An Iterative Algorithm for Context Selection in Adaptive Entropy Coders

Jacques Vaisey and Tong Jin
School of Engineering Science
Simon Fraser University, Burnaby, BC, V5A 1S6, CANADA

{vaisey,tjin}@cs.sfu.ca

ABSTRACT

Context-based adaptive entropy coding is an essential feature of modern image compression algorithms; however, the design of these coders is non-trivial due to the balance that must be struck between the benefits associated with using a large number of conditioning classes, or contexts, and the penalties resulting from data dilution. The problem is especially severe when coding small sub-images where the amount of data available is small. In this paper, we propose an iterative algorithm that begins with a large number of conditioning classes and then uses a clustering procedure to reduce this number to a desired value. This method is in contrast to the more usual approach of defining contexts in an ad-hoc manner. Experiments are conducted on synthetic data sources having varying amounts of memory, as well as on the sub-images resulting from a wavelet decomposition of an image. The results show that our approach to context selection is effective and that the algorithm automatically learns the structure of the data. This technique could be applied to improve the performance of both image and video coders.

1. INTRODUCTION

Adaptive arithmetic entropy coders are used in newer compression standards to achieve rates that are asymptotically close to the source entropy; however, the algorithm takes time to learn the source statistics and the performance suffers when the amount of data to code is small. This issue is significant, since image coders such as JPEG 2000 [1] partition their images into relatively small blocks prior to entropy coding. The problem is compounded by the fact that multimedia sources often contain significant memory (even after a decorrelating transform) and a conditional entropy-coding approach must be taken: separate arithmetic coders are used for each of the possible conditioning classes. The result can be a severe reduction in the amount of data available for learning. To combat this problem, a balance must be struck between the benefit of using many conditioning classes to lower the entropy and the cost associated with data dilution. Intelligent ways must be found to define the reduced set of contexts

In an arithmetic coder, the task of the context model is to accurately estimate the probability mass function (pmf) of the upcoming symbol. This estimation is generally made by counting the number of occurrences of each symbol in each context and computing relative frequencies. The counts can be organized into a histogram for each context, with one entry per symbol. There are many possible ways to form contexts; however, two common techniques are to use finite-state machines [2] and prediction [3][5]. In the former case, the values of previously coded symbols (generally, but not always, a raster scan for image data) are used to index a "state" that then forms the context. The "raw" number of possible contexts is finite, since the spatial extent of the state information is limited by a template; however, the number can also be very large – especially when non-binary alphabets are used. In practice, the number of contexts must be substantially reduced and this is often done by assuming Markov data and only forming the context from spatially adjacent samples. Ad hoc "context quantization" procedures may also be used to reduce the number of contexts further. As an example, see the entropy coding method in JPEG 2000 [1]. The second approach, based on prediction, is to use previously coded symbols, to form a (usually linear) prediction of the current symbol, which can then be quantized to form a finite number of contexts. The CALIC [3] algorithm is an example of a lossless image compression technique that is based on this approach.

In this paper, we present an algorithm based on the "finite-state" approach that reduces the number of contexts by merging those whose pmf is determined to be similar according to a distortion measure. In [4], a related technique called MCEQC was used to design a "context-quantizer" for binary random variables. MCECQ has also been applied to MPEG-4 α-plane sequences. Here we develop the optimal context quantizer problem for non-binary case and use a splitting algorithm similar to the GLA algorithm for the design of vector quantization (VQ) codebooks [6]. We also give a proof for the convergence of the algorithm.

2. PROPOSED ALGORITHM

Our approach is based upon the notion of a histogram quantizer that takes any input histogram (or pmf) and matches it to a finite set of histograms from a "codebook". More precisely, let $T=(T_1,T_2,...,T_K)$ be one of M conditional histograms for a source sequence having K different symbols, with T_k being the conditional probability of symbol k. This set of histograms is termed the *context space* and would typically be defined using the values of already coded symbols that are spatially close to the symbol being coded. An N-level histogram quantizer is a mapping that assigns to each input histogram, T, a reconstruction histogram, $T'=q(T)$ that is drawn from a finite-size codebook of histograms, $A_N=\{R^i, i=1,...,N\}$, where the R^i specify K symbol histograms. The quantizer, q, is completely described by two elements: the reconstruction alphabet A and the partition, of the input histogram space. This partition is defined by the set $S=\{S_i, i=1,...,N\}$, with $S_i=\{T : q(T)=R^i\}$.

In designing the quantizer, we require a distance measure, d, and choose the relative entropy between the histograms T and R_i. This quantity is given by

$$d\left(T,R^i\right)=H\left(T\,\|\,R^i\right)=\sum\nolimits_{k=1}^{K} T_k \log\left(T_k / R_k^i\right) \qquad (1)$$

This measure was also suggested in [4] and is sensible for three reasons: it is strongly related to the entropy; it is a measure of dissimilarity between pmf's; and is non-negative, being zero if and only if $T=R^i$. A minor drawback is that the measure it is not symmetric.

Our adaptation of the GLA is formally stated below:

1) Start with an initial reconstruction histogram codebook, $A_N^{(0)}$; zero $D^{(0)}$ and m; select ε.
2) Determine the N quantization regions defined by

   $$S_i^{(m)}=\{T : d(T,R^i)<d(T,R^j), \forall j\neq i\}, \quad i=1,...,N$$

   and compute the average distortion, $D^{(m)}$, between the input and reconstruction histograms as

   $$D^{(m)}=\sum\nolimits_{i=1}^{N} D_i^{(m)}; \quad D_i^{(m)}=\sum_{T\in S_i}\frac{\Pr(T)}{\Pr(S_i)}d\left(T,R^i\right) \quad (2)$$

3) Stop if $\left|D^{(m-1)}-D^{(m)}\right|/D^{(m)}<\varepsilon$
4) $m=m+1$. Determine the codebooks at iteration m, $A_N^{(m)}$, by computing the average histograms for each $S_i^{(m-1)}$; this is done element by element according to

   $$R_k^i=\sum_{T\in S_i^{(m-1)}}\frac{\Pr(T)}{\Pr\left(S_i^{(m-1)}\right)}T_k$$

   Go to step #1.

Convergence of the Histogram-Quantizer Design

In order to guarantee the convergence of the algorithm, we require that $D^{(m-1)}-D^{(m)}$ be non-negative. It is clear that step (1) above is a nearest neighbor calculation and that it can only lower the distortion; however, we need to prove that step (3) also reduces the distortion. Since the total distortion is made up of a sum of the D_i terms, we can treat these individually. Expanding (2) using (1) gives

$$D_i^{(m)}=\sum_{T\in S_i}\frac{\Pr(T)}{\Pr(S_i)}\sum_k T_k\log T_k-\sum_{T\in S_i}\frac{\Pr(T)}{\Pr(S_i)}\sum_k T_k\log R_k^i$$

Changing R_k^i has no effect on the 1st term and we thus minimize $D_i^{(m)}$ by maximizing the 2nd term. Defining

$$W=\{W_k, k=1,...,N\}; \quad W_k=\sum_{T\in S_i}\frac{\Pr(T)}{\Pr(S_i)}T_k$$

allows us to write the second term as

$$\lambda=\sum_{T\in S_i}\sum_k\frac{\Pr(T)}{\Pr(S_i)}T_k\log R_k^i=\sum_k W_k\log R_k^i$$

Now, inspection shows that both $\sum\nolimits_k W_k=1$ and $\sum\nolimits_k R_k^i=1$, which means that both W and R^i are valid pmf's. Since the relative entropy between two pmf's in non-negative, we have

$$H\left(W\,\|\,R^i\right)\geq 0 \quad\Rightarrow\quad \sum_k W_k\log R_k^i\leq\sum_k W_k\log W_k$$

with equality when $W_k=R_k^i$. We thus see that λ will be maximized if and only if this equality is true. Since this is exactly what step (3) forces, the step can never result in an increased distortion.

Context Determination

Our goal is to now use the histogram quantizer to design locally optimal contexts of a desired size, $\mathfrak{N}$. To achieve this goal, we look to the GLA and adopt the "splitting" version of the algorithm [6]. As presented below, the number of contexts is constrained to be a power of 2; however, this restriction is easily lifted with trivial modifications.

1) Initialization: Let R^1 be the centroid histogram of the M histograms that form the context space.
Set $N = 1$ and define $A_1 = \{R^1\}$.
2) $N = 2N$. To obtain double the number of contexts each set S_i split by forming two new "centroids": R^i itself and the histogram in S_i that is closest to R^i .
3) Run the histogram-quantizer algorithm to produce a system with N contexts
4) If $N \neq \mathfrak{N}$, go to step #1.

3. EXPERIMENTAL RESULTS

We tested the proposed context-selection algorithm using two types of sources with memory. The first is a 1^{st}-order Gauss-Markov source modified to have zero correlation by randomly flipping the sign of each sample with a probability 0.5 after the sequence has been generated. We call this a GM-F source and select it as a test case since we know the correct answer and it demonstrates the power of our approach over context design methods that are based upon linear prediction (which will fail here). Memory without correlation is common in wavelet-transformed images.

In the case of the GM-F source, we let ρ range from 0.1 to 0.9 and generated a 10^7 samples sequence for each ρ . We then applied a 32-level uniform quantizer whose loading factor f_l is set to 4, a value chosen to balance the overload and granular distortion of the quantizer. The context is defined as the two previous samples. For $\rho = 0.9$, the context space contained 774 nonzero histograms out of a possible 1,024. We then started with $\mathfrak{N} = 1$ and increased it in powers of two until no further drop in entropy was obtained. Since the source is 1^{st}-order Markov, there are 32 possible values for the previous symbol; however, the sign flipping operation removes any sign distinction and we thus expect that 16 distinct contexts should be sufficient to describe the source.

The experimental results for the GM-F source with $\rho = 0.9$ are shown in Table 1 and it is indeed seen that there is little point in using more than 16 contexts (8 really). We can get more information regarding what is happening by looking at the conditional histograms themselves and these are shown in Figure 1 for the $\mathfrak{N} = 16$ case. As expected, the histograms are all bi-modal. Indeed, the curves in the

Table 1: Context Merging for the GM-F Source

$\mathfrak{N}$	distortion – as in (2)	entropy [bits/sym]
1	0.5690	4.0617
2	0.2164	3.7091
4	0.0700	3.5628
8	0.0170	3.5098
16	0.0122	3.5039
774	0	3.4927

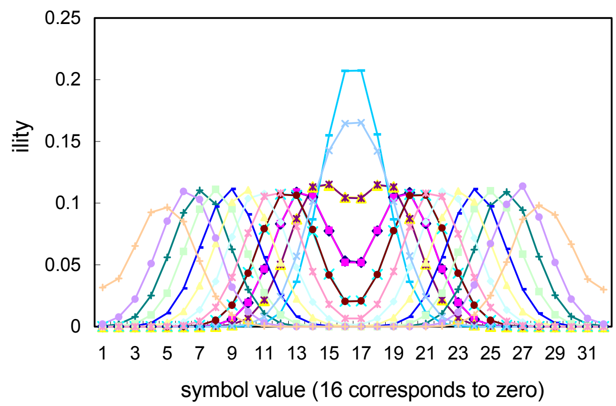


Figure 1: Converged Conditional Histograms for GM-F

figure are essentially identical to the conditional histograms based only on the magnitude of the previous sample. Our algorithm converged in a similar manner for smaller values of ρ, although the benefit to using higher-order contexts decreases with ρ .

The second source is an image processed by a wavelet transform to a depth of three. The image used here is the 512×512 image "baboon" and the filter set is the standard 9-7 configuration [7]. In this case, we have no idea what the optimal context-size will be or what sort of conditional histograms to expect. As with the previous source, we quantized the data with a uniform quantizer and varied $\mathfrak{N}$ in an identical pattern. Quantizers were designed for each subband by determining the difference between the maximum and minimum values of the coefficients and dividing this number of 24, the desired number of levels. This last number was set fairly arbitrarily since our focus is on the entropy coding. The context-space was defined using the four causal nearest neighbors, resulting in a raw count of 313,776, most of which have empty histograms.

Figure 2 plots the entropy as a function of $\mathfrak{N}$ for two subbands: S0, the 64×64 low-pass subband and S8, a 256×256 highpass (LH) image. We can see that the estimated conditional entropy decreases with increasing context-size; however, we also see that the difference between the estimated conditional entropy and the true rate obtained from the arithmetic coder with the proposed context-selection method is increasing due to data dilution. Considering S0 specifically, we see that data dilution begins to have a serious effect when $\mathfrak{N} = 128$. At this point, the overall rate begins rising quickly from its lowest value of 0.99 bits/symbol. We found that S0 had 3020 contexts with non-zero histograms in the context space. Using this number of contexts with a real arithmetic coder resulted in a rate of 4.52 bits/symbol, compared with an "ideal" conditional entropy of 0.46 bits/symbol.

4. FUTURE WORK

An interesting extension to the above study would be to consider the joint design of both quantizers and entropy coders. One possible approach for doing this would be to use rate-distortion optimization techniques to design a quantizer for a specific entropy coder. The entropy coder could then be improved through a re-optimization via histogram-quantization, followed by a re-design of the quantizer. An even more interesting study would be to extend this idea by designing different quantizers for each of the possible contexts – essentially producing a finite-state quantizer and entropy coder with the same procedure.

5. CONCLUSIONS

This paper has presented a method for selecting the contexts for adaptive arithmetic coders. Our method employs a histogram-quantizer to reduce the context space to the desired number of contexts. Similar to VQ, a splitting algorithm was used for initialization. Our experiments have showed that our method has the potential to automatically discern hidden structure in data and that our approach is superior to those based upon linear-prediction (at least for certain types of data). For the wavelet-transformed image, we were able to use the proposed approach to find a (locally) optimal context-selection that takes into account the problem of data dilution. We believe that this method has great potential for improving the performance of image compression algorithms.

REFERENCES

[1] D. Taubman and M .W. Marcellin, *JPEG2000 Image Compression Fundamentals, Standards and Practice*, Kluwer Academic Publishers, 2002.

[2] T. C. Bell, J. G. Cleary and I. H. Witten, *Text Compression*, Prentice Hall, NJ, 1990.

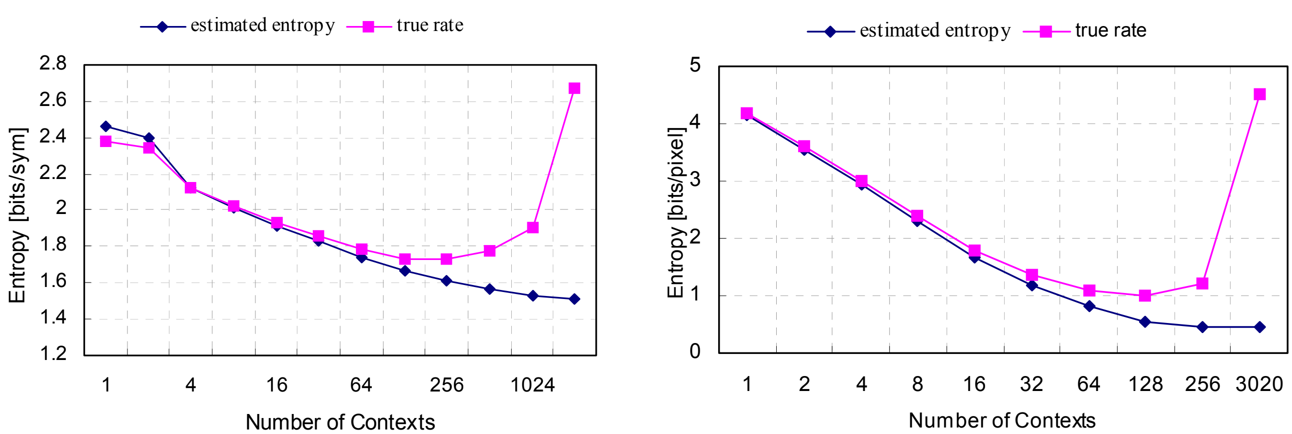
[3] X. Wu and N. Memon, "Context-based, Adaptive, Lossless Image Codec", *IEEE Trans. on Communication,* vol.45, no,4, pp.437-444,April 1997.

[4] X. Wu, Philip A. Chou and X. Xue, "Minimum Conditional Entropy Context Quantization", *IEEE International Symposium on Information Theory,* June. 2000.

[5] Marcelo J. Weinberger, Jorma J. Rissanen, and Ronald B. Arps, "Applications of Universal Context Modeling to Lossless Compression of Grey-Scale Images", *IEEE Trans. Image Processing*, vol.5 no.4, April 1996.

[6] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1991.

[7] M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies, "Image Coding using Wavelet Transform," *IEEE Trans. Image Processing*, vol.1, pp.205--220, 1992.



(a) Subband 8 (highpass) (b) Subband 0 (lowpass)

Figure 2: Context Dilution in Image Subbands