Vector Quantization with Complexity Costs

Joachim Buhmann[*] Hans Kühnel[†]

[*]Lawrence Livermore National Laboratory,
Computational Physics Division, P.O.Box 808, L-270,
Livermore, CA 94550
jb@mozart.llnl.gov

[†] Physik Department, T30, Technische Universität München,
Boltzmann Straße, D–8046 Garching, Fed. Rep. Germany
kuehnel@physik.tu-muenchen.de

published in
IEEE Transactions on Information Theory 39:1133-1145 (1993)

Abstract

Vector quantization is a data compression method where a set of data points is encoded by a reduced set of reference vectors, the codebook. We discuss a vector quantization strategy which *jointly optimizes* distortion errors and the codebook complexity, thereby, determining the size of the codebook. A maximum entropy estimation of the cost function yields an optimal number of reference vectors, their positions and their assignment probabilities. The dependence of the codebook density on the data density for different complexity functions is investigated in the limit of asymptotic quantization levels. How different complexity measures influence the efficiency of vector quantizers is studied for the task of image compression, i.e., we quantize the wavelet coefficients of gray level images and measure the reconstruction error. Our approach establishes a unifying framework for different quantization methods like K-means clustering and its fuzzy version, entropy constrained vector quantization or topological feature maps and competitive neural networks.

Keywords: Vector quantization, complexity costs, maximum entropy estimation, image compression, neural networks

[*]Supported by the German Federal Ministry of Science and Technology (ITR-8800-H1) and by the Air Force Office of Scientific Research (88-0274) (C. von der Malsburg, PI) while working at the Center for Neural Engineering, University of Southern California, Los Angeles, Ca 90089. Present address: Rheinische Friedrich–Wilhelms–Universität, Institut für Informatik II, Römerstraße 164, D-5300 Bonn 1, Fed. Rep. Germany.

[†]Supported by a graduate fellowship of the Technische Universität München.

1 Introduction

Information reduction represents an essential step in many areas of information processing as for example in image and speech processing and pattern recognition in general. It is particularly important in pattern classification problems to uncover the structure of a given data set, to generalize over inessential details or to remove noise. However, in the absence of a parametric model for a data set, data compression techniques are required that preserve the original data as complete as possible. Vector quantization (VQ), a widely used method of lossy information compression, encodes a given set of d-dimensional data vectors $\{\mathbf{x}_i | \mathbf{x}_i \in \Re^d;\ i = 1, \ldots, N\}$ with a much smaller set of *codebook* vectors $\mathbf{Y} = \{\mathbf{y}_\alpha | \mathbf{y}_\alpha \in \Re^d;\ \alpha = 1, \ldots, K\}$ $(K \ll N)$. What strategy should be followed to design and optimize a vector quantization codebook? In this paper we will primarily discuss design strategies to generate application adequate codebooks.[1]

The design of an optimal vector quantizer requires (i) to find a set of reference vectors that represent the data points with minimal residual distortion and (ii) to specify the complexity of the codebook. Both design specifications are mutually dependent and cannot be optimized separately if we aim for an optimally efficient vector quantizer. This paper discusses the strategy to *jointly optimize a distortion measure and a codebook complexity function* for the design of a vector quantizer. We focus on a Lagrangian variation principle to find optimal codebooks for a large class of distortion and complexity measures. The data assignments to reference vectors are estimated in the maximum entropy sense, a strategy first proposed by Rose et al. [38, 39] for K-means clustering. The special case of jointly minimizing suitable distortion costs with constrained entropy of the codebook vectors has been discussed by Chou et al. [9].

The traditional VQ design strategy, as followed in the well-known K–means clustering algorithm [35] or the LBG algorithm by Linde et al. [32], exclusively relies on the distortion errors for placing the reference vectors. The size of the codebook is fixed (K-means) or it is incrementally increased until the distortion error drops below a predefined threshold (LBG). Following the K-means clustering concept lossy data compression with K reference vectors can be mathematically formulated as an optimization of binary assignment variables $M_{i\alpha} \in \{0, 1\}$ with the objective to minimize the sum of all distortion errors introduced by vector quantizing the data. The variables $M_{i\alpha}$ specify to which reference vector a data point i is assigned to, i.e., $M_{i\alpha^*} = 1$ denotes that data point i is uniquely assigned to reference vector $\mathbf{y}_{\alpha^*}$. The uniqueness of assignments implies the constraint $\sum_\alpha M_{i\alpha} = 1$. The K-means cost function $\mathcal{E}^{\mathrm{Km}}$ for vector quantization is given by

$$\mathcal{E}^{\mathrm{Km}}(\{M_{i\alpha}\}) = \sum_{i=1}^{N} \sum_{\alpha=1}^{K} M_{i\alpha} \mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha). \tag{1}$$

The distortion measure $\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha)$ quantifies the quantization error for data point $\mathbf{x}_i$. A widely used distortion measure is the squared Euclidean distance $\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha) \equiv \|\mathbf{x}_i - \mathbf{y}_\alpha\|^r$, $r = 2$. Various other distortion measures have been proposed in the literature [24], but generalizations of the squared Euclidean distortions $(r \neq 2)$ are most common for many

[1] Computational complexity issues like the generation of optimally searchable codebooks in tree-structured vector quantizers [10] are not within the scope of this paper.

signal processing applications, e.g., $r = 1$ supposedly is in accordance with the sensitivity profile of the human visual system [15] and is claimed to be a natural choice for image processing applications. The codebook vectors $\mathbf{y}_\alpha$ are determined by the centroid condition

$$\sum_i M_{i\alpha} \frac{\partial}{\partial \mathbf{y}_\alpha} \mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha) = 0, \quad \alpha = 1, \ldots, K, \tag{2}$$

which is optimal according to rate distortion theory (see [11]). The K Eqs. 2 require that the distortion measure $\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha)$ is differentiable in $\mathbf{y}_\alpha$; for non–differentiable distortions $\mathcal{D}_{i\alpha}(\mathbf{x}_i, \alpha_\alpha)$ we replace Eqs. (2) by the constraints $\mathcal{E}^{\mathrm{Km}}(\{M_{i\alpha}\}) = \min_{\mathbf{z}_\alpha} \left(\sum_{i=1}^{N} \sum_{\alpha=1}^{K} M_{i\alpha} \mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{z}_\alpha) \right)$ (see [32]). Variation of $M_{i\alpha}$ implicitly determines the reference vectors $\mathbf{y}_\alpha$ if we enforce Eqs. (2) strictly. Provided, the optimum set of reference vectors has been found, then each data point is represented by its closest reference vector. It should be mentioned that algorithms which minimize the above cost function under the constraint of unique assignment are also known as “hard clustering” algorithms. In contrast, procedures which return continuous assignments $M_{i\alpha} \in [0, 1]$ of a data point to several “cluster centers”, are called “soft” or “fuzzy clustering” algorithms [17, 4].[2]

How large should the “magic” number K of reference vectors be? It is clear that the result of the optimization of cost function (1) depends on the number of codebook vectors K or, more generally speaking, on the complexity of the codebook. The trivial solution to declare all data points to be reference vectors, $K = N$ and $\mathbf{y}_i = \mathbf{x}_i, i = 1, \ldots, N$, minimizes Eq. (1), but does not achieve any data reduction. Algorithms like K-means clustering assume that the size of the codebook is determined a priori, i.e., that the number K of reference vectors with $K \ll N$ is prespecified. Other procedures like ISODATA [3] stop adding new reference vectors as soon as the residual distortion error falls below a fixed threshold. These approaches can result in suboptimal vector quantizers for particular signal processing tasks since optimization of distortion and intrinsic constraint limits on the codebook are related and do effect each other in general. For example, entropic costs for codebook design have been shown to yield superior quantization performance in speech compression [9, 26].

We, therefore, suggest to extend the cost function (1) by an application dependent complexity measure which has to reflect the inherent costs of too complex vector quantizers. The complexity term limits the number of reference vectors. The admissible class of distortion measures is generalized to include topology preserving vector quantization schemes, also known as source–channel–coding, which reduce the detrimental effect of channel noise on the quantized data set [19, 34, 50]. The new cost function with a discussion of different choices for distortion and complexity measures is introduced in Section 2. A maximum entropy solution for this cost function is derived in Section 3. Our approach is inspired by the work of Rose et al. [38, 39] on K-means clustering. An optimization algorithm and simulation results are summarized in Section 4. The asymptotic level density of vector quantizers in the high complexity limit is studied in Section 5. These calculations determine the functional dependence of the codebook density on the data density and are related to [23, 49]. We

[2] In the spirit of the work on K-means clustering and refering to the close relationship between hard and fuzzy clustering in the maximum entropy framework we will use the terms “cluster center” and “reference vector” synonymously throughout this paper, although we are aware of approaches like hierarchical clustering [16] or Bayesian clustering [25] which are designed to uncover “natural” structures in a data set and which do not necessarily optimize the cost function (1).

also address the question in Section 6 what gain in compression efficiency results from an entropic complexity measure for special data distributions with outliers. As a "real-world" application of the new vector quantization algorithm we discuss the entropy optimized compression of wavelet decomposed images in Section 7 (see also [7]). This quantization scheme preserves psychophysically important image features like edges more accurately than the K-means clustering algorithm. In Section 8 we derive an online algorithm which asymptotically approaches the maximum entropy estimation of the centroids $\mathbf{y}_\alpha$ and the assignment probabilities p_α.

2 Complexity Limited Vector Quantization Costs

Vector quantization is a compromise between precision and simplicity of a data representation. As outlined above, we propose to explicitly reflect this tradeoff in the cost function by adding a complexity term to the usual distortion measure and jointly optimizing both cost terms. This strategy to control the complexity of a codebook is more natural for many applications than to a priori set the number of reference vectors to K, in particular when a "too complex" vector quantizer still constitutes a valid, but inefficient solution as in data transmission and storage. The balance between distortion costs and complexity costs yields an optimal number of reference vectors. The number of clusters naturally depends on the data distribution and on a parameter λ which weights the complexity costs versus the distortion costs. The new vector quantization cost function

$$\mathcal{E}_K(\{M_{i\alpha}\}) = \sum_{i=1}^{N}\sum_{\alpha=1}^{K} M_{i\alpha}\Big(\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha) + \lambda\mathcal{C}(p_\alpha)\Big) \tag{3}$$

only depends on the assignment variables $\{M_{i\alpha}\}$. The reference vector $\mathbf{y}_\alpha$ are defined by the K equations

$$\sum_{i=1}^{N}\sum_{\nu=1}^{K} M_{i\nu}\mathcal{G}_{\alpha\nu}(\mathbf{x}_i, \mathbf{Y}) = 0, \quad \alpha = 1, \ldots, K \tag{4}$$

which are linear in the assignment variables $M_{i\alpha}$. $\mathcal{G}_{\alpha\nu}(\mathbf{x}_i, \mathbf{Y})$ are a set of functions which depend on the codebook $\mathbf{Y}$ and the data point $\mathbf{x}_i$. Variation of $\{M_{i\alpha}\}$ implicitly determines the *reference vector* $\mathbf{y}_\alpha$ and the *assignment probability*

$$p_\alpha = \frac{1}{N}\sum_{i=1}^{N} M_{i\alpha}. \tag{5}$$

In the following we carry out the calculations for general functions $\mathcal{G}_{\alpha\nu}(\mathbf{x}_i, \mathbf{Y})$ and specialize to the centroid condition (2) by setting $\mathcal{G}_{\alpha\nu}(\mathbf{x}_i, \mathbf{Y}) \equiv \delta_{\alpha\nu}\frac{\partial}{\partial \mathbf{y}_\alpha}\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha)$, when necessary. $\delta_{\alpha\nu}$ is the Kronecker symbol ($\delta_{\alpha\nu} = 1$ if $\alpha = \nu$ and $\delta_{\alpha\nu} = 0$ if $\alpha \neq \nu$). We have indexed the quantization costs $\mathcal{E}_K(\{M_{i\alpha}\})$ with the size K of the codebook since we compare VQ solutions with different codebook sizes K. The minimum of $\mathcal{E}_K(\{M_{i\alpha}\})$ for different assignments $\{M_{i\alpha}\}$ and for different numbers of reference vectors K determines the optimal size of the codebook. Too complex codebooks are penalized by large complexity costs $\mathcal{C}(p_\alpha) \equiv \mathcal{C}_\alpha$ which only depends on the assignment probability. Note that the index α in (3) is supposed to sum only over the set of *different* clusters $\alpha \in \{1, \ldots, K\}$. Configurations with degenerate clusters, i.e., $\mathbf{y}_\alpha \equiv \mathbf{y}_\beta$ for $\alpha \neq \beta$ are inadmissible, since they give rise to an overestimation of the "true" complexity of the codebook. We, therefore, have to assure that each cluster index represents a separate cluster. Cluster degeneration does not occur for hard clustering but is a problem for fuzzy clustering [39].

An appropriate complexity measure $\mathcal{C}$ depends on the particular information processing application at hand, i.e., it is a function of the variables $\mathbf{y}_\alpha$, p_α. For reasons of simplicity we limit our analysis to complexity measures depending only on the assignment probabilities p_α and not on the reference vectors $\mathbf{y}_\alpha$. A more general derivation is feasible, if necessary. In data compression we want to encode the data set with minimal distortion error and compress the resulting index set optimally without further losses. In this context a natural complexity measure for a particular cluster α is $\mathcal{C}(p_\alpha) = -\log(p_\alpha)$, which results in an average complexity $\langle \mathcal{C} \rangle = -\sum_\alpha p_\alpha \log p_\alpha$ of the codebook. Note, that this is identical with the entropy of an information source emitting letters of a K-ary alphabet with probabilities $p_\alpha, \alpha = 1, \ldots, K$. $\langle \mathcal{C} \rangle$ sets a lower bound for the minimal possible average codeword length if the messages α are subject to data compaction algorithms like arithmetic coding or Huffman coding [11]. Lagrangian optimization of a vector quantization cost function with entropic complexity costs is extensively discussed in [9]. We will apply entropy optimized quantization to the problem of image compression in Section 7.

Another strategy to limit the number of clusters is defined by complexity costs which penalize small, rarely used reference vectors, i.e., $\mathcal{C}_\alpha = 1/p_\alpha^s, s = 1, 2, \ldots$. Such complexity measures favor equal assignment probabilities since they strongly diverge for $p_\alpha = 0$. The hardware aspect of a vector quantization solution, e.g., scarcity of processing elements or expensive chip area in a `VLSI` implementation is better taken into account by $\mathcal{C}_\alpha = 1/p_\alpha^s$ with its preference of load balanced codebooks than by the entropic complexity measure $\mathcal{C}(p_\alpha) = -\log(p_\alpha)$. The special case $s = 1$ with complexity costs strictly proportional to the number of clusters ($\mathcal{C}_\alpha = 1/p_\alpha \;\Rightarrow \sum_i \sum_\alpha M_{i\alpha}/p_\alpha = NK$) corresponds to the K-means clustering cost function [32]. For $s = 1$ all clusters have the same cost term independent of the number of data points assigned to them. We will use this case as a standard to compare with the entropy optimized vector quantizer in section 6. The limit $s \to \infty$ produces codebooks which are perfectly load–balanced, i.e., all cluster have the same number of data points assigned to ($\lim_{s \to \infty} p_\alpha = 1/K, \forall \alpha$).

The choice of an appropriate distortion measure $\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha)$ is another degree of freedom in the design of a vector quantization algorithm. Usually a generalization of the Euclidean distance is employed to measure quantization errors, i.e.,

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^r. \tag{6}$$

Our formalism, however, is applicable to even more general difference distortion measures. The Minkowski l_p-norms

$$\mathcal{D}(\mathbf{x}, \mathbf{y}) = \left(\sum_{s=1}^{d} |(\mathbf{x})_s - (\mathbf{y})_s|^p \right)^{1/p} \tag{7}$$

offer another alternative to quantify distortion errors [5].

Both measures (6) and (7) leave $\mathcal{E}_K(\{M_{i\alpha}\})$ invariant under permutations of the reference vectors $\mathbf{y}_\alpha$. Such schemes proved to be very sensitive to channel noise in data transmission.

The following block diagram describes the communication process where quantized data are sent through a noisy communication channel:

$$\begin{tikzcd}
 & & \boxed{\text{noise } \mathtt{T}_{\alpha\gamma}} \arrow[d] & & \\
\{\mathbf{x}_i\} \arrow[r] & \boxed{\text{Encoder } \mathbf{x}_i \to \mathbf{y}_\alpha} \arrow[r, "\alpha"] & \boxed{\text{Channel}} \arrow[r, "\gamma"] & \boxed{\text{Decoder}} \arrow[r] & \{\mathbf{y}_\gamma\}
\end{tikzcd}$$

At the encoding stage data point $\mathbf{x}_i$ is assigned to cluster α. The index α, however, gets corrupted by channel noise and arrives as index γ at the receiver's side of the communication channel. Let us denote the transition probability from index α to index γ by $\mathtt{T}_{\alpha\gamma}$ with $\sum_{\gamma=1}^{K} \mathtt{T}_{\alpha\gamma} = 1$. The receiver, consequently, reconstructs the "incorrect" reference vector $\mathbf{y}_\gamma$ as a representation of $\mathbf{x}_i$ instead of the "correct" reference vector $\mathbf{y}_\alpha$. Therefore, the transitions $\mathtt{T}_{\alpha\gamma}$ cause additional distortions which have to be taken into account when we estimate the most likely assignment variables. The channel noise breaks the permutation symmetry of the reference vectors and imposes a topology on the set of indices $\{\alpha | \alpha = 1, \ldots, K\}$. Codeword assignments which take the characteristics of the channel noise into account yield superior results [19, 50]. Such a procedure is also known as source–channel–coding. Following the same line of thought Luttrell [34] established a connection to a class of topological vector quantization algorithms known as self-organizing feature maps [30, 37].

Mathematically, we replace the distortion costs $\mathcal{D}_{i\alpha}$ in (3) by the averaged distortion errors

$$\langle\!\langle \mathcal{D}_{i\alpha} \rangle\!\rangle \equiv \sum_{\gamma} \mathtt{T}_{\alpha\gamma} \mathcal{D}_{i\gamma}(\mathbf{x}_i, \mathbf{y}_\gamma). \tag{8}$$

Note, that the centroid condition generalizes to $\sum_i \sum_\gamma M_{i\gamma} \mathtt{T}_{\gamma\alpha} \frac{\partial}{\partial \mathbf{y}_\alpha} \mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha) = 0$. It has to be emphasized that the resulting cost function for topological vector quantization as well as the generalized centroid condition are still linear in the assignment variables $\{M_{i\alpha}\}$ and, therefore, the minimization is tractable like the original optimization problem (3). The transition probability $\mathtt{T}_{\alpha\gamma}$ has to be determined on the basis of the channel noise characteristics and on the basis of the chosen code for the reference vector indices. The special case of a tridiagonal transition matrix $\mathtt{T}_{\alpha\alpha} = 1 - \eta; \mathtt{T}_{\alpha,\alpha\pm1} = \eta/2; \mathtt{T}_{\gamma\alpha} = 0 \; \forall |\alpha - \gamma| > 1$ defines a linear chain with nearest neighbor transitions. Topology preserving vector quantization reduces to non-topological clustering if we set $\mathtt{T}_{\alpha\gamma} = 1$ for $\alpha = \gamma$ and $\mathtt{T}_{\alpha\gamma} = 0$ otherwise.

3 Reestimation Equations of $\{\mathbf{y}_\alpha\}$ and $\{p_\alpha\}$

Our main objective is to jointly minimize the distortion and complexity costs in (3) with respect to the variables $M_{i\alpha}$. We can interpret the minimization of the vector quantization cost function (3) as a search for a set of assignment variables $\{M_{i\alpha}\}$ which yield the quantization costs $\mathcal{E}_K(\{M_{i\alpha}\}) = \langle \mathcal{E}_K \rangle$. The inference principle which selects the most stable set of assignment variables with respect to random fluctuations in the assignment process is the maximum entropy principle [43]. Furthermore, the maximum entropy principle is the least biased inference method, being *maximally noncommittal with respect to missing data* as Jaynes formulated it [27, 28]. Consequently, a search strategy for assignments $\{M_{i\alpha}\}$ based on the maximum entropy principle promises robustness and fast convergence. The success of simulating annealing [29] and of neural optimization algorithms [18, 41, 42, 47] supports this design philosophy for optimization algorithms. Another substantial advantage of maximum entropy inference is the inherent parallelism of the method which facilitates the mapping of resulting algorithms to parallel hardware or VLSI implementations on chips.

In this section we will calculate the maximum entropy distribution of the assignment variables $\{M_{i\alpha}\}$ for quantization costs $\mathcal{E}_K(\{M_{i\alpha}\}) = \langle \mathcal{E}_K \rangle$. Subsequent minimization of $\langle \mathcal{E}_K \rangle$ yields the global minimum of the vector quantization costs $\mathcal{E}_K(\{M_{i\alpha}\})$. As a consequence of the maximum entropy principle the distribution function of the assignment variables is the Gibbs distribution

$$P(\{M\}) = \exp\left(-\beta(\mathcal{E}_K - \mathcal{F}_K)\right). \tag{9}$$

The parameter β, often refered to as the inverse of the "computational temperature" T, is a Lagrange multiplier, chosen such that the average quantization costs amount to $\langle \mathcal{E}_K \rangle$. The factor $\exp(\beta \mathcal{F}_K)$ normalizes the Gibbs. $\mathcal{F}_K$ is called the free energy in statistical physics and it is given by

$$\mathcal{F}_K = \min_{\substack{\{p_\alpha\},\{\mathbf{y}_\alpha\},\\ \{\hat{p}_\alpha\},\{\hat{\mathbf{y}}_\alpha\}}} \left\{ -N \sum_\alpha \hat{p}_\alpha p_\alpha - \frac{1}{\beta} \sum_i \log \sum_\alpha \exp\left(-\beta \left[\mathcal{E}_{i\alpha} - \sum_\nu \hat{\mathbf{y}}_\nu \mathcal{G}_{\nu\alpha}(\mathbf{x}_i, \mathbf{Y}) + \hat{p}_\alpha \right] \right) \right\}. \tag{10}$$

The equation (10) is derived in appendix A. The parameters $\{\mathbf{y}_\alpha\}$, $\{p_\alpha\}$ are the order parameters of the vector quantization problem. The respective conjugate fields $\{\hat{\mathbf{y}}_\alpha\}$, $\{\hat{p}_\alpha\}$ have to be introduced to strictly enforce the definition of the reference vectors $\frac{1}{N}\sum_i \sum_\nu M_{i\nu}\mathcal{G}_{\alpha\nu}(\mathbf{x}_i, \mathbf{Y}) = 0$ and the definition of the assignment probabilities $p_\alpha = \frac{1}{N}\sum_i M_{i\alpha}$. The differential quantization costs $\mathcal{E}_{i\alpha}$ in Eq. (10) are

$$\mathcal{E}_{i\alpha} = \sum_\gamma \mathtt{T}_{\alpha\gamma} \mathcal{D}_{i\gamma}(\mathbf{x}_i, \mathbf{y}_\gamma) + \lambda \mathcal{C}(p_\alpha). \tag{11}$$

The minimization in Eq. (10) is performed under the constraint $\mathbf{y}_\alpha \neq \mathbf{y}_\beta$, $\forall \alpha \neq \beta$, i.e., all configurations with degenerate reference vectors are rejected. The parameters $\mathbf{y}_\alpha^{\mathrm{min}}$, p_α^{min} which define the free energy $\mathcal{F}_K$ are the expectation values of the random variables $\mathbf{y}_\alpha$, p_α defined in (4,5). Since we study the limit of many data points ($N \to \infty$) and since we assume that the free energy and variables derived from it are self-averaging we will use the same notation for the random variables $\mathbf{y}_\alpha$, p_α and their expectation values.

Assuming that $\mathcal{D}_{i\alpha}$ and $\mathcal{G}_{\alpha\nu}$ are differentiable functions we derive the order parameter equations

$$p_\alpha = \frac{1}{N} \sum_{i=1}^{N} \langle M_{i\alpha} \rangle, \tag{12}$$

$$0 = \frac{1}{N} \sum_{i=1}^{N} \sum_{\nu=1}^{K} \mathcal{G}_{\alpha\nu}(\mathbf{x}_i, \mathbf{Y}) \langle M_{i\nu} \rangle, \tag{13}$$

$$\hat{p}_\alpha = \lambda \frac{d\mathcal{C}}{dp_\alpha} p_\alpha, \tag{14}$$

$$0 = \sum_{i=1}^{N} \sum_{\nu=1}^{K} \langle M_{i\nu} \rangle \Big(\mathtt{T}_{\nu\alpha} \frac{\partial}{\partial \mathbf{y}_\alpha} \mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha) - \sum_{\mu=1}^{K} \hat{\mathbf{y}}_\mu \frac{\partial}{\partial \mathbf{y}_\alpha} \mathcal{G}_{\mu\nu}(\mathbf{x}_i, \mathbf{Y}) \Big) \tag{15}$$

with the abbreviation

$$\langle M_{i\alpha}\rangle = \frac{\exp\left[-\beta(\mathcal{E}_{i\alpha} + \hat{p}_\alpha - \sum_\nu \hat{\mathbf{y}}_\nu \mathcal{G}_{\nu\alpha}(\mathbf{x}_i, \mathbf{Y})\right]}{\sum_{\mu=1}^K \exp[-\beta(\mathcal{E}_{i\mu} + \hat{p}_\mu - \sum_\nu \hat{\mathbf{y}}_\nu \mathcal{G}_{\nu\mu}(\mathbf{x}_i, \mathbf{Y}))]}, \tag{16}$$

$\mathcal{E}_{i\alpha}$ being defined in (11).

The equations (12-15) can be simplified if we specialize the functions $\mathcal{G}_{\alpha\nu}(.)$ to be the generalized centroid conditions, i.e., $\mathcal{G}_{\alpha\nu}(\mathbf{x}_i, \mathbf{Y}) = \mathtt{T}_{\nu\alpha}\frac{\partial}{\partial \mathbf{y}_\alpha}\mathcal{D}_{i\alpha}$. Given such a definition of $\mathcal{G}_{\alpha\nu}(.)$ the first term in the brackets of (15) vanishes due to equation (13). The second term in the bracket of (15) can be rewritten as

$$0 \;=\; \hat{\mathbf{y}}_\alpha \frac{1}{N}\sum_i \sum_\gamma \langle M_{i\gamma}\rangle \mathtt{T}_{\gamma\alpha}\frac{\partial^2}{\partial {\mathbf{y}_\alpha}^2}\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha). \tag{17}$$

If we assume that the distortion measure $\mathcal{D}_{i\alpha}$ is a convex function of $\mathbf{y}_\alpha$ then the only solution of (17) is $\hat{\mathbf{y}}_\alpha = 0$. In cases where the reference vectors $\mathbf{y}_\alpha$ are not the centroids of the respective clusters α, e.g., self-organizing feature maps, we find solutions with $\hat{\mathbf{y}}_\alpha \neq 0$. The equations (14) and (15) can be inserted into (12) and (13) which reduces the order parameter equations to two systems of K transcendental equations

$$p_\alpha \;=\; \frac{1}{N}\sum_{i=1}^N \langle M_{i\alpha}\rangle, \tag{18}$$

$$0 \;=\; \frac{1}{N}\sum_{i=1}^N \sum_{\gamma=1}^K \langle M_{i\gamma}\rangle \mathtt{T}_{\gamma\alpha}\frac{\partial}{\partial \mathbf{y}_\alpha}\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha), \tag{19}$$

with the fuzzy assignment variables

$$\langle M_{i\alpha}\rangle = \frac{\exp\left[-\beta(\mathcal{E}_{i\alpha} + \lambda \frac{d\mathcal{C}}{dp_\alpha}p_\alpha)\right]}{\sum_{\mu=1}^K \exp[-\beta(\mathcal{E}_{i\mu} + \lambda \frac{d\mathcal{C}}{dp_\mu}p_\mu)]}. \tag{20}$$

The resulting free energy is

$$\mathcal{F}_K = \min_{\{\mathbf{y}_\alpha\},\{p_\alpha\}}\left\{-N\lambda\sum_\alpha p_\alpha^2\frac{d\mathcal{C}}{dp_\alpha} - \frac{1}{\beta}\sum_i \log\sum_\alpha \exp\left(-\beta\left[\mathcal{E}_{i\alpha} + \lambda p_\alpha \frac{d\mathcal{C}}{dp_\alpha}\right]\right)\right\}. \tag{21}$$

$\langle M_{i\alpha}\rangle$ can be interpreted as the probability that data point i is assigned to reference vector α. Consequently, p_α measures the percentage of data points assigned to α, hence p_α is the assignment probability. $\mathbf{y}_\alpha$ is the generalized centroid of cluster α. Equations (18) and (19) demonstrate explicitly, that the placement of reference vectors as well as their assignment probabilities depend on the particular complexity measure. It has to be emphasized that a solution of (18) and (19) ensures only an extremal value of the free energy $\mathcal{F}_K$, but does not guarantee that we have found the global minimum of (10) or, equivalently, the maximum entropy estimation of $\{M_{i\alpha}\}$. Depending on the data distribution the minimization in Eqs. (10,21) might be a highly non-convex optimization problem with many local minima.

The statistical mechanics approach for the K-means clustering cost function (1) which corresponds to the case $\lambda = 0$ and fixed K has first been discussed in [38, 39]. Note, that this case corresponds to the specific choice of $\mathcal{C} = 1/p_\alpha$ where the conjugate potential $\hat{p}_\alpha = -\lambda/p_\alpha = -\lambda\mathcal{C}_\alpha$ exactly cancels the complexity term, adding essentially a constant cost term (chemical potential) λK to the free energy $\mathcal{F}$. Rose et al. advocated a deterministic annealing approach for K-means clustering in the spirit of [29]. They did not include complexity costs to limit the number of codebook vectors but proposed to start with many codebook vectors at high temperature, to lower the temperature to a finite value and to use the resulting set of different vectors as the codebook. This procedure, however, is plagued by the problem of degenerated reference vectors, i.e., $\mathbf{y}_\alpha = \mathbf{y}_\gamma$ for $\alpha \neq \gamma$ are solutions at high temperature. The degeneration of reference vectors affects the data assignments and, thereby, the values of the reference vectors.

The order parameter equations (18), (19) form a system of $K(d+1)$ transcendental equations which have to be solved simultaneously for a particular vector quantization problem. The fuzziness[3] of the assignment process in a maximum entropy sense is expressed by the gradual membership functions $\langle M_{i\alpha} \rangle$. For very high temperature ($\beta \rightarrow 0$) the data point i is assigned with equal probability $1/K$ to all clusters and the reference vectors $\mathbf{y}_\alpha$ are identical to the center of mass of the data distribution. p_α in (18) can be interpreted in an intuitive way as the fuzzy cluster probability which adds up the fuzzy assignment variables $\langle M_{i\alpha} \rangle$ of all data points $\mathbf{x}_i$. $\mathbf{y}_\alpha$ in (19) is the centroids of all points assigned to cluster α, weighted with the corresponding membership probability. In the limit of zero temperature we recover the hard clustering case. A datapoint $\mathbf{x}_i$ is assigned to cluster α if and only if $\mathcal{E}_{i\alpha} + \lambda \frac{d\mathcal{C}}{dp_\alpha} p_\alpha < \mathcal{E}_{i\mu} + \lambda \frac{d\mathcal{C}}{dp_\mu} p_\mu,\ \forall \mu \neq \alpha$.

4 An Algorithm for the Design of Complexity Optimized Codebooks

Finding the maximum entropy estimation of the cost function (3) is a $\mathcal{NP}$-hard problem [20] plagued by the characteristic non–convexity of the free energy. In the following we propose an iterative cluster splitting algorithm which approximates the maximum entropy solution and which determines the optimal size K of the codebook. An iteration of the following two step procedure is required:

(i) We have to calculate the free energy $\mathcal{F}_K$ for fixed K, which requires to search for the global minimum in (10,21).

(ii) We have to vary K to determine the optimal number of reference vectors.

A variety of different methods exist to perform the minimization required in step (i). The algorithm described below is based on the reestimation of $\mathbf{y}_\alpha, p_\alpha$, i.e., to solve the transcendental equations (18), (19) and to evaluate the respective free energy. In case that $\mathcal{D}_{i\alpha}$ and $\mathcal{G}_{\alpha\nu}$ are not differentiable we have to minimize (10,21) directly. The problem of local minima can be alleviated by *simulated annealing*, an optimization strategy which was introduced by Kirkpatrick [29] and which proved to be successful in a large variety of difficult non-convex optimization problems like the traveling salesman problem. The reestimation equations (18),

[3] Note, that fuzziness should be understood as a partial membership of data points in clusters. Our approach is still a probabilistic one and is not related to fuzzy set theory as proposed by Zadeh [48].

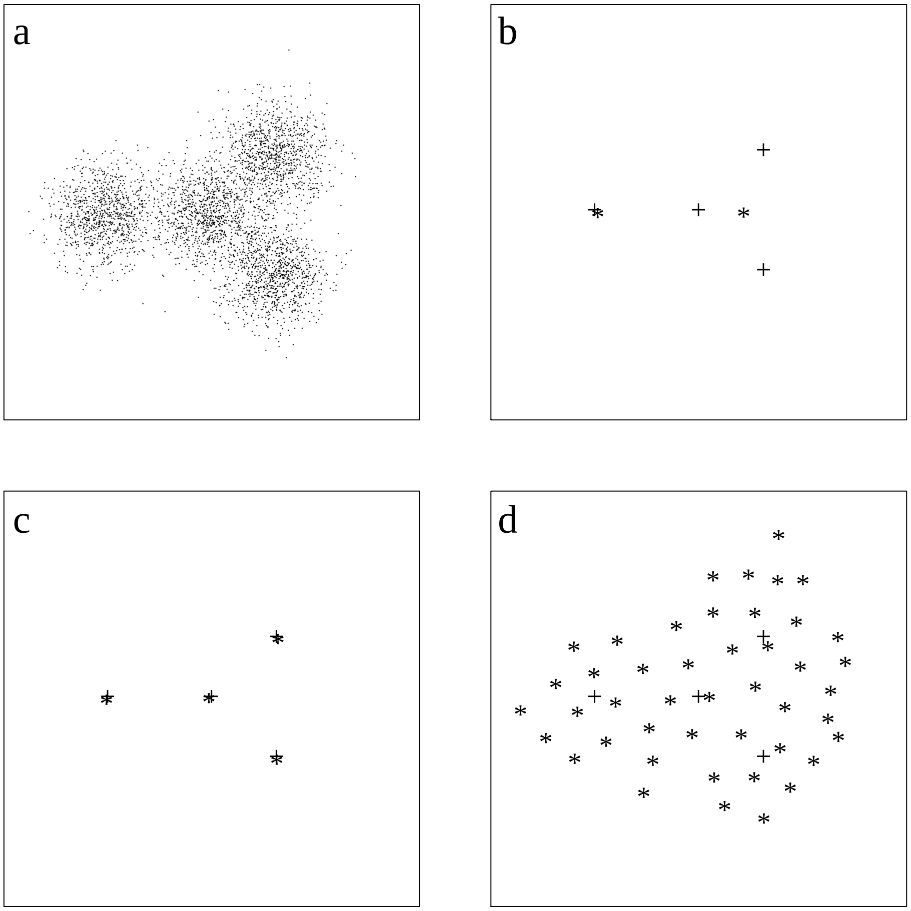

Figure 1: (a) Data distribution consisting of 4000 points, generated by four normally distributed sources of equal variance. Centers of the sources are marked by plus signs "+" (b–d). Zero temperature solutions of entropy optimized quantization $\mathcal{C}_\alpha = -\log p_\alpha$ are shown for the complexity weights $\lambda = 5.0, 2.5, 0.4$ in (b,c,d). The locations of the reference vectors are marked by asterisks "*".

(19) explicitly show the functional dependence on the computational temperature which is slowly decreased in simulated annealing. It has to be mentioned that step (i) does not require a complete cooling schedule from high temperature to zero temperature since cluster splitting and cooling are coupled [38]. We have abstained from temperature variations in the simulation examples presented below since a $T = 0$ strategy yielded sufficiently good estimates of the global minimum of $\mathcal{F}_K$. The second step (ii) is necessary[4] since we do not have a direct estimate of the optimal codebook size K.

The following algorithm implements the two step procedure under the constraint that no two reference vectors are degenerate. The heuristics to split only a single cluster, could be

[4] A more general approach with the optimal K being an order parameter has to be based on the grand canonical partition function (GCPF) instead of the canonical partition function. The GCPF sums over all possible codebook sizes $K = 1, \ldots, N$. Unfortunately, the summation over K is analytically intractable. A variation of the codebook size K can be interpreted as a discrete approximation of the GCPF.

replaced by a more sophisticated splitting strategy, e.g., larger portions of the codebook or even the whole codebook as in the LBG-algorithm could be split simultaneously. Thereby, the computational complexity would be drastically reduced, but we might have to merge clusters again to find the optimal codebook size. The cluster splitting can be interpreted as an optimization step of the cost function $\mathcal{E}_K$ along the K coordinate. The algorithm employed in the simulations in this section and in the sections 6 and 7 comprises the following steps:

```
0.) Initialization: Compute the centroid of the data set and place the
    first codebook vector there ($K = 1$).

1.) Determine a random order of codebook vectors $\mathbf{y}_\alpha$ with list index $\kappa \equiv 1$.

2.) Split codebook vector $\kappa$, i.e., $\mathbf{y}_{K+1} \leftarrow \mathbf{y}_\kappa + \mathbf{z}$, $\mathbf{y}_\kappa \leftarrow \mathbf{y}_\kappa$, $p_{K+1} \leftarrow p_\kappa/2$,
    $p_\kappa \leftarrow p_\kappa/2$, (z is a ``small'' random vector with $\|\mathbf{z}\| \ll \|\mathbf{y}_\kappa\|$).

3.) Reestimate $p_\alpha, \mathbf{y}_\alpha$ using Eqs. (18,19).

4.) If $\mathcal{F}_{K+1}^{\text{new}} < \mathcal{F}_K^{\text{old}}$ then accept new codebook with $K+1$ vectors, increment
    $K \leftarrow K+1$ and goto step 1.

5.) If $\mathcal{F}_{K+1}^{\text{new}} \geq \mathcal{F}_K^{\text{old}}$ or $\exists \alpha : p_\alpha = 0$ or $\exists \alpha, \beta : \mathbf{y}_\alpha = \mathbf{y}_\beta$ then reject new codebook and
    increment $\kappa$. If $\kappa > K$ then quit, else goto 2.
```

The resulting solution is a minimum of $\mathcal{F}_K$ under single cluster splitting, although it is not guaranteed to be a global minimum since there might exist configurations with lower free energy than the one found by our algorithm. The iterative cluster splitting is necessary since we might find a local minimum $\mathcal{F}_K$ with K reference vectors where the global minimum has a codebook size smaller than K. The random ordering of codebook vectors α^* in step 1 can be replaced by an application dependent heuristics, e.g., to order the clusters according to size. The proposed algorithm is a generalization of the vector quantization algorithm discussed by Chou et al. [9] in two respects: (i) we optimize the codebook size K by a systematic search for the lowest free energy $\mathcal{F}_K$; (ii) temperature variation in the spirit of simulating annealing allows us to avoid suboptimal local minima of $\mathcal{F}_K$.

In the first quantization example (Fig. 1) we have chosen a two dimensional data distribution, generated by four normally distributed sources of equal variance (Fig. 1a). Three zero temperature solutions for the complexity measure $\mathcal{C} = -\log p_\alpha$, calculated for different values of λ, are shown in Fig. 1b-d. For very high complexity costs the algorithm finds only two cluster centers. The plus signs indicate the centers of the data sources. For $\lambda = 2.5$ four reference vectors are positioned at the centers of the data sources. In the limit of very small complexity costs the best optimization of the cost function found by our iterative cluster splitting algorithm densely covers the data distribution. The specific choice of the logarithmic complexity measure causes the homogeneous density of reference vectors which is known to yield the smallest entropy per codebook vector in the limit of asymptotic quantization levels [9, 23]. This demonstrates how different complexity measures can drastically modify the codebook structure. We will analyse this observation in section 5.

5 Asymptotic Quantization Level Density

Different complexity measures influence the distributions of reference vectors, even if we choose the same data distribution and the same distortion measure. In this section, we study the asymptotic level density for codebooks with high complexity, i.e., dense quantization levels, and we derive the dependency of the cluster density $\Upsilon(\mathbf{x})$ on the probability density $\Pi(\mathbf{x})$ for the case of hard clustering. The analysis follows Zador's [49], Gersho's [22] and Yamada et al.'s [46] line of reasoning, although we use a variational approach to determine the density of quantization levels as a function of the probability density instead of searching bound for the distortion costs. The variational approach is required since we treat the case of general complexity costs which balance the monotonous decrease of distortion costs with increasing codebook vector density. In the special case of K-means clustering or entropic complexity costs the solutions reduce to the classic results of Zador [49] and of Gish and Pierce [23]. Our results are restricted to rth-power Euclidean distortions $\mathcal{D}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^r$, although the analysis can be generalized to the class of difference distortion measure [46].

We denote the codebook size by K. S_α, $\alpha = 1,\ldots,K$ are the partitions of $\Re^d$ which define the quantizer. Following [33] we define the codebook density function as

$$g_K(\mathbf{x}) = \frac{1}{KV(S_\alpha)}, \quad \text{if} \quad \mathbf{x} \in S_\alpha, \tag{22}$$

where $V(S_\alpha)$ is the volume of partition S_α. In the limit of dense quantization levels we expect $g_K(\mathbf{x})$ to closely approximate a continuous density function $\Upsilon(\mathbf{x}) \approx 1/(KV(S_\alpha))$. The total distortion costs can be written as

$$\langle \mathcal{D} \rangle = \sum_{\alpha=1}^{K} \int_{S_\alpha} \mathcal{D}(\mathbf{x},\mathbf{y}_\alpha)\Pi(\mathbf{x})d\mathbf{x} \approx \sum_{\alpha=1}^{K} \Pi(\mathbf{y}_\alpha) \int_{S_\alpha} \mathcal{D}(\mathbf{x},\mathbf{y}_\alpha)d\mathbf{x}. \tag{23}$$

The approximation holds for a smoothly varying probability density $\Pi(\mathbf{x}) \approx \Pi(\mathbf{y}_\alpha)$ for $\mathbf{x} \in S_\alpha$. Following Gersho [22] we make the basic assumption that for large K the partitions S_α approximate the optimal polytope S^* for dimension d and distortion measure $\mathcal{D}(\mathbf{x},\mathbf{y})$. Given $\mathcal{D}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|^r$ we substitute

$$\int_{S_\alpha} \mathcal{D}(\mathbf{x},\mathbf{y}_\alpha)d\mathbf{x} = I(d,r)(V(S_\alpha))^{(d+r)/d}, \tag{24}$$

$I(d,r) = \int_{S^*} \mathcal{D}(\mathbf{x},\mathbf{y}_\alpha)d\mathbf{x}/(V(S^*))^{(d+r)/d}$ being the normalized inertia of the optimal polytope S^*. Inserting (24) into (23) the resulting total distortion costs are given by

$$\langle \mathcal{D} \rangle = \sum_{\alpha=1}^{K} \Pi(\mathbf{y}_\alpha) I(d,r)(V(S_\alpha))^{(d+r)/d}. \tag{25}$$

The assignment probability of codebook vector α in the limit of asymptotic quantization levels is a function of the probability density and the codebook vector density, i.e., p_α is given by

$$p_\alpha \;=\; \int_{S_\alpha} \Pi(\mathbf{x})d\mathbf{x} \approx \frac{\Pi(\mathbf{y}_\alpha)}{K\Upsilon(\mathbf{y}_\alpha)} \tag{26}$$

Inserting (26) into $\mathcal{C}_\alpha$ we obtain for the total complexity

$$\langle \mathcal{C} \rangle \;=\; \sum_{\alpha=1}^{K} \Pi(\mathbf{y}_\alpha) V(S_\alpha) \mathcal{C}\left(\frac{\Pi(\mathbf{y}_\alpha)}{K \Upsilon(\mathbf{y}_\alpha)} \right). \tag{27}$$

In the continuum limit of asymptotic quantization levels the summation $\sum_{\alpha=1}^{K} V(S_\alpha)$ is approximated by the integral $\int d\mathbf{y}$ and the total quantization costs for complexity optimized vector quantization are measured by the functional

$$\mathcal{E} = \langle \mathcal{D} \rangle + \lambda \langle \mathcal{C} \rangle = \int d\mathbf{y} \Pi(\mathbf{y}) \left[I(d,r)(K\Upsilon(\mathbf{y}))^{-r/d} + \lambda \mathcal{C}\left(\frac{\Pi(\mathbf{y})}{K\Upsilon(\mathbf{y})} \right) \right]. \tag{28}$$

The optimal distribution of codebook vectors is determined by the function $\Upsilon(\mathbf{x})$ which minimizes the functional (28). Therefore, we vary this cost functional with respect to $\Upsilon(\mathbf{x})$ to derive the dependence of the codebook density $\Upsilon(\mathbf{x})$ on the probability density $\Pi(\mathbf{x})$. The resulting Euler-Lagrange equation is

$$\frac{r}{\lambda d} \frac{I(d,r)}{\Pi} (K\Upsilon)^{1-r/d} + \left[\frac{\partial \mathcal{C}(p)}{\partial p} \right]_{p=\Pi/K\Upsilon} = 0. \tag{29}$$

For the cases of load balancing (lb) and entropic complexity (ent), $\mathcal{C}(p) = 1/p^s$ and $\mathcal{C}(p) = -\log p$ respectively, the solutions of (29) are the densities

$$\Upsilon_{\text{lb}}(\mathbf{x}) = \frac{1}{K} \left(\frac{r}{\lambda d s} I(d,r) \Pi^s(\mathbf{x}) \right)^{d/(ds+r)} \tag{30}$$

and

$$\Upsilon_{\text{ent}}(\mathbf{x}) = \frac{1}{K} \left(\frac{r}{\lambda d} I(d,r) \right)^{d/r}. \tag{31}$$

The (in)dependence of Υ on Π for the K-means density (Eq. 30 for $s = 1$) and the entropic density (31) have been derived in the classic papers of Gish and Pierce [23] and Zador [49].

6 Performance Comparison between K-means and Entropy Optimized Clustering

The quantization results shown in Fig. 1 and the analysis of the asymptotic quantization level density (30,31) demonstrate that the complexity measure drastically influences the placement of the codebook vectors. How does it influence the performance of vector quantizers? We will address this question twofold: (i) In this section we quantize an artificial, d-dimensional data distribution using the K-means and the entropic complexity measure. (ii) In Section 7 we compress wavelet decomposed images and measure the reconstruction error after quantization with the K-means and the entropic complexity costs. We restrict the discussion of both compression applications to the zero temperature limit since the influence of the complexity measure is the focus of this study and not so much the performance gain by temperature variations.

The independence of the codebook vector density from the data density for entropy encoding indicates that sparsely populated areas of data space are more accurately represented by an entropy encoding scheme than by other quantization strategies. To verify this hypothesis we select an artificial, d-dimensional data distribution which consists of two regions with uniform densities Π_{high} and Π_{low} respectively and vanishing density outside. The respective volumes are V_{high} and V_{low}. Π_{low} is assumed to be small. Such a limit is interesting for many pattern recognition applications, e.g., image processing where the psychophysically important edge information is sparse. The functional relationship that an increase in the frequency of data features decreases their information content favors quantization schemes which minimize the error in sparsely populated (outlier) regions.

Let us compare the compression efficiencies of K-means clustering (Km) and of entropy coding (ent) for the two cases of (i) identical overall distortions costs and (ii) identical distortion costs in the outlier region with low data density Π_{low}. The rth-power Euclidean distortion costs $\mathcal{D}_{\text{single}}$ caused by a single codebook vector with partition volume V_{single} amount to

$$\mathcal{D}_{\text{single}} = \Pi\, I(d,r) V_{\text{single}}^{(d+r)/d}. \tag{32}$$

The volume of a single cluster in the high- and low-density parts of the distribution is given by $V_{\text{single}} = V_{\text{high/low}}/K_{\text{high/low}}$ respectively, $K_{\text{high/low}}$ being the number of codebook vector in the respective regions. Using (30) and (31) we find in the K-means clustering case

$$K_{\text{high,low}} = K_{\text{Km}} \frac{\Pi_{\text{high,low}}^{d/(d+r)} V_{\text{high,low}}}{\Pi_{\text{high}}^{d/(d+r)} V_{\text{high}} + \Pi_{\text{low}}^{d/(d+r)} V_{\text{low}}}, \tag{33}$$

and for entropy coding complexity

$$K_{\text{high,low}} = K_{\text{ent}} \frac{V_{\text{high,low}}}{V_{\text{high}} + V_{\text{low}}}. \tag{34}$$

Using these expressions we find the distortion costs

$$\begin{align}
\mathcal{D}_{\text{Km}} &= \mathcal{D}_{\text{Km,high}} + \mathcal{D}_{\text{Km,low}} \nonumber \\
&= I(d,r) K_{\text{Km}}^{-r/d} \left(\Pi_{\text{high}}^{d/(d+r)} V_{\text{high}} + \Pi_{\text{low}}^{d/(d+r)} V_{\text{low}}\right)^{r/d} \left(\Pi_{\text{high}}^{d/(d+r)} V_{\text{high}} + \Pi_{\text{low}}^{d/(d+r)} V_{\text{low}}\right) \tag{35} \\
&= I(d,r) K_{\text{Km}}^{-r/d} \left(\Pi_{\text{high}}^{d/(d+r)} V_{\text{high}} + \Pi_{\text{low}}^{d/(d+r)} V_{\text{low}}\right)^{(r+d)/d}, \tag{36} \\
\mathcal{D}_{\text{ent}} &= \mathcal{D}_{\text{ent,high}} + \mathcal{D}_{\text{ent,low}} \nonumber \\
&= I(d,r) K_{\text{ent}}^{-r/d} \left(V_{\text{high}} + V_{\text{low}}\right)^{r/d} \left(\Pi_{\text{high}} V_{\text{high}} + \Pi_{\text{low}} V_{\text{low}}\right) \nonumber \\
&= I(d,r) K_{\text{ent}}^{-r/d} \left(V_{\text{high}} + V_{\text{low}}\right)^{r/d}. \tag{37}
\end{align}$$

Our first criterion, that the overall distortion are identical for both vector quantizers ($\mathcal{D}_{\text{ent}} \equiv \mathcal{D}_{\text{Km}}$), establishes a relationship between the required codebook sizes, i.e.,

$$\left(V_{\text{high}} + V_{\text{low}}\right) \frac{K_{\text{Km}}}{K_{\text{ent}}} = \left[\Pi_{\text{high}}^{d/(d+r)} V_{\text{high}} + \Pi_{\text{low}}^{d/(d+r)} V_{\text{low}}\right]^{(r+d)/r}. \tag{38}$$

The second, more stringent condition that the outlier error is the same for both VQs, i.e, ($\mathcal{D}_{\text{ent,low}} \equiv \mathcal{D}_{\text{Km,low}}$) constrains the sizes of the codebooks to

$$\left(V_{\text{high}} + V_{\text{low}}\right) \frac{K_{\text{Km}}}{K_{\text{ent}}} = \Pi_{\text{low}}^{-d/(d+r)} \left[\Pi_{\text{high}}^{d/(d+r)} V_{\text{high}} + \Pi_{\text{low}}^{d/(d+r)} V_{\text{low}}\right]. \tag{39}$$

The cluster probabilities in the high- and low-density regions are given by $p_\alpha = \Pi_{\text{high/low}} V_{\text{high/low}} / K_{\text{high/low}}$ respectively. The codebook entropy in the K-means clustering case is therefore

$$H_{\text{Km}} = \log K_{\text{Km}} - \frac{r}{d+r}\Big(\Pi_{\text{high}} V_{\text{high}} \log \Pi_{\text{high}} + \Pi_{\text{low}} V_{\text{low}} \log \Pi_{\text{low}}\Big) - \log\left[\Pi_{\text{high}}^{d/(d+r)} V_{\text{high}} + \Pi_{\text{low}}^{d/(d+r)} V_{\text{low}}\right], \tag{40}$$

while in case of entropy coding complexity we find

$$\mathcal{H}_{\text{ent}} = \log \frac{K_{\text{ent}}}{V_{\text{high}} + V_{\text{low}}} - \Pi_{\text{high}} V_{\text{high}} \log \Pi_{\text{high}} - \Pi_{\text{low}} V_{\text{low}} \log \Pi_{\text{low}}. \tag{41}$$

The difference of these entropies is

$$\begin{aligned} \mathcal{H}_{\text{Km}} - \mathcal{H}_{\text{ent}} &= \log\left(\frac{K_{\text{Km}}}{K_{\text{ent}}}(V_{\text{high}} + V_{\text{low}})\right) - \log\left[\Pi_{\text{high}}^{d/(d+r)} V_{\text{high}} + \Pi_{\text{low}}^{d/(d+r)} V_{\text{low}}\right] \\ &\quad + \frac{d}{d+r}\left[\Pi_{\text{high}} V_{\text{high}} \log \Pi_{\text{high}} + \Pi_{\text{low}} V_{\text{low}} \log \Pi_{\text{low}}\right]. \end{aligned} \tag{42}$$

Demanding identical overall distortion errors we find using (38)

$$\mathcal{H}_{\text{Km}} - \mathcal{H}_{\text{ent}} = \frac{d}{r} \log\left[\Pi_{\text{high}}^{d/(d+r)} V_{\text{high}} + \Pi_{\text{low}}^{d/(d+r)} V_{\text{low}}\right] + \frac{d}{d+r}\left[\Pi_{\text{high}} V_{\text{high}} \log \Pi_{\text{high}} + \Pi_{\text{low}} V_{\text{low}} \log \Pi_{\text{low}}\right]. \tag{43}$$

For identical outlier distortion we insert Eq. (39) in Eq. (42) and we derive the entropy difference

$$\mathcal{H}_{\text{Km}} - \mathcal{H}_{\text{ent}} = -\frac{d}{d+r} \log \Pi_{\text{low}} + \frac{d}{d+r}\left[\Pi_{\text{high}} V_{\text{high}} \log \Pi_{\text{high}} + \Pi_{\text{low}} V_{\text{low}} \log \Pi_{\text{low}}\right]. \tag{44}$$

A specific choice for the parameters of a data distribution is $V_{\text{high}} = 1$, $V_{\text{low}} = (2\Pi_{\text{low}})^{-1/(d+r)}$ and $\Pi_{\text{high}} = 1 - 2^{-d/(d+r)} \Pi_{\text{low}}^{r/(d+r)}$. This scaling behavior of the low-density region ensures that its distortion error does not vanish if we take the limit $\Pi_{\text{low}} \to 0$.

Calculations of the relative gain by entropy encoding $\chi \equiv (H_{\text{Km}} - H_{\text{ent}})/H_{\text{ent}}$ in the case of identical overall distortion show improvements of up to 20% for $K_{\text{ent}} = 20$. The gain, however, increases dramatically if we require identical distortion error in the outlier region. The difference in encoding costs diverges logarithmically in the limit $\Pi_{\text{low}} \to 0$. The ratio of the codebook sizes (39) diverges as well in that limit. The divergence means that we need much larger codebooks for K-means clustering to describe sparse data than for entropy encoding.

The approximations introduced to derive (43) and (44) are asymptotically valid. They should therefore serve only as a crude estimate of the potential efficiency gain. To test the quality of our estimate of χ in the case of identical distortion error in the outlier region, we have quantized one-dimensional step distributions with two different densities $\Pi_{\text{low}} = 0.1$ and $\Pi_{\text{low}} = 0.05$ in the low-density part of the distribution and with the other parameters adjusted as stated above. The simulation results and the theoretical estimates (in brackets) are shown in Table 1.

The simulation results are consistent with the χ values derived from (43) and (44). Deviations from the theoretical value are caused by locally optimal arrangements of reference vectors which are not globally optimal. We conclude that it pays off to introduce a problem adequate complexity measure for vector quantizer design when too large distortions in outlier regions cannot be tolerated, an issue that will be of importance in the next section.

(I) $\Pi_{\rm low} = 0.1,\ \Pi_{\rm high} = 0.83,\ V_{\rm high} = 1,\ V_{\rm low} = 1.71,\ N = 5000$

λ	$\mathcal{D}_{\rm ent}\ \times 10^{-5}$	$\mathcal{D}_{\rm Km}\ \times 10^{-5}$	$K_{\rm ent}$	$K_{\rm Km}$	χ
0.0020	17.7–22.4	19.4–23.0	23–25	30–32 (33)	15.8–17.5% (21.4%)
0.0010	7.6–8.5	8.4–8.7	35–37	48–50 (48)	18.2–20.9% (18.8%)
0.0005	3.4–3.9	3.5–3.9	51–53	69–73 (70)	16.3–17.3% (16.7%)

(II) $\Pi_{\rm low} = 0.05,\ \Pi_{\rm high} = 0.89,\ V_{\rm high} = 1,\ V_{\rm low} = 2.15,\ N = 5000$

λ	$\mathcal{D}_{\rm ent}\ \times 10^{-5}$	$\mathcal{D}_{\rm Km}\ \times 10^{-5}$	$K_{\rm ent}$	$K_{\rm Km}$	χ
0.0020	8.4–11.3	9.2–12.4	29–31	42–46 (45)	29.0–36.0% (32.0%)
0.0010	4.6–5.3	4.8–5.6	40–42	59–61 (60)	27.9–28.8% (28.9%)
0.0005	2.2–2.6	2.2–2.6	56–59	81–87 (87)	23.1–24.9% (25.7%)

Table 1: Comparison of quantization results obtained by K-means clustering and logarithmic complexity. χ is the gain in quantization efficiency applying the constraint of equal outlier distortions. The values in brackets are the theoretical estimates according to (43) and (44). Table I and II show results for one-dimensional distributions and squared Euclidean ($r = 2$) distortion costs and for two different data densities $\Pi_{\rm low}$ in the low-density region of the distribution. The density in the high-density region and the size V of the high- and low-density regions are chosen as described in the text. For every λ value in Table I (II) four (eight) different configurations were generated using logarithmic complexity costs.

7 Compression of Wavelet-decomposed Images

Lossy image compression is a "*real-world*" information processing problem which is well suited to study the influence of different complexity measures on the performance of vector quantization algorithms. Image compression based on orthogonal [45, 36, 15] or nonorthogonal [14, 44] wavelet decompositions has witnessed increasing popularity in recent years. The wavelet data format is supposedly optimal for natural image representation since the coefficients are statistically independent if we average over a large set of natural images [21, 44]. Quantization of wavelet coefficients yields a psychophysically more pleasing image quality [15] than quantization schemes which are based on raw pixel blocks or on the discrete cosine transform of pixel blocks.

In this section we use the image compression task to compare the efficiency of the K-means clustering scheme ($\mathcal{C}_\alpha = 1/p_\alpha$) with entropy optimized vector quantization ($\mathcal{C}_\alpha = -\log p_\alpha$). The compression results have been published in part [7]. The images are decomposed with the wavelet algorithm of Mallat [36] which uses quadrature mirror filtering for the subband decomposition. The partitioning of the Fourier plane by Mallat's algorithm is shown in Fig. 2a. The filter functions $\Psi^\rho, \rho = 1, 2, 3$ decompose an image $\mathcal{I}$ in a lowpass filtered image $A_{2^{-1}}\mathcal{I}$ and a bandpassed image signal represented by the wavelet coefficients $D^\rho_{2^{-1}}\mathcal{I},\ \rho = 1, 2, 3$. These wavelet coefficients essentially are a combination of lowpass filtering in x-direction and bandpass filtering in y-direction (Ψ^1) or vice versa (Ψ^2) or bandpass filtering in both directions (Ψ^3). The coefficients of a two band wavelet decomposition can be conveniently arranged in matrix form as shown in Fig. 2b.

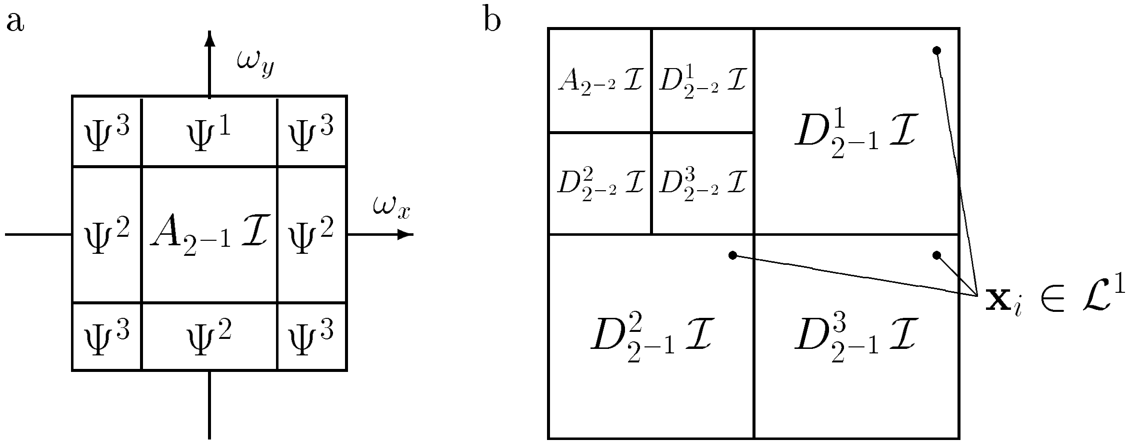
a
ω_y
ω_x
Ψ^3
Ψ^1
Ψ^3
Ψ^2
$A_{2^{-1}}\,\mathcal{I}$
Ψ^2
Ψ^3
Ψ^2
Ψ^3
b
$A_{2^{-2}}\,\mathcal{I}$
$D^1_{2^{-2}}\,\mathcal{I}$
$D^2_{2^{-2}}\,\mathcal{I}$
$D^3_{2^{-2}}\,\mathcal{I}$
$D^1_{2^{-1}}\,\mathcal{I}$
$D^2_{2^{-1}}\,\mathcal{I}$
$D^3_{2^{-1}}\,\mathcal{I}$
$\mathbf{x}_i \in \mathcal{L}^1$

Figure 2: (a) Approximate frequency support of the wavelet filters Ψ^1, Ψ^2 and Ψ^3 in Fourier space. (b) Wavelet coefficients of image $\mathcal{I}$ arranged in matrix form: The upper index $\nu \in \{1, 2, 3\}$ of $D^{\nu}_{2^{-\rho}}\,\mathcal{I}$ refers to the respective filter function Ψ^{ν}, the lower index ρ denotes the reduction factor. $A_{2^{-1}}\,\mathcal{I}$ and $A_{2^{-2}}\,\mathcal{I}$ are the lowpass filtered images after one or two reductions steps, respectively. Filter coefficients from a particular image position i are combined to a three-dimensional vector $\mathbf{x}_i$.

To compare the quantization efficiency of entropy optimized VQ with K-means clustering we transform 8 bit gray level images (size 128×128) of human faces using a two band wavelet transformation. The resulting coefficients are combined to three dimensional vectors for the bandpassed signals and scalar values for the lowpassed image. In each frequency band we combine the three wavelet coefficients $(D^{\nu}_{2^{-\rho}}\mathcal{I})\,(i)$, $\nu = 1, 2, 3$ at position i (notation as in [36]) to a three-dimensional vector $\mathbf{x}_i$ (see Fig. 2). The index ρ denotes the reduction factor $2^{-\rho}$ of the bandpass filters. On the highest frequency level ($\rho = 1$) this data set, refered to as $\mathcal{L}^1 = \Big\{\mathbf{x}_i = \Big((D^1_{2^{-1}}\mathcal{I})\,(i), (D^2_{2^{-1}}\mathcal{I})\,(i), (D^3_{2^{-1}}\mathcal{I})\,(i)\Big)\Big\}$, is of size 4096 vectors. The intermediate ($\rho = 2$) frequency set $\mathcal{L}^2 = \Big\{\mathbf{x}_i = \Big((D^1_{2^{-2}}\mathcal{I})\,(i), (D^2_{2^{-2}}\mathcal{I})\,(i), (D^3_{2^{-2}}\mathcal{I})\,(i)\Big)\Big\}$ has size 1024. A lowpass filtered image of size 32×32 forms the set $\mathcal{A}^2 = \Big\{x_i = \Big(A_{2^{-2}}\mathcal{I}\Big)(i)\Big\}$ of 1024 scalar values.

The training data for the K-means algorithm and the complexity optimized vector quantizer are the union of $\mathcal{L}^2$-sets, the union of $\mathcal{L}^1$-sets and the union of $\mathcal{A}^2$ sets, all taken from 10 different face images. During the training stage we solve Eqs. (18,19) for each of these training sets and for both complexity measures under comparison. This procedure yields three codebooks for entropy optimized quantization and three codebooks for K-means clustering. All distortions are measured as the squared Euclidian distance $\mathcal{D}_{i\alpha}(\mathbf{x}_i, \mathbf{y}_\alpha) = (\mathbf{x}_i - \mathbf{y}_\alpha)^2$. The complexity weights λ in the K-means clustering case are adjusted such that the compression ratio of both quantization schemes is the same, i.e., we require that K-means clustering and entropy optimized clustering compress the training data with the same resulting entropy. The error between an original test image and its reconstruction from the quantized wavelet representation allows us to quantify the efficiency of the complexity measures $\mathcal{C}_\alpha = 1/p_\alpha$ and $\mathcal{C}_\alpha = -\log p_\alpha$ for image compression. The reconstruction error is measured as the absolute difference between the original pixel value $\mathcal{I}(i)$ and its reconstructed value $\mathcal{I}^c(i)$, i.e.,

$\frac{1}{N}\sum_{i=1}^{N}|\mathcal{I}(i) - \mathcal{I}^c(i)|$. Such an error measure supposedly is compatible with the sensitivity of the visual system [15].

In the first series of compression experiments we set $\lambda = 5$ for entropy optimized vector quantization of $\mathcal{L}^1$ and $\mathcal{L}^2$ and $\lambda = 0.5$ for entropy optimized scalar quantization of $\mathcal{A}^2$. The resulting codebooks for $\mathcal{L}^1$, $\mathcal{L}^2$, $\mathcal{A}^2$ had sizes $K = 112, 123, 146$, respectively. We have not optimized the bit allocation per frequency level for our specific training set but used the same complexity parameter $\lambda = 5$ for $\mathcal{L}^1$ and $\mathcal{L}^2$ quantization. Furthermore, we introduced a small constant cost term of $32 \times d$ bits per reference vector to reflect the costs for transmitting or storing the codebook. That cost term, which is independent of the data volume and whose relative contribution to the total quantization costs vanishes for $N \to \infty$, prevents very sparsely populated reference vectors with less than 5 data vectors assigned to. The codebooks for K-means clustering were determined such that they quantized the training data generating the same entropy as the codebooks for entropy optimized quantization. (The resulting entropies of the test data might differ for both quantization schemes since single images do not exactly match the statistics of the training set). The K-means clustering codebooks for $\mathcal{L}^1$, $\mathcal{L}^2$, $\mathcal{A}^2$ had sizes $K = 12, 40, 104$. Note that the codebooks for K-means clustering are much smaller than the codebooks for entropy optimized quantization. This computational disadvantage of entropy optimized quantization is outweighted by the considerably smaller error rate of compressed images. It also has to be mentioned that not all codebook vectors are used in the compression experiments due to differences between the statistics of individual images and the statistics of the training set.

The following table 2 summarizes four different compression experiments. The first three test images were taken from the same gallery as the training images, the fourth face image (terry) was generated under different lighting conditions with another camera setting. The entropy optimized quantizers produce a 10%-25% smaller error for comparable or superior compression ratios than K-means clustering. Furthermore, face images which have been compressed by optimizing the complexity $\mathcal{C}_\alpha = -\log p_\alpha$, look sharper and more natural than the results from K-means clustering. Psychophysically important image features like edges generate large wavelet amplitudes in the high frequency band. The corresponding outlier regions of $\mathcal{L}^1$ and $\mathcal{L}^2$ are more accurately sampled by entropy optimized quantization than by K-means clustering. Differences in the peak signal to noise ratios (SNR) are listed in the table to make a comparison with published compression results [40] possible.

Images of a compression and reconstruction experiment are shown in Fig. 3. We have used a complexity parameter $\lambda = 50$ to design entropy optimized codebooks for $\mathcal{L}^1$, $\mathcal{L}^2$ and $\lambda = 5$ to design an entropy optimized codebook for $\mathcal{A}^2$. The respective codebook sizes are $K = 11, 15, 50$ for $\mathcal{L}^1$, $\mathcal{L}^2$, $\mathcal{A}^2$. The resulting compression ratio is 24.5. The corresponding K-means clustering codebooks with the same resulting entropy on the training set have sizes $K = 4, 5, 28$ for $\mathcal{L}^1$, $\mathcal{L}^2$, $\mathcal{A}^2$. The high complexity costs have been chosen to demonstrate the different types of reconstruction errors caused by the two vector quantization schemes. Image (a) is the original, the first row shows reconstructions of compressed, wavelet decomposed image using entropy optimized vector quantization (b) or K-means clustering (c). The second row (d,e) shows the differences between the original image and the compressed and reconstructed images (b,c), respectively. According to our efficiency criterion entropy optimized compression is 36.8% more efficient than K-means clustering for a compression factor 24.5. The peak SNR values for (b,c) are 30.1 and 27.1, respectively. The considerable higher

face	pixel	entropy of			average	peak	compr.	efficiency
images	entropy	$\mathcal{L}^1$	$\mathcal{L}^2$	A^2	error	SNR	ratio	
doro	5.9	1.57	1.01	1.47	2.83	35.91	14.8	11.0% (4.2%)
		1.77	1.05	1.44	3.14	34.45	13.4	
brix	4.6	1.88	1.05	1.31	2.46	35.75	12.9	22.4% (7.2%)
		1.81	1.05	1.28	3.01	33.35	13.2	
yildir	6.2	1.76	1.04	1.76	3.08	35.06	13.3	17.8% (6.3%)
		1.82	1.07	1.82	3.63	32.99	13.0	
terry	7.0	2.39	1.18	1.55	3.60	33.69	10.4	29.4% (11.9%)
		2.17	1.15	1.49	4.66	30.11	11.3	

Table 2: Summary of four compression trials: The upper numbers in the stacks refer to entropy optimized quantization, the lower numbers result from K-means clustering experiments. The average error is the average, absolute difference between the original pixel value $\mathcal{I}(i)$ and its reconstructed value $\mathcal{I}^c(i)$, i.e., $\frac{1}{N}\sum_{i=1}^{N}|\mathcal{I}(i)-\mathcal{I}^c(i)|$. The efficiency is defined as the relative difference in reconstruction error between the two quantization methods. The efficiency values in brackets are the gains measured by peak SNR changes.

error near edges in the reconstruction based on K-means clustering (e) demonstrates that entropy optimized clustering of wavelet coefficients preserves psychophysically important image features like edges more faithfully than conventional compression schemes.

8 Online Vector Quantization

The described vector quantization procedures are *batch* algorithms since they require all data points $\{\mathbf{x}_i\}$ to be available at the same time, a very restrictive constraint for applications dealing with a large data volume. *Online* vector quantization algorithms process a data stream sequentially by updating $\mathbf{y}_\alpha$ and p_α in an iterative fashion [35, 30]. But how do we have to change the reference vectors and the cluster probabilities after data point $\mathbf{x}_N$ has been processed? This question is normally studied in stochastic approximation theory. We will estimate the most likely changes of the codebook $\{\mathbf{y}_\alpha\}$ by making a Taylor expansion of the reestimation equations (18,19), i.e., we assume that the changes $\Delta p_\alpha \equiv p_\alpha^{(N)} - p_\alpha^{(N-1)}, \Delta\mathbf{y}_\alpha \equiv \mathbf{y}_\alpha^{(N)} - \mathbf{y}_\alpha^{(N-1)}$ are of order $\mathcal{O}(1/N)$. The upper indizes (N) and $(N-1)$ denote the estimates of p_α, $\mathbf{y}_\alpha$ at the time steps $t = N$ and $t = N-1$, respectively. The update rule for $p_\alpha, \mathbf{y}_\alpha$ after $N-1$ data points have been processed is

$$\Delta p_\alpha \equiv p_\alpha^{(N)} - p_\alpha^{(N-1)} \quad = \quad \frac{1}{N}\left(\langle M_{N,\alpha}\rangle - p_\alpha^{(N-1)}\right), \tag{45}$$

$$\Delta \mathbf{y}_\alpha \equiv \mathbf{y}_\alpha^{(N)} - \mathbf{y}_\alpha^{(N-1)} = -\frac{\sum_\gamma \langle M_{N,\gamma}\rangle \mathtt{T}_{\gamma\alpha} \frac{\partial}{\partial \mathbf{y}_\alpha} \mathcal{D}_{N,\alpha}(\mathbf{y}_\alpha^{(N-1)}, \mathbf{x}_N)}{\sum_{i=1}^{N} \sum_\gamma \langle M_{i,\gamma}\rangle \mathtt{T}_{\gamma\alpha} \frac{\partial^2}{\partial {\mathbf{y}_\alpha}^2} \mathcal{D}_{i,\alpha}(\mathbf{y}_\alpha^{(N-1)}, \mathbf{x}_i)}. \tag{46}$$

To derive equation (45,46) we have expanded the equations (18,19) up to linear terms in $\Delta p_\alpha, \Delta \mathbf{y}_\alpha$, keeping $\langle M_{N,\alpha}\rangle$ fixed. The changes proportional to

$$\begin{aligned} \frac{\partial}{\partial \mathbf{y}_\alpha} \langle M_{i\alpha}\rangle &= -\frac{1}{T} \langle M_{i\alpha}\rangle \left(1 - \langle M_{i\alpha}\rangle\right) \frac{\partial}{\partial \mathbf{y}_\alpha} \mathcal{D}_{i\alpha} \\ \frac{\partial}{\partial \mathbf{y}_\nu} \langle M_{i\alpha}\rangle &= \frac{1}{T} \langle M_{i\alpha}\rangle \langle M_{i\nu}\rangle \frac{\partial}{\partial \mathbf{y}_\nu} \mathcal{D}_{i\nu}, \quad \nu \neq \alpha \end{aligned}$$

disappear in the hard clustering limit exponentially fast, i.e., $\mathcal{O}(\exp(-C/T)/T)$ for $T \to 0$ with C being a positive constant independent of T. In the case of squared Euclidean distortion costs ($\mathcal{D}_{i\alpha} = \|\mathbf{x}_i - \mathbf{y}_\alpha\|^2$) equation (46) reduces to the form

$$\Delta \mathbf{y}_\alpha = \frac{1}{N \sum_\gamma \mathtt{T}_{\gamma\alpha} p_\gamma^N} \sum_\gamma \mathtt{T}_{\gamma\alpha} \langle M_{N,\gamma}\rangle \left(\mathbf{x}_N - \mathbf{y}_\alpha^{(N-1)}\right) \tag{47}$$

which corresponds to MacQueens update rule for K-means clustering [35] and which is similar to learning in competitive neural networks. $\mathbf{y}_\alpha^{(N-1)}$ is moved towards the most recent data point $\mathbf{x}_N$ proportionally to the error $(\mathbf{x}_N - \mathbf{y}_\alpha^{(N-1)})$ and proportional to $\mathbf{x}_N$'s effective membership $\sum_\gamma \mathtt{T}_{\gamma\alpha} \langle M_{N,\gamma}\rangle$ in cluster α. In addition to that, the update formula (47) weights any change in $\mathbf{y}_\alpha^{(N)}$ by the effective number of data points $N \sum_\gamma \mathtt{T}_{\gamma\alpha} p_\gamma^{(N)}$ which are already assigned to cluster α. The learning rate $1/(N \sum_\gamma \mathtt{T}_{\gamma\alpha} p_\gamma^{(N)})$ treats different clusters according to their history. That generalizes conventional topological feature maps which suggest the same learning rate for all clusters [30]. The question if there exists any faster learning rate schedule $c/(N \sum_\gamma \mathtt{T}_{\gamma\alpha} p_\gamma^{(N)})$ with $c > 1$ than the schedule proposed by (47) is still open although numerical simulations [12, 13] suggest that it is advisable to choose $c > 1$ to speed up the convergence.

Online optimization of the codebook size K relies on a heuristics for cluster merging and cluster creation. We have explored the following heuristics for cluster creation: A data point $\mathbf{x}_i$ initializes a new reference vector $K + 1$ with $\mathbf{y}_{K+1} = \mathbf{x}_i$ if the costs of assigning $\mathbf{x}_i$ to an already existing cluster exceeds the complexity costs of the new codebook vector $K + 1$, i.e., $\mathcal{C}_{K+1} < \min_\alpha (\|\mathbf{x}_i - \mathbf{y}_\alpha\|^r + \lambda \mathcal{C}_\alpha)$. We found in a series of quantization experiments that this strategy causes a slight overestimation of the optimal codebook size but the resulting codebooks have comparable quantization costs to codebooks found in batch optimization.

9 Discussion

Vector quantization and data clustering have a wide spectrum of engineering applications ranging from data compression for transmission and storage purposes to image and speech processing for pattern recognition tasks [1]. When we assign partitions of a data set to a reduced set of reference vectors we have to make a compromise between the simplicity and precision of the resulting representation, e.g., between the size of the codebook and the distortion error due to data quantization. The key point of our paper is to explicitly express this compromise in a cost function that comprises both complexity and distortion costs. Joint optimization of various complexity and distortion terms results in an intrinsic limitation of the number of reference vectors, in contrast to other currently known approaches like K-means clustering or the LBG algorithm.

We have applied the maximum entropy principle and the formalism of statistical mechanics to estimate optimal solutions of our vector quantization problem. There exist several distinct advantage of the statistical mechanics approach for the design of vector quantizers:

1. Maximizing the entropy allows us to study a variety of different distortion measures and complexity measures and to derive a system of reestimation equations for the reference vectors and the assignment probabilities.

2. The resulting algorithm to determine the optimal codebook maps naturally to a simulated annealing approach where a slow, controlled decrease of the computational temperature produces nearly optimal solutions. The computational temperature controls the degree of noise in the data assignments, which essentially interpolates between a hard and a fuzzy clustering solution.

3. The structure of the reestimation equations suggests hardware implementations in analog VLSI as they are known from neural network research. Complexity optimized vector quantization maps onto a two layer neural network with a winner-take-all architecture as discussed in [6]. Hardware implementations of such network architectures have been successfully tested [2, 31].

Our study of entropy optimized quantization of wavelet transformed gray level images revealed the interesting fact that entropy optimized codebooks reproduce sparse image features like edges more faithfully than the conventional K-means clustering approach. Those image features, however, possess a high information content due to their rare occurence and compression errors there impair any subsequent information processing step much more than errors in other parts of data space. The sharp appearence of the reconstruction in Fig. 3b supports this finding.

In a related study we have extended the maximum entropy approach for vector quantization to the case of supervised data clustering [8, 6]. An additional cost term is used to penalize partitionings of data space which are in conflict with a priori known class knowledge. The respective algorithm is implemented by a three layer neural network with classification units in the third layer [6]. We consider the similarity of the discussed algorithms and the underlying reestimation equations for codebook parameters with neural network systems as a very fruitful direction for future research which might not only produce better information processing systems but might also lead to a more fundamental understanding of perception.

Acknowledgement

It is a pleasure to thank H. Sompolinsky, T. Tishby and P. Tavan for helpful discussions and two anonymous reviewers for useful suggestions to improve the manuscript. JB acknowledges the hospitality of the Aspen Center for Physics where this work was initiated during a workshop on Physics of Neural Networks. Part of this work has been performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

A Derivation of the Free Energy $\mathcal{F}_K$

The factor $\exp \beta \mathcal{F}_K$ in the Gibbs distribution (9) can be written as

$$\begin{aligned} \mathcal{Z} &= \exp(-\beta \mathcal{F}_K) = \sum_{\{M\}} \int_{-\infty}^{\infty} \prod_{\nu=1}^{K} dp_\nu d\mathbf{y}_\nu \; \delta(Np_\nu - \sum_i M_{i\nu}) \\ & \quad \delta(\sum_i \sum_\mu M_{i\mu} \mathcal{G}_{\nu\mu}(\mathbf{x}_i, \mathbf{Y})) \; \exp(-\beta \sum_{i\alpha} M_{i\alpha} \mathcal{E}_{i\alpha}). \end{aligned} \tag{48}$$

$\mathcal{Z}$ is known as the partition function in statistical mechanics and plays the role of a generating function. The differential quantization costs $\mathcal{E}_{i\alpha}$ in Eq. (48) are defined by Eq. (11). The sum $\sum_{\{M\}}$ runs over all K^N legal configurations $\{ M_{i\alpha} | \, i = 1, \ldots, N; \alpha = 1, \ldots, K\}$, i.e., over all configurations with $M_{i\alpha} \in \{0, 1\}$ and $\sum_\alpha M_{i\alpha} = 1$. The product of delta functions strictly enforces the constraints $p_\alpha = \sum_{i=1}^{N} M_{i\alpha}/N$ and $\sum_i \sum_\nu M_{i\nu} \mathcal{G}_{\alpha\nu}(\mathbf{x}_i, \mathbf{Y}) = 0$.

The integral representation of δ-functions allows us to rewrite the partition function (48) as

$$\begin{aligned} \mathcal{Z} &= \sum_{\{M\}} \int_{-\infty}^{\infty} \prod_\nu dp_\nu d\mathbf{y}_\nu \int_{-\imath\infty}^{\imath\infty} \prod_\nu d\hat{p}_\nu d\hat{\mathbf{y}}_\nu \Bigg(\exp(-\beta \sum_{i\alpha} M_{i\alpha} \mathcal{E}_{i\alpha}) \\ & \quad \exp \sum_\alpha \hat{p}_\alpha (Np_\alpha - \sum_i M_{i\alpha}) \exp \sum_\alpha \hat{\mathbf{y}}_\alpha (\sum_i \sum_\mu M_{i\mu} \mathcal{G}_{\alpha\mu}(\mathbf{x}_i, \mathbf{Y})) \Bigg). \end{aligned} \tag{49}$$

After summing over all legal configurations of assignment variables $M_{i\alpha}$ in (49) we derive

$$\begin{aligned} \mathcal{Z} &= \int_{-\infty}^{\infty} \prod_\nu dp_\nu d\mathbf{y}_\nu \int_{-\imath\infty}^{\imath\infty} \prod_\nu d\hat{p}_\nu d\hat{\mathbf{y}}_\nu \exp(\beta N \sum_\alpha \hat{p}_\alpha p_\alpha) \\ & \quad \prod_{i=1}^{N} \sum_{\alpha=1}^{K} \exp \left[-\beta \left(\mathcal{E}_{i\alpha} + \hat{p}_\alpha - \sum_{\nu=1}^{K} \hat{\mathbf{y}}_\nu \mathcal{G}_{\nu\alpha}(\mathbf{x}_i, \mathbf{Y}) \right) \right] \\ &= \left(\frac{\beta N}{2\pi} \right)^{2K} \int_{-\infty}^{+\infty} \prod_{\alpha=1}^{K} dp_\alpha d\mathbf{y}_\alpha \int_{-\imath\infty}^{+\imath\infty} \prod_{\alpha=1}^{K} d\hat{p}_\alpha d\hat{\mathbf{y}}_\alpha \exp\Big(-\beta \mathcal{F}'_K(\mathbf{y}_\alpha, \hat{\mathbf{y}}_\alpha, p_\alpha, \hat{p}_\alpha) \Big) \end{aligned} \tag{50}$$

with the exponent being

$$\mathcal{F}'_K = -N \sum_\alpha \hat{p}_\alpha p_\alpha - \frac{1}{\beta} \sum_i \log \left(\sum_\alpha \exp \left(-\beta \left[\mathcal{E}_{i\alpha} - \sum_\nu \hat{\mathbf{y}}_\nu \mathcal{G}_{\nu\alpha}(\mathbf{x}_i, \mathbf{Y}) + \hat{p}_\alpha \right] \right) \right). \tag{51}$$

The function $\mathcal{F}'_K(\mathbf{y}_\alpha, \hat{\mathbf{y}}_\alpha, p_\alpha, \hat{p}_\alpha)$ is extensive, i.e., it scales as $\mathcal{O}(N)$. The integral in (50) is dominated by the global minimum of the function (51) which is called the free energy. Necessary conditions for the global minimum of $\mathcal{F}'_K$ are $\dfrac{\partial \mathcal{F}'_K}{\partial \hat{p}_\alpha} = \dfrac{\partial \mathcal{F}'_K}{\partial p_\alpha} = 0$, $\dfrac{\partial \mathcal{F}'_K}{\partial \mathbf{y}_\alpha} = \dfrac{\partial \mathcal{F}'_K}{\partial \hat{\mathbf{y}}_\alpha} = 0$, which are used to derive the reestimation equations (12–15).

References

[1] S. C. Ahalt, A. K. Krishnamurthy, P. Chen, and D. E. Melton. Competitive learning algorithms for vector quantization. *Neural Networks*, 3:277–290, 1990.

[2] A. G. Andreou, K. A. Boahen, P. O. Pouliquen, A. Pavasović, R. E. Jenkins, and K. Strohbehn. Current mode subthreshold MOS circuits for analog VLSI neural systems. *IEEE Transactions on Neural Networks*, 2:205–213, 1991.

[3] G. Ball and D. Hall. A clustering technique for summarizing multivariate data. *Behavioral Sciences*, 12:153–155, 1967.

[4] J. C. Bezdek. A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(1):1–8, 1980.

[5] L. Bobrowski and J. C. Bezdek. C-means clustering with the l_1 and l_∞ norms. *IEEE Transactions on Systems, Man and Cybernetics*, 21(3):545–554, 1991.

[6] J. Buhmann and H. Kühnel. Complexity optimized data clustering by competitive neural networks. *Neural Computation*, 5(1):(in press), 1992.

[7] J. Buhmann and H. Kühnel. Complexity optimized vector quantization: A neural network approach. In J. Storer, editor, *Data Compression Conference '92*, pages 12–21, Los Alamitos, CA, 1992. IEEE Computer Society Press.

[8] J. Buhmann and H. Kühnel. Unsupervised and supervised data clustering with competitive neural networks. In *IJCNN International Conference on Neural Networks, Baltimore*, pages IV–796 – IV–801. IEEE, 1992.

[9] P. A. Chou, T. Lookabaugh, and R. M. Gray. Entropy-constrained vector quantization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37:31–42, 1989.

[10] P. A. Chou, T. Lookabaugh, and R. M. Gray. Optimal pruning with applications to tree-structured source-coding and modeling. *IEEE Transactions on Information Theory*, 35(2):299–315, 1989.

[11] T. M. Cover. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

[12] C. Darken and J. Moody. Fast adaptive K-means clustering: Some empirical results. In *International Joint Conference on Neural Networks, San Diego*, volume II, pages 233–238. IEEE, June 17-21, 1990 1990.

[13] C. Darken and J. Moody. Towards faster stochastic gradient search. In *Neural Information Processing Systems 4*, San Mateo, California, 1992. Morgan Kaufmann.

[14] J. Daugman. Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1169–1179, 1988.

[15] R. A. DeVore, B. Jawerth, and B. J. Lucier. Image compression through wavelet transform coding. *IEEE Transactions on Information Theory*, 38:719–746, 1992.

[16] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[17] J. C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. Cybernetics*, 3(3):32–57, 1974.

[18] R. Durbin and D. Willshaw. An analogue approach to the travelling salesman problem using an elastic net method. *Nature*, 326:689–691, 1987.

[19] N. Farvardin. A study of vector quantization for noisy channels. *IEEE Transactions on Information Theory*, 36(4):799–809, 1990.

[20] T. Feder and D. H. Greene. Optimal algorithms for approximate clustering. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing (STOC)*, pages 434–444, 1988.

[21] D.J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12):2379–2394, 1987.

[22] A. Gersho. Asymptotically optimal block quantization. *IEEE Transactions on Information Theory*, 25:373–380, July 1979.

[23] H. Gish and J. N. Pierce. Asymptotically efficient quantizing. *IEEE Transactions on Information Theory IT*, 14:676–683, 1968.

[24] R. M. Gray. Vector quantization. *IEEE Acoustics, Speech and Signal Processing Magazine*, pages 4–29, April 1984.

[25] R. Hanson, J. Stutz, and P. Cheeseman. Bayesian classification theory. Technical Report FIA-90-12-7-01, NASA Ames Research Center, 1991.

[26] Y. Hussain and N. Farvardin. Adaptive block transform coding of speech based on LPC vector quantization. *IEEE Transactions on Signal Processing*, 39(12):2611–2620, 1991.

[27] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[28] E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70:939–952, 1982.

[29] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[30] T. Kohonen. *Self–organization and Associative Memory*. Springer, Berlin, 1984.

[31] J. Lazzaro, R. Ryckebusch, M. A. Mahowald, and C. A. Mead. Winner-take-all networks of $O(n)$ complexity. In *Neural Information Processing Systems 1*, pages 703–711, San Mateo, California, 1989. Morgan Kaufmann.

[32] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Transactions on Communications COM*, 28:84–95, 1980.

[33] S. P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[34] S. P. Luttrell. Hierarchical vector quantisation. *IEE Proceedings*, 136:405–413, 1989.

[35] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[36] S. G. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:674–693, 1989.

[37] H. Ritter, T. Martinetz, and K. Schulten. *Neural Computation and Self-organizing Maps*. Addison Wesley, New York, 1992.

[38] K. Rose, E. Gurewitz, and G. Fox. A deterministic annealing approach to clustering. *Pattern Recognition Letters*, 11(11):589–594, 1990.

[39] K. Rose, E. Gurewitz, and G. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948, 1990.

[40] T. Senoo and B. Girod. Vector quantization for entropy coding of image subbands. Preprint, 1992.

[41] P.D. Simic. Statistical mechanics as the underlying theory of "elastic" and "neural" optimizations. *Network*, 1:89–103, 1990.

[42] P.D. Simic. Constrained nets for graph matching and other quadratic assignment problems. *Neural Computation*, 3:268–281, 1991.

[43] Y. Tikochinsky, N.Z. Tishby, and R. D. Levine. Alternative approach to maximum–entropy inference. *Physical Review A*, 30:2638–2644, 1984.

[44] B. Wegmann and C. Zetsche. Statistical dependence between orientation filter outputs used in an human vision based image code. In M. Kunt, editor, *SPIE Proceedings of the Visual Communications and Image Processing'90*, volume 1360, pages 909–923, 1990.

[45] P. H. Westerink, D. E. Boekee, J. Biemond, and J.W. Woods. Subband coding of images using vector quantization. *IEEE Transactions on Communications*, 36:713–719, 1988.

[46] Y. Yamada, S. Tazaki, and R. Gray. Asymptotic performance of block quantizers with difference distortion measures. *IEEE Transactions on Information Theory*, 26:6–14, 1980.

[47] A. L. Yuille. Generalized deformable models, statistical physics, and matching problems. *Neural Computation*, 2(1):1–24, 1990.

[48] L. Zadeh. Fuzzy sets. *Inform. Contr.*, 8:338–353, 1965.

[49] P. L. Zador. Asymptotic quantization error of continuous signals and the quantization dimension. *IEEE Transactions on Information Theory*, 28(2):139–149, 1982.

[50] K. Zeger and A. Gersho. Pseudo-gray coding. *IEEE Transactions on Communications*, 38(12):2147–2158, 1990.

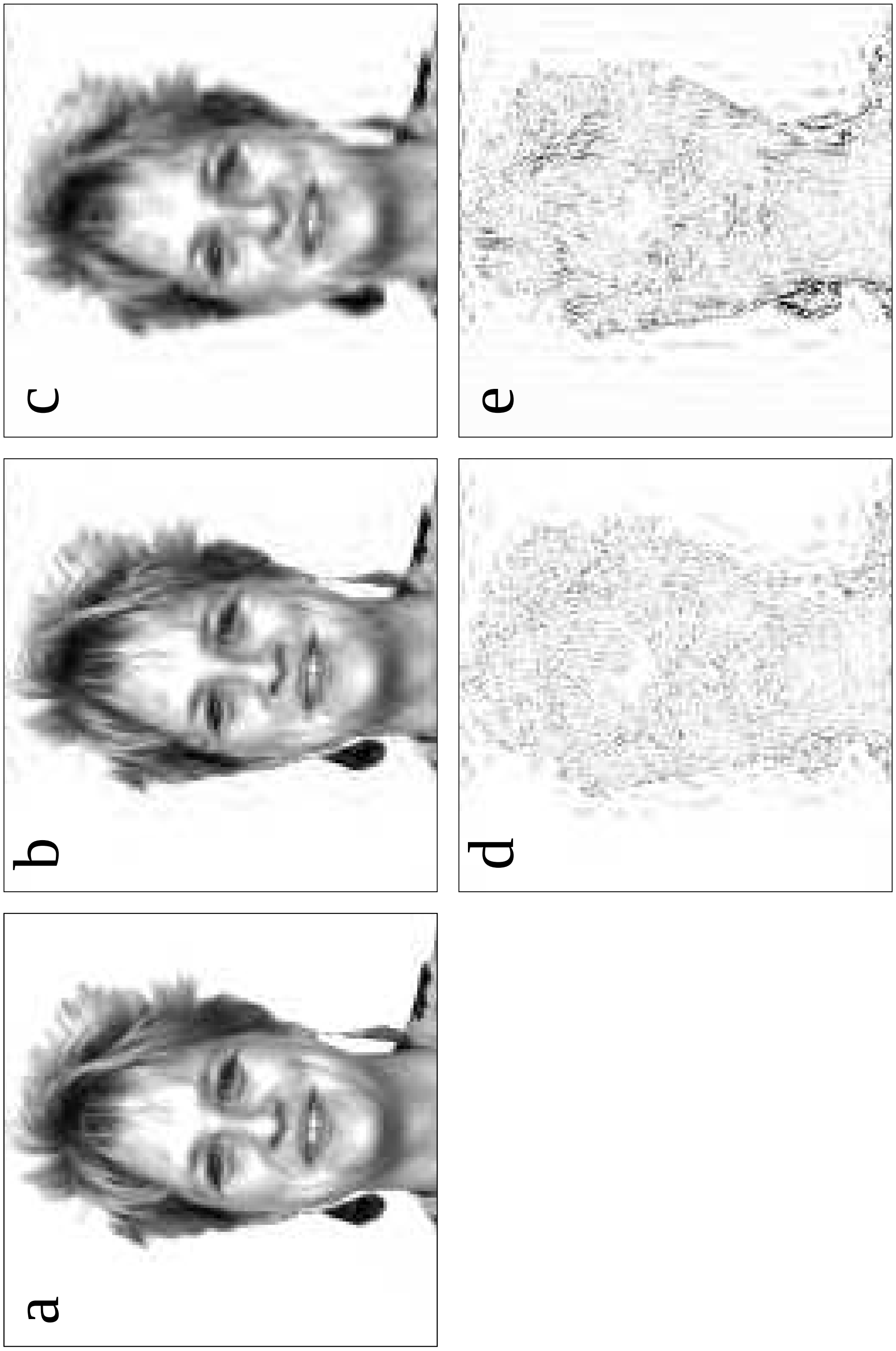

Figure 3: Quantization of a 128×128, 8bit, gray-level image. (a) Original picture. (b) Image reconstruction from wavelet coefficients which were quantized with entropic complexity. (c) Reconstruction from wavelet coefficients quantized by K-means clustering. (d) Reconstruction error of image (b). (e) Reconstruction error of image (c). Black is normalized in image (d) and (e) to a deviation of 92 gray values. Note the large errors near edges in (e).