Chapter 4

Adaptive Modulus Quantization

As demonstrated in the previous chapter and in other works [45] [3], overcomplete signal decomposition using matching pursuits can be used to efficiently code motion residual images in a hybrid video coder. All of these sytems use a simple quantizer with fixed stepsize, and a simple rate control mechanism based solely on setting the number of atoms. While this type of system demonstrates good performance on the MPEG test set, the quantizer cannot be considered optimal, and is particularly unsuitable for use at higher bit rates. This chapter investigates improvements to the quantizer design with an emphasis on optimal frame-adaptive schemes. Practical 2-pass and 1-pass quantizer design algorithms are developed, and these adaptive schemes are shown to outperform the fixed quantizer used in earlier works, especially at high bit rates where the gain is as high as 1.7 dB.

Quantizer design for overcomplete transforms such as matching pursuit is more intricate than the equivalent problem for complete orthogonal transforms. To see why, consider the case for an orthogonal transform such as DCT or orthogonal wavelets. Quantization of such transforms has the important property that the error incurred in the quantization of one basis coeficient is orthogonal to all of the other basis functions. Introducing quantization error for one coeficient thus has no effect on the computation of the other coeficients. For this reason, the quantizer for orthogonal transforms need not be known in advance, and may in fact be designed and applied based on the already-computed transform coeficients. The situation is much different for non-orthogonal matching pursuit decomposition. Since the quantization error in a single basis element is generally not orthogonal to the other basis functions, this quantization error must be considered when computing the remaining coeficients. Quantization is thus performed in-the-loop. This means that each matching pursuit coefficient or modulus is quantized as it is found, and the resulting quantized basis function is subtracted from the remaining signal energy before the next atom can be found. In this way, the quantization error from one matching pursuit stage may be progressively coded in future stages. The set of atoms chosen thus depends on the design of the quantizer. Likewise the optimal quantizer design depends on the modulus statistics from the chosen atoms. This results in a chicken-and-egg problem which is absent in coeficient quantizer design for orthogonal transforms.

For simplicity, previously published matching pursuit compression systems have used a quantizer in which both the form and the stepsize are fixed in advance heuristically. As mentioned before, rate control is then based solely on the number of atoms. However, the rate-distortion tradeoff for the system is in reality sensitive to both the number of atoms and the quantizer design. Fixing the quantizer thus removes a degree of optimization freedom.

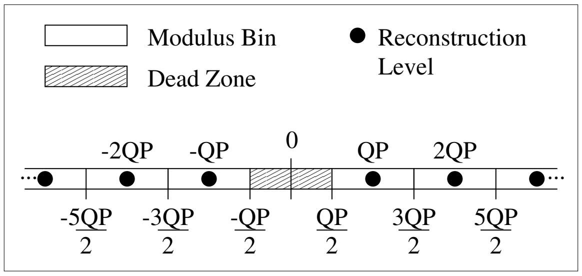
In this chapter, we improve the efficiency of the matching pursuit video coding algorithm by applying optimal adaptive modulus quantization at the frame level. In section 4.1, we review the fixed quantizer used in earlier works, and show that the usefulness of this quantizer is limited to low bit rates. A modified fixed quantizer design is shown to allow comparison at higher rates without changing the low bit rate performance documented in previous works. Section 4.2 then considers the problem of optimal scalar quantization for the system. An adaptive dead zone subtraction method is applied to the modulus values before quantization, and a uniform threshhold quantizer (UTQ) with an adaptive step size is shown to be optimal for quantization of the resulting source. The problem is then reduced to finding the best combination of dead zone size DZ, quantization step size QP, and number of coded atoms N in a rate-distortion sense. A 2-pass algorithm is defined to solve this problem and is shown to have near optimal performance. Since the 2-pass algorithm is computationally expensive, a compromise 1-pass algorithm is developed in Section 4.3. This algorithm is shown to perform nearly as well as the 2-pass scheme, but with about the same complexity as the simple fixed quantizer. A comparison of results is presented in Section 4.4, along with final conclusions.

4.1 Fixed Quantizer Design

Previously published matching pursuit video coders used a midtread linear symmetric quantizer with midpoint reconstruction, as shown in Figure 4.1. For the results shown in [45, 3] the quantization stepsize was fixed at $QP = 30$, a value heuristically chosen to perform well on the MPEG-4 test sequence set.

Since the focus of this chapter is adaptive quantizer design, we now consider this

Figure 4.1: Simple midtread linear symmetric quantizer with fixed stepsize $QP = 30$, used in previously published M.P. systems.



fixed quantizer for use as a comparison base for our later adaptive designs. The advantages and disadvantages of this design may be summarized as follows:

Advantages:

A1. Fixed quantizer requires no knowledge or pre-analysis of the source.

A2. Provides for simple and effective rate control based on number of coded atoms N.

A3. Performs well for the standard MPEG-4 test sequences and rates.

Disadvantages:

D1. Fixed quantizer removes a degree of optimization freedom. That is, for a given frame, a different quantizer with a different number of coded atoms may produce a lower distortion at the same target bit rate.

D2. Large fixed dead zone imposes a hard limit on attainable bit rate and image quality.

Of the two listed disadvantages, note that *D1* is intrinsic to any fixed quantizer design. That is, no fixed quantizer will be suitable for all frames at all possible bit rates. It is thus appropriate that this disadvantage be present in our fixed quantizer comparison model. However, *D2* is really a flaw in the specific quantizer design shown in Figure 4.1.

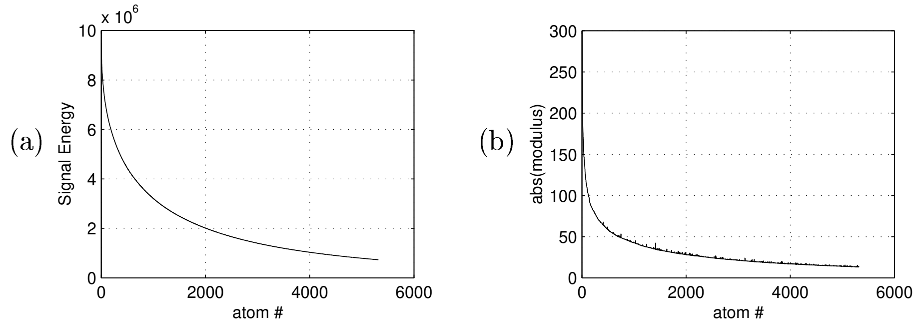
This flaw will be explained and characterized in the remainder of this section. Sections 4.1.1 and 4.1.2 will then develop a modified fixed quantizer design which eliminates this flaw while preserving the listed advantages.

To see why *D2* is true, consider the matching pursuit decomposition of a single frame of the Stefan sequence as shown in Figure 4.2. The sequence was coded at 30 fps at a constant bit rate of 3 Mbit/s. This particular frame reached the desired rate using 5318 atom stages. Figure 4.2a shows the remaining signal energy $\|\vec{f_i}\|^2$ after each atom iteration i. Figure 4.2b shows the absolute modulus values $|p_i|$ before quantization. Note that the signal energy is strictly decreasing, reflecting that each atom results in some positive energy reduction. The modulus values generally decrease with iteration, although not monotonically. If a given frame is coded to an arbitrarily high bit rate using the fixed quantizer of Figure 4.1, at some stage i the modulus value p_i will satisfy $|p_i| \leq (QP/2)$, and thus will fall into the dead zone of the quantizer. Since the resulting atom quantizes to zero, the matching pursuit update equation (2.3) gives us $\vec{f_i} = \vec{f}_{i-1}$. Because the residual energy was unaffected by stage i, the "Find Atoms" block shown in Figure 2.1 will find the exact same atom for stage $i+1$, and the algorithm will become deadlocked, unable to code additional nonzero coeficients. This effectively imposes a hard limit on both rate and quality, and makes the specific quantizer design of Figure 4.1 unusable at high bit rates. Specifically, it restricts the matching pursuit decomposition to operate only at rates low enough such that all moduli p_i satisfy:

$$|p_i| > QP/2 \tag{4.1}$$

Our experience has shown this condition to be violated for SIF sequences such as Mobile

Figure 4.2: Matching pursuit decomposition of first residual frame of the Stefan sequence. (a) Remaining signal energy vs. iteration. (b) Absolute modulus vs. iteration.



and Stefan when coded in the 1 to 2 Mbit/s range. It may also be violated at much lower rates for particular frames with low residual energy, which occur during low motion periods of the sequence. The absolute modulus values for such frames tend to be much lower than frames with higher levels of motion.

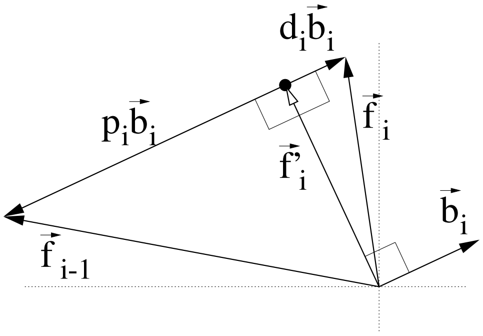
Restriction (4.1) is specific to the fixed quantizer design of Figure 4.1. In the next section, we develop the comparable restriction for a more general fixed quantizer design. A modified design which may be used at higher bit rates is then developed in Section 4.1.2. This modified design will then be used as the comparison base for the adaptive quantizers developed in Section 4.2.

4.1.1 Restriction for a general fixed quantizer

Consider a general quantizer $Q(\cdot)$, and allow the quantization error to be represented by $d_i = p_i - Q(p_i)$. Suppose $\vec{f_i}$ is the signal remaining after the atom for iteration i has been quantized and subtracted from the source energy. Then the Energy Reduction $\overline{ER}$ for stage i can be defined as:

$$\overline{ER}_i \triangleq \left\|\vec{f}_{i-1}\right\|^2 - \left\|\vec{f_i}\right\|^2 \tag{4.2}$$

Figure 4.3: Two dimensional vector representation of a single quantized matching pursuit stage, used in the proof of Equation (4.3).



We claim that this single-stage Energy Reduction can be represented as:

$$\overline{ER}_i = {p_i}^2 - {d_i}^2 \tag{4.3}$$

Proof: Since p_i is the projection of $\vec{f}_{i-1}$ onto basis element $\vec{b}_i$, the difference $\vec{f}_i' = \vec{f}_{i-1} - p_i\vec{b}_i$ will be orthogonal to $\vec{b}_i$, as shown in Figure 4.3. $\vec{f}_i'$ would be the remaining signal energy for the current stage in the absence of quantization, and is also orthogonal to the quantization error $d_i\vec{b}_i$. We can thus apply the pythagorean theorem twice to get:

$$\left\|\vec{f}_{i-1}\right\|^2 = \left\|p_i\vec{b}_i\right\|^2 + \left\|\vec{f}_i'\right\|^2 \tag{4.4}$$

$$\left\|\vec{f}_i\right\|^2 = \left\|\vec{f}_i'\right\|^2 + \left\|d_i\vec{b}_i\right\|^2 \tag{4.5}$$

Since $\vec{f}_i$ is the remaining signal energy with quantization, we can combine the two above equations to see that the single stage energy reduction is:

$$\left\|\vec{f}_{i-1}\right\|^2 - \left\|\vec{f}_i\right\|^2 = \left\|p_i\vec{b}_i\right\|^2 - \left\|d_i\vec{b}_i\right\|^2 \tag{4.6}$$

$$= \left(|p_i|^2 - |d_i|^2\right)\left\|\vec{b}_i\right\|^2 \tag{4.7}$$

which is equal to $p_i^2 - d_i^2$ since $\vec{b}_i$ is unit-norm.□

To avoid the deadlocking problem seen in the earlier fixed quantizer design, we require positive energy reduction, specifically that $\overline{ER}_i > 0$, for each stage i. For an even-symmetric quantizer design, consider without loss of generality the $p_i > 0$ case. By combining Equation (4.3) with the definition for d_i, we arrive at the following restriction in order to avoid deadlocking:

$$0 < Q(p_i) < 2p_i \tag{4.8}$$

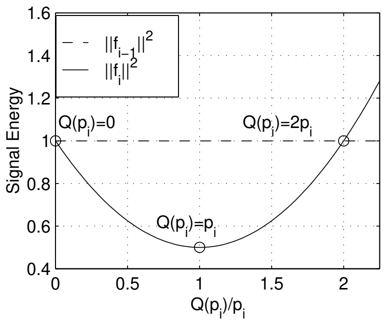
A graphical interpretation of this restriction is shown in Figure 4.4a. The figure shows the remaining signal energy before and after stage i as a function of $Q(p_i)/p_i$. The dotted line represents $||f_{i-1}||^2$, the energy before the atom is coded, and the solid line shows $||f_i||^2$, the energy after the atom is coded. The plot is a direct consequence of Equation (4.3). The maximum energy reduction of ${p_i}^2$ occurs in the absence of quantization error, that is when $Q(p_i)/p_i = 1$. When this ratio falls outside the range $(0, 2)$, the coded atom has a destructive effect in that the remaining signal energy is increased. By designing a modified fixed quantizer to operate within this restricted range (4.8) for a wide range of interesting bit rates, we can avoid the deadlocking problem seen earlier.

4.1.2 Modified fixed quantizer design

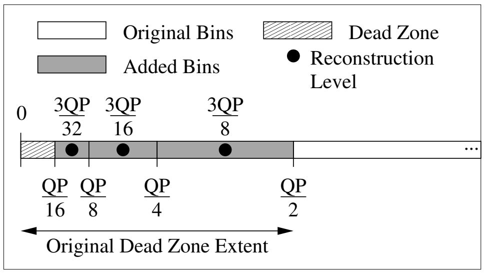
Here we show a heuristically modified fixed quantizer design which allows high bit rate operation without changing the earlier documented low bit rate performance. Recall that the main disadvantage *D2* of the original fixed quantizer design was a large fixed dead zone which resulted in the deadlocking problem at high rates. We modify this design by splitting the dead zone into additional bins to be used at high bit rates when the modulus

Figure 4.4: (a) Remaining signal energy before and after the current stage as a function of $Q(p_i)/p_i$. Positive energy reduction requires $Q(p_i) \in [0, 2p_i]$. (b) Modification of the fixed quantizer design. Original dead zone has been divided into several additional quantization bins.

(a)



(b)



values become small. The resulting nonlinear quantizer is shown in Figure 4.4b. Midpoint reconstruction is used, and even symmetry allows us to show only the positive half. Each new bin is formed by splitting the remaining dead zone in half. Reducing the dead zone in this way allows more non-zero atoms to be coded, and thus enables the encoder to work at higher bit rates without deadlocking. Any number of such bins may be added in order to achieve arbitrarily high rates. We use three added bins as shown in the figure, reducing the dead zone to $[0, QP/16]$. Our experiments show that this is sufficient to allow operation without deadlocking for the SIF MPEG-4 test sequences at 3 Mbit/s, the highest rate tested in this paper. Note however that this strategy merely postpones the deadlocking problem. That is, at some rate beyond 3 Mbit/s, the condition $p_i > QP/16$ will be violated, and deadlocking will again occur.

It is important to note that the added bins are consistent with the general quantizer restriction (4.8), and so positive energy reduction, that is $\overline{ER}_i > 0$, is guaranteed for all moduli which fall into these bins. At low bit rates where moduli are larger than $QP/2$, the quantizer is identical to the earlier design, and so low bit rate performance is not affected. This is true for the standard MPEG-4 sequences and rates, for which the earlier restriction (4.1) holds.

A remaining question is how to entropy code the new high rate bins. The simplest approach is to add a new VLC code for each new bin and retrain the system as before. This method increases the size of the modulus VLC table, and will be denoted "NewVLC". We also present an alternate approach from our MPEG standards work [35]. Suppose the size of the modulus VLC table remains unchanged and instead a single *shift* codeword is sent at the frame level to specify how many of the new bins were used during coding. For three added bins, *shift* takes a value in $\{0, 1, 2, 3\}$, for an overhead of 2 bits per frame. The VLC table is then interpreted so that the first code in the table points to the lowest bin actually used in coding the current frame. The VLC table is effectively shifted down by *shift* bins. This idea is based on the empirical observation that the modulus distribution is effectively a shifted exponential distribution, and so the lowest bin used typically has a high probability of occurrence. We defer a detailed look at the modulus distribution until Section 4.2.1.

To clarify the difference between the two schemes, the VLC mappings are shown as Table 4.1. Note that $shift = 0$ at low bit rates; in this case both the quantizer and VLC structure match those of the original fixed quantizer design shown in Figure 4.1. The 2-bit per frame *shift* codeword is insignificant, and so low bit rate performance is equivalent to earlier published results [45, 43].

Since "NewVLC" and "ShiftVLC" are based on the same quantizer design, they differ only by the resulting cost per atom. To compare them on this basis, the two schemes were implemented and separately trained on a set of sequences outside the MPEG4 test

Table 4.1: Variable Length Code (VLC) mappings for NewVLC and ShiftVLC coding schemes.

NewVLC Scheme		
Index	Recon	Bin Type
0	$3 \cdot QP/32$	Added
1	$3 \cdot QP/16$	Added
2	$3 \cdot QP/8$	Added
3	QP	Original
4	$2 \cdot QP$	Original
5	:	:

ShiftVLC Scheme				
Index	$shift = 0$	$shift = 1$	$shift = 2$	$shift = 3$
0	QP	$3 \cdot QP/8$	$3 \cdot QP/16$	$3 \cdot QP/32$
1	$2 \cdot QP$	QP	$3 \cdot QP/8$	$3 \cdot QP/16$
2	$3 \cdot QP$	$2 \cdot QP$	QP	$3 \cdot QP/8$
3	$4 \cdot QP$	$3 \cdot QP$	$2 \cdot QP$	QP
4	$5 \cdot QP$	$4 \cdot QP$	$3 \cdot QP$	$2 \cdot QP$
5	:	:	:	:

set. The QCIF sequences from the MPEG-4 set were then run for a comparison, with the resulting cost per modulus summarized in the top section of Table 4.2. The first three columns show that at these rates the two schemes produce a nearly identical cost per modulus. The remaining columns show that the ShiftVLC system uses a slightly lower percentage of its atom budget on modulus codewords, but the difference is very small. The two schemes are thus nearly identical at low bit rates where the additional “high-rate” bins aren’t used. The difference is much more apparent at higher rates, as shown in the bottom section of Table 4.2. Here the QCIF Mother sequence is coded with bit rates from 24 to 96 kbit/s. At the highest rates, ShiftVLC shows a performance advantage over the simpler NewVLC system, saving up to 1.6 bits per modulus code and using a smaller percentage of atom bits for modulus codewords. Because of this advantage, ShiftVLC will be used for comparison to the adaptive systems developed in the next sections.

4.2 Optimal Adaptive Quantization

This section considers the problem of designing an optimal adaptive modulus quantizer for the matching pursuit video encoder. Section 4.2.1 examines the modulus distribu-

Table 4.2: Average modulus VLC cost and modulus bits as a percentage of total atom bits.

	Bits/Modulus			% Bits to Modulus		
Sequence	NewVLC	ShiftVLC	Diff	NewVLC	ShiftVLC	Diff
Container10	3.30	3.29	-0.01	14.97	14.90	-0.07
Hall10	4.36	4.36	0.00	18.79	18.78	-0.01
Mother10	3.10	3.11	0.01	12.26	12.21	-0.05
Container24	2.38	2.38	0.00	12.61	12.65	0.04
Silent24	3.62	3.63	0.01	15.73	15.71	-0.02
Mother24	2.34	2.34	-0.01	11.21	11.13	-0.07
Coast48	3.44	3.44	0.00	16.26	16.23	-0.03
Foreman48	3.25	3.25	0.00	15.38	15.32	-0.06
Mother24	2.34	2.34	-0.01	11.21	11.13	-0.07
Mother48	2.12	2.14	0.02	11.34	11.42	0.08
Mother72	2.93	2.29	-0.64	15.65	12.70	-2.95
Mother96	4.18	2.60	-1.58	21.37	14.62	-6.75

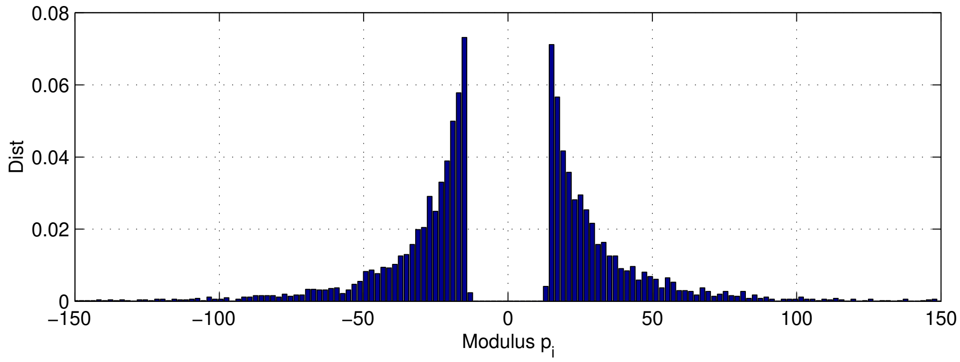
tion and shows that an adaptive dead zone subtraction technique can be used to reduce the energy content before quantization. The optimal form and stepsize of the quantizer for the resulting source will be developed in Sections 4.2.2 and 4.2.3, respectively, and a practical 2-pass algorithm will be shown to have near-optimal performance.

4.2.1 Adaptive Dead Zone Subtraction

A sample modulus distribution for the matching pursuit video coding system is shown in Figure 4.5. This is a normalized histogram of the moduli for a single frame of the Stefan sequence coded to 3 Mbit/s. In-loop quantization[1] is used in coding, but the values before quantization are used to produce the continuous distribution shown. The figure illustrates some properties that generally hold for the modulus distribution. Because the distribution is symmetric about zero, the quantized moduli may be coded as magnitude

[1]Figure 4.5 uses ShiftVLC as the in-loop quantizer. This choice is arbitrary, since the distribution changes very little if a different quantizer is used, or even no quantizer at all. This is further investigated in Figure 4.9.

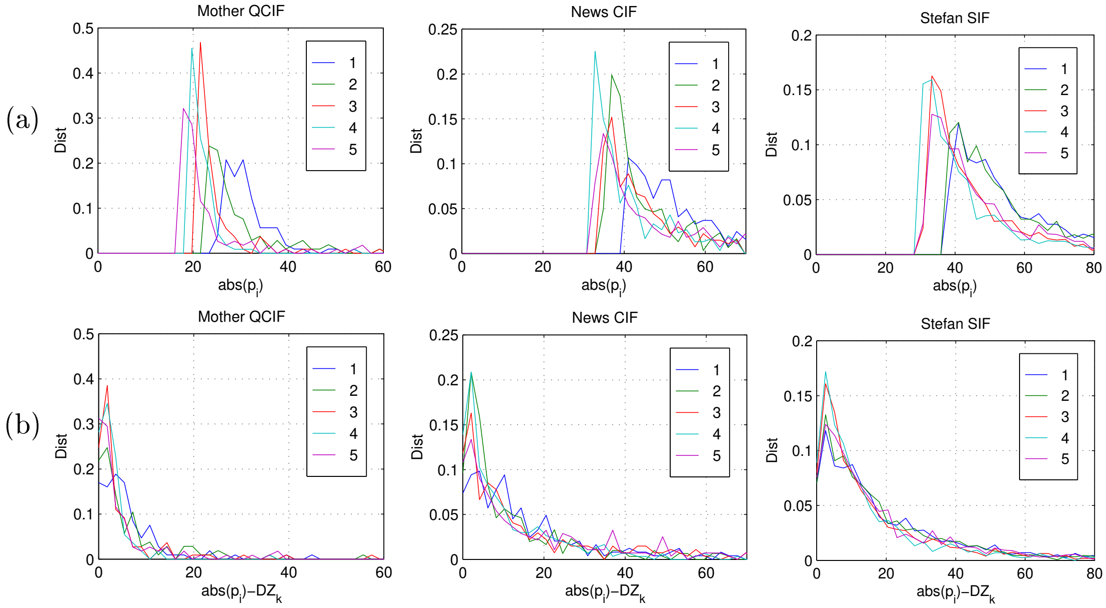
Figure 4.5: Modulus distribution for a single frame of Stefan (SIF) coded at about 3 Mbit/s.



plus a sign bit with no loss of efficiency. This reduces the number of quantizer bins and VLC codes by half, and has an added benefit for adaptive quantizer design, since the data per bin available for run-time analysis is effectively doubled. A second interesting property is the presence of a large dead zone. Note that the cutoff around the dead zone is very sharp, with high probabilities near the dead zone decision border.

We now show how the modulus distribution changes from frame to frame. Figure 4.6a shows absolute modulus distributions for coded frames of three different MPEG-4 test sequences. For each sequence, five frames are coded sequentially without quantization and the resulting per-frame modulus distributions are superimposed. Since the in-loop quantizer is omitted, exact bit rates cannot be computed. Instead, the number of atoms per frame is matched to the corresponding standard MPEG-4 test run. The resulting plots show a poor match between individual distribution curves of different frames within the same sequence. This suggests the need for a frame adaptive quantizer design. Close examination of the plots shows that the shape of the distribution is fairly consistent, but that this shape shifts horizontally from frame to frame. This is seen most clearly at high bit rates

Figure 4.6: Absolute modulus distributions for single matching pursuit residual frames coded sequentially without quantization. (a) Sequences coded without DZ-subtraction. (b) Same sequences with framewise DZ-subtraction. Note the number of atoms per frame match the standard MPEG-4 sequence runs of Mother at 24 kbit/s, News at 112 kbit/s, and Stefan at 1 Mbit/s.



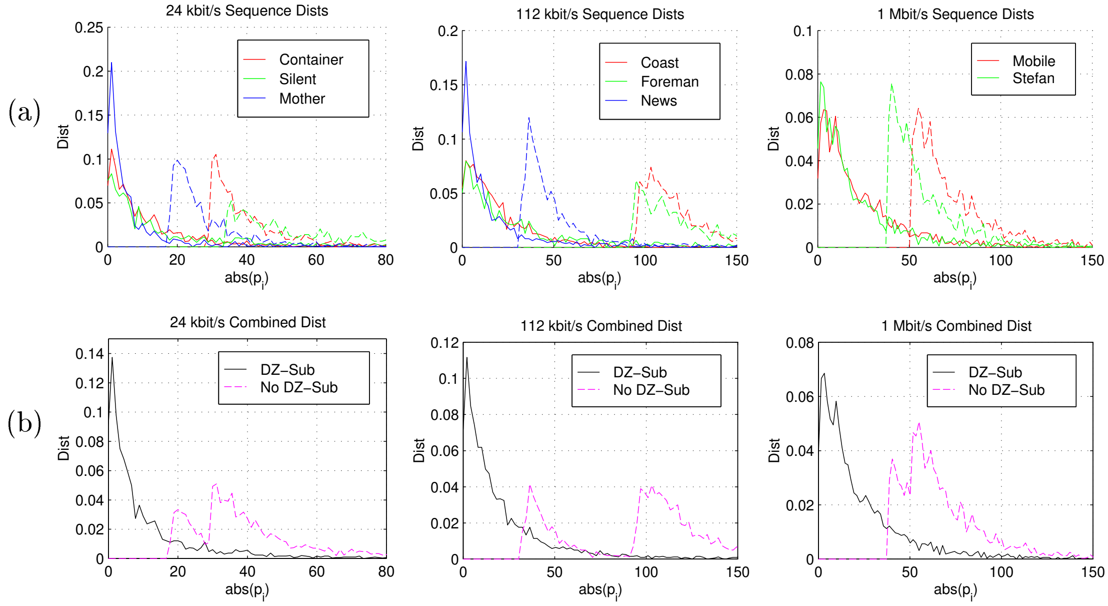
such as Stefan, where a large number of atoms per frame allows more reliable computation of the distribution curves.

Because the distribution shifts between frames, we propose a simple transformation to the original modulus source. Define DZ_k as the minimum absolute modulus value found in encoding frame k. A "DZ-Subtraction" operation is then performed on the modulus source, that is DZ_k is subtracted from all absolute modulus values within the frame. The distributions of the resulting DZ-Subtracted source are more consistent from frame to frame, as illustrated in Figure 4.6b.

The effect is even more drastic when modulus distributions are combined over multiple sequences or across multiple bit rates. Figure 4.7 shows the combination of multiple sequences at the same rate. Each plot is generated by encoding all of the MPEG-4 test sequences from Table 1.2 that correspond to the given bit rate categories (24 kbit/s, 112 kbit/s, and 1 Mbit/s). The sequences from each category were coded without quantization using the same conditions as in the previous figure. Modulus data was then taken from enough frames to produce about 3000 atoms per bit rate category. For example, the left-most plots combine the 24 kbit/s QCIF encodings of Container, Silent and Mother with 10 frames used per sequence. Each trace in Figure 4.7(a) shows the modulus distribution combined across the frames of a single sequence. Note that the dotted lines, which represent the distributions without DZ-subtraction, match poorly across sequences. A much better match is seen in the solid lines, which show the distributions from each sequence after DZ-subtraction. When all sequences from each category are combined, the resulting distributions are shown in Figure 4.7(b). In each of these plots, a sharp, exponential-like distribution is produced by the DZ-subtraction method as shown by the solid line. This shape is well-matched to the single-frame distributions seen in Figure 4.6b. These distributions are even consistent across bit rates. The distributions without DZ-subtraction are much less consistent, as shown by the dotted line. This is because the original modulus source varies greatly across frames, sequences, and bit rates.

Figure 4.8 shows the effect of combining distributions across bit rate categories. To generate the figure, all of the standard MPEG-4 test sequences were coded in the same manner as used in the previous two figures, and enough frames were taken from each of the five bit rate categories (10, 24, 48, 112, and 1024 Kbit/s) to produce about 3000 atoms from each category. This was done to prevent the higher bit rate sequences, having more coded atoms per frame, from dominating over the lower bit rate data. Figure 4.8a shows

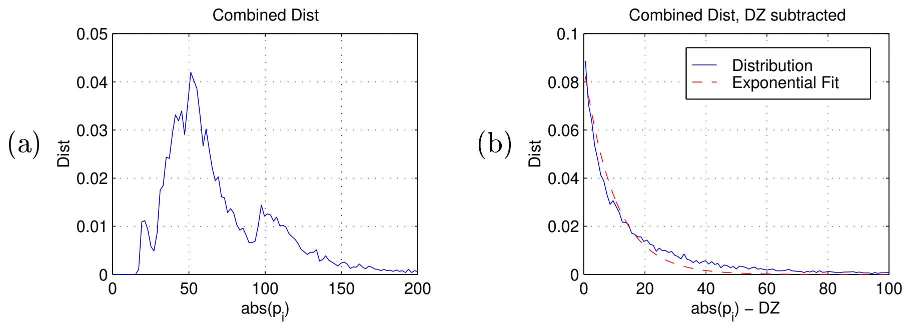
Figure 4.7: Combined modulus distributions for each of three MPEG-4 test rate categories. (a) Separate distribution curves for each sequence in the given bit rate category. Solid lines show runs with DZ-subtraction. Dotted lines show runs without DZ-subtraction. (b) Combined distribution curve for all sequences in the given rate category.



a simple combination of the absolute moduli, while Figure 4.8b shows the DZ-subtracted result. Without the DZ-subtraction, the combined distribution is somewhat random and matches neither the individual frame distributions from Figure 4.6a nor the combined distributions for the various rate categories shown in Figure 4.7. With DZ-subtraction, the final combined distribution is again a sharp exponential, which well matches the corresponding DZ-subtracted distributions in Figures 4.6b and 4.7. To verify the exponential form of the distribution, a least squares best fit exponential curve is also shown in Figure 4.8b.

We compare the information content of the two distributions of Figure 4.8 by

Figure 4.8: Modulus distributions combined across all MPEG-4 sequences and rates. Enough coded frames were taken to provide approximately 3000 atoms from each of the five bit rate categories (10, 24, 48, 112, and 1024 Kbit/s). (a) Without dead-zone subtraction. (b) With frame-wise dead zone subtraction. Dotted line shows best exponential fit.



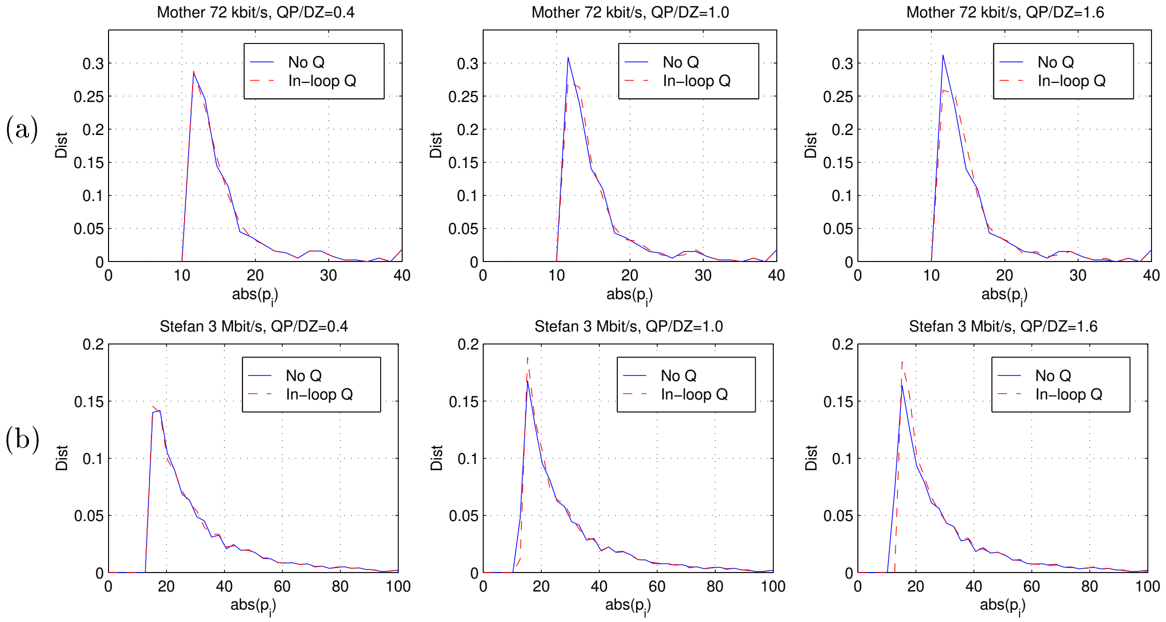
computing the Shannon differential entropy, $h(X)$, for each distribution.[2] Simple numerical integration [4] gives a value $h(X) = 6.914$ for the combined distribution in Figure 4.8a, and a value $h(X) = 5.682$ for the DZ-subtracted distribution in Figure 4.8b. Suppose the modulus source is to be quantized using the optimal entropy constrained scalar quantizer as defined in [23]. Dead-zone subtraction will then improve the rate-distortion performance $r(d)$ of the quantizer by the difference of 1.279 bits per symbol, a savings of about 18 percent. The dead zone parameter DZ_k must be transmitted once per frame, but this is a negligible cost even at low bit rates.

The above experiments were done in the absence of quantization. We now show that the addition of in-loop quantization does not significantly change the observed distribution. Figure 4.9 shows single-frame modulus distributions for two sample frames coded with and without in-loop quantization. The top row of plots, (a), shows a frame from

[2]Shannon differential entropy reflects the information content of a continuous source, similar to entropy for discrete valued sources [31]. It is defined in terms of the probability density function $p_x(x)$ as:

$$h(X) = \int_{-\infty}^{\infty} p_x(x) \log_2 p_x(x) dx$$

Figure 4.9: The effect of in-loop quantization on the absolute modulus distribution for single coded frames. (a) Mother QCIF at 72 kbit/s. (b) Stefan SIF at 3 Mbit/s. Three different quantizers with progressively larger stepsize to deadzone ratios are used for each frame.



Mother QCIF at 24 kbit/s, while the bottom row, (b), shows a frame of Stefan SIF at 3 Mbit/s. For each plot, the solid line depicts matching pursuit without quantization, and the dotted line shows in-loop quantization. Clearly the amount of error introduced by the quantizer will determine the amount of change seen in the resulting distribution. For this reason, three quantizers are used with varying stepsizes. The quantizer is uniform with an expanded dead zone, which is equivalent to DZ-subtraction followed by UTQ. The dead zone size DZ is set to be the minimum absolute modulus value seen in the non-quantized encoding. The step size QP is then set so that the ratio QP/DZ takes on values of 0.4, 1.0 and 1.6. Note that this stepsize to dead-zone ratio provides a convenient description of the quantizer, and so will be used in the remainder of this chapter.

Scanning the plots of Figure 4.9 from left to right, it is clear that for ratio values of 1.0 and below, in-loop quantization produces no noticeable change to the absolute modulus distribution. For more coarse quantization with $QP/DZ = 1.6$, some change is visible, but the distributions with and without quantization are still nearly identical. We can assume that within this range of ratio values, in-loop quantization has little effect on the modulus distribution or the effective dead zone size. Under this assumption, the skeleton of a 2-pass adaptive quantization algorithm can be seen. A first pass pre-codes the residual signal without quantization to estimate the number of atoms and the size of the dead zone DZ_k. The second pass uses dead-zone subtraction with in-loop quantization of the resulting moduli. The details of this algorithm, along with a solution for the optimal quantizer form and stepsize, will be developed in the next two sections.

4.2.2 Optimal Quantizer for the M.P. System

In this section, we develop the optimal form of the quantizer to use in conjunction with the DZ-subtraction process. To clarify the problem, we first introduce some notation. Consider the abstracted view of the encoder shown in Figure 4.10. The large box represents the entire system, which takes the current original and previous reconstructed images as input, performs both motion compensation and matching pursuit based residual coding, and outputs the coded bitstream and the current reconstruction to be used in encoding the next stage. Overall distortion D is defined as the mean square error in the coded image. Overall rate R is the number of bits used to code the current frame, which is the sum of rates to header R_{hdr}, motion vectors R_{mot}, and atoms R_{atm}. The atom budget may be further divided into bits used for position coding R_P, basis indices R_I, and quantized modulus R_Q.

Figure 4.10: Abstraction of the encoder, showing relation between overall rate and distortion (R,D) and that of the internal modulus quantizer (r,d).

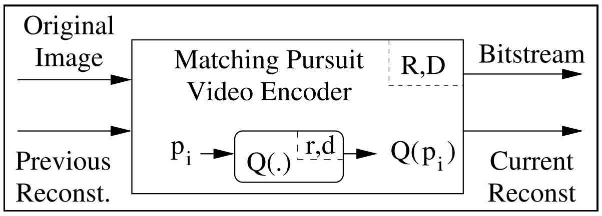
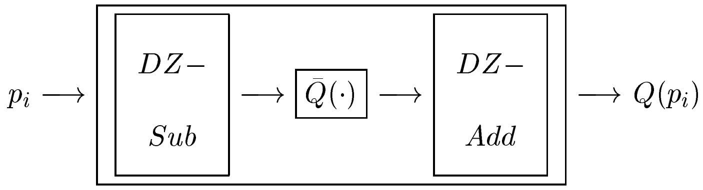


Figure 4.11: Definition of quantizer $Q(\cdot)$, which combines DZ-subtraction and quantization.



In summary,

$$R = R_{hdr} + R_{mot} + R_{atm} \tag{4.9}$$

$$= R_{hdr} + R_{mot} + (R_P + R_I + R_Q) \tag{4.10}$$

The modulus quantizer for the system is $Q(\cdot)$, shown as the smaller block in Figure 4.10. The quantization process also has an associated rate and distortion. Quantizer rate r is defined as the average number of bits spent to encode each modulus value, and quantizer distortion is:

$$d = mean((p_i - Q(p_i))^2) = mean(d_i^2)$$

We define the quantization process to include the DZ-subtraction method developed in the previous section. Specifically, we consider a 2-pass encoding algorithm in which a first pass without quantization provides an estimate of the dead-zone parameter DZ_k. The second pass then uses in loop quantizer $Q(\cdot)$. This quantizer performs DZ-subtraction, applies quantizer $\bar{Q}(\cdot)$ to the resulting moduli, and then adds DZ_k to the result to obtain the quantized value. This is illustrated in block form as Figure 4.11. A magnitude VLC code and separate sign-bit are transmitted for each modulus. The floating point representation of DZ_k is sent in the frame header to allow quantizer tracking at the decoder.

The specific problem is to design quantizer $\bar{Q}(\cdot)$ such that the overall distortion D for the matching pursuit encoder is minimized at a given overall rate R. Essentially the design of $\bar{Q}(\cdot)$ allows a tradeoff between modulus precision and number of coded atoms. A quantizer may code modulus values more precisely at the cost of increasing R_Q, however this decreases the total number of atoms which may be coded at an overall target rate R. We thus seek an optimal balance between quantizer precision and number of coded atoms in order to minimize distortion for a given rate.

In the previous section, we showed that the DZ-subtracted modulus source to which $\bar{Q}(\cdot)$ is to be applied has an exponential distribution. The optimal entropy constrained scalar quantizer (ECSQ) for this distribution is a UTQ [31, 63]. That is to say, using a UTQ for $\bar{Q}(\cdot)$ will provide the best rate-distortion performance for the DZ-subtracted moduli, and will in turn minimize d for any given quantizer rate r. Through the use of some simplifying assumptions, we will show in the remainder of this section that the UTQ is optimal in terms of overall (R, D) for the matching pursuit system. We begin by presenting a pair of assumptions to relate (r, d) for the quantizer to to (R, D) for the overall system. These are presented without proof, but empirical evidence is provided to show that each holds for the matching pursuit video system.

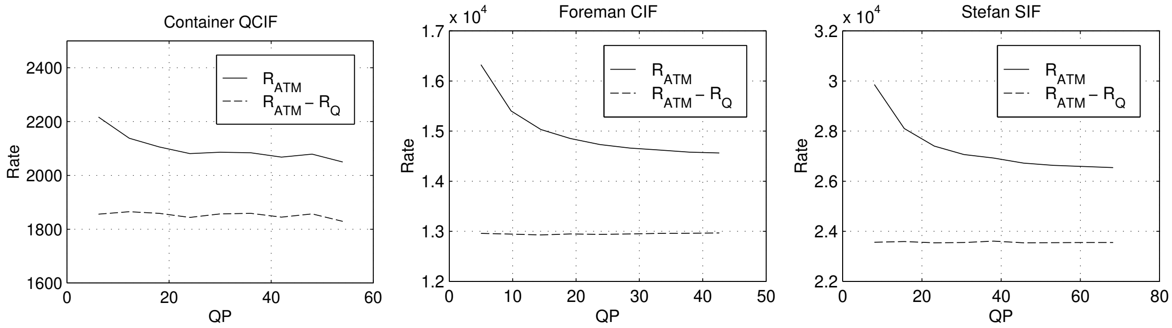
Assumption 4.1 *For a given number of atoms N, R_Q may vary as the quantizer changes. The other bit rate terms and their sum $R - R_Q = (R_{hdr} + R_{mot} + R_P + R_I)$ will remain* *roughly constant with changes in the quantizer.*

The first part of the assumption is obvious, since $R_Q = Nr$, and r is certainly affected by changes in the design of $\bar{Q}(\cdot)$. As for the other bit rate terms, R_{hdr} and R_{mot} may be calculated before atom coding begins and thus are unaffected by the quantizer design. R_P depends on the spatial density of atoms, and thus should be approximately constant for a constant N. R_I depends on the set of basis shapes coded during the matching pursuit process. This set is mostly a function of the basis dictionary and the residual signal to be coded. While the set of coded basis functions does depend on the in-loop quantizer design, the average basis index codeword length should be approximately constant for sufficiently large N.

We are mainly interested in the last part of the assumption, that $R - R_Q$ is roughly constant with changes in the quantizer. Empirical evidence is provided in Figure 4.12. Three different sample frames with a wide range of coding complexities and effective bit rates are shown. Each frame is coded to a constant number of atoms using the 2-pass DZ-subtracted quantizer described earlier. For $\bar{Q}(\cdot)$, we use a UTQ with a variable stepsize QP. The total bit rate to atoms R_{atm} and the rate without quantization $R_{atm} - R_Q$ are plotted against stepsize QP. The plots show an increase in the total rate R_{atm} for the finer stepsizes, but the rate without quantization is roughly constant as the quantizer changes. This implies that the resulting change in rate is entirely within the R_Q term, and that the sum of the remaining terms is approximately constant.

Assumption 4.2 *Suppose a frame is coded using an initial quantizer $\bar{Q}(\cdot)$ to produce N atoms, and that the rate and distortion are (r, d) for the resulting modulus quantizer and*

Figure 4.12: Bit rates used to code a set number of atoms as quantizer stepsize QP changes. Total rate R_{atm} changes with QP, while the rate excluding quantization $R_{atm} - R_Q$ stays approximately constant. Sample coded frames are from: (a) Container QCIF coded to 110 atoms, (b) Foreman CIF coded to 735 atoms, and (c) Stefan SIF coded to 1419 atoms.



(a) (b) (c)

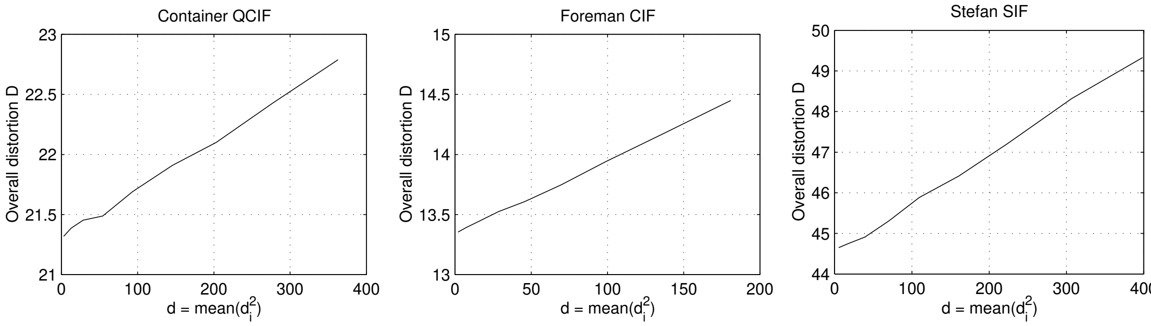
(R, D) for the overall system. A second quantizer $\bar{Q}'(\cdot)$ produces rate and distortion pairs (r', d') and (R', D') for the same number of atoms N. If $d' < d$, then $D' < D$.

Although this assumption seems reasonable, it is not strictly true. Consider the single stage energy reduction expression given in Equation (4.3). The total energy reduction for N stages is thus given as:

$$\overline{ER}_{total} = \sum_{i=1}^{N} p_i^2 - \sum_{i=1}^{N} d_i^2 \tag{4.11}$$

Suppose a change in quantizer $\bar{Q}(\cdot)$ results in a decrease in quantizer distortion d. This reduces the right summation term and thus works to increase the total energy reduction. However, since quantization is done in-loop, the quantizer change may also affect the chosen atoms and thus the left summation term which is based on the set $\{p_i, i = 1...N\}$ may also change with $\bar{Q}(\cdot)$. Define the set of coded basis functions in coding order as the "atom path." We refer to changes in the left summation term as "atom path noise," since these arise due to changes in the atom path due to in-loop quantization noise. Since the atom

Figure 4.13: Overall image distortion D plotted against mean square distortion of the modulus quantizer d for various quantization stepsizes. Conditions shown are: (a) Container QCIF coded to 110 atoms, (b) Foreman CIF coded to 735 atoms, and (c) Stefan SIF coded to 1419 atoms.



(a) (b) (c)

path noise is a complicated function of the basis set and the signal to be coded, it is possible that in switching from one quantizer to another, the resulting atom path noise may be larger than the legitimate change in the distortion term. Hence a decrease in d could cause the overall distortion D to increase. This phenomenon has been observed experimentally, and is particularly likely for small changes in d.

Although the assumption is not strictly true, it does approximately hold for our system when the change $(d - d')$ is reasonably large. This is demonstrated in Figure 4.13. Sample frames are coded each to a constant number of atoms N as the UTQ stepsize is varied. The same sample frames and stepsize values are used as in Figure 4.12. Each stepsize value produces a mean quantizer distortion d and an overall distortion D which are plotted on the horizontal and vertical axes, respectively. Since each plot is strictly increasing, the assumption is shown to hold.

We are now ready to show that the optimal ECSQ for the DZ-subtracted modulus distribution is also optimal in terms of (R, D) for the overall system.

Claim 4.1 *For a given frame, define the optimal scalar quantizer for the DZ-subtracted modulus source as the quantizer which minimizes overall distortion D for the given rate R. If $\bar{Q}_{opt}(\cdot)$ is the optimal quantizer, then it must take the form of the optimal ECSQ for the DZ-subtracted modulus distribution produced by the encoder. That is, for a given rate r, $\bar{Q}_{opt}(\cdot)$ should minimize d.*

Proof: $\bar{Q}_{opt}(\cdot)$ is the optimal scalar quantizer for the system, which therefore codes the given residual signal to rate R using some number of atoms N to achieve the minimum overall distortion D. Take (r, d) to be the quantizer rate and distortion produced by $\bar{Q}_{opt}(\cdot)$. For the moment, assume that $\bar{Q}_{opt}(\cdot)$ is not the optimal ECSQ for the DZ-subtracted modulus distribution produced in coding. Then there exists a quantizer which achieves a lower scalar quantization error at the same average rate, that is

$$\exists \bar{Q}'(\cdot) \text{ s.t. } r' = r \text{ and } d' < d$$

Note that (r', d') is the rate-distortion point produced by quantizer $\bar{Q}'(\cdot)$. We are implicitly assuming that the DZ-subtracted modulus distribution is unaffected by application of either in-loop quantizer. This is a reasonable assumption considering the results shown earlier in Figure 4.9.

Assumption 4.1 shows that this new quantizer $\bar{Q}'(\cdot)$ will hit the target rate R with the same number of atoms as the optimal quantizer, since for coding N atoms,

$$R'_Q = Nr' = Nr = R_Q$$

and the other contributions $(R_{hdr} + R_{mot} + R_P + R_I)$ are approximately constant with change in quantizer.

Since the number of atoms N is constant and $d' < d$, Assumption 4.2 implies that $D' < D$. This is a contradiction, since $\bar{Q}_{opt}(\cdot)$ is given to be the optimal quantizer for the system. Therefore our momentary assumption must be incorrect, and $\bar{Q}_{opt}(\cdot)$ must be the optimal ECSQ for the given modulus distribution.$\square$

We conclude that $\bar{Q}_{opt}(\cdot)$ should be the optimal ECSQ for the DZ-subtracted modulus source, and thus takes the form of a UTQ. However, the stepsize QP must still be optimized for the given frame and target rate. Changing the stepsize effectively traverses the rate distortion curve $r(d)$ for the quantizer. If the overall system is coding the frame to a specific rate R, then we are effectively trading quantizer precision for the number of atoms. For example, reducing the stepsize will reduce the quantizer distortion but allow fewer atoms to be coded at the same overall rate R. Choosing the optimal combination of step size and number of atoms is the topic of the next section.

4.2.3 Optimal Stepsize

In the previous section, the optimal quantizer $\bar{Q}_{opt}(\cdot)$ for application to the DZ-subtracted modulus source was shown to be a UTQ. Although the form of the quantizer is known, the step size has not been determined, and a given target rate may be achieved through many combinations of quantizer step size and number of coded atoms. We now look at the problem of choosing the optimal combination, that is the step size and number of coded atoms which minimize D for a given target rate R.

Figure 4.14 shows the proposed 2-pass quantization algorithm in greater detail. At the start of matching pursuit residual coding, motion compensation has been performed and so R_{hdr} and R_{mot} have already been computed. The first pass decomposes residual frame $\vec{f}$ using matching pursuit without quantization. The goal of this pass is to determine the number of atoms N_k needed to hit the target rate R and also approximate the dead zone parameter DZ_k for use in the second quantized pass. To determine the stopping point, a progressive estimate of R_{atm} is needed, since the exact value cannot be computed in the absence of quantization. The estimated value $\tilde{R}_{atm(i)}$ is computed in the following way. After each stage, the i atoms found so far are quantized as a post-operation and coded in order to estimate the rate. The quantizer for rate estimation is based on DZ-subtraction using the minimum $|p_i|$ found so far as the value for DZ. This is followed by uniform quantizer $\bar{Q}(\cdot)$ with stepsize QP. For the moment, consider QP to be arbitrarily set by the user. Pass 1 then continues until the following stopping criteria is reached:

$$R_{hdr} + R_{mot} + \tilde{R}_{atm(i)} > R \tag{4.12}$$

At this point Pass 1 ends, and the number of atoms N_k and minimum absolute modulus DZ_k are preserved for use in Pass 2.

The step size QP_k is the only degree of optimization freedom in the 2-pass algorithm described above. It is thus possible, although costly, to find the optimal QP_k using an exhaustive linear search. This is done for a few sample frames in Figure 4.15. For each coded frame, the overall distortion D for the 2-pass algorithm is plotted against step size QP. A distinct minimum is found in each case, however the optimal step size varies greatly with sequence and bit rate.

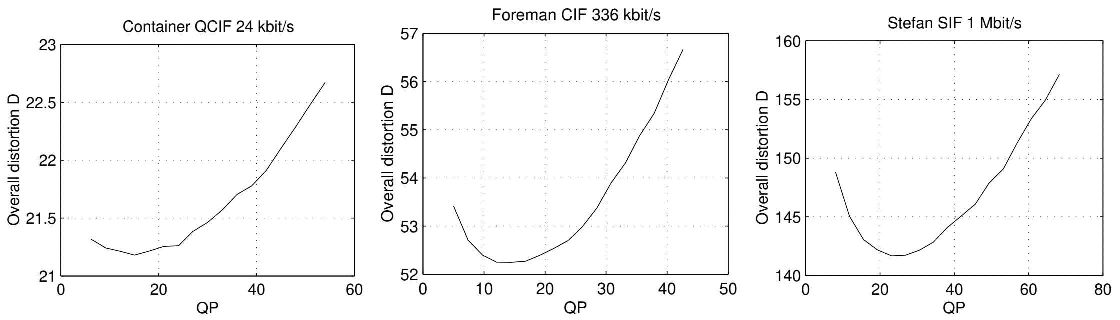
We have found experimentally that a much more consistent value is the ratio between the optimal step size QP_k and the associated dead zone value DZ_k. This is

Figure 4.14: Basic 2-pass adaptive quantization algorithm for the matching pursuit system.

For each frame k:

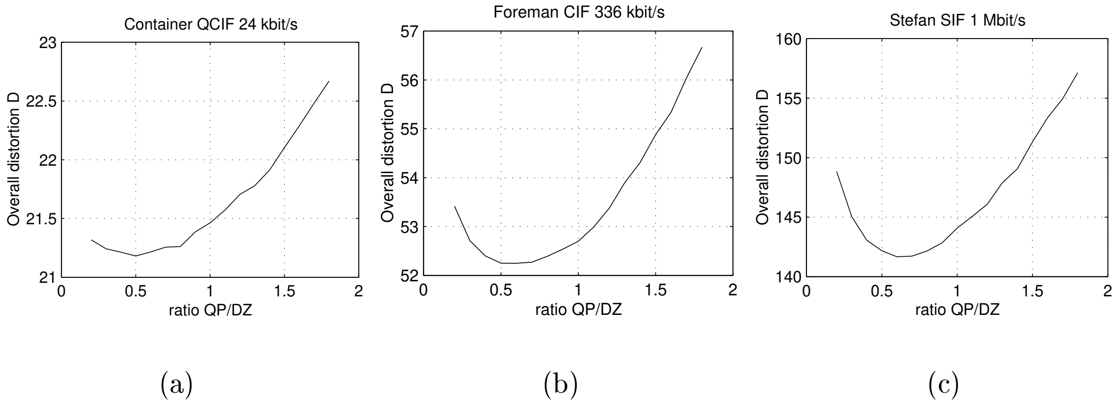
- Compute header budget (R_{hdr}).
- Perform motion estimation; Compute motion budget (R_{mot}).
- Create motion residual frame $\vec{f}$.
- **Pass 1:** Decompose $\vec{f}$ without quantization to determine the N_k needed to hit target rate R and resulting dead zone parameter DZ_k.
- **Pass 2:** Decompose $\vec{f}$ to N atoms using in-loop quantization. The quantizer is defined by (DZ_k, QP_k).
- Transmit **Pass 2** atoms (R_{atm}).
- Increment k; repeat.

Figure 4.15: Search for best quantizer step QP for single coded frames. Overall image distortion D is plotted against quantization step QP for the 2-pass algorithm. (a) Container QCIF coded to 24 kbit/s, (b) Foreman CIF coded to 336 kbit/s, and (c) Stefan SIF coded to 1 Mbit/s.



(a) (b) (c)

Figure 4.16: Search for best ratio QP/DZ for single coded frames. Overall image distortion D is plotted against quantization step QP for the 2-pass algorithm. (a) Container QCIF coded to 24 kbit/s, (b) Foreman CIF coded to 336 kbit/s, and (c) Stefan SIF coded to 1 Mbit/s.



demonstrated in Figure 4.16. In this figure, the same three sample frames are coded multiple times with the 2-pass algorithm, each coding reflecting a different quantizer ratio QP/DZ. Note that the optimal ratio is much more consistant than the optimal step size, falling near 0.6 for each of the three sample frames. This suggests that we choose QP in our 2-pass algorithm so that the quantizer ratio is near this optimal value. To maintain a constant ratio QP_k/DZ_k in the 2-pass algorithm of Figure 4.14, we modify the rate estimation quantizer used to compute $\tilde{R}_{atm(i)}$ during Pass 1. As before, DZ is computed as the minimum $|p_i|$ found for the i atoms found so far. QP is then calculated to maintain the desired quantizer ratio. When stopping criteria (4.12) is reached, current values N_k, DZ_k and QP_k are passed on for use in the second quantized pass.

Table 4.3 shows the result of the optimal ratio search proceedure for a wider range of sample frames and bit rates. A sample frame from each of the standard MPEG test sequences is coded at the usual test rate and three times that rate. Two additional high bit rate sequences, FunFair and Tunnel, are also coded. The two pass algorithm using a

Table 4.3: Search for best ratio QP/DZ for single coded frames. Sample frames coded at one and three times the standard MPEG test bit rate. Optimal ratio value and resulting PSNR are shown for each of 32 test points.

	Base Rate				3x Base Rate			
Sequence	DZ	QP	Ratio	PSNR	DZ	QP	Ratio	PSNR
Container10	54.32	27.16	0.5	31.33	23.15	11.58	0.5	35.96
Hall10	54.96	21.99	0.4	31.85	23.75	9.50	0.4	36.29
Mother10	43.92	13.18	0.3	34.19	18.89	9.44	0.5	38.34
Container24	30.12	15.06	0.5	34.87	14.15	8.49	0.6	38.89
Silent24	39.10	15.64	0.4	32.60	17.84	7.14	0.4	36.90
Mother24	25.95	7.78	0.3	37.18	10.85	6.51	0.6	41.79
Coast48	40.97	20.49	0.5	31.36	19.58	11.75	0.6	35.44
Foreman48	45.05	27.03	0.6	32.92	14.89	11.91	0.8	38.93
News48	67.63	40.58	0.6	32.46	24.04	12.02	0.5	37.67
Coast112	97.84	29.35	0.3	27.60	40.22	24.13	0.6	32.21
Foreman112	106.95	53.47	0.5	30.09	24.08	14.45	0.6	36.97
News112	40.09	28.06	0.7	35.47	14.75	10.33	0.7	40.65
Mobile1024	51.27	30.76	0.6	28.00	20.66	18.59	0.9	33.87
Stefan1024	38.52	23.11	0.6	31.85	13.36	13.36	1.0	38.54
FunFair1024	46.78	37.42	0.8	30.54	16.78	16.78	1.0	36.45
Tunnel1024	34.09	17.05	0.5	31.58	13.24	9.27	0.7	37.35

linear search for best ratio QP_k/DZ_k is again performed, and the results summarized in the table. Note that the optimal ratio ranges from 0.3 to 1.0, with an average of 0.57 over all 32 tested frames. Although this does not match the earlier value of 0.6 in every case, the shape of the curves in Figure 4.16 suggests that the overall distortion is not sensitive to small deviations from the true optimal ratio. This is verified for the test set in Table 4.4, which shows the PSNR loss due to using constant ratio values of 0.6 or 1.0 instead of the optimal ratio found by linear search. For a constant ratio value of 0.6, the loss is extremely small, less than .02 dB difference for all sequences coded up to 1 Mbit/s. For the 3 Mbit/s sequences, the loss is a bit higher, with values in the .05 dB to .13 dB range. For such high coded bit rates, a constant ratio value of 1.0 is a better choice, as shown in the last column of the table.

Table 4.4: Performance loss for same 32 sample frames using constant ratio values of 0.6 and 1.0. PSNR differences are relative to the best ratio found using linear search.

	Base Rate			3x Base Rate		
Sequence	Best Ratio	Ratio=0.6 PSNR Diff	Ratio=1.0 PSNR Diff	Best Ratio	Ratio=0.6 PSNR Diff	Ratio=1.0 PSNR Diff
Container10	0.5	0.0042	0.0448	0.5	0.0097	0.0486
Hall10	0.4	0.0152	0.0776	0.4	0.0145	0.0971
Mother10	0.3	0.0197	0.0853	0.5	0.0115	0.0793
Container24	0.5	0.0071	0.0576	0.6	0.0	0.0572
Silent24	0.4	0.0139	0.0842	0.4	0.0153	0.1265
Mother24	0.3	0.0023	0.0723	0.6	0.0	0.0547
Coast48	0.5	0.0141	0.0727	0.6	0.0	0.0641
Foreman48	0.6	0.0	0.0616	0.8	0.0201	0.0288
News48	0.6	0.0	0.0642	0.5	0.0097	0.0704
Coast112	0.3	0.0099	0.0638	0.6	0.0	0.0662
Foreman112	0.5	0.0061	0.0456	0.6	0.0	0.0373
News112	0.7	0.0006	0.0262	0.7	0.0085	0.0714
Mobile1024	0.6	0.0	0.0634	0.9	0.0573	0.0239
Stefan1024	0.6	0.0	0.0736	1.0	0.1287	0.0
FunFair1024	0.8	0.0073	0.0273	1.0	0.0711	0.0
Tunnel1024	0.5	0.0069	0.0866	0.7	0.0124	0.0575

We thus conclude that an adaptive 2-pass algorithm using dead zone subtraction and a constant value of step size to dead zone ratio achieves nearly optimal performance for the matching pursuit video coding system. For our experiments, a ratio value of 0.6 will be used for all rates up to 1 Mbit/s and a value of 1.0 will be used for higher rates. Note that our quantizer uses simple midpoint reconstruction. The next section investigates the effectiveness of the more optimal centroid reconstruction [31] on the 2-pass algorithm. Our experiments suggest, however, that this method provides little or no gain.

4.2.4 Centroid Reconstruction

The results presented so far all use a UTQ with midpoint reconstruction. For completeness, we present in this section a comparison between these results and an equivalent system using centroid reconstruction. Essentially this means that the reconstruction level for each quantizer bin is defined to be the average of the source values which map to that bin. Centroid reconstruction is a necessary condition for an optimal quantizer [31], and for a given source should reduce the quantizer distoration d without affecting the rate r. However, gains for centroid over midpoint reconstruction are often small [63].

Since quantization is done in-loop, an estimate of the centroid must be known before coding begins. Because the dead-zone subtracted modulus distribution is a relatively stable exponential as shown in Figure 4.8, it is possible to predict the bin centroids for the current frame from the previous frame modulus values. Since the quantizer stepsize varies from frame to frame, we represent the reconstruction level as a fractional bin value (FBV), essentially mapping the continuous positions within each bin linearly to the interval $[0, 1)$. Specifically, if the reconstruction value c lies on a bin defined by $[a, b)$, then the FBV representation of c is:

$$FBV = (c - a)/(b - a)$$

To avoid erroneous centroid values from insufficient data, the reconstruction for each bin in the current frame is set according to the following rule:

if: the previous frame was coded using M.P. and the corresponding quantizer bin had at least ten associated atoms,
then: the reconstruction FBV for the bin in the current frame is set to the FBV of the centroid computed from data in the previous frame.
else: the reconstruction FBV of the bin in the current frame is set to 0.5, equivalent to midpoint reconstruction.

The improvement due to the predictive centroid method is summarized in Table 4.5. Average PSNR gains are shown for the standard MPEG test sequences and rates. There is no improvement seen, as the average gain is 0.01 dB, and the difference is randomly

Table 4.5: Performance gain for centroid reconstruction over midpoint reconstruction.

Sequence	PSNR-Diff	Sequence	PSNR-Diff
Container10	-0.08	Container24	-0.01
Hall10	0.13	Silent24	0.08
Mother10	-0.04	Mother24	0.03
Coast48	-0.02	Coast112	0.05
Foreman48	-0.01	Foreman112	0.03
News48	0.02	News112	0.05
Mobile1024	-0.04	Stefan1024	0.01

positive or negative from sequence to sequence. As noted in the justification of **Claim 3.1**, Section 4.2.2, this is the expected result for a quantizer change in which the improvement in average distortion d is small. The resulting change in overall distortion D may be either positive or negative, depending on the contribution from atom path noise. Since centroid reconstruction does not improve the results, the simplicity of midpoint reconstruction will be used in the final system.

4.3 Single-Pass Adaptive Solution

The main disadvantage of the two-pass algorithm is the computational cost of the first pass. This initial pass performs a pre-analysis of the residual image in order to determine DZ and QP to achieve near-optimal quantizer performance. This doubles the complexity of the encoder, and so a less expensive estimate of these parameters is highly desirable. In this section, we present a method for predicting DZ for the current frame from known previous frame values. Prediction of DZ is sufficient, since QP follows from the constant quantizer ratio.

Consider the 2-pass algorithm and suppose that the encoded modulus values for frame k are $\{\dot{p}_{(k,i)}, i = 1...N\}$ from the first pass and $\{\ddot{p}_{(k,i)}, i = 1...N\}$ from the second pass. Recall that dead zone parameter DZ_k is normally determined from Pass 1 results as:

$$DZ_k = min\{|\dot{p}_{(k,i)}|\} \tag{4.13}$$

To eliminate the analysis pass, we make two observations. First, quantization has little effect on the modulus distribution, as demonstrated earlier in Figure 4.9. This allows us to approximate:

$$min\{|\dot{p}_{(k,i)}|\} \approx min\{|\ddot{p}_{(k,i)}|\} \tag{4.14}$$

If we also assume that the dead zone parameter changes slowly from frame to frame, then we have the approximation:

$$DZ_k \approx DZ_{k-1} \tag{4.15}$$

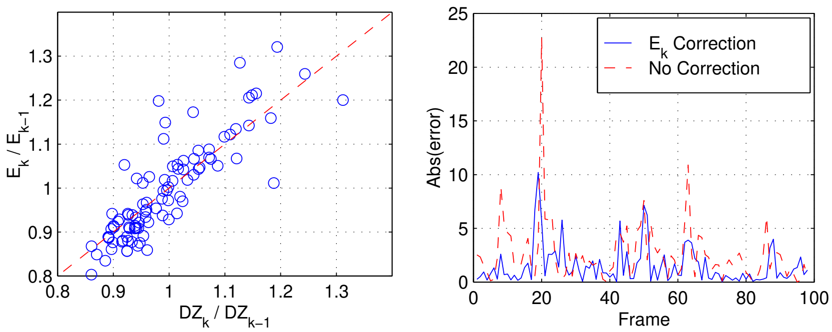
By combining these with Equation (4.13), we may approximate DZ_k as:

$$DZ_k \approx min\{|\ddot{p}_{(k-1,i)}|\} \tag{4.16}$$

The dead zone parameter for the current frame is thus defined by the Pass 2 decomposition of the previous frame, and so the analysis pass is eliminated altogether.

This scheme works to some extent, however it is not always true that DZ_k changes slowly from frame to frame. To compensate for this, we may consider the initial energy $E = ||\vec{f_0}||^2$ of the current frame. If the residual for frame k has a higher initial energy E_k, all else equal, we would expect the distribution of $\{|p_i|, i = 1...N\}$ for the frame to shift to the right and effectively increase the dead zone size. Figure 4.17a confirms this relationship for the Mother sequence coded using the 2-pass algorithm at 24 kbit/s. Each coded frame is represented by a circle, with the ratio of current to previous frame dead zone values on

Figure 4.17: (a) Justification of the energy correction for DZ prediction. Horizontal axis shows ratio of current to previous frame DZ values as found by the full 2-pass adaptive algorithm. Vertical axis shows ratio of current to previous frame residual energy before atom coding. (b) Absolute error in predicting DZ_k for each frame of Mother at 24 kbit/s. Dotted line shows error for simple prediction. Solid line shows reduced error when energy corrected prediction is used.



(a) (b)

the horizontal axis and the corresponding ratio of initial frame energies on the vertical. A correlation between the two ratios can be seen, and the dotted line $y = x$ suggests the following energy-corrected prediction for the current frame dead zone:

$$DZ_k = DZ_{k-1} \cdot \left(\frac{E_k}{E_{k-1}}\right) \tag{4.17}$$

The effectiveness of these prediction methods is evaluated in Figure 4.17b. Data is from the 2-pass encoding of Mother at 24 kbit/s, and the value of DZ_k is predicted with energy correction according to Equation (4.17) and without correction as in Equation (4.16). The energy corrected prediction is seen to be better on average, however it is not guaranteed to be better for individual frames. If we compute mean square prediction error in DZ_k, then energy correction reduces the error from 15.03 to 5.45.

The improved DZ prediction method is thus used to define a one-pass adaptive quantization algorithm. This algorithm is similar to the two-pass algorithm of Figure 4.14,

Figure 4.18: The 1-pass adaptive quantization algorithm for the matching pursuit system.

```
For each frame k:
    • Compute header budget (R_hdr).
    • Perform motion estimation; Compute motion budget (R_mot).
    • Create motion residual frame f⃗.
    ▷ If (previous frame data exists)
        Set DZ_k according to Equation (4.17).
    ▷ Else
        Set DZ_k = 15.
    • Set QP_k according to DZ_k and desired ratio.
    • Decompose f⃗ using in-loop quantization.
        Quantizer is defined by (DZ_k, QP_k).
        Dead zone is divided into high-rate (ShiftVLC) bins.
        Code until progressive rate estimate equals target rate.
    • Transmit resulting atoms (R_atm).
    • Increment k; repeat.
```

except Pass 1 is omitted and the dead zone parameter is predicted according to Equation (4.17). If there is no previous coded frame, the dead zone parameter is arbitrarily set to a reasonable value $DZ = 15$. The resulting 1-pass algorithm is summarized in Figure 4.18.

There is one potential problem of combining DZ-subtraction with a 1-pass algorithm. Without the benefit of a pre-analysis pass, the predicted DZ_k may be large enough to allow the single encoder pass to produce atoms which violate the condition $|p_i| > DZ_k$. When this occurs, we revert back to the dead zone splitting technique used for the modified fixed quantizer of Figure 4.4b. That is, ShiftVLC-like bins are thus added to allow such small moduli to be coded. Note that these were not necessary for the full 2-pass algorithm, since the approximation in (4.14) holds very well in practice.

The resulting 1-pass system can thus be thought of as a combination of the fixed ShiftVLC and adaptive 2-pass algorithms. The advantages of each algorithm are retained, as will be observed in the next section.

4.4 Comparison of Results

This section compares the three quantization schemes presented in the previous sections on the basis of PSNR performance and complexity. The schemes are summarized as:

ShiftVLC: Fixed quantization using midtread linear quantizer with $QP = 30$ and three added high-rate quantization bins.

2-Pass: Two-pass adaptive quantization. The first analysis pass sets the DZ and QP for the second quantized pass.

1-Pass: Single-pass adaptive quantization. Similar to 2-Pass, but analysis pass is omitted and DZ and QP are predicted from known values in the previous frame. Three high-rate quantization bins are added for safety, similar to the ShiftVLC case.

All three encoders were implemented and trained on a set of sequences outside the standard MPEG-4 test set. Coding was done on luminance only at a constant bit rate. Each ten-second sequence has a single leading intra frame, which was encoded using DCT followed by an in-the-loop deblocking filter. All other frames were encoded using motion compensated matching pursuit with one of the associated quantization schemes.

The PSNR results are summarized in Tables 4.6 and 4.7. Table 4.6 shows average PSNR results for standard MPEG4 test sequences and rates. The rightmost three columns are the most interesting, since they show the PSNR differences between the three

Table 4.6: Average Luminance PSNR for the standard MPEG4 sequences and rates.

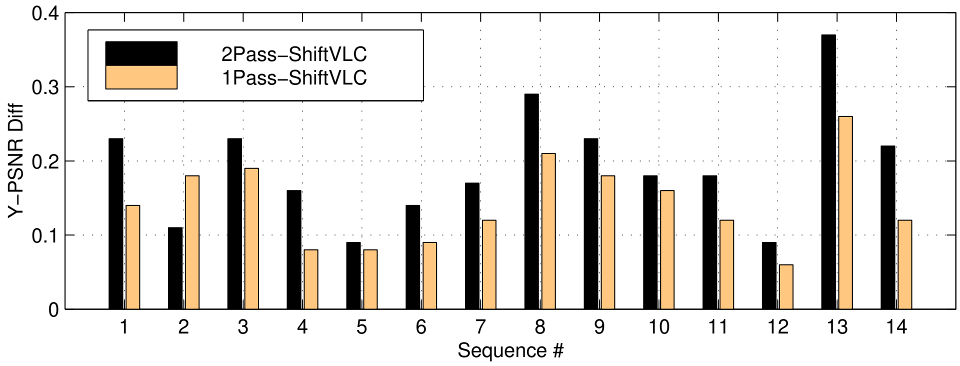
Sequence	Shift VLC	Two Pass	One Pass	2Pass-1Pass	2Pass-ShiftVLC	1Pass-ShiftVLC
Container10	31.27	31.50	31.41	0.09	0.23	0.14
Hall10	32.12	32.23	32.29	-0.07	0.11	0.18
Mother10	33.28	33.52	33.47	0.05	0.23	0.19
Container24	34.39	34.55	34.46	0.08	0.16	0.08
Silent24	32.27	32.36	32.35	0.01	0.09	0.08
Mother24	36.54	36.69	36.63	0.05	0.14	0.09
Coast48	30.04	30.21	30.16	0.05	0.17	0.12
Foreman48	31.88	32.18	32.10	0.08	0.29	0.21
News48	33.02	33.25	33.20	0.05	0.23	0.18
Coast112	26.42	26.60	26.58	0.02	0.18	0.16
Foreman112	30.94	31.12	31.07	0.05	0.18	0.12
News112	36.06	36.15	36.12	0.03	0.09	0.06
Mobile1024	27.00	27.37	27.26	0.11	0.37	0.26
Stefan1024	32.04	32.25	32.15	0.10	0.22	0.12

schemes. The first of these columns shows that the 1-Pass algorithm comes very close in performance to the 2-Pass algorithm, almost always within 0.1 dB. The second and third difference columns show the improvement of the adaptive algorithms over the fixed quantizer. These differences are also shown graphically in Figure 4.19. A consistent improvement over ShiftVLC is seen across all tested sequences. For the 2-Pass case, the improvement ranges from 0.09 to 0.37 dB with an average value of .19 dB. The 1-Pass algorithm also shows consistant improvement over ShiftVLC, with an average gain of .14 dB.

Performance at higher bit rates is shown in Table 4.7. Results are shown for five of the MPEG4 test sequences chosen to represent a range of encoding conditions. For each sequence, the standard MPEG-4 test rate is used as a base, and several multiples of this rate are also encoded.[3] Again it is useful to focus on the rightmost three columns which

[3]Note that the Foreman48 series and the Foreman112 series differ by image size as well as bit rate. The sequences based on Foreman48 are encoded at QCIF resolution, while those based on Foreman112 are encoded at CIF. See Table 1.2 for the base sequence conditions.

Figure 4.19: Average Y-PSNR gains over ShiftVLC for the 1-Pass and 2-Pass algorithms. See Table 1.2 for sequence indices.



describe the PSNR differences. The first of these columns shows that 1-Pass performance is again close to that of the full 2-Pass algorithm, typically less than 0.10 dB. The second and third difference columns effectively show how the improvement over ShiftVLC due to the adaptive schemes varies with increasing bit rate. These differences are also shown graphically in Figure 4.20. It is seen that the adaptive schemes perform much better than the fixed quantizer at higher bit rates, and that for three of the five sequences the improvements fall in the 1 to 2 dB range for the highest tested bit rates. The largest improvements are 1.68 dB for the QCIF Foreman sequence coded at 192 kbit/s, and 1.40 dB for the SIF Stefan sequence at 3 Mbit/s. For these high rates, the visual improvement is mainly seen as a reduction of ringing around sharp edges.

Figures 4.21 and 4.22 show a more detailed look at PSNR vs. Frame for specific sequences. Four sequences coded at standard MPEG-4 test rates are depicted in Figure 4.21. In each case, it is evident that the adaptive schemes show a consistent improvement over the fixed quantizer, and that the 1-Pass algorithm has performance only slightly below that of the 2-Pass algorithm. Higher encoding rates for the same sequences are shown in

Table 4.7: Average Luminance PSNR for ascending bit rates.

Sequence	Shift VLC	Two Pass	One Pass	2Pass-1Pass	2Pass-ShiftVLC	1Pass-ShiftVLC
Container10	31.27	31.51	31.41	0.09	0.23	0.14
Container20	34.19	34.36	34.26	0.11	0.18	0.07
Container30	35.91	35.94	35.89	0.05	0.04	-0.02
Container40	36.74	37.01	36.94	0.07	0.27	0.20
Mother24	36.54	36.69	36.64	0.05	0.15	0.09
Mother48	38.77	39.63	39.52	0.11	0.86	0.75
Mother72	40.26	41.58	41.50	0.08	1.31	1.24
Mother96	41.51	42.93	42.83	0.10	1.42	1.32
Foreman48	33.15	33.36	33.27	0.09	0.21	0.12
Foreman96	37.09	37.24	37.16	0.08	0.15	0.07
Foreman144	38.38	39.53	39.39	0.14	1.15	1.01
Foreman192	39.55	41.22	41.04	0.19	1.68	1.49
Foreman112	30.96	31.14	31.08	0.06	0.18	0.13
Foreman224	35.06	35.16	35.09	0.07	0.11	0.03
Foreman336	36.93	36.99	36.90	0.08	0.06	-0.03
Foreman448	37.92	38.24	38.18	0.06	0.33	0.27
Stefan1024	32.04	32.26	32.16	0.10	0.22	0.12
Stefan2048	35.99	36.35	36.23	0.12	0.36	0.24
Stefan3072	37.75	39.14	38.92	0.22	1.40	1.17

Figure 4.20: Average Y-PSNR gains over ShiftVLC for five sequences at various bit rates.

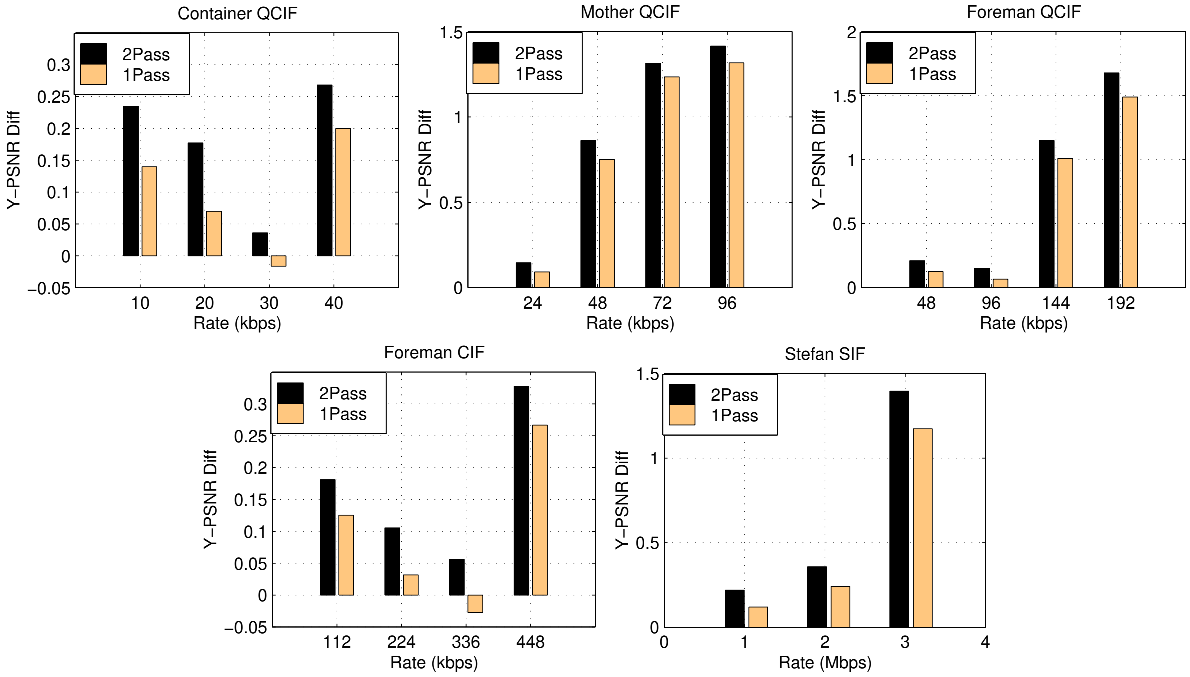


Figure 4.21: Luminance PSNR vs. frame for four sample sequences at standard MPEG4 test rates.

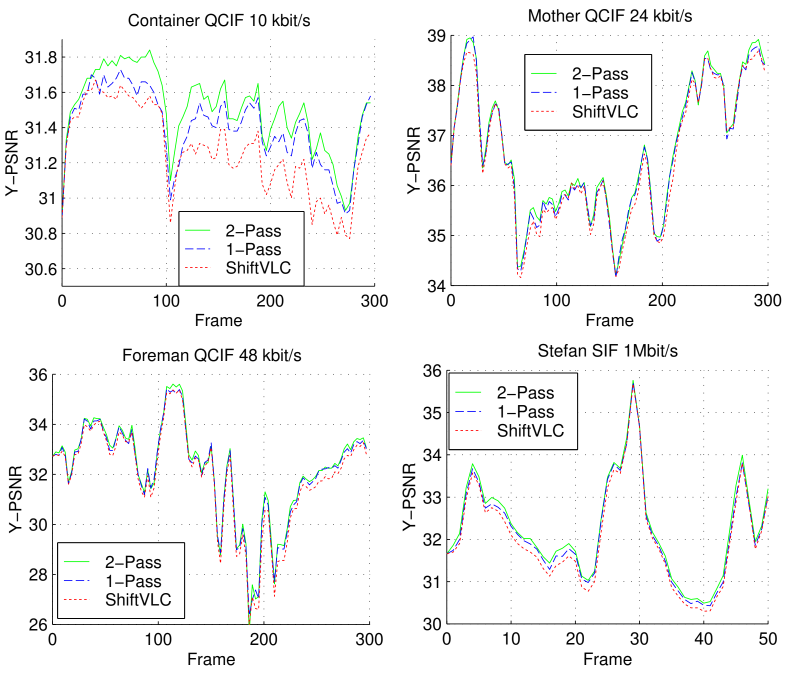
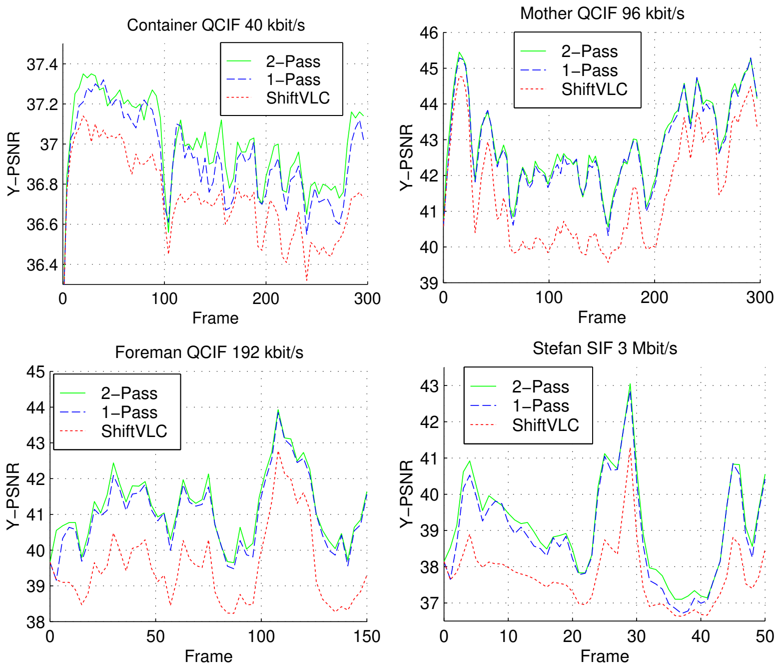


Figure 4.22. It is clear from these plots that the gain for the adaptive schemes is much higher at these rates, and that again the 1-Pass and 2-Pass schemes are nearly equivalent.

4.5 Conclusions

Frame adaptive quantizer designs have been shown to outperform the fixed quantizer used in earlier matching pursuit video coding work. Our design uses a combination of adaptive dead zone subtraction and a UTQ with adaptive stepsize. A 2-pass algorithm was shown to choose a nearly optimal balance between the quantizer precision and the number of coded atoms, and hence comes close to minimizing overall coding distortion D for a given target rate R. A 1-pass algorithm was shown to have nearly the same performance, but with complexity similar to that of the original fixed quantizer design. Small improvements

Figure 4.22: Luminance PSNR vs. frame for four sample sequences at higher bit rates.



of .1 to .4 dB are shown for the MPEG sequences at standard rates, and larger gains of up to 1.7 dB can be shown for the same sequences at higher bit rates.