Context Quantization by Kernel Fisher Discriminant

Mantao Xu[♀], Xiaolin Wu[*][1], and Pasi Fränti[♀]

[♀]Department of Computer Science
University of Joensuu
Joensuu, Finland
{franti, xu}@cs.joensuu.fi

*Department of Electrical and Computer Engineering
McMaster University
Hamilton, Ontario, Canada L8S 4K1
xwu@mail.ece.mcmaster.ca

Abstract - *Optimal context quantizers for minimum conditional entropy can be constructed by dynamic programming in the probability simplex space. The main difficulty, operationally, is the resulting complex quantizer mapping function in the context space in which the conditional entropy coding is conducted. To overcome this difficulty we propose new algorithms for designing context quantizers in the context space based on multi-class Fisher discriminant and the kernel Fisher discriminant. In particular, the kernel Fisher discriminant can describe linearly-nonseparable quantizer cells by projecting input context vectors onto a high-dimensional curve, in which these cells become better separable. The new algorithms outperform the previous linear Fisher discriminant method for context quantization. They approach the minimum empirical conditional entropy context quantizer designed in the probability simplex space, but with a practical implementation that employs a simple scalar quantizer mapping function rather than a large look-up table.*

Key words: context quantization, entropy coding, Fisher discriminants, image compression.

[1] Supported by NSERC and NSF. Currently visiting University of Joensuu on a Nokia Fellowship.

1. Introduction and Problem Formulation

A key and important task in compressing a discrete sequence X_0, X_1, X_2, … is the estimation of conditional probabilities $P(X_t \mid X^{t-1})$, where $X^{t-1} = X_0, X_1, X_2, \ldots\ X_{t-1}$ is the prefix or context of X_t. Given a class of source models, the model order or the number of parameters must be carefully chosen in the principle of minimum description length or universal source coding. The pioneer solution to the problem is Rissanen's algorithm Context [1], which dynamically selects a variable-order subset of the past samples in X^{t-1}, called the context, C_t. The algorithm structures the contexts of different orders by a tree and it can be shown to be, under certain assumptions, universal in terms of approaching minimum adaptive code length for a class of finite memory sources. A more recent and increasingly popular universal source coding technique is context tree weighting [2]. The idea is to weight the probability estimates associated with different branches of a context tree to obtain a better estimate of $P(X_t \mid X^{t-1})$.

Although the tree-based context modeling techniques have had remarkable success in text compression, applying them to image compression poses a great difficulty. The context tree can only model a sequence not a two-dimensional signal like images. In order to apply the context tree-based techniques to image coding one needs to schedule the pixels (or transform coefficients) of an image into a linear sequence as proposed by the authors of [3][4]. Recently, Mrak *et al.* investigated how to optimize the ordering of the context parameters within the context trees [5]. But any linear ordering of pixels will inevitably destroy the intrinsic two-dimensional sample structures of an image. This is why most image/video image compression algorithms choose a priori two-dimensional context model with fixed complexity, based on domain knowledge such as correlation structure of the pixels and typical input image size, and estimate only the model parameters. For instance, the JBIG standard for binary image compression uses the contexts of a fixed size causal template [6]. The actual coding is implemented by sequentially applying arithmetic coding based on the estimated conditional probabilities.

Estimating the conditional probabilities $P(X_t|C_t)$ directly using count statistics from past samples can incur severe context dilution problem if the number of symbols in the context is large or/and if the symbol alphabet is large with respect to the length of input signal, which is the case for image/video compression. Context quantization is a common technique to overcome this difficulty [7][8][9]. For examples, the state-of-the-art lossless image compression algorithm CALIC [10] and the JPEG 2000 entropy coding algorithm EBCOT [11] quantize the context, C_t into a relatively small number M of conditioning states, and estimate $P(X_t|Q(C_t))$, $1 \leq Q(\cdot) \leq M$, instead of $P(X_t|C_t)$, where Q denotes a context quantizer.

Context quantization is a form of vector quantization because context C is a random vector in the d-dimensional context space E^d (i.e., the context model has order d). Naturally, the objective of optimal context quantization should be minimization of the conditional entropy $H(X| Q(C))$. Although the convexity of the entropy function H implies $H(X|Q(C)) \geq H(X|C)$, we would like to make $H(X|Q(C))$ as close to $H(X|C)$ as possible for a given M, or minimize the Kullback-Leibler distance:

$$D(Q) = H(X \mid Q(C)) - H(X \mid C)\,. \tag{1}$$

Note that in the above H refers to the true source entropy not actual code length which should include the model cost. Although Kullback-Leiber distance (relative entropy) is not strictly a distance metric for its violation of symmetry and triangular inequality, the standard practice is to use it as a non-negative “distortion” of context quantizer Q.

The problem of context quantization in minimizing Kullback-Leibler distance was first studied by Wu [7] and then by Chen [12] for the application of wavelet image compression. Greene *et al.* also developed optimal context quantization algorithm for compression of binary images [13]. Recently, Forchhammer *et al.* proposed a context quantizer design algorithm under the criterion of minimal adaptive code length, and applied it to lossless video coding. A more theoretical treatment of the problem can be found in [8].

The existing context quantizer design algorithms can be classified into two approaches: those that form coding contexts directly in the context space of conditioning events (or the feature space in the terminology of classification and pattern recognition) like [7] and [12], and those that form coding contexts in the probability simplex space [8][9][13]. In the context space one can apply the generalized Lloyd method [13] to design context quantizer by clustering raw contexts of a training set according to Kullback-Leiber distance, which was the idea in [12]. But this iterative approach of gradient descent can not guarantee the globally optimal solution. If the random variable *X* to be coded is binary, then the VQ problem of context quantization can be converted to a scalar quantization problem in the probability simplex space of *P*(*X*). This change of space makes it possible to design globally optimal context quantizer by dynamic programming [8][9][13]. For the sake of rigor we remind the reader that the above mentioned optimality is with respect to the statistics of the chosen training data. In practice, if the statistics of an input image mismatches those of the training set then the coding performance becomes of course suboptimal. Nevertheless, designing optimal context quantizer still has practical significance because situations exist where suitable training set can be found. Furthermore, an off-line optimized context quantizer can be used in conjunction to adaptive arithmetic coding to compensate for any coding loss due to the mismatch of statistics.

Regardless of what space is chosen to design context quantizer, an input context (feature) vector $\mathbf{c}$ (a realization of the random variable C) has to be mapped to a coding state (a context quantizer cell) when it comes to actual context-based coding using $P(X|Q(\mathbf{c}))$. In this regard, both design approaches face a common operational difficulty of complex quantizer mapping function $Q(\mathbf{c})$. Unlike in conventional VQ, the cells (coding states) of optimal context quantizer are not convex or even connected in the context space. Since the quantizer mapping function $Q(\mathbf{c})$ is highly unstructured and complex in the context space of $\mathbf{c}$, its description seems only possible via table look-up. Unfortunately, the table size required by $Q(\mathbf{c})$ grows exponentially in the order of the context. To circumvent this problem the previous authors resorted to prequantization of raw contexts **c**, i.e., limiting the resolution of the context space [12], or the technique of product quantization [13]. Another technique is the projection by linear Fisher discriminant [7]. However, all these techniques compromise optimality. In this paper we reexamine the problem of optimal context quantization and strive to approach the minimal empirical conditional entropy of X under the constraint of a simple quantizer mapping function $Q(\mathbf{c})$. We have made a measured progress in meeting the objective by designing context quantizers using Kernel Fisher discriminant.

The presentation of this paper is organized as follows. Section 2 characterizes the structure of the cells of context quantizer in both probability simplex space and context space, and exposes the complexity of quantizer mapping function. The main results of this research, i.e., the context quantizer design algorithms based on multi-class linear Fisher discriminant and kernel Fisher discriminant, are presented in Section 3. The details of the design algorithm by using kernel Fisher discriminant are given in Section 4. Section 5 presents some experimental results, and the conclusion follows in Section 6.

2. Structure and Complexity of Quantizer Mapping

A context quantizer Q partitions a d-dimensional context space E^d into M subsets:

$$A_m=\{\mathbf{c} \mid \mathrm{Q}(\mathbf{c}) = m\},\ m=1, \ldots, M. \tag{2}$$

The criterion of minimizing the Kullback-Leibler distance in context quantizer design leads to complex structures and shapes of quantizer cells, which are in general not convex or even connected [8]. However, the associated sets of probability mass functions (*pmf*s)

$$B_m = \{P_{X|\mathrm{C}}(\cdot|\mathbf{c}) \mid \mathbf{c}\in A_m\},\ \ m=1, \ldots, M, \tag{3}$$

are simple convex sets in the probability simplex space of X, owing to a necessary condition for minimum conditional entropy quantizer Q [9].

If X is a binary random variable, then the probability simplex is one-dimensional. In this case, the quantizer cells B_m are simple intervals. Let Z=$P_{X|C}(1|\mathbf{c})$ (the conditional probability of $X = 1$ as a function of context **c**) be a random variable, then the conditional entropy $H(X|Q(\mathbf{c}))$ of a context quantizer Q can be expressed by

$$\begin{aligned} &H(X \mid Q(\mathbf{c})) = \sum_{m=1}^{M} P\{Z \in (q_{m-1}, q_m]\} H(X \mid Z \in (q_{m-1}, q_m]) \\ &0 = q_0 < q_1 < \cdots < q_{M-1} < q_M \ = 1 \end{aligned} \tag{4}$$

where the quantizer thresholds $\{q_m\}_{m=1}^{M-1}$ partition the unit interval into M contiguous cells $\{B_m\}_{m=1}^{M}$. Thus the minimal condition entropy context quantizer (MCECQ) can be reduced to a scalar quantization problem in Z, even though the context **c** is drawn from a d-dimensional vector space. The globally optimal solution of the problem

$$(q_1^*, q_2^*, \cdots, q_{M-1}^*) = \underset{0<q_1<\cdots<q_{M-1}<1}{\operatorname{argmin}} \sum_{m=1}^{M} P\{Z \in (q_{m-1}, q_m]\} H(X \mid Z \in (q_{m-1}, q_m]) \tag{5}$$

can be obtained using dynamic programming. Greene *et al.* showed that the MCECQ design problem can be solved in $O(NM)$ time, where N is the number of raw, i.e. unquantized contexts, thanks to a so-called concave Monge property of the objective function (4) [13].

Once Z is scalar quantized for minimal empirical conditional entropy of a training set, the optimal MCECQ cells A_m are formed implicitly by

$$A_m = \{\mathbf{c} \mid P_{X|C}(1 \mid \mathbf{c}) \in (q_{m-1}^*, q_m^*]\} \tag{6}$$

However, $P(X|C)$ is seldom known exactly in practice. Otherwise one would directly drive an entropy coder with $P(X|C)$. Instead, a training set is used to estimate $P(X|C)$. Wu *et al.* [8] showed that the partition of the context space E^d by MCECQ cells, A_m, is generally very complex in shape and structure, resulting highly irregular quantizer mapping function $Q(\mathbf{c})$. An example of the distribution of A_m in the context space is given in Fig. 1. Only when $P_{C|X}$ ($\mathbf{c}|X$=0) and $P_{C|X}$ ($\boldsymbol{c}|X$=1) are of Kotz-type d-dimensional elliptical distributions, the MCECQ cells A_m are bounded by quadratic surfaces [8]. Consequently, the implementation of an arbitrary quantizer mapping function Q becomes an operational difficulty in using MCECQ in practice, which is the main issue that motivated this research.

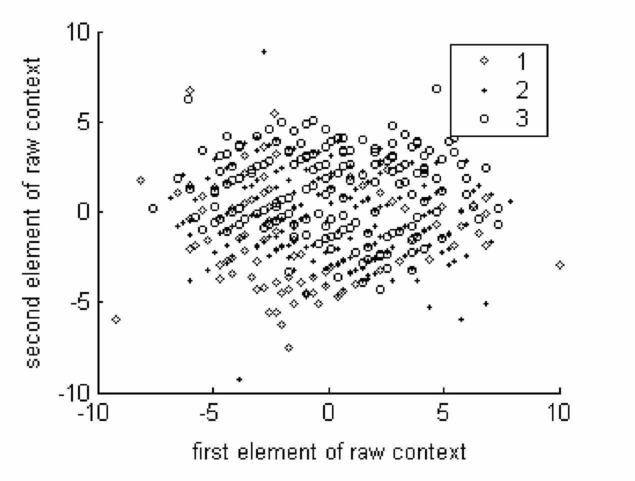


Figure 1: An example distribution of MCECQ cells A_m in context space, for M=3 and the source of least significant bits of DPCM errors of image *cameraman*. The x and y axes represent values of the first two elements in raw context (the two directional gradients $I(i, j\text{-}1)$ - $I(i, j\text{-}2)$ and $I(i\text{-}1, j)$ - $I(I\text{-}2, j)$ as given in (20) and (21)). The symbols ◊, +, and o in the scatter plot are respectively the raw contexts of cells A_1, A_2, and A_3.

The simplest way of implementing Q is to use a look-up table. But since $|C|$, the number of all possible raw contexts, grows exponentially in the order of contexts, building a huge table of $|C|$ entries for Q is clearly impractical. Hashing techniques can be used to avoid excessive memory use of the Q table by exploiting the fact that the actual number of different raw contexts appearing in an input image is much smaller than $|C|$. But this saving of memory is at the expense of increased time of quantizer mapping operation when collision in table access occurs. To achieve constant execution time of the quantizer mapping function the size of hashing table has to be larger than the number of distinct raw contexts by a sufficient factor. In case of image coding, the table size needs to be comparable to the image size since many raw contexts have very low frequency of occurrence.

A common technique to simplify the quantizer mapping function Q is through projection. Wu proposed a suboptimal context quantizer design algorithm based on Fisher's linear discriminant [7]. The idea was to project the training context vectors in the direction $\boldsymbol{y}$ such that the two marginal posterior distributions of $P_{C|X}(\mathbf{y}\cdot\mathbf{c}|X=0)$ and $P_{C|X}(\mathbf{y}\cdot\mathbf{c}|X=1)$, $\mathbf{c}\in E^d$, have maximum separation. Then a dynamic programming algorithm was used to form a convex M-partition of the corresponding one-dimensional projection space to minimize the conditional entropy:

$$H(X\,|\,Q(\mathbf{y}\cdot\mathbf{c})) \tag{7}$$

in which the intervals $(q_{m-1}, q_m]$, $1 \le m \le M$, define the context quantizer Q. In this design approach the context quantizer Q is a scalar one in the projection direction $\mathbf{y}$, i.e., a subspace of the original context space E^d. Although the projection approach is suboptimal, it simplifies the quantizer mapping function to $Q(\mathbf{c}) = m$ if and only if $\mathbf{y}\cdot\mathbf{c} \in (q_{m-1}, q_m]$, which has operational advantages in practice [7].

3. Improved Design Algorithms of Fisher Discriminants

The progress made by this paper is to combine the advantages of the two MCECQ design approaches in the probability simplex space and in the projection context space of Fisher's discriminant. Namely, we seek to attain simultaneously the optimality of MCECQ in probability simplex space and the simplicity of quantizer mapping in the projection space.

3.1 Multi-class Linear Fisher Discriminant

In [7], a linear Fisher discriminant was used to separate the two posterior distributions of $P_{C|X}(\mathbf{c}|X=0)$ and $P_{C|X}(\mathbf{c}|X=1)$, which is a two-class clustering problem. However, the success of this approach is limited to cases where $P_{C|X}(\mathbf{c}|X=0)$ and $P_{C|X}(\mathbf{c}|X=1)$ are linearly separable to certain degree. But for more difficult, linearly non-separable shapes of context cells a departure from [7] is needed. We seek to separate the *M* optimal MCECQ cells formed in the probability simplex space via a suitable, non-linear projection of the context space. The goal is to apply the discriminant classifier to form a convex partition in the projection subspace that best matches the optimal partition of B_m's in the probability simplex space. The multi-class Fisher discriminant [15] lends us a tool to design a classifier that approximates the optimal partition of contexts in the probability simplex space by an optimized partition in a projection subspace. The separation of input classes (i.e., the partition of A_m's formed by MCECQ in the context space) in projection direction **y** can be measured by the so-called F-ratio validity index, $J(\mathbf{y})$, defined as the ratio of between-class variance versus within-class variance:

$$J(\mathbf{y}) = \frac{\sum_{j=1}^{M} n_j (\mathbf{y}^T(\mathbf{m}_j - \bar{\mathbf{x}}))^2}{\sum_{i=1}^{N} (\mathbf{y}^T(\mathbf{x}_i - \mathbf{m}_{\pi(i)}))^2} \tag{8}$$

where $\pi(i)$ is the class label of each sample x_i and $\bar{\mathbf{x}}$ is the mean vector of all raw context samples. The multi-class linear Fisher discriminant is the maximization of F-ratio validity index in (8), i.e.,

$$\mathbf{y} = \arg\max_{\mathbf{v}} \frac{\mathbf{v}^{\mathrm{T}}\mathbf{S}_B\mathbf{v}}{\mathbf{v}^T\mathbf{S}_W\mathbf{v}} \tag{9}$$

where **v** represents a discriminant vector in raw context space. $\mathbf{S_B}$ and $\mathbf{S_W}$ in (9) are the between-class covariance matrix and the within-class covariance matrix respectively:

$$\begin{aligned} \mathbf{S}_B &= \sum_{j=1}^{M} n_j (\mathbf{m}_j - \bar{\mathbf{x}})(\mathbf{m}_j - \bar{\mathbf{x}})^{\mathrm{T}}, \\ \mathbf{S}_W &= \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{m}_{\pi(i)})(\mathbf{x}_i - \mathbf{m}_{\pi(i)})^T \end{aligned} \tag{10}$$

where $\mathbf{m}_{\mathrm{j}}$ and n_{j} are the mean vector and sample size of class *j* in context space respectively. After the projection direction **y** is determined by (9), one can still apply dynamic programming to the projected samples **y**·**c** to optimize context quantizer the same way as in (7).

3.2 Kernel Fisher Discriminant

The multi-class linear Fisher discriminant outperformed the two-class linear Fisher discriminant in terms of designing context quantizers of shorter code length in our experiments (see Section 5). But the contexts of different MCECQ cells (input classes for the Fisher discriminant) are not linearly separable in the context space as shown in [8]. A superior alternative is to use a non-linear classifier of higher discriminating power. Encouraged by the success of the kernel-based learning machines, such as support vector machine, kernel principal component analysis and kernel Fisher discriminant analysis (KFD) in many other classification and learning applications [16][17][18][19][20], we propose a new design technique of context quantizers by using the multi-class kernel Fisher discriminant. The multi-class kernel Fisher discriminant has been intensively studied as a generalization of discriminant analysis using kernel approach [21][22]. As an extension of Fisher discriminant, the kernel one is known for its high discriminating powers on the input clusters of complex structures. The kernel discriminant first maps the source feature vectors (or context vectors in MCECQ design) into some new feature space F in which different classes are better separable. A linear discriminant is computed to separate input classes in F. Implicitly, this process constructs a non-linear classifier of high discriminating power in the original feature space. In our application of context quantization, the objective of the kernel discriminant is, given an M input partition $A_m=\{\mathbf{c}: Q(\mathbf{c}) = m\}$, $1 < m < M$, to find a projection direction $\boldsymbol{y}$ in a new feature space F such that different A_m's are most separable in $\mathbf{y}$. A dynamic programming algorithm is then applied to design an MCECQ in $\mathbf{y}$. The resulting MCECQ in F implicitly constructs a context quantizer in the context space E^d.

Let $\Phi(\mathbf{c})$ be the nonlinear mapping from context space to some high-dimensional Hilbert space F. Our goal is to find the projection line $\boldsymbol{y}$ in F such that the F-ratio validity index $J(\mathbf{y})$

$$J(\mathbf{y}) = \frac{\mathbf{y}^T \mathbf{S}_B^{\Phi} \mathbf{y}}{\mathbf{y}^T \mathbf{S}_W^{\Phi} \mathbf{y}} \tag{11}$$

is maximized, where $\mathbf{S}_B^\Phi$ and $\mathbf{S}_W^\Phi$ are the between-class and within-class covariance matrices. Since the space F is of very high or even infinite dimensions, the function $\mathbf{\Phi}(\mathbf{c})$ is infeasible. A technique to overcome this difficulty is the Mercer kernel function $k(\mathbf{x}, \mathbf{y}) = (\mathbf{\Phi}(\mathbf{x})\cdot\mathbf{\Phi}(\mathbf{y}))$, which is the dot product in Hilbert feature space F. A popular choice for the kernel function k that has been proved useful (e.g. in support vector machines) is the *Gaussian* **RBF** (radial basis function), $k(\mathbf{x},\mathbf{y}) = exp(-\|\mathbf{x}\text{-}\mathbf{y}\|^2/2\sigma)$. It is known that under some mild assumptions on $\mathbf{S}_B^\Phi$ and $\mathbf{S}_W^\Phi$, any solution $\mathbf{y}\in F$ maximizing (11) can be written as the linear span of all mapped context samples [19]:

$$\mathbf{y} = \sum_{j=1}^{N} \alpha_j \mathbf{\Phi}(\mathbf{c}_j) \tag{12}$$

As a result, the F-ratio $J(\mathbf{y})$ can be reformulated as:

$$J(\mathbf{y}) = \frac{\mathbf{y}^{\mathrm{T}} \mathrm{S}_{\mathrm{B}}^{\Phi} \mathbf{y}}{\mathbf{y}^{\mathrm{T}} S_W^{\Phi} \mathbf{y}} = \frac{\alpha^T \mathbf{A} \alpha}{\alpha^T \mathbf{B} \alpha} \tag{13}$$

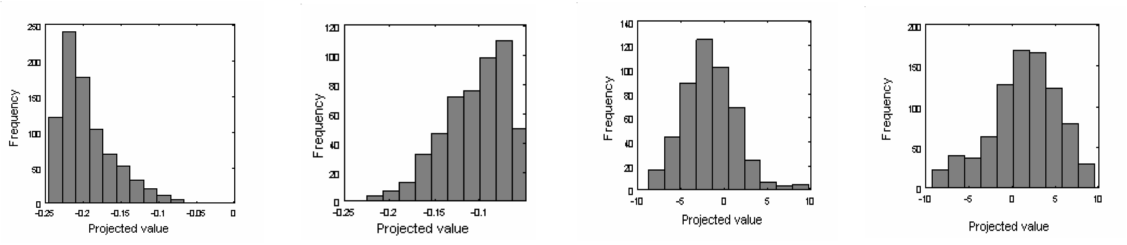
where $\mathbf{A}$ and $\mathbf{B}$ are $N\times N$ matrices:

$$\begin{aligned} \mathbf{A} &= \sum_{j=1}^{M} n_j (\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_j)(\bar{\boldsymbol{\mu}} - \boldsymbol{\mu}_j)^T, \\ \mathbf{B} &= \mathbf{K}\mathbf{K}^T - \sum_{j=1}^{M} n_j \boldsymbol{\mu}_j \boldsymbol{\mu}_j^T \end{aligned} \tag{14}$$

where $\mathbf{K}$ is the kernel matrix, $\mathbf{K}_{ij}$=$\mathbf{\Phi}(\mathbf{c}_i)\cdot\mathbf{\Phi}(\mathbf{c}_j)$, $\boldsymbol{\mu}_j = \mathbf{K}\cdot\mathbf{1}_j / n_j$, $\bar{\boldsymbol{\mu}} = \mathbf{K}\cdot\mathbf{1}/N$, $\mathbf{1}_j\in(0,1)^N$ are membership vectors corresponding to class labels, and $\mathbf{1}$ is the vector of all ones. The projection of a test context $\boldsymbol{c}$ onto the discriminant is given by the inner product

$$(\mathbf{y}, \Phi(\mathbf{c})) = \sum_{j=1}^{N} \alpha_j k(\mathbf{c}, \mathbf{c}_j) \tag{15}$$

where $k(\mathbf{x},\mathbf{y}) = exp(-\|\mathbf{x}\text{-}\mathbf{y}\|^2/2\sigma)$ is the **RBF** kernel function. The superior discriminating power of KFD over the linear Fisher discriminant (LFD) method of [7] for MCECQ design is illustrated in Fig. 2. The plots are for the context vectors in the modeling of the least significant bit of the test image cameraman. By comparing the histograms of the projected MCECQ cells A_1 and A_2 from *Cameraman* image (for case of $M = 2$) for the two methods respectively, one can easily see that KFD offers significantly better separation of A_1 and A_2 than LFD. Note that the projection of KFD is in general non-linear unlike the classic LFD.



Projected A_1 by KFD. Projected A_2 by KFD. Projected A_1 by LFD. Projected A_2 by LFD.

Figure 2. Comparison of the kernel and linear Fisher discriminants in the separablility of two MCECQ cells in the projection subspace .

Computationally, the KFD problem is to find the leading eigenvector of $\mathbf{B}^{-1}\mathbf{A}$. As the dimension of F is higher than the number of source samples N, and B is a highly singular $N\times N$ matrix obtained from only N source samples, some form of regularization is necessary. The simplest solution is to add either the identity or kernel matrix $\mathbf{K}$ to matrix $\mathbf{B}$, namely matrix B is replaced by $\mathbf{B}_\beta = \mathbf{B} + \beta\mathbf{I}$. This makes the problem numerical more stable because the within-class matrix $\mathbf{B}$ becomes more positive definite for large β. It is also roughly equivalent to add independent noises to each of the kernel bases.

4. Implementation of KFD for Context Quantization

In the above formulations, matrices $\mathbf{B}$ and $\mathbf{A}$ are too large in size in practice. Maximizing (13) takes $O(N^3)$ time since it needs to solve the $N\times N$ matrix eigenvalue problem. This complexity is too high for large N. More importantly in context quantization application, we are not able to use all the basis functions corresponding to all raw training contexts. Solving the kernel Fisher discriminant for two classes can be cast to a quadratic optimization problem [18][19]. However, this scheme can not be directly applied to estimating the multi-class kernel Fisher discriminant.

The possible solution applicable to any choice of **A** and **B** is to restrict the discriminator **y** to be in a subspace of *F*, as proposed in [19][20]. Instead of using (12), we express **y** in the subspace:

$$\mathbf{y} = \sum_{j=1}^{l} \alpha_j \Phi(\mathbf{c}_j) \tag{16}$$

where $l << N,$ and samples $\mathbf{c}_j$ could be either selected from all raw training context samples or estimated by some clustering algorithms. Without loss of generality, if we choose each $\mathbf{c}_j$ in (16) from the training set, $1 \le j \le l$, then:

$$J(\mathbf{y}) = \frac{\boldsymbol{\alpha}^{\mathrm{T}}(l)\mathbf{A}(l)\boldsymbol{\alpha}(l)}{\boldsymbol{\alpha}^{\mathrm{T}}(l)\mathbf{B}(l)\boldsymbol{\alpha}(l)} \tag{17}$$

where $\alpha(l)$ is l-dimensional vector, and $\mathbf{A}(l)$ is an $l \times l$ matrix:

$$\mathbf{A}(l) = \sum_{j=1}^{M} n_j (\overline{\boldsymbol{\mu}}(l) - \boldsymbol{\mu}(l)_j)(\overline{\boldsymbol{\mu}}(l) - \boldsymbol{\mu}(l)_j)^{\mathrm{T}} \tag{18}$$

and $\mathbf{B}(l)$ is an $l \times l$ matrix:

$$\mathbf{B}(l) = \mathbf{K}(l)\mathbf{K}(l)^T - \sum_{j=1}^{M} n_j \boldsymbol{\mu}(l)_j \boldsymbol{\mu}(l)_j^T \tag{19}$$

with $\mathbf{K}(l)$ being an $l \times N$ sub-matrix of **K**, $\boldsymbol{\mu}(l)_j = \mathbf{K}(l) \cdot \mathbf{1}_j / n_j$ and $\overline{\boldsymbol{\mu}}(l) = \mathbf{K}(l) \cdot \mathbf{1} / N$.

Given the dimension *l* of the subspace of *F*, the partial expansion (16) presents a greedy approximation of the optimal KFD solution, which was described in [19][20] and studied theoretically as the reduced set method for supported vector machines in [23]. This approximation can be incrementally improved by adding a raw context sample or a new context base one at a time to the existing expansion, i.e., incrementing the dimensionality *l* by one at a time. Such incremental expansion can be done in a greedy fashion as follows. For each iteration we first randomly select a subset *U* of the remaining training set, then we conduct an exhaustive search in *U*, instead of in the whole remaining training set, for the training context **c** that maximizes (17) after **c** being added to (16). The proper size of U was shown to be 59 in order to obtain nearly as good a performance as if the search was through the entire remaining training set

[24]. Since $l<<N$, incrementing the kernel expansion (16) by one base context merely takes $O(59\times N\times l)$ time. Consequently, the approximation of the kernel discriminant in l-dimensional subspace of F has $O(59\times N\times l^2)$ time complexity, which is drastically lower than $O(N^3)$. The pseudo code of this practical approximation algorithm of kernel Fisher discriminant for context quantization is presented in Fig. 3.

```
input:     $C = \{\mathbf{c}_1, \mathbf{c}_2, \dots \mathbf{c}_N\}$: a set of raw training contexts.
           $l_{\max}$: the maximum number of expansion coefficients.
           $T$: stopping threshold in relative entropy.
           $M$ : the number of context quantizer cells.

output:    $I$ : the set of bases in the linear spanning as in (16).
           $\boldsymbol{\alpha} = \{\alpha_j | 1 \le j \le l_{max}\}$: kernel Fisher discriminant coefficients as in (16).
           $P(1|Q(\Phi(\mathbf{c})\cdot \mathbf{y}) = j)$, $1 \le j \le M$: the empirical conditional probabilities in context cells in the projection
             subspace.
           $(q_{j-1}, q_j]$: context quantizer intervals in the projection subspace.

Function   ContextVQKFD ($C$, $T$, $l_{max}$, $M$)
           $D_{\text{opt}}(Q) \leftarrow$ solve the MCECQ problem by dynamic programming in probability simplex space.
           $l \leftarrow 0$; $I \leftarrow \varnothing$; $\delta \leftarrow \infty$.
           while $\delta > T$ and $l < l_{\max}$
               $S \leftarrow$ randomly pick 59 elements from $C \setminus I$
               $l \leftarrow l + 1$
               $J_{\text{KFD}} \leftarrow$ initialize the KFD F-ratio as 0
               for $\mathbf{z} \in S$ do
                   $I^* \leftarrow I \cup \{\mathbf{z}\}$;
                   Update covariance matrices $\mathbf{A}(l)$ and $\mathbf{B}(l)$ as in (18) and (19) for $I^*$;
                   $\boldsymbol{\alpha}^* \leftarrow$ leading eigenvector of matrix $\mathbf{A}^{-1}(l)\mathbf{B}(l)$;
                   $J^* \leftarrow$ update F-ratio of $\mathbf{A}(l)$ and $\mathbf{B}(l)$ for $\boldsymbol{\alpha}^*$;
                   If $J^* > J_{\text{KFD}}$ then
                       $J_{\text{KFD}} \leftarrow J^*$; $\mathbf{c}_l \leftarrow \mathbf{z}$; $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha}^*$
                   end if
               end for
               $I \leftarrow I \cup \{\mathbf{c}_l\}$;
               $C_{\text{proj}} \leftarrow$ project all contexts $\mathbf{c} \in C$ into the projection direction by $\mathbf{y} \cdot \mathbf{\Phi}(\mathbf{c}) = \sum_{1 \le j \le l} \alpha_j k(\mathbf{c}, \mathbf{c}_j)$;
               obtain $(q_{j-1}, q_j]$, $P(1|Q(\Phi(\mathbf{c})\cdot \mathbf{y}) = j)$, $D_{\text{KFD}}(Q)$ by solving the MCECQ problem by dynamic
                 programming in projection subspace $C_{\text{proj}}$;
               $\delta \leftarrow D_{\text{KFD}}(Q) - D_{\text{opt}}(Q)$.
           end while
           return $I$, $\boldsymbol{\alpha}$, $(q_{j-1}, q_j]$ and $P(1|Q(\Phi(\mathbf{c})\cdot \mathbf{y}) = j)$.
```

Figure 3: Pseudocode of context quantization by kernel Fisher discriminant

We build the context quantizer in three steps. In the first step, we apply the dynamic programming algorithm to design MCECQ in the probability simplex space. This produces the MCECQ cells B_m that constitutes the input classes of KFD. In the second step, we map B_m back to A_m in the context space, and use the kernel Fisher discriminant to find a projection direction in F (corresponding to a curve in the context space) in which MCECQ cells A_m have maximum separation. In the final step, we compute all projection values of training contexts and put them into a sorted list. Since each class in projection direction in general is not convex, in order to make the underlying classification problem tractable and more importantly make the quantizer mapping function simple, the dynamic programming is used again to construct a convex partition of the projection subspace that minimizes the conditional entropy $H(X|\mathbf{y}\cdot\mathbf{\Phi}(\mathrm{c}) \in (q_{m\text{-}1}, q_m])$, where $\mathbf{y}\cdot\mathbf{\Phi}(\mathbf{c}) = \sum_{1 \le j \le l} \alpha_j k(\mathbf{c}, \mathbf{c}_j)$.

5. Experimental Results

We implemented the proposed context quantizers and evaluated them in DPCM predictive lossless coding of gray scale images. The prediction residuals are coded by binary arithmetic coding that uses context states optimized by the proposed algorithms. The binary random variables to be coded are the binary decisions in resolving the value of the prediction residual. In particular we are interested in two binary sources: the signs of DPCM prediction errors on grey scale images, and the least significant bits of the DPCM prediction errors. These binary sources are among the most difficult to compress with their self entropy being maximum (*1* bit per sample), and thus present great challenges to context-based entropy coding. Consequently, they serve as good, demanding test cases for the performance of different context quantizers.

The causal context in which the current pixel $I(i, j)$ is coded consists of three gradients in a local window $\mathbf{c} = (c_1,c_2,c_2)$:

$$\begin{aligned} c_1 &= I(i, j-1) - I(i, j-2) \\ c_2 &= I(i-1, j) - I(i-2, j) \\ c_3 &= I(i-1, j) - I(i, j-1) \end{aligned} \qquad (20)$$

The reason for choosing (c_1,c_2,c_2) as feature vector in context modeling is because they capture the variance and signal the presence of edge structure in the image signal while keeping the dimensionality of the feature space low. We did not use higher order context model to avoid overfitting in the coding phase. Even this three-dimensional feature space generates a very large number of raw contexts, namely 512^3. A scalar prequantization scheme:

$$Q_k(c_i) = \begin{cases} j, & \text{if } \; c_i \in [2^j - 1, 2^{j+1} - 1),\, 0 \le j < k \\ -j, & \text{if } \; c_i \in (-2^{j+1} + 1, \text{-}2^j + 1],\, 0 < j < k \\ k, & \text{if } \; c_i \ge 2^k - 1 \\ -k, & \text{if } \; c_i \le -2^k + 1 \end{cases} \tag{21}$$

is used to reduce the number of raw contexts to a manageable level of $(2\times k+1)^3$ (k was chosen to be 6 in our experiments). Since the gradient is the difference of adjacent samples, it obeys geometrical distribution for natural images. The above scalar prequantization merges the raw contexts into equally probable regions.

The training set of raw contexts was generated out of 23 images that were samples from two archives of benchmark gray scale images on the Internet [25][26]. The test set consisting of images *airplane*, *barb*, *boat*, *cameraman*, *couple*, *crowd, girl, lena*, *peppers*, *tiffany*, is disjoint from the training set. The model parameters (β,σ) to construct the kernel discriminants for the two training sets are chosen as (0.0076, 4.16) and (0.0043, 5.33) respectively, which can be estimated by applying the cross-validation [27][28] estimation of the minimized misclassification rate or desirable minimum conditional entropy. Either the encoding or decoding of each binary symbol by a KFD context quantizer needs projecting a context to the discriminant direction in $O(l)$ time according to (16). Thus, the encoding or decoding complexity of a KFD context quantizer is $O(l\times N)$, where N is the length of input sequence.

We compare three context quantizers of Fisher discriminant type reviewed and developed in this paper. Namely, LFD-I: the two-class linear Fisher discriminant scheme of [7], LFD-II: the multi-class linear Fisher discriminant scheme discussed in Section 3.1, and KFD: the MCECQ design algorithm based on kernel Fisher discriminant developed in Section 3.2 and Section 4. All the three context quantizer design algorithms output convex quantizer cells in the context space with simple quantizer mapping functions. As a performance benchmark we also include the ideal results, i.e., the conditional entropy rates of MCECQ quantizer in the probability simplex space, against which the testing results of the three practical methods are measured. These rates were obtained by MCECQ designed for the sample statistics of each individual test image. Clearly, these rates serve as a theoretical lower bound with respect to the context model in question, since they are the best achievable in the ideal situation when the training data and input image have identical statistics and as though the quantizer mapping function, regardless how complex, could be precisely implemented in practice.

Figures 4 and 5 plot the average bit rates achieved by the three MCECQ design methods in the context space, LFD-I, LFD-II and KFD*,* on coding the sign and the least significant bit of DPCM errors for the ten test images. The bit rates are presented as functions of the number of context quantizer cells. As lower bounds for the achievable bit rates by any convex partition of the context space, we also include in the figures the corresponding average conditional entropy rates of optimal MCECQs designed in the probability simplex space as explained above. It can be observed from our experimental results, as expected, that LFD-II outperforms LFD-I, and KFD outperforms the two variants of linear discriminant type, because KFD has higher discriminating power than the other two with its capability of forming more complex quantizer cells. In fact, the KFD method achieves the bit rates that are less than 0.5% away from the lower bound.

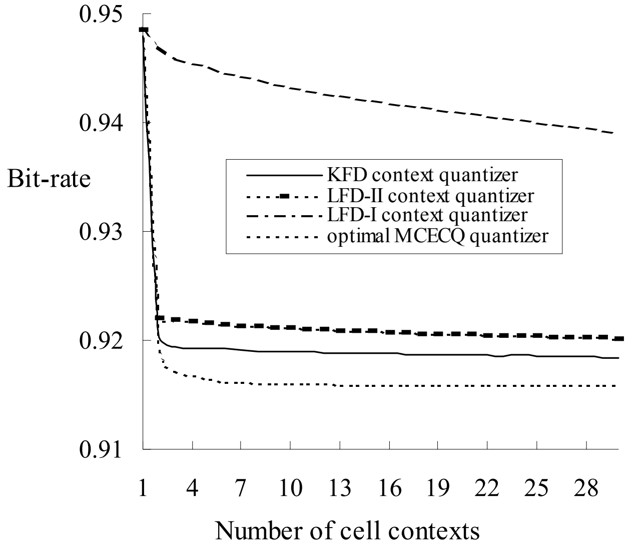


Figure 4: Average bit rates achieved by the four context quantizers on coding the sign of DPCM error pixel in bits/sample.

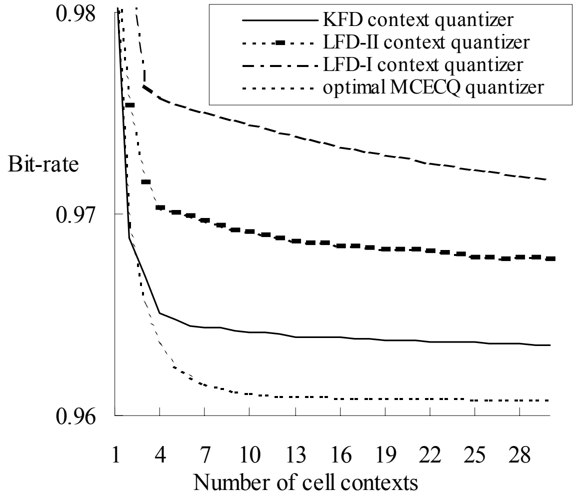


Figure 5: Average bit rates achieved by the four context quantizers on coding the least significant bit of DPCM error pixel in bits/sample.

We apply the three context quantizers designed from the training set to encode the signs and the least significant bits of DPCM errors from 10 test images outside of the training set. All three context quantizer have 12 cells, in other words the conditional entropy coding is carried out in 12 coding states. Tables 1 and 2 show the actual code lengths obtained by the three context quantizers. Not surprisingly, the kernel Fisher discriminant in general outperforms the two linear ones.

Table 1: Bit rates of signs of DPCM errors for different methods.

Image	*Lower bound*	*LFD-I*	*LFD-II*	*KFD*
airplane	0.903412	0.919363	0.906214	0.906678
barb	0.903873	0.939119	0.907764	0.907621
boat	0.925852	0.943870	0.928001	0.926753
cameraman	0.892693	0.909089	0.896163	0.895801
couple	0.914312	0.921110	0.916768	0.917992
crowd	0.932237	0.948894	0.936294	0.935742
girl	0.914502	0.945503	0.919236	0.919035
lena	0.917931	0.944266	0.921174	0.921701
peppers	0.957923	0.985236	0.961451	0.960999
tiffany	0.928765	0.949091	0.932569	0.932043

Table 2: Bit rates of least significant bits of DPCM errors for different methods.

Image	*Lower bound*	*LFD-I*	*LFD-II*	*KFD*
airplane	0.959321	0.982544	0.972848	0.968383
barb	0.983122	0.994972	0.991413	0.987895
boat	0.978024	0.990999	0.986999	0.980999
cameraman	0.946543	0.971188	0.958581	0.949875
couple	0.893815	0.909343	0.903141	0.900728
crowd	0.953596	0.957038	0.953710	0.957381
girl	0.979238	0.992968	0.986548	0.983537
lena	0.986358	0.992127	0.991570	0.989302
peppers	0.991213	0.994391	0.991873	0.993025
tiffany	0.979252	0.991235	0.987100	0.982065

Table 3: Bit rates of lossless image compression by different methods.

Image	*KFD*	*LFD-I*	*LFD-II*	*JPEG-LS*
airplane	4.530	4.795	4.727	4.582
barb	4.830	5.083	5.060	4.862
boat	4.843	5.092	5.028	4.907
cameraman	4.244	4.519	4.450	4.314
couple	3.603	3.730	3.701	3.658
crowd	4.932	5.181	5.132	5.048
girl	4.050	4.206	4.157	4.125
lena	4.492	4.685	4.648	4.581
peppers	4.758	4.918	4.879	4.847
tiffany	4.350	4.504	4.486	4.435

Table 3 presents the lossless bit rates of the ten test images achieved by adaptive binary arithmetic coding that uses the modeling contexts designed by the proposed MCECQ methods for each binary decision. As a reference in comparison the bit rates of the JPEG-LS lossless image coding standard are also listed in the table. The comparison is fair and meaningful because JPEG-LS uses the same context template as in our experiments but it employs a heuristic context quantization scheme [29]. The proposed KFD-based context quantizer has a small improvement over JPEG-LS. The small margin between the two methods indicates that the heuristic context quantizer of JPEG-LS is already very good compared with a heavily optimized one. We envision this work to be a useful algorithmic tool to evaluate the quality of more practical context quantizers.

6. Conclusions

We proposed new algorithms for designing context quantizers toward minimum conditional entropy based on multi-class Fisher discriminant and the kernel Fisher discriminant. We succeeded in approaching the lower bound of the achievable bit rates with a practical implementation that employs a simple scalar quantizer mapping function rather than a large look-up table.

7. References

[1] J. Rissanen, "A universal data compression system", *IEEE Trans. Info. Theory*, vol. 29, no. 5, pp. 656-664, Sept. 1983.

[2] F. Willems, Y. Shtarkov, T. Tjalkens, "The context-tree weighting method: basic properties", *IEEE Trans. Inf. Theory*, vol. 41, no. 3, pp. 653-664, 1995.

[3] N. Ekstrand, "Lossless compression of grayscale images via context tree weighting", *Proc. of IEEE Data Compression Conf.*, pp. 132-139, Apr. 1996.

[4] M. Arimura, H. Yamamoto and S. Arimoto, "A bitplane tree weighting method for lossless compression of gray scale images", *IEICE Trans. Fundamentals*, vol. E80-A, no. 11, pp. 2268-2271, Nov. 1997.

[5] M. Mrak, D. Marpe and T. Wiegand, “A context modeling algorithm and its application in video compression” *Proc. 2003 Int. Conf. on Image Processing*, Barcelona, Spain, Sept. 2003.

[6] CCITT Draft Recommendation T.82, ISO/IEC Draft International Standard 11544, *Coded Representation of Picture and Audio Information – Progressive Bi-level Image Compression*, Apr. 1992.

[7] X. Wu, “Context quantization with fisher discriminant for adaptive embedded wavelet image coding”, *Proc. 1999 IEEE Data Compression Conf.*, pp.102-111, Mar. 1999.

[8] X. Wu, P. A. Chou and X. Xue, “Minimum conditional entropy context quantization”, *Proc. of 2000 IEEE Int'l. Symp. Inform. Theory*, 2000, pp. 43, 2000.

[9] S. Forchhammer, X. Wu and J.D. Andersen, “Lossless image data sequence compression using optimal context quantization”, *IEEE Transaction on Image Processing*, vol. 13, no. 4, pp. 509-517, Apr. 2004.

[10] X. Wu and N. Memon, “Context-based, adaptive, lossless image codec”, *IEEE Trans. on Communications*, vol. 45, no. 4, pp. 437-444, April 1997.

[11] D. Taubman, “High performance scalable image compression with EBCOT”, *IEEE Trans. Image Processing*, vol. 9, pp. 1158-1170, July 2000.

[12] J. Chen, “Context modeling based on context quantization with application in wavelet image coding”, *IEEE Transaction on Image Processing*, vol. 13, no. 1, Jan. 2004.

[13] D. Greene, F. Yao, T. Zhang, “A linear algorithm for optimal context clustering with application to bi-level image coding”, *Proc. 1998 Int’l. Conf. Image Processing,* pp. 508-511, 1998.

[14] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. New York: Kluwer Academic Publishers, 1992.

[15] D. H. Foley and J. W. Sammon, “A optimal set of discriminant vectors”, *IEEE Transactions on Computers*, vol. 3, no. 24, pp. 281-289, 1975.

[16] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.-R. Müller**,** “Fisher discriminant analysis with kernels**”**, In Y.-H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Neural Networks for Signal Processing IX*, pages 41-48. IEEE, 1999.

[17] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A.J. Smola, and K.-R. Müller, “Invariant feature extraction and classification in kernel spaces”, In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pp. 526-532. MIT Press, 2000.

[18] S. Mika, G. Rätsch, and K.-R. Müller, “A mathematical programming approach to the Kernel Fisher algorithm”, In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pp. 591-597, MIT Press, 2001.

[19] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola and K. R. Müller, "Constructing Descriptive and Discriminative Nonlinear Features: Rayleigh Coefficients in Kernel Feature Spaces", *IEEE Transactions on PAMI.*, vol. 25 , no. 5 pp. 623 – 633, 2003.

[20] S. Mika, A.J. Smola, and B. Schölkopf, "An improved training algorithm for kernel fisher discriminants", In T. Jaakkola and T. Richardson, editors, *Proceedings AISTATS 2001*, pp. 98-104, San Francisco, CA, 2001.

[21] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach", *Neural Computation*, 12(10): 2385-2404, 2000.

[22] P. Navarrete and J. Ruiz del Solar**,** "On the Generalization of Kernel Machines", *First International Workshop on Pattern Recognition with Support Vector Machine - Lecture Notes in Computer Science 2388 Springer 2002*, pp. 24-39, Niagara Falls, Canada, August 10, 2002.

[23] B. Schölkopf, S. Mika, C.J.C. Burges, P. Knirsch, K.-R. Müller, G. Rätsch, and A.J. Smola, "Input space vs. feature space in kernel-based methods", *IEEE Transactions on Neural Networks*, 10(5): pp. 1000-1017, September 1999

[24] A. J. Smola and B. Schölkopf, "Sparse Greedy Matrix Approximation for Machine Learning", In Pat Langley editor, *Proceedings of the Seventeenth International Conference on Machine Learning* (*ICML 2000*), pp. 911-918, Stanford University, Stanford, CA, USA, June 29 - July 2, 2000.

[25] ftp://links.uwaterloo.ca:/pub/BragZone/

[26] http://www.cipr.rpi.edu/resource/stills/index.html

[27] G.C. Cawley and N.L.C. Talbot**,** "Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers", *Pattern Recognition,* 36(11): pp. 2585-2592, November 2003.

[28] G. Fung, M. Dundar, J. Bi, and B. Rao, "A fast iterative algorithm for fisher discriminant using heterogeneous kernels", In Carla E. Brodley editor, *Proceedings of the Twenty-first International Conference* (*ICML 2004*), Banff, Alberta, Canada, July 4-8, 2004.

[29] CCITT Draft Recommendation T.87, ISO/IEC Final Draft International Standard FDIS14495-1, *Information technology – Lossless and near-lossless compression of continuous-tone still images*, 1998.