A RATE-DISTORTION OPTIMAL CODING ALTERNATIVE TO MATCHING PURSUIT

Tom Ryen[†]*, Guido M. Schuster*[††]*, and Aggelos K. Katsaggelos*[†††]

[†] Stavanger University College, Department of Electrical and Computer Engineering,
P.O.Box 2557 Ullandhaug, 4091 Stavanger, Norway. E-mail: tom.ryen@tn.his.no
[††] HSR Hochschule für Technik Rapperswil, Abteilung Elektrotechnik,
Oberseestrasse 10, 8640 Rapperswil, Switzerland. E-mail: guido.schuster@hsr.ch
[†††] Northwestern University, Department of Electrical and Computer Engineering,
Evanston, Illinois 60208-3118, USA. E-mail: aggk@ece.nwu.edu

ABSTRACT

This paper presents a method to find the operational rate-distortion optimal solution for an overcomplete signal decomposition. The idea of using overcomplete dictionaries, or frames, is to get a sparse representation of the signal. Traditionally, suboptimal algorithms, such as Matching Pursuit (MP), is used for this purpose. When using frames in a lossy compression scheme, the major issue is to find the best possible rate-distortion (RD) tradeoff. Given the frame and the Variable Length Code (VLC) table embedded in the entropy coder, the solution of the problem of establishing the best RD tradeoff has a very high complexity. The proposed approach reduces this complexity significantly by structuring the solution approach such that the dependent quantizer allocation problem reduces into an independent one. It is important to note that this large reduction in complexity is achieved without sacrificing optimality.

The optimal rate-distortion solution depends on the VLC table embedded in the entropy coder. Thus, VLC optimization is part of this work. We show experimentally that the new approach outperforms Rate-Distortion Optimized Matching Pursuit, previously proposed in [1].

1. INTRODUCTION

A widely used method in lossy compression is transform coding. The idea in it is to decorrelate the data and compact the energy of the signal in few coefficients. Low frequency coefficients are subject to fine quantization while high frequency coefficients are subject to course quantization, resulting in a number of zero coefficients. The small number of nonzero coefficients results in a *sparse* representation. An entropy coder is used as the last step in the compression scheme. A sparse representation is preferable as an input, since it can be represented with fewer bits, due to its low entropy.

In recent years, the use of overcomplete dictionaries, or *frames*, has received a lot of attention in lossy compression. A frame is a set of column vectors, just as a transform, but with a larger number of vectors than the number of elements in each vector, thus the name *overcomplete*. The basic idea of using a frame instead of a transform is that we have more vectors to choose from and thus a better chance of finding a small number of vectors whose linear combination match the signal vector well. One disadvantage of using a frame instead of a transform, is the complexity of finding an optimal sparse representation from an overcomplete set of vectors. It is shown to be an NP-complete problem [2]. Thus, more practical but suboptimal vector selection algorithms have been developed, such as Matching Pursuit (MP) [3], Orthogonal Matching Pursuit (OMP) [4] and Fast Orthogonal Matching Pursuit (FOMP) [5]. A drawback with these methods is that even if we had a optimal selection of continuous valued weight coefficients, an independent scalar quantization of each coefficient would be suboptimal.

The object of MP, OMP and FOMP is to minimize the distortion subject to a sparsity constraint, e.g., a given number of nonzero coefficients per signal block. In a compression scheme it would be better to have the bit rate as the constraint, since in lossy compression the rate-distortion tradeoff is the main object.

The major issue in *rate-distortion optimization* is to find the best tradeoff between rate and distortion. Rate-Distortion Theory (RDT) [6] has been used in many application, such as in Video compression [7, 8], Shape Coding [9] and Compression of electrocardiogram (ECG) data [10]. The central entity in RDT is the Rate-Distortion Function (RDF), which is the lower bound of the distortion that is obtainable with a given bit rate available. When the *Variable Length Code* (VLC) table embedded in the entropy coder table is known, we can find the *Operational Rate-Distortion Function* (ORDF) [7]. When using frame coding, each combination of coefficients will result in a rate R and a distortion D, which can be viewed as an (R, D)-point in a rate-distortion diagram. An (R, D)-point is a part of the ORDF if there is no other (R, D)-points with a smaller distortion using the same or a smaller rate. A simple example of a rate-distortion diagram is shown in Fig. 1. All (R, D)-points are indicated as plus signs. The circled ones are the members of the ORDF, while the solid line represent the ORDF's *convex hull*. A much used method to find members of the convex hull is the *Lagrangian Multiplier Method* [7, 8]. This method is essential in the way we formulate and solve our problem in the next two sections.

This paper presents a new method to find the optimal rate-distortion tradeoff for a compression scheme using overcomplete dictionaries. A formulation of the optimization problem is presented in Section 2. In Section 3 we present a solution method which finds the optimal solution of the rate-distortion tradeoff for a one-dimensional signal, given the frame and the VLC. Experiment results from the use of this method with an AR(1) process are presented in Section 4. Here we also make a comparison between our method and a MP algorithm based on rate-distortion optimization.

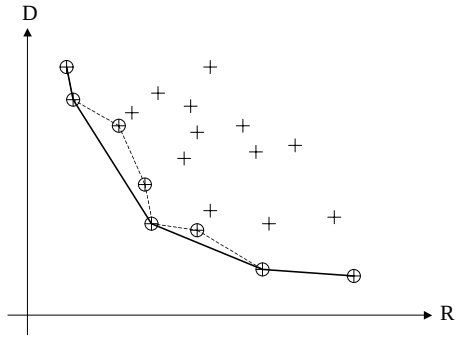

Fig. 1. Operational rate-distortion function (ORDF). All (R, D)-points are indicated as plus signs. Members of ORDF are circled. The solid line is the convex hull for the ORDF.

2. PROBLEM FORMULATION

Consider a one-dimensional signal, $\mathbf{x}$, divided in L blocks, each consisting of N samples. The l-th signal block, $\mathbf{x}_l$, is a column vector of length N. Consider a frame, $\mathbf{F}$, with dimension $N \times K$, where $K > N$. $\mathbf{x}_l$ can be written as a combination of column vectors $\mathbf{f}_k$ of $\mathbf{F}$, that is,

$$\mathbf{x}_l = \mathbf{F}\,\mathbf{w}_l = \sum_{k=1}^{K} w_{l,k}\mathbf{f}_k, \tag{1}$$

where $\mathbf{w}_l$ is the continuous valued coefficient vector for the l-th signal block and $w_{l,k}$ is the k-th element in $\mathbf{w}_l$. The frame's column vectors are all of unit length. In a compression scheme, we deal with a sparse and quantized coefficient vector, $\tilde{\mathbf{w}}_l$, which will give us the reconstructed signal vector, $\tilde{\mathbf{x}}_l = \mathbf{F}\,\tilde{\mathbf{w}}_l$. We define both the bit rate and the distortion for each block to be independent, i.e., the total bit rate, R, and the total distortion, D, is the sum of the rate and the distortion for each block, respectively. The distortion for block l, D_l, is defined by

$$D_l(\tilde{\mathbf{w}}_l) = \left\|\mathbf{x}_l - \tilde{\mathbf{x}}_l\right\|^2 = \left\|\mathbf{x}_l - \mathbf{F}\,\tilde{\mathbf{w}}_l\right\|^2. \tag{2}$$

When using frames, we expect sparse coefficient vectors, i.e., a large number of the K elements in $\tilde{\mathbf{w}}_l$ are zero. Therefore it is convenient to use a Run-lenght coder (RLC) as a part of the entropy coder. The RLC counts the number of zeros between each nonzero coefficient. After the last nonzero coefficient in each block, we use an End Of Block (EOB) symbol to indicate the start of the next block. Each nonzero coefficient has a value taken from a finite set. We use two different VLC tables, one for the coefficient *values* and one for the *runs* between the nonzero coefficients. We can now define the rate for block l, R_l, as

$$R_l(\tilde{\mathbf{w}}_l) = \sum_{k \in nz} (R_{l,k}^{val} + R_{l,k}^{run}) + R_l^{EOB}, \tag{3}$$

where nz is the set of indices for the nonzero coefficients, $R_{l,k}^{val}$ and $R_{l,k}^{run}$ the number of bits used to code the *value* and the *run* for the k-th coefficient, respectively, and R_l^{EOB} the number of bits needed to transmit the EOB symbol.

Our goal is to find the minimum total distortion subject to a given bit budget, R_{budget}. We can formulate our initial optimization problem as

$$\begin{aligned} \min_{\tilde{\mathbf{w}}_l} \quad & \sum_{l=1}^{L} D_l(\tilde{\mathbf{w}}_l) \\ \text{s.t.} \quad & \sum_{l=1}^{L} R_l(\tilde{\mathbf{w}}_l) = R_{budget}. \end{aligned} \tag{4}$$

This is an Integer Optimization Problem due to the discrete valued $\tilde{\mathbf{w}}_l$ and nonlinear due to Eq. (2). We relax the problem by using the *Lagrangian Multiplier Method*. That is, the following unconstrained optimization is performed

$$\begin{aligned} &\min_{\tilde{\mathbf{w}}_l} \Big(\sum_{l=1}^{L} D_l(\tilde{\mathbf{w}}_l) + \lambda \sum_{l=1}^{L} R_l(\tilde{\mathbf{w}}_l) \Big) \\ &= \sum_{l=1}^{L} \Big[\min_{\tilde{\mathbf{w}}_l} \Big(D_l(\tilde{\mathbf{w}}_l) + \lambda R_l(\tilde{\mathbf{w}}_l) \Big) \Big], \end{aligned} \tag{5}$$

for $\lambda \in \mathbb{R}^+$. After Eq. (5) is solved optimaly the appropriate λ needs to be chosen so that the constraint in Eq. (4) is satisfied. By using the formulation in Eq. (5) with several different λ-values, we can find segments of the ORDF's convex hull. This is of great value in a quality study of the compression scheme. Even though the signal blocks are independent, there are still dependencies between the coefficients, both in rate and distortion. The complexity of the minimization in Eq. (5) is high, since for every block we need to search between all possible ways to place M nonzero coefficient in a coefficient vector of length K, where $M = 1$, $M = 2$, and so on. The total number of combinations with M nonzero coefficients is $\binom{K}{M}$. In addition, each nonzero coefficient can take on a given number of different *value* symbols, C. For each combination, the number of solutions is C^M. Thus, the total number of different solutions is $\sum_{M=0}^{M_{max}} \binom{K}{M} C^M$, where M_{max} is the largest number of nonzero coefficients we choose to use. In all practical cases, $M_{max} << K$, due to the sparse representation idea. A typical number of elements in the coefficient vector is $K = 32$. Consider that all nonzero coefficient can take on $C = 30$ different values. For $M = \{1, 2, 3, 4\}$, $\binom{K}{M} = \{32, 496, 4960, 35960\}$ and $C^M = \{30, 900, 27000, 810000\}$. A way to get rid of the latter combinatorial explosion is presented in the next section.

3. PROPOSED SOLUTION METHOD

As we have shown in the previous paragraph, the complexity of an exhaustive search is extremely large. In this section we introduce the core contribution of this work, which is the way we structure the optimization problem. First, we pick a particular combination of M nonzero coefficients. The problem is now reduced to finding an optimal solution for this particular selection. If the M vectors would be orthonormal, then the total distortion would be the sum of the coordinate distortions. However, this is not the case since these M vectors are not necessarily orthonormal and so the optimization problem is still that of dependent quantizers, and thus very complex. We now force this orthonormal condition on the problem by using a *QR decomposition* [11], which results in a new space where the total distortion is simply the sum of the coordinate distortions. Hence in that space, there are no dependencies among the coefficients and therefore the problem is now an independent quantizer allocation problem, which is much faster to solve as the

set of optimal quantizers for each coefficient is also the optimal solution to the overall problem.

For a given a combination of M nonzero coefficients, the rate for the *run* symbols is known. Yet, we still don't know the distortion nor the rate for the *value* symbols. It is always the case that $M < N$, due to the sparse representation idea. Let us define a new matrix, Φ, which is formed by the column vectors of $\mathbf{F}$ corresponding to the nonzero coefficients. Φ will have dimensions $N \times M$ and represent an *undercomplete* set of vectors. Let $\mathbf{v}_l = [v_{l,1}, \cdots, v_{l,M}]^T$ and $\tilde{\mathbf{v}}_l = [\tilde{v}_{l,1}, \cdots, \tilde{v}_{l,M}]^T$ be the corresponding subsets of $\mathbf{w}_l$ and $\tilde{\mathbf{w}}_l$, respectively. Since $M < N$, the best reconstruction of signal vector $\mathbf{x}_l$ when continuous valued coefficients are used is given by

$$\hat{\mathbf{x}}_l = \Phi \mathbf{v}_l = \Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{x}_l, \tag{6}$$

due to the Best Approximation Theorem [11]. The error is orthogonal to any vector spanned by the column vectors in Φ. Using the Pythagorean theorem the distortion in block l can be written as

$$D_l = \|\mathbf{x}_l - \tilde{\mathbf{x}}_l\|^2 = \|\mathbf{x}_l - \hat{\mathbf{x}}_l\|^2 + \|\hat{\mathbf{x}}_l - \tilde{\mathbf{x}}_l\|^2. \tag{7}$$

This is illustrated in Fig. 2. The dots in the figure represent possible values the reconstructed vector, $\tilde{\mathbf{x}}_l$, can take. $(\mathbf{x}_l - \hat{\mathbf{x}}_l)$ will always be orthogonal to $(\hat{\mathbf{x}}_l - \tilde{\mathbf{x}}_l)$ regardless to the value and direction of $\tilde{\mathbf{x}}_l$. The first term in Eq. (7) is a constant, since both $\mathbf{x}_l$ and Φ are known. We can now focus on the last term of the equation. The column vectors in Φ are not necessarily orthogonal. Thus, we still have dependencies between the coefficients. By using QR decomposition, we can get an orthogonal version of Φ. We can write $\Phi = \mathbf{QR}$, where $\mathbf{Q}$ is an $N \times N$ matrix with all orthonormal vectors, and $\mathbf{R}$ an $N \times M$ upper triangular matrix. Let us define new coefficient vectors related to the orthonormal basis as $\mathbf{v}_l^o = \mathbf{R}\mathbf{v}_l$ and $\tilde{\mathbf{v}}_l^o = \mathbf{R}\tilde{\mathbf{v}}_l$. We can write

$$\begin{aligned} \|\hat{\mathbf{x}}_l - \tilde{\mathbf{x}}_l\|^2 &= \|\mathbf{QR}\mathbf{v}_l - \mathbf{QR}\tilde{\mathbf{v}}_l\|^2 \\ &= (\mathbf{v}_l^o - \tilde{\mathbf{v}}_l^o)^T\mathbf{Q}^T\mathbf{Q}(\mathbf{v}_l^o - \tilde{\mathbf{v}}_l^o) \\ &= (\mathbf{v}_l^o - \tilde{\mathbf{v}}_l^o)^2 = \sum_{n=1}^{M}(v_{l,n}^o - \tilde{v}_{l,n}^o)^2. \end{aligned} \tag{8}$$

The distortion, D_l, can now be written as a constant added to the sum of *independent* coefficient distortions. From the last term of Eq. (8) we note that it is not necessary to sum the $(N - M)$ last coefficients, since they will all be zero, due to the upper triangular matrix $\mathbf{R}$. We will from now on use the coefficient vector $\tilde{\mathbf{v}}_l^o = [\tilde{v}_{l,1}^o, \cdots, \tilde{v}_{l,M}^o]^T$ as the decision variables. For a given combination of M nonzero coefficients, we need to find the combination's Φ, $\mathbf{Q}$ and $\mathbf{R}$, and then minimize with respect to $\tilde{\mathbf{v}}_l^o$

$$\begin{aligned} & \|\mathbf{x}_l - \Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{x}_l\|^2 \\ + \quad & \lambda\Big(\sum_{m=1}^{M}(R_{l,m}^{run}) + R_l^{EOB}\Big) \\ + \quad & \sum_{n=1}^{M}\min_{\tilde{v}_{l,n}^o}\big((v_{l,n}^o - \tilde{v}_{l,n}^o)^2 + \lambda R_{l,n}^{val}\big). \end{aligned} \tag{9}$$

The problem is now much faster to solve, since only C comparisons are necessary, compared to the C^M comparisons needed in our original problem. To find the optimal rate-distortion tradeoff, we must solve Eq. (9) for all $\binom{K}{M}$ combinations for $M = \{1, \ldots, M_{max}\}$, and store the minimum of all solutions. When working with low bit rate compression, the maximum number of nonzero coefficients per block is low. Thus, M_{max} could be a small number, and the time used to find the optimal solution could be substantially lower than with a higher M_{max}. It should be mentioned that the decoder needs a QR decomposition depending on each coefficient vector's nonzero coefficient indices, in addition to the knowledge of the frame and the VLC tables used in the encoder.

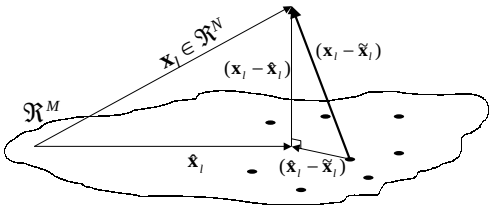

Fig. 2. A visualization of the orthogonality between the minimum error vector and the subspace spanned by the column vectors in Φ. The dots is possible values $\tilde{\mathbf{x}}_l$ can take on.

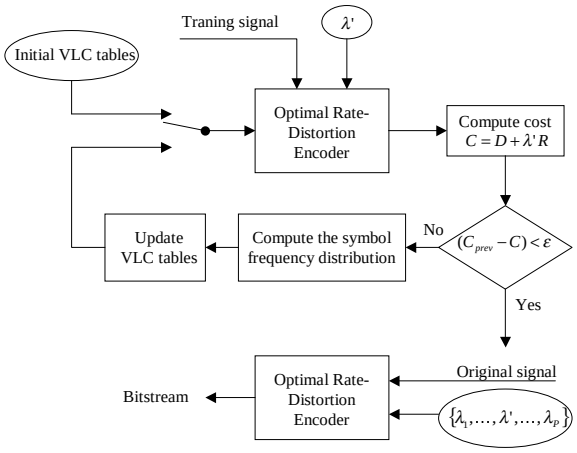

Fig. 3. Optimization of the VLC tables is done on a training signal prior to the coding of the original signal.

4. EXPERIMENTAL RESULTS

The optimality of the operational rate-distortion solution described in the previous section depends on the given frames and VLC tables. The design of frames [12] is not considered in this work. As far as the VLC tables are concerned, we follow the approach in [9] and [10], illustrated in Fig. 3. That is, we optimize the VLC tables on a training signal, $\mathbf{x}_{tr}$, before using them for our test signal, $\mathbf{x}$. The training signal should be of the same class as the test signal. Based on an initial VLC and a specific λ-value, λ', the algorithm iterates on the optimization procedure presented in previous section. For each iteration, the Symbol Frequency Distribution (SFD) is calculated. The VLC tables are updated based on a weighting function between the probability density function for the previous VLC tables and the current SFD. The iteration stops when the total cost reduction is under a specified value, ϵ. After the last iteration, the coding of the original signal starts and a set of λ-values is used, in order to find several points on the ORDF's convex hull.

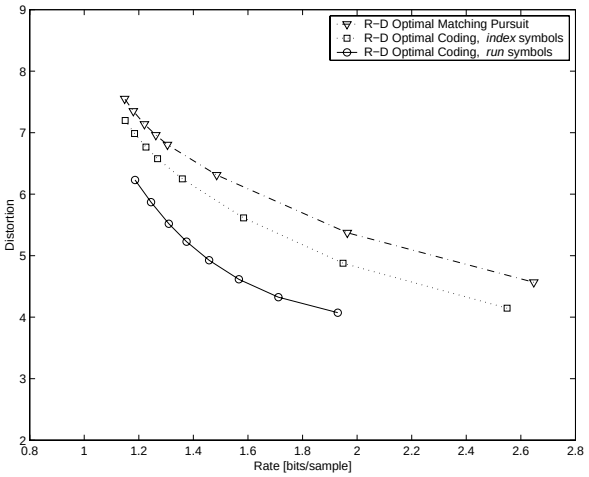

Fig. 4. Frame *Fr0*. Comparisons between RDOMP, the proposed method using *index* symbols and the proposed method using *run* symbols. The points in all curves represent members of the ORDF's convex hull.

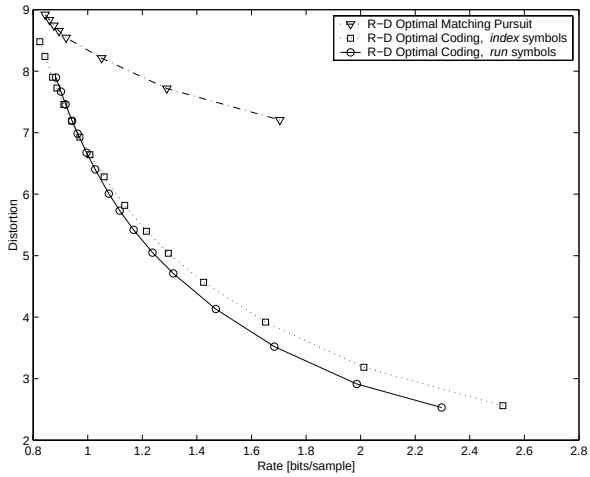

Fig. 5. Frame *Fr2*. Comparisons between RDOMP, the proposed method using *index* symbols and the proposed method using *run* symbols. The points in all curves represent members of the ORDF's convex hull.

We show next some of our results from using the developed algorithm to code a Gaussian AR(1) process with $\rho = 0.95$, $L = 512$, $N = 16$, $K = 32$. Two different frames are used: *Fr0* is built up by putting a 16×16 DCT together with a 16×16 Haar transform. *Fr2* is a designed frame for using Orthogonal Matching Pursuit on an AR(1) signal [12]. The number of different *value* symbols is 30, and the number of *run* symbols is 32, as the length of the coefficient vector. We will compare our method with Rate-Distortion Optimized Matching Pursuit (RDOMP) [1]. The vector selection is not based on the minimization of the distortion, D, but the minimization of $D + \lambda R$. Since the algorithm is based on a *one by one* selection procedure of the coefficients, we cannot use Run-length coding in rate-distortion optimization. Instead, the coefficient *indices* are directly coded. The VLC optimization scheme in Fig. 3 is used in both cases.

The results from the comparison are shown in Figs. 4 and 5. We can see that the proposed method outperforms RDOMP independently of which frame we use. The benefit of using a well designed frame is demonstrated in Fig. 5, where *Fr2*, used together with our new approach, has a good rate-distortion tradeoff. To show the effect of using Run-length coding, we perform an experiment with the proposed method using *index* symbols instead of *run* symbols. The result is plotted as dotted curves in Figs. 4 and 5. The run-length coding provides a benefit, especially when the number of nonzero coefficients is increasing. The new approach is better than RDOMP in all cases, regardless of the use of *run* or *index* symbols.

5. CONCLUSION

An optimal and efficient method for frame based coding is presented in this paper. The optimality is in the ORD sense and the efficiency is achieved by using a QR decomposition which results in a new set of independent decision variables. This involved a considerably reduction in complexity, and made it possible to find the optimal solution in a reasonable amount of time. Experiments show that this method outperforms RD Optimized Matching Pursuit. We also see the benefit of using a well designed frame, and Run-length coding as part of the entropy coder.

6. REFERENCES

[1] M. Gharavi-Alkhansari, "A model for entropy coding in matching pursuit," in *IEEE Proc. ICIP '98*, Chicago, USA, Nov. 1998, pp. 778–782.

[2] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, pp. 227–234, Apr. 1995.

[3] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[4] G. Davis, *Adaptive Nonlinear Approximations*, Ph.D. thesis, New York University, Sept. 1994.

[5] M. Gharavi-Alkhansari and T. S. Huang, "A fast orthogonal matching pursuit algorithm," in *IEEE Proc. ICASSP '98*, Seattle, USA, May 1998, pp. 1389–1392.

[6] T. Berger, *Rate distortion theory: A mathematical basis for data compression*, Prentice Hall, 1971.

[7] G. M. Schuster and A. K. Katsaggelos, *Rate-Distortion Based Video Compression*, Kluwer Academic Publishers, Boston, 1997.

[8] A. Ortega and K. Ramchandran, "Rate-distortion methods for image and video compression," *IEEE Signal Processing Magazine*, pp. 23–50, Nov 1998.

[9] G. Melnikov, G. M. Schuster, and A. K. Katsaggelos, "Shape coding using temporal correlation and joint vlc optimization," *IEEE Trans. Circuits and Systems for Video Technology*, pp. 744–754, Aug 2000.

[10] R. Nygaard, G. Melnikov, and A. K. Katsaggelos, "A rate distortion optimal ecg coding algorithm," *IEEE Trans. Biomedical Engineering*, pp. 28–40, Jan 2000.

[11] H. Anton, *Elementary Linear Algebra*, John Wiley and Sons, Inc., New York, 7th edition, 1994.

[12] K. Engan, *Frame Based Signal Representation and Compression*, Ph.D. thesis, Norwegian University of Science and Technology/ Stavanger University College, Sept. 2000.