = 0.58 bits/pel, Chernoff dist. = 0.0872/pel, MSE = 184 /pel.