

Competitive Learning Algorithms for Robust Vector Quantization

Thomas Hofmann and Joachim M. Buhmann

Abstract—The efficient representation and encoding of signals with limited resources, e.g., finite storage capacity and restricted transmission bandwidth, is a fundamental problem in technical as well as biological information processing systems. Typically, under realistic circumstances, the encoding and communication of messages has to deal with different sources of noise and disturbances. In this paper, we propose a unifying approach to data compression by *robust vector quantization*, which explicitly deals with channel noise, bandwidth limitations, and random elimination of prototypes. The resulting algorithm is able to limit the detrimental effect of noise in a very general communication scenario. In addition, the presented model allows us to derive a novel competitive neural networks algorithm, which covers topology preserving *feature maps*, the so-called *neural-gas algorithm*, and the *maximum entropy soft-max rule* as special cases. Furthermore, *continuation methods* based on these noise models improve the codebook design by reducing the sensitivity to local minima. We show an exemplary application of the novel robust vector quantization algorithm to image compression for a teleconferencing system.

Keywords—Vector quantization, neural networks, competitive learning, deterministic annealing, robust encoding, image compression, teleconferencing

I. INTRODUCTION

Vector quantization [1], [2] is a central topic in data compression and signal processing, which deals with the problem of encoding an information source by means of a finite size codebook. In many cases, the probability density function of the source is not known a priori, but a sample set of training data vectors $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^d : 1 \leq i \leq n\}$ is available. *Adaptive codebook design* deals with the task of finding a codebook $\mathcal{Y} = \{\mathbf{y}_\alpha \in \mathbb{R}^d : 1 \leq \alpha \leq m\}$, with codebook vectors or *prototypes* \mathbf{y}_α , and an encoding mapping $c : \mathbb{R}^d \rightarrow \{1, \dots, m\}$, such that the induced expected distortion with respect to a (given) differentiable distortion measure \mathcal{D} is minimized. This is often achieved by minimizing the *empirical risk* or *average distortion*

$$\mathcal{H}^{\text{Vq}}(c, \mathcal{Y}; \mathcal{X}) = \frac{1}{n} \sum_{i=1}^n \mathcal{D}(\mathbf{x}_i, \mathbf{y}_{c_i}), \quad c_i \equiv c(\mathbf{x}_i) . \quad (1)$$

The described problem setting is of theoretical interest and also has important applications, e.g., in speech and image compression [1], [3], [4], [5]. Since c partitions the signal space, \mathcal{H}^{Vq} is equivalent to a clustering problem with \mathbf{y}_α being the representative for cluster $\mathcal{C}_\alpha = \{\mathbf{x}_i : c_i = \alpha\}$. The most prominent design technique is the LBG algorithm [6],

T. Hofmann is with the Center for Biological and Computational Learning, Massachusetts Institute of Technology, Cambridge, MA. E-mail: hofmann@ai.mit.edu

J.M. Buhmann is with the Institut für Informatik III, Rheinische Friedrich-Wilhelms-Universität, Bonn, Germany. E-mail: jb@cs.uni-bonn.de

which is basically an extension of the generalized Lloyd method [7]. Close relatives from cluster analysis are the ISODATA method [8] and the K -means algorithm [9].

Competitive learning algorithms very similar to on-line adaptive vector quantization techniques have been discussed extensively in the *neural networks* literature [10], [11], [12], [13], where the prototypes \mathcal{Y} correspond to synaptic weights of *neurons*, e.g., the center of their receptive field in the input signal space. This close link between adaptive codebook design and unsupervised learning algorithms has often been ignored, but is today well established [14], [15], [16]. A common trait of these algorithms is a competitive stage which precedes each learning step and decides to what extent a neuron may adapt its weights to a new stimulus \mathbf{x} . The simplest and most prototypical learning rule is *Winner-Take-All* (WTA) learning [17], where adaptation is restricted to the single neuron which best matches \mathbf{x} . WTA learning assures that different neurons are tuned to different parts of the input space.

In the context of neural networks a lot of emphasis has been put on learning algorithms which not only produce sparse representations of the data, but possess some additional features, like: generating topographic organization of neurons [11], [18], avoiding unused ('dead') units [19], accelerating the learning phase [12], avoiding local minima and overfitting [13], adding complexity costs [16], learning in non-stationary environments [20], balancing adaptation frequencies [21], and incorporating learning history [22]. A common means to achieve many generalizations of this kind is by *soft competition*. By weakening the WTA restriction, soft competition enables more than a single neuron to adapt on presentation of a pattern \mathbf{x}_{n+1} . Each neuron has got an individual relative adaptation strength $\rho_\alpha^{(n+1)} \equiv \rho_\alpha(\mathbf{x}_{n+1})$. The general form of the soft competition learning rule considered in this paper can be written as

$$\mathbf{y}_\alpha^{(n+1)} = \mathbf{y}_\alpha^{(n)} - \frac{1}{2} \rho_\alpha^{(n+1)} \frac{\partial}{\partial \mathbf{y}_\alpha^{(n)}} \mathcal{D}(\mathbf{x}_{n+1}, \mathbf{y}_\alpha^{(n)}) . \quad (2)$$

One of our main goals is to show, how three important neural network algorithms, (i) the self-organizing *feature map* (SOFM) [11], [18], (ii) the '*neural-gas*' (NG) algorithm [12], and (iii) the maximum entropy learning rule (ME) [23], [16] can be derived within the framework of *robust encoding* under different sources of *noise* and bandwidth variations or limitations. It has previously been noticed that *source-channel coding* may lead to a topological ordering of prototypes very similar to the SOFM [24], [14], [25], [13]. The major contribution of this work is to demonstrate how generalized versions of the NG algorithm and

ME learning fit into this framework.

For several reasons, it is important to understand how different soft competition schemes are related to the empirical risk minimization framework naturally applied in adaptive vector quantization. First, many competitive learning algorithms have been applied to vector quantization and data compression tasks, e.g., [26]. In our opinion, this is not satisfying without knowledge of the underlying objective function and the corresponding encoding scenario. Second, such knowledge has the advantage not only to guide algorithm design in potential applications, but also to supply a sound information-theoretic foundation. In particular, this insight will foster a more fundamental understanding how different learning schemes can be distinguished and how they can be systematically combined. Third, since artificial neural networks are often used as models for biological neural systems, such an analysis establishes a connection between engineering problems and possible mechanisms developed by biological evolution of neural systems.

The rest of the paper is organized as follows: Section II-V successively develop the proposed robust vector quantization model as well as the corresponding algorithms for adaptive codebook design. The noise-free vector quantization scenario (Section II) is extended to cover channel noise (Section III), random eliminations of prototypes (Section IV), and bandwidth limitations (Section V). Section VI deals with the combination and unification of these three models. In Section VII we address the problem of continuation methods. Finally, Section VIII summarizes results for compression of wavelet transformed video sequences from a teleconferencing application.

II. VECTOR QUANTIZATION FOR NOISELESS COMMUNICATION CHANNELS

A. Codebook Design for Noiseless Channels

We first consider the simplest case of adaptive codebook design for a noiseless channel, i.e. \mathbf{x} is encoded by $\mathbf{y}_{c(\mathbf{x})}$, the index $c(\mathbf{x})$ is sent through the channel, it is received without corruption, and $\mathbf{y}_{c(\mathbf{x})}$ is restored. This communication scenario is captured by the objective function \mathcal{H}^{VQ} in Eq. (1). The LBG algorithm [6] alternates the minimization of \mathcal{H}^{VQ} with respect to the encoding function c and with respect to the codebook \mathcal{Y} , which results in the following set of equations¹,

$$c_i = \arg \min_{\alpha} \mathcal{D}(\mathbf{x}_i, \mathbf{y}_{\alpha}), \quad \sum_{i=1}^n \delta(\alpha, c_i) \frac{\partial \mathcal{D}(\mathbf{x}_i, \mathbf{y}_{\alpha})}{\partial \mathbf{y}_{\alpha}} = 0. \quad (3)$$

δ denotes Kronecker's delta function. From rate distortion theory the first type of equation is known as the *nearest neighbor rule*, while the second is called *centroid condition* [27]. Squared Euclidean distortions² imply the optimal choice of the codebook vectors as the center of mass of the associated data $\mathbf{y}_{\alpha} = \sum_i \delta(\alpha, c_i) \mathbf{x}_i / \sum_i \delta(\alpha, c_i)$.

¹Ties as events of measure zero are broken according to index order.

²We will mainly focus on this choice of \mathcal{D} , however we will continue to use the more general notation \mathcal{D} whenever this is possible.

B. On-line Learning for Noiseless Channels

Minimization of an empirical risk function like \mathcal{H}^{VQ} by LBG or any other *batch learning* algorithm requires that all training data \mathcal{X} are given prior to the optimization process. For many interesting applications, it is more adequate to consider an on-line setting, where code vector estimates are adapted sequentially with the presentation of new data. To obtain on-line learning rules, we evaluate the difference between the centroid conditions for n and $n+1$ data, which is in the Euclidean case given by

$$\mathbf{y}_{\alpha}^{(n+1)} - \mathbf{y}_{\alpha}^{(n)} = \frac{\delta(\alpha, c_{n+1}^{(n+1)})}{n_{\alpha}^{(n+1)}} \mathbf{x}_{n+1} + \sum_{i=1}^n \tilde{w}_{i\alpha} \mathbf{x}_i \quad (4)$$

$$\tilde{w}_{i\alpha} \equiv \frac{\delta(\alpha, c_i^{(n+1)})n_{\alpha}^{(n)} - \delta(\alpha, c_i^{(n)})n_{\alpha}^{(n+1)}}{n_{\alpha}^{(n)}n_{\alpha}^{(n+1)}}. \quad (5)$$

Here $n_{\alpha}^{(n)} \equiv \sum_{i=1}^n \delta(\alpha, c_i^{(n)})$ is the encoding frequency for \mathbf{y}_{α} and the superscripts denote encoding functions and codebook vectors for n and $n+1$ data, respectively. Since we do not want to recalculate the encoding of past data in an on-line setting, we make the approximation $c_i^{(n+1)} \approx c_i^{(n)}$ and after dropping all unnecessary superscripts we arrive at the on-line rule

$$\mathbf{y}_{\alpha}^{(n+1)} = \mathbf{y}_{\alpha}^{(n)} + \rho_{\alpha}^{(n+1)}(\mathbf{x}_{n+1} - \mathbf{y}_{\alpha}^{(n)}), \quad \rho_{\alpha}^{(n)} \equiv \frac{\delta(\alpha, c_n)}{n_{\alpha}^{(n)}}. \quad (6)$$

Eq. (6) corresponds to WTA learning with an individual learning rate for each prototype. The rate is inverse proportional to the number of assigned data vectors, which implies that prototypes representing only a small fraction of the past data keep a higher adaptivity than prototypes with high encoding frequencies. It is straightforward to add a forgetting mechanism with a time constant τ in order to be able to deal with non-stationary sources. In this case $n_{\alpha}^{(n)}/n$ simply becomes a running average.

The above derivation of a learning rule from an on-line optimization principle is theoretically sound and yields asymptotically the correct relative update weights. However, the convergence rate might be too slow, and it is advantageous to include an additional learning gain to accelerate the adaptation. In our simulations we have therefore utilized learning rate schedules, based on the 'First-Search-Then-Converge' heuristic [28], $\rho_{\alpha}^{(n)} \equiv \delta(\alpha, c_n)/(1 + n_{\alpha}^{(n)}/\rho_0)$. By choosing $\rho_0 > 1$, the plasticity of the weights is increased. A similar modification applies to the update rules derived in the sequel.

III. VECTOR QUANTIZATION FOR NOISY COMMUNICATION CHANNELS

A. Source-Channel Coding

An important extension of the vector quantization problem known as *source-channel coding* is to consider a noisy transmission channel in the codebook design phase. We restrict the discussion to the case of discrete memoryless channels with known noise characteristics. Denote by $s_{\nu|\alpha}$

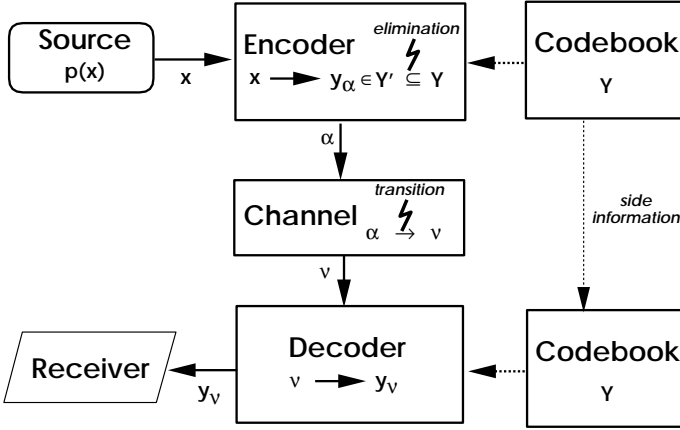


Fig. 1. Vector quantization with channel noise and random eliminations of codebook vectors.

the elements of the *transmission matrix* \mathcal{S} , i.e. the probabilities of receiving index ν after sending α through the channel. The objective functions \mathcal{H}^{VQ} generalizes to

$$\mathcal{H}^{\text{svq}}(c, \mathcal{Y}; \mathcal{X}) = \frac{1}{n} \sum_{i=1}^n \sum_{\nu=1}^m s_{\nu|c_i} \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu). \quad (7)$$

It has been noticed that source-channel coding may lead to a topological ordering of prototypes [24], [25], [14], [13]. If $s_{\nu|\alpha}$ is high, it is advantageous to place \mathbf{y}_α and \mathbf{y}_ν close to each other in order to limit the detrimental effect of a reconstruction from corrupted indices. This is directly expressed by the stationary equations

$$c_i = \arg \min_{\alpha} \mathcal{D}_{i\alpha}^{\mathcal{S}}, \quad \sum_{i=1}^n s_{\nu|c_i} \frac{\partial \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu)}{\partial \mathbf{y}_\nu} = 0, \quad (8)$$

where $\mathcal{D}_{i\alpha}^{\mathcal{S}} \equiv \sum_{\nu=1}^m s_{\nu|\alpha} \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu)$. Setting $s_{\nu|\alpha} \equiv \delta(\nu, \alpha)$ one obtains Eq. (3), as expected. The above mechanism can also be applied, if the goal is not to communicate information via noisy channels, but to impose an a priori specified topology in the index space of prototypes as in the SOFM. In this case the transmission matrix exactly specifies the topology and connection strength between neurons.

B. Generalized SOFM Learning Rule

The relationship to the SOFM can be made more explicit by deriving the on-line learning equations corresponding to \mathcal{H}^{svq} . Proceeding along the same lines as in the noise-free case, we obtain a generalization of Eq. (6) [16] with $\rho_{\alpha}^{(n)} \equiv s_{\alpha|c_n}/n_{\alpha}^{(n)}$, where c_n is now given by Eq. (8) instead of the WTA rule in Eq. (3) and $n_{\alpha}^{(n)} \equiv \sum_{i=1}^n s_{\alpha|c_i}$. The above soft competitive learning rule modifies the original SOFM learning [18] by defining the ‘winner’ as the neuron with minimal *expected distortion*, a property which also depends on the weights of neighboring neurons. In addition, each neuron has an individual learning rate as in the noise-free case.

IV. ROBUST VECTOR QUANTIZATION

A. Prototype Elimination Model

A second fundamental extension of the basic vector quantization model deals with random eliminations of codebook vectors and is called *robust vector quantization* [29]. The encoding/transmission scheme for robust vector quantization, including channel noise is depicted in Fig. 1.

In this communication model the codebook design addresses the problem, that certain prototypes may not be available at encoding time t due to a temporary codebook reduction $\mathcal{Y}^{(t)} \subseteq \mathcal{Y}$. In other words, the alphabet of the utilized code is dynamically reduced by randomly discarding certain indices α without altering the reconstruction codebook \mathcal{Y} .

In encoding applications codebook reductions might be necessitated by rapidly varying bandwidth limits, a problem also known as variable-rate vector quantization [4]. Usually, variable-rate codebooks are designed to cover a large range of possible bandwidths, e.g., by storing hierarchies of codebooks of different size. This strategy is problematic for adaptive vector quantization, since a large set of codebooks has to be adapted, creating severe performance problems especially in an on-line setting. Contrary to this computationally expensive strategy, only one *universal* codebook has to be learned in the proposed robust vector quantization model. The codebook is designed to be most robust with respect to small fluctuations in the availability of codebook vectors. This model is expected to yield satisfactory results as long as the bandwidth variations are sufficiently small.

In the biological context, prototype vectors correspond to neurons representing a specific part of the input signal space, and eliminations of prototypes thus correspond to single neuron defects. To achieve robustness is clearly desirable for biological systems where defective units cannot be immediately replaced. The fault-tolerance property aims at redundancy which contrasts the specialization achieved by WTA learning: specialization is only a valuable property if the system performance does not depend too heavily on single neurons. As a consequence, important parts of the signal space will be covered by more neurons than would be necessary under ideal operating conditions.

B. The Generalized ‘Neural-Gas’ Model

The random elimination induces encoding uncertainty about the available code alphabet, which is fundamentally different from the decoding uncertainty due to a noisy transmission channel. For every data vector we thus introduce a ranking r_i as a bijective mapping on $\{1, \dots, m\}$. $r_i(u)$ denotes the prototype index having rank u , i.e. the u -th choice for encoding \mathbf{x}_i , while the inverse function $\bar{r}_i(\alpha) \equiv r_i^{-1}(\alpha)$ denotes the rank of \mathbf{y}_α . For the data encoding at time t the \mathbf{y}_α with the lowest rank r among the available part $\mathcal{Y}^{(t)} \subseteq \mathcal{Y}$ of the codebook is used to encode \mathbf{x}_i , more formally $c_i \equiv \arg \min_{\{\nu: \mathbf{y}_\nu \in \mathcal{Y}^{(t)}\}} \bar{r}_i(\nu)$.

Consider the case of an independent elimination noise ϵ_α for \mathbf{y}_α , i.e. prototype \mathbf{y}_α has a probability of $(1 - \epsilon_\alpha)$ of

being available at encoding time and the joint probability for sets of prototypes factorizes. In this case the objective function for robust vector quantization is given by³

$$\mathcal{H}^{\text{rvq}}(r, \mathcal{Y}; \mathcal{X}) = \frac{1}{n} \sum_{i=1}^n \sum_{\alpha=1}^m p_{\alpha|r_i} \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha), \quad (9)$$

$$p_{\alpha|r_i} \equiv \frac{1 - \epsilon_\alpha}{1 - \prod_{\nu=1}^m \epsilon_\nu} \prod_{u=1}^{\bar{r}_i(\alpha)-1} \epsilon_{r_i(u)}. \quad (10)$$

$p_{\alpha|r_i}$ can be interpreted as the encoding probability for \mathbf{y}_α given the ranking r_i . In the case of uniform elimination probabilities $\epsilon_\alpha \equiv \epsilon$ the objective function simplifies to the neural-gas model \mathcal{H}^{ng} [12], where $p_{\alpha|r_i} \propto \bar{e}^{\bar{r}_i(\alpha)}$. For \mathcal{H}^{ng} the weight for the contribution of $\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha)$ thus depends only on the rank of \mathbf{y}_α , which makes the neural gas model computationally simpler, but also less flexible to model elimination noise.

The stationary conditions for \mathcal{H}^{rvq} are given by

$$\mathcal{D}(\mathbf{x}_i, \mathbf{y}_{r_i(u)}) \leq \mathcal{D}(\mathbf{x}_i, \mathbf{y}_{r_i(v)}), \quad \text{whenever } u < v, \quad (11)$$

$$\sum_{i=1}^n p_{\alpha|r_i} \frac{\partial \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha)}{\partial \mathbf{y}_\alpha} = 0. \quad (12)$$

Eqs. (11,12) can be used in an iterative update rule which naturally generalizes the LBG or K -means algorithm. Again, it is straightforward to derive the corresponding on-line equations, the adaptation strength being $\rho_\alpha^{(n)} = p_{\alpha|r_n} / \sum_{i=1}^n p_{\alpha|r_i}$.

V. BANDWIDTH LIMITATIONS AND MAXIMUM ENTROPY LEARNING

A. The Distortion-Rate Function

In Section III, we have assumed the availability of a communication channel, possibly noisy, but with a capacity which perfectly matches the required bandwidth. In Section IV, a code vector elimination model was introduced which is of interest for robust encoding under rapidly *varying* bandwidth limitations. Here, we consider the problem of codebook design for a *fixed* rate R . This constraint requires a constructive approximation of the *distortion-rate function* [30], [31]. Let us introduce association probabilities $c_{i\alpha}$, which are probabilistic variants of the encoding function, $c_{i\alpha}$ denoting the probability of using \mathbf{y}_α in reconstructing \mathbf{x}_i . We consider $\{1, \dots, n\}$ to be our discrete source alphabet and $\{1, \dots, m\}$ to be our reconstruction alphabet⁴. The empirical distortion-rate function for a fixed codebook \mathcal{Y} is then given by

$$\mathcal{H}^{\text{drf}}(\mathcal{Y}; \mathcal{X}, R) = \min_{c \in \mathcal{C}(R)} \frac{1}{n} \sum_{i=1}^n \sum_{\alpha=1}^m c_{i\alpha} \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha). \quad (13)$$

³In case that all prototypes are eliminated the process is restarted, which results in an additional normalization by $1 - \prod_{\nu=1}^m \epsilon_\nu$.

⁴The finite source alphabet reflects the empirical approximation of the source density. The discretization of the reconstruction alphabet can also be justified in the case of continuous source alphabets, as shown rigorously in [31].

Besides non-negativity and normalization, $\mathcal{C}(R)$ is restricted to probabilistic encodings for which the mutual information between the optimal codeword prior distribution p and the encoding probabilities c does not to exceed the rate limit R . An efficient way to compute c (and p) for given \mathcal{Y} is known as the Blahut-Arimoto algorithm [32], [33], which is a special case of the alternating minimization of Csiszár and Tuszáný [34]. The stationary equations

$$c_{i\alpha} = \frac{p_\alpha e^{-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha)/\lambda}}{\sum_{\nu=1}^m p_\nu e^{-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu)/\lambda}}, \quad p_\alpha = \frac{1}{n} \sum_{i=1}^n c_{i\alpha}, \quad (14)$$

are solved by re-calculating c and p in alternation until convergence. Here λ is a Lagrange parameter to enforce the rate constraint, and the mapping $\lambda(R) : [0; \infty] \rightarrow [0; R_{\text{max}}]$ is known to be monotonically decreasing [30]. Further minimization of $\mathcal{H}^{\text{drf}}(\mathcal{Y}; \mathcal{X}, R)$ w.r.t. \mathcal{Y} results in the generalized centroid equations, $c_{i\alpha}$ replacing the Kronecker delta in Eq. (3). For $\lambda \rightarrow 0$ the rate R achieves its maximum and the encoding probabilities approach the Boolean values of the nearest neighbor rule. This limit corresponds to the case of a sufficient channel capacity, $R_{\text{max}} = -\sum_{\alpha=1}^m p_\alpha \log p_\alpha$.

B. Maximum Entropy Soft-Competition

Minimization of the distortion-rate function in principle allows us to design a codebook which asymptotically obtains the minimal distortion in encoding \mathcal{X} at a given rate. The design of an optimal code relies on the prior knowledge of the encoding frequencies p_α . If we disregard this knowledge, being maximally ignorant instead by assuming $p_\alpha = 1/K$, the cross-entropy condition simplifies to a constraint on the negative entropy. In addition, the alternating minimization in Eq. (14) reduces to a single step, since constant p_α cancel out. The resulting equations are equivalent to the ones obtained by applying the maximum entropy principle to the empirical risk in Eq. (1), following an optimization technique known as *deterministic annealing* (DA) [23], [35], [16]. In the context of DA, the Lagrange parameter λ is identified with the *computational temperature* T . The ME stationary equations for the Euclidean case are thus simply given by

$$c_{i\alpha} = \frac{e^{-(\mathbf{x}_i - \mathbf{y}_\alpha)^2/T}}{\sum_{\nu=1}^m e^{-(\mathbf{x}_i - \mathbf{y}_\nu)^2/T}}, \quad \mathbf{y}_\alpha = \frac{\sum_{i=1}^n c_{i\alpha} \mathbf{x}_i}{\sum_{i=1}^n c_{i\alpha}}. \quad (15)$$

Similarly the adaptation strength for the on-line equations is given by $\rho_\alpha^{(n)} = c_{n\alpha} / \sum_{i=1}^n c_{i\alpha}$.

VI. UNIFIED ROBUST VECTOR QUANTIZATION MODEL

So far, we have independently introduced three extensions of the basic codebook design problem. It remains to be seen how the different models could be combined.

A. Channel Noise and Prototype Elimination

The combination of prototype elimination with a noisy communication model for a fixed transmission matrix \mathcal{S} is

straightforward. The stationary conditions in Eq. (11,12) are modified in the following way:

$$\mathcal{D}_{r_i(u)}^S \leq \mathcal{D}_{r_i(v)}^S, \quad \sum_{i=1}^n \sum_{\alpha=1}^m s_{\nu|\alpha} p_{\alpha|r_i} \frac{\partial \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu)}{\partial \mathbf{y}_\nu} = 0. \quad (16)$$

Similarly, adaptation weights for on-line optimization are obtained which combine the SOFM with the NG algorithm.

B. Channel Noise and the Maximum Entropy Principle

The generalization of Eq. (15) to noisy channels is achieved without further conceptual difficulties (cf. [16]) by

$$c_{i\alpha} = \frac{e^{-\frac{1}{T} \sum_{\nu} s_{\nu|\alpha} (\mathbf{x}_i - \mathbf{y}_\nu)^2}}{\sum_{\mu} e^{-\frac{1}{T} \sum_{\nu} s_{\nu|\mu} (\mathbf{x}_i - \mathbf{y}_\nu)^2}}, \quad (17)$$

$$\mathbf{y}_\nu = \frac{\sum_i \sum_{\alpha} s_{\nu|\alpha} c_{i\alpha} \mathbf{x}_i}{\sum_i \sum_{\alpha} s_{\nu|\alpha} c_{i\alpha}}. \quad (18)$$

As can be seen, the weights in the above generalized centroid condition combine the probabilistic effects of the entropy and the channel noise by a multiplication of the association probabilities $c_{i\alpha}$ with the transmission matrix.

C. Prototype Elimination and the Maximum Entropy Principle

The maximum entropy version of the prototype elimination model turns out to be the hardest problem. Denote elimination events by $e \subseteq \{1, \dots, m\}$, i.e. e is the available subset of the reconstruction alphabet. Given elimination probabilities ϵ_α , each event e has a probability

$$P_e = \frac{\prod_{\alpha \in e} (1 - \epsilon_\alpha) \prod_{\alpha \notin e} \epsilon_\alpha}{1 - \prod_{\alpha=1}^m \epsilon_\alpha}. \quad (19)$$

For fixed \mathcal{Y} and elimination pattern e we compute the optimal association probabilities $c_{i\alpha}^e$ by setting $c_{i\alpha}^e = 0$ whenever $\alpha \notin e$. Conceptually, the source code is optimized conditioned on the instantaneous availability of codebook vectors. However, we have to find one *universal* codebook for all events. This implies that for given association probabilities $\{c_{i\alpha}^e\}$, \mathcal{Y} has to minimize the expected distortion, which yields centroid equations with weights $\sum_e P_e c_{i\alpha}^e$. The problem in evaluating these weights is the summation over 2^m index sets e and the required availability of the corresponding probabilities $\{c_{i\alpha}^e\}$. Since we are not aware of any efficient way to compute these weights, we focus in the following on approximations.

A systematical way to obtain an approximation in the $T \rightarrow 0$ limit is to consider probability distributions with truncated support, i.e. to restrict the association probabilities $c_{i\alpha}^e$ such that $|\{c_{i\alpha}^e > 0 : \alpha \in e\}| \leq g \forall i$ and $\forall e$. In the maximum entropy case it is sufficient to consider the support set being the g closest prototypes to \mathbf{x}_i since, otherwise, the expected distortion for constant entropy could be decreased by renumbering the prototypes. The following approximation scheme has a complexity growing like $\binom{m}{g}$.

We discuss the $g = 2$ case, generalizations for $g > 2$ are straightforward. Assume r_i is the optimal ranking given by Eq. (11). An approximative solution is then given by

$$\sum_{e \subseteq \{1, \dots, m\}} P_e c_{i\alpha}^e \approx \sum_{\nu=1}^m \frac{w_{\alpha\nu}^i e^{-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha)/\lambda}}{e^{-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha)/\lambda} + e^{-\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu)/\lambda}} \quad (20)$$

where $w_{\alpha\nu}^i \equiv (1 - \epsilon_\alpha)(1 - \epsilon_\nu) \prod_{\mu \in I_{\alpha\nu}^i} \epsilon_\mu$ is the probability that \mathbf{y}_α and \mathbf{y}_ν are the two closest available prototypes, i.e. $I_{\alpha\nu}^i \equiv \{\mu : \bar{r}_i(\mu) \leq \max\{\bar{r}_i(\alpha), \bar{r}_i(\nu)\}\} - \{\alpha, \nu\}$. The approximation error is always bounded by $\lambda(\log m - \log g)$.

As a second limiting case consider $\epsilon_\alpha \rightarrow 0 \forall \alpha$. Higher order products $\prod_{\alpha \in I} \epsilon_\alpha$ vanish at least with an order of $|I|$ relative to $\max_{\alpha \in I} \epsilon_\alpha$. It is thus reasonable to restrict the sum to a subset of elimination events e , i.e. approximating P_e for most events e by $P_e \approx 0$. Obviously, it is again the ranking r_i which guides this systematic approximation. Defining a truncation level h we propose the approximation

$$\sum_e P_e c_{i\alpha}^e \approx C_i \sum_{e \supseteq \{\nu : \bar{r}_i(\nu) > h\}} P_e c_{i\alpha}^e. \quad (21)$$

Effectively, this approximation ignores possible eliminations of prototypes with a rank higher than h relative to \mathbf{x}_i . The factor $C_i \geq 1$ rescales the probabilities appropriately. The number of events in the sum is reduced to 2^h instead of 2^m , and for $h = m$ the approximation becomes exact. For the 2^h elimination events considered, we simply compute the association probabilities $c_{i\alpha}^e$ according to Eq. (14) or Eq. (15). The systematic underestimation of the expected distortions which result from this approximation are roughly bounded by $\max_{I \subseteq \{1, \dots, m\}, |I|=h} (\prod_{\alpha \in I} \epsilon_\alpha) \times \max_{\{(\alpha, \nu)\}} (\mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha) - \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\nu))$.

D. The General Robust Vector Quantization Model

The general robust vector quantization model which unifies the maximum entropy principle with channel noise and random prototype eliminations is governed by the objective function

$$\mathcal{H}^{\text{uni}}(\mathcal{Y}; \mathcal{X}) = \sum_{e \subseteq \{1, \dots, m\}} \mathcal{F}_e, \quad (22)$$

$$\mathcal{F}_e = \sum_{i=1}^n \min_{q \in \mathcal{P}} \left[\sum_{\alpha=1}^m \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha) \sum_{\nu \in e} s_{\alpha|\nu} q_\nu + T \sum_{\nu \in e} q_\nu \log q_\nu \right]. \quad (23)$$

\mathcal{F}_e has an interpretation in terms of statistical physics, where it is known as the *free energy* [35], [16]. Hence \mathcal{H}^{uni} is an average of free energies \mathcal{F}_e over elimination events e .

The $T \rightarrow 0$ truncated ME approximation is equivalent to restricting the minimization inside of \mathcal{F}_e to association probabilities with limited support, while the second approximation restricts the sum over events e . This demonstrates that both approximations minimize a cost function and thus yield convergent update schemes. If we denote by $\{c_{i\nu}^e\}$ the optimal choice of $\{q_\nu\}$ in the definition of \mathcal{F} for fixed index i and event e , the stationary equations corre-

TABLE I

PERFORMANCE (MEASURED AS MSE) OF DESIGN ALGORITHMS FOR BATCH LEARNING ON THE FIRST 4 FRAMES OF ‘MISS AMERICA’ (174 × 144 PIXELS) BASED ON 21 RUNS WITH RANDOMIZED INITIALIZATIONS. ALL RESULTS ARE FOR THE DETAIL- X SUBBAND IMAGES WITH

$$\epsilon_\alpha = \epsilon = 0.0, 0.02, 0.1.$$

Method	Subband	N	d	K	MSE, $\epsilon = 0.0$	MSE, $\epsilon = 0.02$	MSE, $\epsilon = 0.1$
LBG	Level 1, x	1584	16	16	76.8 ± 4.9	93 ± 14	191 ± 48
(\mathcal{H}^{vq})	Level 2, x	396	16	32	469.2 ± 23.1	648 ± 107	1315 ± 322
	Level 3, x	396	4	32	313.7 ± 18.5	681 ± 102	2087 ± 789
Deterministic	Level 1, x	1584	16	16	75.4 ± 2.3	88 ± 12	140 ± 22
Annealing	Level 2, x	396	16	32	448.2 ± 15.7	633 ± 88	1277 ± 276
(\mathcal{H}^{vq})	Level 3, x	396	4	32	290.8 ± 11.0	434 ± 51	786 ± 103
Robust K-means	Level 1, x	1584	16	16	76.2 ± 2.9	93 ± 10	146 ± 30
(\mathcal{H}^{ng})	Level 2, x	396	16	32	466.0 ± 18.8	626 ± 19	1116 ± 306
	Level 3, x	396	4	32	305.4 ± 16.7	450 ± 42	1085 ± 154
Truncated DA	Level 1, x	1584	16	16	74.9 ± 2.5	79 ± 4	89 ± 5
for Robust VQ	Level 2, x	396	16	32	431.7 ± 12.2	508 ± 14	776 ± 182
(\mathcal{H}^{ng})	Level 3, x	396	4	32	286.3 ± 10.1	419 ± 35	642 ± 46

sponding to the unified model \mathcal{H}^{uni} are given by

$$\sum_{i=1}^n \left(\sum_{\nu=1}^m s_{\alpha|\nu} \sum_{e \subseteq \{1, \dots, m\}} P_e c_{i\nu}^e \right) \frac{\partial}{\partial \mathbf{y}_\alpha} \mathcal{D}(\mathbf{x}_i, \mathbf{y}_\alpha) = 0. \quad (24)$$

Eq. (24) naturally combines the encoding uncertainty and the transmission uncertainty. The channel noise is incorporated via transmission matrix coefficients $s_{\alpha|\nu}$, the prototype elimination probabilities via P_e , and the effect of the maximum entropy ‘smoothing’ is represented by the association probabilities $c_{i\nu}^e$. The competitive learning rule for the unified robust vector quantization model thus takes the form

$$\rho_\alpha^n = \frac{1}{n_\alpha^{(n)}} \sum_{\nu=1}^m s_{\alpha|\nu} \sum_{e \subseteq \{1, \dots, m\}} P_e c_{i\nu}^e, \quad (25)$$

where $n_\alpha^{(n)} \equiv \sum_i \sum_\nu s_{\alpha|\nu} \sum_e P_e c_{i\nu}^e$.

The following is a pseudo-code implementation for a batch version of the unified robust vector quantization algorithm with $g = 2$ truncation. The outer loop performs *annealing* on the temperature T (cf. Section VII):

```

INITIALIZE all  $\mathbf{y}_\alpha$  randomly, temperature  $T \leftarrow T_{\text{start}}$ ;
WHILE  $T > T_{\text{final}}$ 
  REPEAT
    FOR  $i = 1, \dots, n$ 
      FOR  $\alpha = 1, \dots, m$ ; FOR  $\nu = 1, \dots, m$ ;  $\nu \neq \alpha$ 
        calculate weights  $w_{\alpha\nu}^i$ 
        calculate ME factors according to Eq. (20)
      FOR  $\alpha = 1, \dots, m$ 
        sum over index  $\nu$  to approximate  $\sum_e P_e c_{i\nu}^e$ 
      FOR  $\alpha = 1, \dots, m$ 
        update  $\mathbf{y}_\alpha$  according to the centroid rule Eq. (24)
  UNTIL converged
  add small amplitude Gaussian random noise to all  $\mathbf{y}_\alpha$ 
   $T \leftarrow \eta \cdot T$ ,  $0 < \eta < 1$ 
END

```

TABLE II

BLOCK/CODEBOOK SIZES, MAXIMUM BITS PER PIXEL (MBPP), AND AVERAGE BIT RATES (‘MISS AMERICA’) FOR THE ON-LINE EXPERIMENTS.

	Level 1			Level 2			Level 3		
	dtl x	dtl y	dtl xy	dtl x	dtl y	dtl xy	dtl x	dtl y	dtl xy
Blocks	16	16	-	4	4	16	4	4	4
Codebook	64	64	0	128	128	256	256	256	128
mbpp	6/16	6/16	0	7/4	7/4	4/16	8/4	8/4	7/4
⊗ bpp	4.6/16	3.7/16	0	3.7/4	3.4/4	3.9/16	4.6/4	4.4/4	2.5/4

VII. CONTINUATION METHODS AND DETERMINISTIC ANNEALING

There exists another important utilization of robust vector quantization as part of a *continuation method* [36]. The key idea in continuation methods is to optimize an objective function \mathcal{H} by tracking solutions of a family of objective functions \mathcal{H}_a in the limit of $a \rightarrow 0$, where $\mathcal{H}_0 = \mathcal{H}$. These procedures primarily try to avoid unfavorable local minima by smoothing \mathcal{H} with increasing a .⁵ In fact, all of the discussed competitive learning rules have been utilized in this context: (i) In [18] it was proposed to let the neighborhood strength shrink in a combined schedule with the learning rate. More recently, this idea has been discussed under the title of *noisy channel relaxation* in the context of source-channel coding [38]. A conceptual problem of these techniques in the noise-free case is the need to specify a complete noise model including a topology on the index space $\{1, \dots, m\}$. (ii) In the original NG algorithm [12] the noise level was reduced according to the exponential law $\epsilon^{(n+1)} = r\epsilon^{(n)}$, $0 < r < 1$, since the authors wanted to accelerate learning and did not search for robust data representations. (iii) A continuation method based on the

⁵Typically, \mathcal{H}_a is convex for large enough $a \geq a_0$ and the global minimum of \mathcal{H}_a can be found easily. Continuation methods with this features are generalizations of a method known as *graduated non-convexity* (GNC) in computer vision [37].

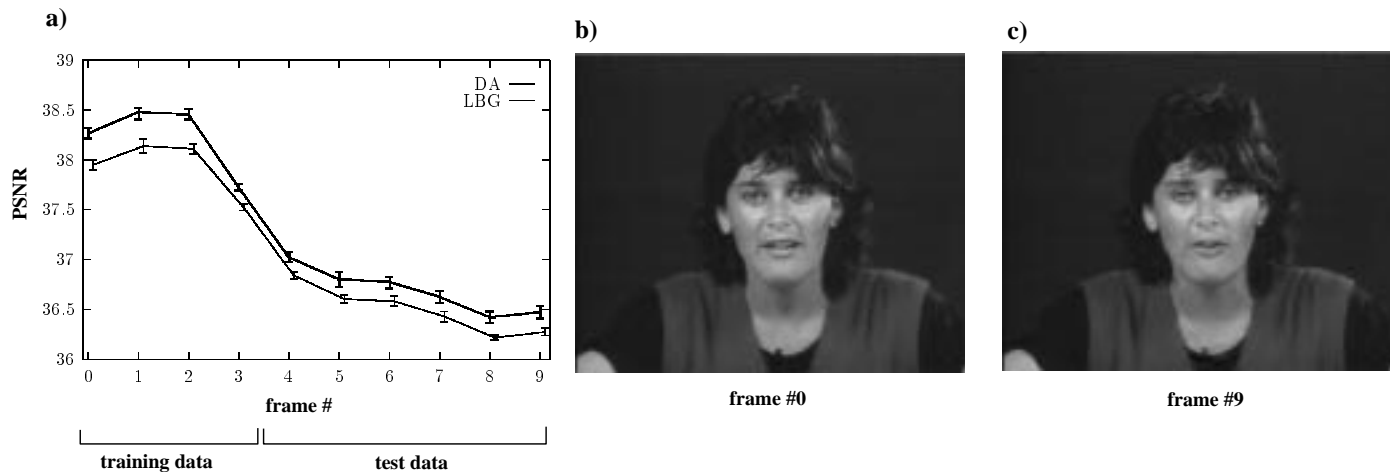


Fig. 2. Performance for codebook design based on the first four frames of the ‘Miss America’ sequence (354×288 pixels). The tests have been performed on the first ten frames. (a) PSNR curve for LBG and DA for the noise-free case (average over 10 runs). (b) and (c) reconstructed frame from the training and test set, quantized with the DA multi-resolution codebook.

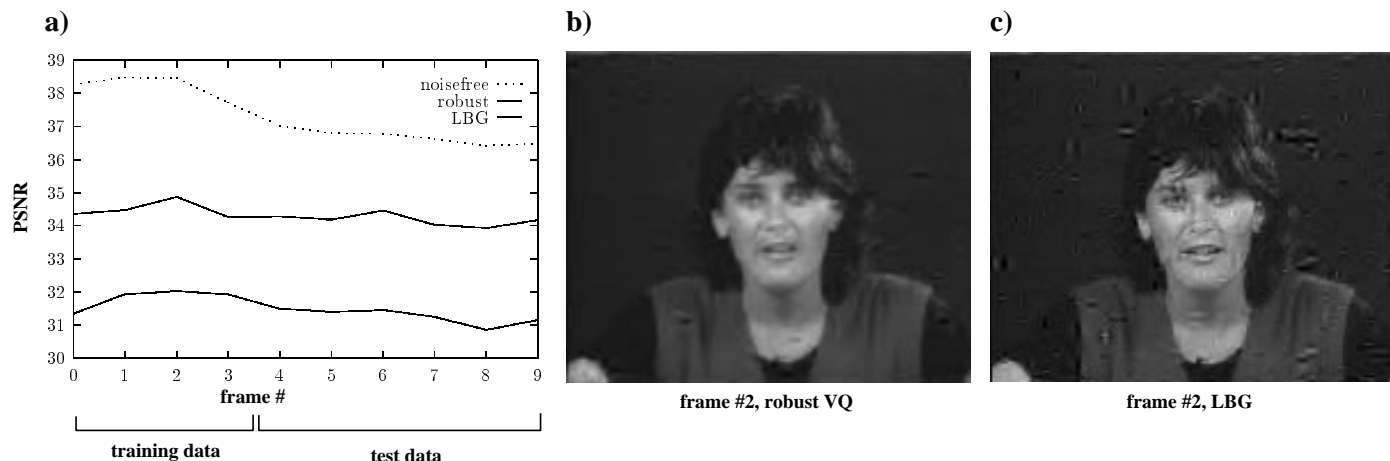


Fig. 3. Performance of robust vector quantization on the ‘Miss America’ sequence (354×288 pixels). (a) PSNR in dB for robust (middle) and LBG (lower) batch codebook design (upper PSNR curve for noise-free DA added as a reference). (b) and (c) example frame encoded with a robust and LBG codebook, respectively.

maximum entropy principle is known as *deterministic annealing* (DA) [39], [23], which is a deterministic variant of *simulated annealing* (SA) [40]. With SA it shares the intuitive motivation from statistical physics to introduce a *temperature* scale T and to track solutions from high to low temperatures ($T \rightarrow 0$ limit). In general, DA incorporates stochastic smoothing by optimizing over a probabilistic solution space, which may result in a significant speed-up compared to SA since no Monte Carlo sampling is required. DA is a very general heuristic optimization technique which has been applied to many different combinatorial optimization problems, including the problem of adaptive codebook design [23], [35], [16].

Continuation methods like DA are optimization heuristics which have empirically shown to result in a satisfactory tradeoff between solution quality and algorithmic efficiency. However, due to the lack of theoretical insight about the geometry of solution curves, performance guarantees can not be given in the general case. In our opinion, DA methods for combinatorial and mixed combinatorial optimization

problems are preferable due to the canonical way how entropy-based ‘noise’ control is implemented. However, one has to be aware of the fact, that the superiority of one continuation method over an alternative technique has not been proven so far and might not even be universally true, but could depend on the specific problem and data.

VIII. RESULTS

We have tested the batch and on-line version of the presented vector quantization algorithm on wavelet-transformed video sequences from a teleconferencing application.⁶ Since severe bandwidth limitations as well as noisy transmission channels are a typical problem especially for wireless teleconferencing, we consider this experimental setup as a realistic scenario for robust vector quantization.

The wavelet transformation and the grouping scheme of

⁶All compression experiments have been performed on single images. The bit rates are state-of-the-art for still compression but can be substantially improved with elaborate motion compensation [41].

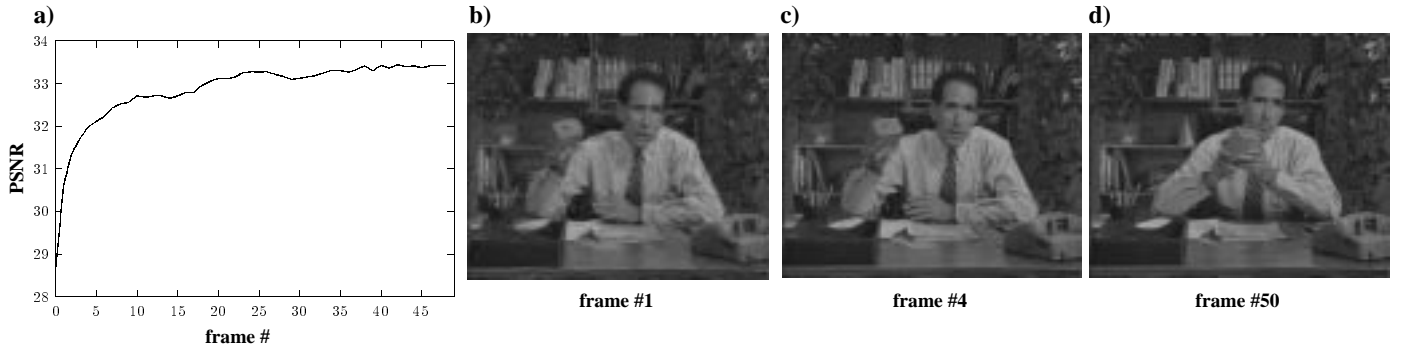


Fig. 4. Performance of on-line WTA learning on the first 50 frames of the ‘Salesman’ sequence: (a) PSNR curve, (b) - (d) reconstructed example frames.

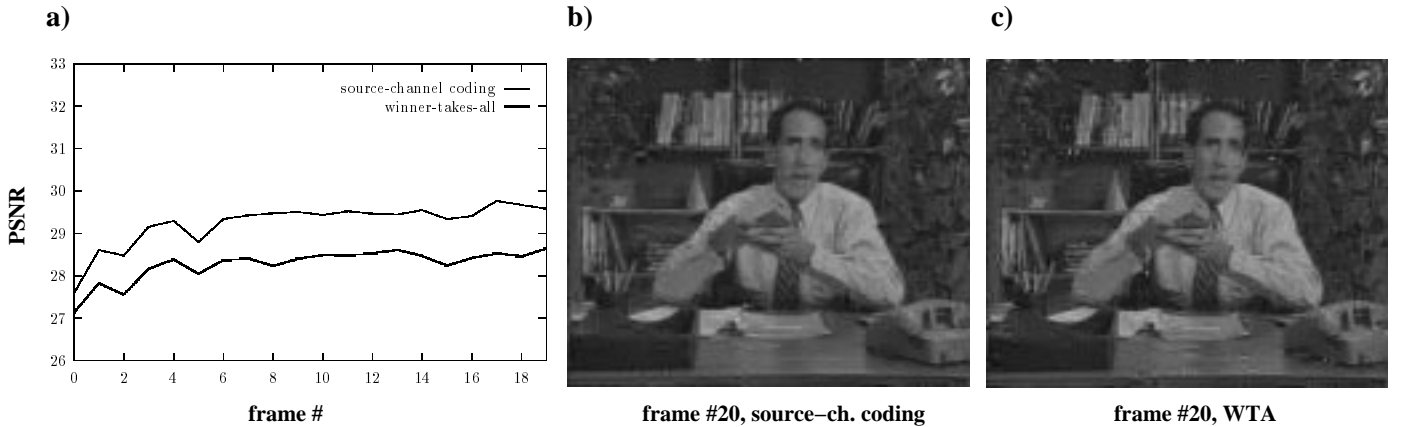


Fig. 5. (a) PSNR for on-line adaptation according to the source-channel cost function \mathcal{H}^{SVQ} and for the WTA algorithm. (b) example frame for codebook design by source-channel coding, (c) example frames for WTA. The channel noise was a bit flip probability of $p = 0.01$.

wavelet coefficients into blocks is performed according to Antonini et al. [3]. All image sequences have been preprocessed by a three-level biorthogonal wavelet transformation [42]. Data vectors were generated by grouping neighboring wavelet coefficients in blocks of size 4×4 and 2×2 for each subband separately. This results in a multi-resolution codebook with independent codebooks for all detail signals.⁷ Table I shows the result of a series of repeated experiments on the ‘Miss America’ sequence with uniform elimination noise. Two facts are remarkable: (i) DA significantly improves the codebook optimization for all subbands and all elimination probabilities ϵ . Furthermore, the variance over different runs is reduced, compared to the corresponding $T = 0$ algorithm. (ii) The robust codebook design clearly reduces the sensitivity with respect to code vector eliminations. The overall best result was obtained with the annealed robust vector quantization algorithm. For $\epsilon = 0$ the annealing was performed by decreasing T and ϵ in a joint schedule.

The results of further experiments with block and codebooks sizes as given by Table II are depicted in Fig. 2. The first experiment for the noise-free case demonstrates, that codebooks designed by DA not only yield superior results on the training data, but also on new test data, as

⁷The residuum on the third level was predictively coded in combination with a scalar quantization.

compared to the LBG algorithm. However, all codebooks obey a significant difference of about 2 dB in PSNR between the training and the test error, which is due to data overfitting. The second experiment in Fig. 3 on the same data demonstrates the advantages of a robust codebook design. The elimination probability was uniformly set to 30%, $\epsilon_\alpha = \epsilon = 0.3$. The channel noise was independent bit-noise with a 1% error rate. Notice that for a codebook of size $K = 256$ this results in a transmission error probability of about 8%. The PSNR curves for typical codebooks show an improvement of approximately 3 dB on both, training and test data. This means that 50% of the PSNR loss due to noise are compensated by the robust design procedure. The improvement is clearly visible in the depicted example frames. The images encoded with robust codebooks are still of good quality. In contrast, the LBG design yields very noisy reconstructions with distortions which are typically observed for wavelet encoded images.

For the on-line competitive learning algorithms, we have taken the first 50 frames of the standard sequence ‘Salesman’ (354×288 pixels). The ‘Salesman’ sequence is known as a difficult sequence in teleconferencing, since the background is highly structured and thus possesses a significant high frequency contents. The acceleration gain ρ_0 was set uniformly to $\rho_0 = 30.0$ in all experiments. Fig. 4 shows a typical on-line learning run over 50 frames with

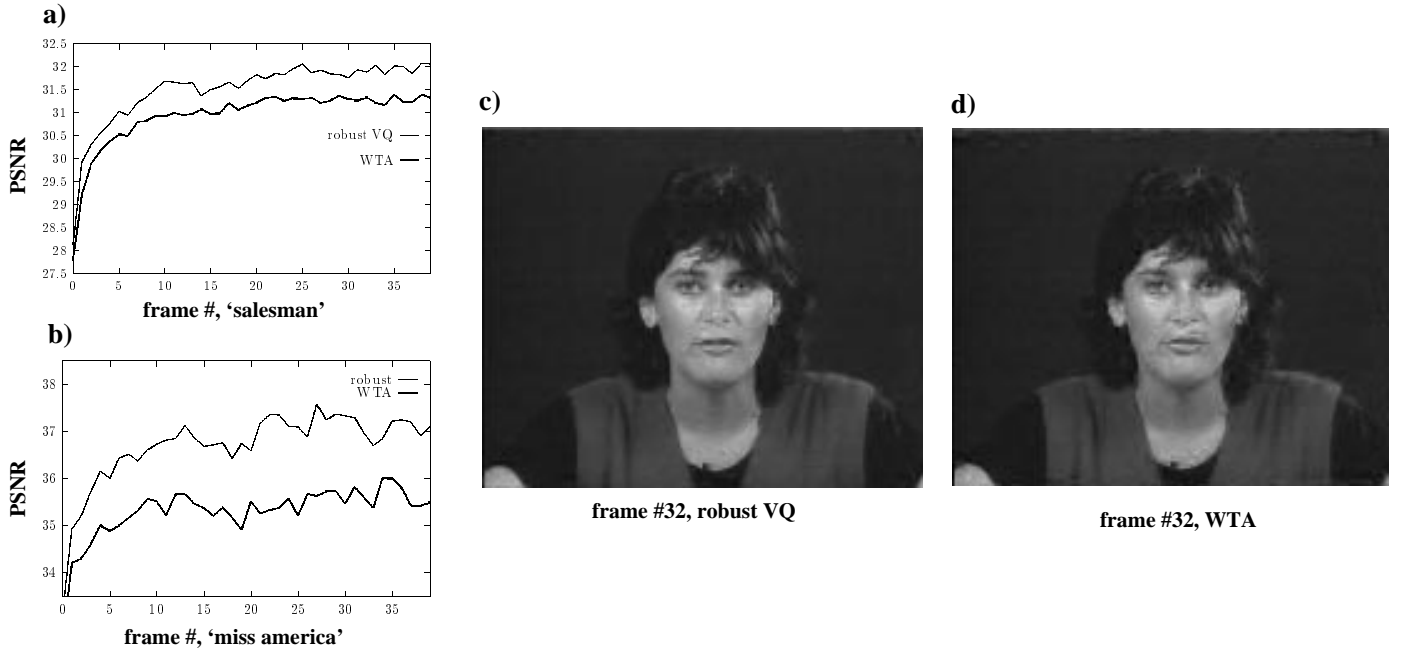


Fig. 6. PSNR for on-line robust learning and the Winner-Takes-All (WTA) algorithm: (a) PSNR on the ‘Salesman’ sequence with $\epsilon = 0.3$, (b) PSNR for ‘Miss America’ with $\epsilon = 0.5$, (c) example of a reconstructed frame for both methods

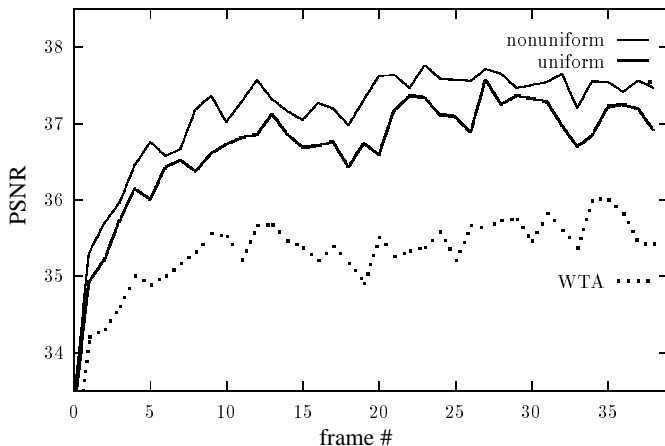


Fig. 7. Comparison in PSNR between robust on-line learning with a uniform elimination probability $\epsilon_\alpha = \epsilon = 0.5$ and with structured elimination model, with $q = 0.685$, $\epsilon \in \{q, q^2, q^3, 0\}$. The upper two curves correspond to robust on-line learning, while the lower gives the WTA result as a reference.

a $T = 0$ Winner-Take-All update rule. The increasing PSNR curve demonstrates a fast adaptation to the statistics of the source. After 20 – 25 frames the convergent phase is reached, the final quality being 33.8 dB in PSNR.

We have furthermore investigated the effect of bit flip channel noise on the distortion induced by different on-line learning rules. We have documented an example run for 1% bitnoise in Fig. 5. The improvement by source-channel coding is around 1 dB in PSNR. The PSNR improvement is clearly visible in the reconstructed frames, since the source-channel optimization significantly suppresses the effect of noise. To study the robustness properties against elimination noise in an on-line setting, we have performed a series

of experiments on the ‘Salesman’ and the ‘Miss America’ sequence (see Fig. 6). For the ‘Salesman’ sequence the improvement by robust on-line learning is about 0.8 dB at an elimination probability of $\epsilon = 0.3$. On the ‘Miss America’ sequence the gain is approximately 2 dB for $\epsilon = 0.5$.

Finally, we have investigated the full robust model with elimination probabilities which depend on the prototype index α according to the following scheme: The elimination probabilities for half of the codebook vectors are high, $\epsilon_\alpha = q$, and are recursively reduced, until a ‘core codebook’ is obtained, which has zero elimination probability. As opposed to a uniform elimination probability, this will partially break the permutation symmetry of codewords. Prototypes with a very low elimination probability will be deployed differently from those possessing a high probability of not being available at encoding time. Our experiments clearly indicate that the codebook design can take advantage of this design knowledge if compared to a uniform elimination model with the same average elimination probability. Furthermore, this setup is realistic for fast on-line rate control, since it is not recommendable to eliminate codebook vectors completely at random. The PSNR curves in Fig. 7 show, that the performance degradation of the Winner-Take-All rule is similar for uniform and non-uniform elimination probabilities, while the robust vector quantizer can take advantage of the structured elimination model to further increase its performance.

CONCLUSION

We have presented a robust vector quantization model with three distinctive types of limitations on the data encoding and transmission: (i) noisy transmission channel, (ii) limited availability of codebook vectors, and (iii) rate

constraints. As we have shown, each of these limitations is related to prominent competitive learning algorithms: (i) the self-organizing feature map, (ii) a generalized version of the ‘neural-gas’ algorithm, and (iii) the maximum entropy learning rule. The established correspondences result in an information–theoretic characterization of these competitive learning schemes and offer a unified framework for their systematic combination. The unified model is suitable to handle the different limitations simultaneously and is, therefore, applicable to robust codebook design in many realistic encoding scenarios. On–line adaptivity can be achieved with the discussed neural net update rules. Moreover, we have stressed the fact that artificially introducing ‘noise’ and reducing the noise level in the course of optimization can significantly improve the codebook design. All our comparisons with standard techniques show that the increase in computational complexity is by far outweighed by the reduction of reconstruction loss.

Acknowledgment: It is a pleasure to thank H. Klock and A. Polzer for stimulating discussions and significant help with the teleconferencing experiments. This work was supported by the German Federal Ministry of Education and Science BMBF and a MIT Postdoctoral Fellowship.

REFERENCES

- [1] R. M. Gray, “Vector quantization,” *IEEE Acoustics, Speech and Signal Processing Magazine*, pp. 4–29, April 1984.
- [2] A. Gersho and R. M. Gray, *Vector Quantization and Signal Processing*, Kluwer Academic Publisher, Boston, 1992.
- [3] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, “Image coding using wavelet transform,” *IEEE Trans Image Processing*, vol. 1, no. 2, pp. 205–220, 1992.
- [4] T. Lookabaugh, E.A. Riskin, P.A. Chou, and R.M. Gray, “Variable rate vector quantization for speech, image, and video compression,” *IEEE Trans Communications*, vol. 41, no. 1, pp. 186–199, 1993.
- [5] W.P. Li and Y.P. Zhang, “Vector-based signal processing and quantization for image and video compression,” *Proceedings of the IEEE*, vol. 83, no. 2, pp. 317–335, 1995.
- [6] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE Trans Communications*, vol. 28, pp. 84–95, 1980.
- [7] S.P. Lloyd, “Least squares quantization in PCM,” *IEEE Trans Information Theory*, vol. 28, pp. 129–137, 1982, reprint of 1957 paper.
- [8] G.B. Ball and D.J. Hall, “A clustering technique for summarizing multivariate data,” *Behavioral Science*, vol. 12, pp. 153–155, 1967.
- [9] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.
- [10] S. Grossberg, “On learning and energy-entropy dependence in recurrent and nonrecurrent signed networks,” *J. Statist. Phys.*, vol. 1, pp. 319–350, 1969.
- [11] T. Kohonen, “Self-organizing formation of topologically correct feature maps,” *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.
- [12] T. Martinez, S.G. Berkovich, and K.J. Schulten, “Neural-gas network for vector quantization and its application to time-series prediction,” *IEEE Trans Neural Networks*, vol. 4, no. 4, pp. 558–569, 1993.
- [13] J. M. Buhmann and H. Kühnel, “Complexity optimized data clustering by competitive neural networks,” *Neural Computation*, vol. 5, pp. 75–88, 1993.
- [14] S.P. Luttrell, “Hierarchical vector quantization,” *IEE Proceedings*, vol. 136, pp. 405–413, 1989.
- [15] T. Kohonen, “The self-organizing map,” *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, Sept. 1990.
- [16] J. M. Buhmann and H. Kühnel, “Vector quantization with complexity costs,” *IEEE Trans Information Theory*, vol. 39, no. 4, pp. 1133–1145, July 1993.
- [17] J. Hertz, A. Krogh, and R. G. Palmer, *Introduction to the Theory of Neural Computation*, Addison Wesley, New York, 1991.
- [18] T. Kohonen, *Self-organization and Associative Memory*, Springer, Berlin, 1984.
- [19] D.E. Rumelhart and D. Zipser, “Feature discovery by competitive learning,” *Cognitive Science*, vol. 9, pp. 75–112, 1985.
- [20] G.A. Carpenter and S. Grossberg, “A massively parallel architecture for a self-organizing neural pattern recognition machine,” *Computer, Vision, Graphics, and Image Processing*, vol. 37, pp. 54–115, 1987.
- [21] S.C. Ahalt, P.K. Chen, and D.E. Melton, “Competitive learning algorithms for vector quantization,” *Neural Networks*, vol. 3, no. 3, pp. 277–290, 1990.
- [22] B. Kosko, “Unsupervised learning in noise,” *IEEE Trans Neural Networks*, vol. 1, pp. 44–57, Mar. 1990.
- [23] K. Rose, E. Gurewitz, and G. Fox, “Statistical mechanics and phase transitions in clustering,” *Physical Review Letters*, vol. 65, no. 8, pp. 945–948, 1990.
- [24] H. Kumazawa, M. Kasahara, and T. Namekawa, “A construction of vector quantizers for noisy channels,” *Electronics and Engineering in Japan*, vol. 67B, no. 4, pp. 39–47, 1984.
- [25] M. Farvardin, “A study of vector quantization for noisy channels,” *IEEE Trans Information Theory*, vol. 36, pp. 799–809, 1990.
- [26] A.K. Krishnamurthy, S. C. Ahalt, D.E. Melton, and P. Chen, “Neural networks for vector quantization of speech and images,” *IEEE Journal on Selected Areas in Communications*, vol. 8, no. 8, pp. 1449–1457, 1990.
- [27] A. Gersho, “On the structure of vector quantizers,” *IEEE Trans Information Theory*, vol. 28, no. 2, pp. 157–166, 1982.
- [28] Ch. Darken and J. Moody, “Note on learning rate schedules for stochastic optimization,” in *Advances in Neural Information Processing Systems*, 1991, vol. 3.
- [29] J. M. Buhmann and T. Hofmann, “Robust vector quantization by competitive learning,” in *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Munich*, 1997.
- [30] T. Berger, *Rate Distortion Theory*, Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1971.
- [31] K. Rose, “A mapping approach to rate-distortion computation and analysis,” *IEEE Trans Information Theory*, vol. 40, no. 6, pp. 1939–1952, 1994.
- [32] R.E. Blahut, “Computation of channel capacity and rate distortion function,” *IEEE Trans Information Theory*, vol. 18, pp. 460–473, 1972.
- [33] S. Arimoto, “An algorithm for calculating the capacity of an arbitrary discrete memoryless channel,” *IEEE Trans Information Theory*, vol. 18, pp. 14–20, 1972.
- [34] I. Csiszár and G. Tusnády, “Information geometry and alternating minimization procedures,” *Statistics and Decisions, Supplement Issue 1*, pp. 205–237, 1984.
- [35] K. Rose, E. Gurewitz, and G. Fox, “Vector quantization by deterministic annealing,” *IEEE Trans Information Theory*, vol. 38, no. 4, pp. 1249–1257, 1992.
- [36] E.L. Allgower and K. Georg, *Numerical Continuation Methods. An Introduction*, vol. 13 of *Springer Series in Computational Mathematics*, Springer-Verlag, Berlin Heidelberg, 1990.
- [37] A. Blake and A. Zisserman, *Visual Reconstruction*, MIT Press, 1987.
- [38] S. Gadkari and K. Rose, “Noisy channel relaxation for VQ design,” in *Proceedings International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996, pp. 2048–2051.
- [39] J. Hopfield and D. Tank, “Neural computation of decisions in optimisation problems,” *Biological Cybernetics*, vol. 52, pp. 141–152, 1985.
- [40] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi, “Optimization by simulated annealing,” *Science*, vol. 220, no. 4598, pp. 671–680, 1983.
- [41] H. Klock, A. Polzer, and J. M. Buhmann, “Region-based motion compensated 3d-wavelet transform coding of video,” in *Proceedings of the International Conference on Image Processing, Santa Barbara*, 1997.
- [42] S. Mallat, “A theory for multidimensional signal decomposition: the wavelet representation,” *IEEE Trans Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.