Johann Wolfgang Goethe-Universität
Frankfurt am Main

Fachbereich Informatik

The Principal Independent Components of Images

B. Arlt, R. Brause

arlt@informatik.uni-frankfurt.de
brause@informatik.uni-frankfurt.de

INTERNAL REPORT 1/98

Fachbereich Informatik
Robert-Mayer-Straße 11-15
60054 Frankfurt am Main

Abstract

Classically, encoding of images by only a few, important components is done by the Principal Component Analysis (PCA). Recently, a data analysis tool called Independent Component Analysis (ICA) for the separation of independent influences in signals has found strong interest in the neural network community. This approach has also been applied to images. Whereas the approach assumes continuous source channels mixed up to the same number of channels by a mixing matrix, we assume that images are composed by only a few image primitives. This means that for images we have less sources than pixels. Additionally, in order to reduce unimportant information, we aim only for the most important source patterns with the highest occurrence probabilities or biggest information called „Principal Independent Components (PIC)".

For the example of a synthetic picture composed by characters this idea gives us the most important ones. Nevertheless, for natural images where no a-priori probabilities can be computed this does not lead to an acceptable reproduction error. Combining the traditional principal component criteria of PCA with the independence property of ICA we obtain a better encoding. It turns out that this definition of PIC implements the classical demand of Shannon's rate distortion theory.

Keywords: *Principal Component Analysis PCA, Independent Component Analysis ICA, Principal Independent Component Analysis PICA, Rate Distortion Theory*

1 Introduction

One of the most interesting and ambitious properties of artificial neural networks is grounded in the active information processing of real world data: the unsupervised analysis of signals.

1.1 Principal components and PCA

An interesting approach has been developed throughout the recent years: the linear transformation of the input space to the base of principal components which minimizes the mean squared error when dropping some of the transformed channels. This transformation called 'Principal Component Analysis' (PCA) and obtained by aligning the base vectors to the directions of maximal variance, is identical to a discrete Karhunen-Loève or Hotelling transformation.

Here, we decompose the n signals $(x_1,\ldots,x_n)^T \equiv \mathbf{x}$ by a linear transform

$$\mathbf{y} = \mathbf{W}\mathbf{x} \qquad \text{with } \mathbf{y} = (y_1,\ldots,y_n)^T \quad (1)$$

such that a subset $\mathbf{y'} = (y_1,\ldots,y_m)^T$ of $m < n$ components used with the matrix $\mathbf{W}_m^{-1}$ (consisting of m columns of $\mathbf{W}^{-1}$) to reconstruct the original signals by

$$\mathbf{x'} = \mathbf{W}_m^{-1}\mathbf{y'}$$

obtains the smallest mean squared error

$$\langle(\mathbf{x}-\mathbf{x'})^2\rangle = \min$$

in the reconstruction process. It is well known that this is the case for the projections of the input $\mathbf{x}$ on the m eigenvectors with the biggest eigenvalues $\lambda_1,\ldots,\lambda_m$ of the covariance matrix

$$\mathbf{C}_{xx}=\langle(\mathbf{x}-\langle\mathbf{x}\rangle)(\mathbf{x}-\langle\mathbf{x}\rangle)^T\rangle$$

Thus, the variance of a component y_i is given by

$$\lambda_i = \langle(y_i-\langle y_i\rangle)^2\rangle = \sigma_i^2,$$

and the rows of $\mathbf{W}$ meet the conditions for orthonormality

$$\mathbf{w}_i^T\mathbf{w}_i = 1 \text{ and } \mathbf{w}_i^T\mathbf{w}_j = 0 \text{ for } i \neq j \qquad (2)$$

We see that the whole signal is decomposed by a non-scaling linear transformation into different directions $\mathbf{w}_i$. To obtain the smallest error of reconstruction, we use the directions with the biggest variances. So, the components (and the corresponding directions or base vectors) are ordered according to a criterion. The selected m ones are called the 'principal components'.

Many neural networks have already been proposed which let their associated weight vectors converge to the base of principal components, the eigenvectors of the input covariance matrix, by proper learning rules, see e.g. [OJA92], [BRA93a].

For images, the search for the principal components (called "transform image coding") can be organized as a local process. Thus, a whole picture can be encoded in parallel by many neurons on a sensory plane with local interactions (e.g. lateral inhibition), using only the self-organized principal components [BRA96] obtained by analog circuits [BRA94].

1.2 Independent components and ICA

The approach of PCA is only optimal for the performance measure of the mean squared error and assumes no specific information about the higher order statistics of the observed signals. If we want to maximize other measures of information processing, for instance the information capacity of the encoding coefficients (i.e. the output signals of the transforming system), we have to obtain other properties.

Here, the mutual information $H(y_1;y_2;\ldots;y_n)$ between the output channels is a good measure for an efficient output coding. The output information $H(y_1,y_2)$ of two channels y_1 and y_2

$$H(y_1,y_2) = H(y_1) + H(y_2) - H(y_1;y_2)$$

becomes maximal if for constant channel information $H(y_i)$ the mutual information becomes minimal. This is the case if

$$H(y_1,y_2) = H(y_1) + H(y_2)$$

which means for the probability density functions (pdf)

$$p(y_1,y_2) = p(y_1)p(y_2)$$

Thus, the demand for minimal transinformation is identical with the demand for independent channel pdf ("factorial code"). For *n* channels this means

$$p(\mathbf{x}) = p(x_1)p(x_2)\cdots p(x_n) \qquad (3)$$

Let us assume that all observed signals $\mathbf{x} = (x_1,\ldots,x_n)^T$ are derived from a linear mixture of *n* unknown independent source signals $\mathbf{s} = (s_1,\ldots,s_n)^T$ with an unknown mixing matrix **M** with rows $\mathbf{m}_i$

$$\mathbf{x} = \mathbf{Ms}, \qquad x_i = \mathbf{m}_i\mathbf{s} \qquad (4)$$

How can the original source signals be reconstituted? Another linear transformation with a matrix **B**

$$\mathbf{y} = \mathbf{Bx} = \mathbf{BMs} \qquad (5)$$

might obtain the sources if

$$\mathbf{y} = \mathbf{s} \quad \Leftrightarrow \quad \mathbf{BM} = \mathbf{I} \qquad (6)$$

the demixing matrix **B** becomes the inverse of **M.**

The problem of finding the demixing matrix is known as the problem of "blind separation of sources" or "Independent Component Analysis" (ICA) and is a fast growing topic in neural network research, see e.g. [ACY96], [BUR92],[COM94], [DEO96], [HYO96].

The independent signals are obtained by using objective functions (called 'contrast functions' [COM94]). One of them is the demand for minimal transinformation between the signals and can be used to obtain learning rules for the unknown base vectors of the inverse transformation **B** of ICA, see [ACY96].

There are several conditions involved in the demixing process in order to get the source signals (see [COM94]):

- The mixing matrix **M** must be regular to have the inverse $\mathbf{B}=\mathbf{M}^{-1}$ to exist with $\mathbf{Bx} = \mathbf{BMs} = \mathbf{M}^{-1}\mathbf{Ms} = \mathbf{s}$. This means that we have to have the same number *n* of sources as of observed mixtures.
- The source is determined regardless of the order (index) of the channels in **s**. This is due to the fact that the crucial condition for independence, the factorization $p(\mathbf{s}) = p(s_1)p(s_2)\cdots p(s_n)$ of the probability distribution function (pdf) by the marginal pdfs, is still valid for $p(\mathbf{s}) = p(s_1)p(s_n)\cdots p(s_2)$ or any other permutation of the indices.
- In eq.(4), the same mixture **x** is produced if we scale a source s_i by a factor c_i and the corresponding column M_i of **M** by a factor $1/c_i$. Thus, without further knowledge, we cannot determine the scale of the source signals: the ICA is an "ill-posed problem".
- For two Gaussian sources s_1 and s_2 a simple decorrelation procedure (PCA) gives us independent sources. Nevertheless, it is well known that the PCA decorrelation is done by an orthogonal matrix composed by the eigenvectors of $\mathbf{C}_{xx}$, see eq.(2). Since we assume **M** to be generally not orthogonal (i.e. it does perform more than a rotation), we cannot demix the signals just by a rotation: the demixing is not correct. The operation of separating the signals into s_1 and s_2 is not unique; without any further information the ambiguity for Gaussian signals cannot be resolved. For additional Gaussian sources, this problem aggravates. This means for successful demixing at most one source can have a pdf with Gaussian characteristic.

Thus, we cannot expect to recover the exact source signals **s** but only their scaled and permutated versions

$$\mathbf{y} = \mathbf{DPs}$$

with a diagonal scaling matrix **D** and a permutation matrix **P**. This relaxes the conditions on the demixing matrix **B** in eq.(5) to

$$\mathbf{BM} = \mathbf{DP} \qquad (7)$$

Here, **B** is in general not equal to $\mathbf{M}^{-1}$ although in the following we still call **B** "the inverse matrix of **M**" and **y** "the source signals".

In order to enable a solution it is convenient to assume that the recovered source signals y_i have unit variance σ_i^2 since **D** is unknown. Furthermore we assume that the y_i are centered, i.e. $\langle\mathbf{y}\rangle \equiv \mathbf{0}$. This requires the demixing process to center the observed signals **x** as well for their average $\langle\mathbf{x}\rangle$ might be non-zero. Consequently, we get the relation

$$\mathbf{y} = \mathbf{B}\,(\mathbf{x}-\langle\mathbf{x}\rangle) = \mathbf{BM}\,(\mathbf{s}-\langle\mathbf{s}\rangle) = \mathbf{DP}\,(\mathbf{s}-\langle\mathbf{s}\rangle) \qquad (8)$$

The standard ICA procedure consists mainly of the following stages (shown in Fig.1).

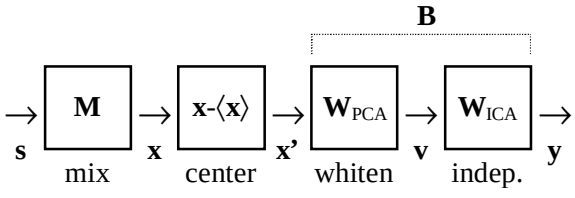

Fig.1 *The processing stages in ICA*

The observed signals **x** are diminished by their first and second moments: They are centered, decorrelated and whitened to unit variance by a linear transform with a matrix $\mathbf{W}_{PCA}$, and then separated by their higher moments in the last stage by another linear transform $\mathbf{W}_{ICA}$. The latter which uses the preprocessed input is often referred as "the ICA matrix". So far we have

$$\mathbf{y} = \mathbf{B}\,(\mathbf{x}-\langle\mathbf{x}\rangle) = \mathbf{W}_{ICA}\mathbf{W}_{PCA}\,(\mathbf{x}-\langle\mathbf{x}\rangle) \qquad (9)$$

with $\mathbf{B} = \mathbf{W}_{ICA}\mathbf{W}_{PCA}$.

If we use a PCA process for the decorrelation process in $\mathbf{W}_{PCA}$ we also can additionally scale the rows $\mathbf{w}_i$ of $\mathbf{W}_{PCA}$ which are the eigenvectors of $\mathbf{C}_{xx}$ by their eigenvalues

$$\mathbf{w}_i \rightarrow \mathbf{w}_i\,\lambda_i^{-1/2} \text{ such that } \mathbf{w}_i^2 = \lambda_i^{-1}$$

This normalizes the variance of $\mathbf{v}$ because we have

$$\langle\mathbf{v}^2\rangle = \langle(\mathbf{w}_i^T\mathbf{x'})^2\rangle = \mathbf{w}_i^T\langle\mathbf{x'x'}^T\rangle\mathbf{w}_i = \mathbf{w}_i^T\mathbf{C}_{xx}\mathbf{w}_i$$

$$= \mathbf{w}_i^T\mathbf{w}_i\,\lambda_i = 1$$

The whitening process gives us an advantage: For whitened, decorrelated input of $\langle\mathbf{vv}^T\rangle = \mathbf{I}$ the ICA matrix $\mathbf{W}_{ICA}$ is orthogonal, i.e. just a rotation of the base of the input space. This can be easily shown: With $\mathbf{v} \equiv \mathbf{W}_{PCA}\mathbf{x'}$ and the assumptions of centered and independent sources having unit variance (i.e. $\langle\mathbf{y}\rangle \equiv \mathbf{0}$ and $\langle\mathbf{yy}^T\rangle = \mathbf{I}$), we get

$$\mathbf{I} = \langle\mathbf{yy}^T\rangle = \mathbf{W}_{ICA}\langle\mathbf{vv}^T\rangle\mathbf{W}_{ICA}^T = \mathbf{W}_{ICA}\mathbf{W}_{ICA}^T$$

Thus, the inverse matrix $\mathbf{W}_{ICA}^{-1}$ is identical to the transposed matrix $\mathbf{W}_{ICA}^T$ which implies that $\mathbf{W}_{ICA}$ has to be orthogonal.

The classical ICA encoding system above can be trained using separate layers of neural networks. The first stage is obtained by learning the expectation value as an offset in order to center the input:

$$\mathbf{x}_0(t+1) = \mathbf{x}_0(t) + 1/t\;\;(\mathbf{x}(t)-\mathbf{x}_0(t))$$

For the second stage standard PCA learning rule can be used, see e.g. [OJA92], coupled by a rescaling described above. Otherwise, special whitening learning rules can be used, see [SIL91],[PLUM93],[BRA98]. For the third stage, the ICA layer, one of the ICA learning rules may be taken, e.g. [HYO96].

Now, for encoding pictures by a decomposition with the most important, independent components we will run into trouble. Let us assume that we have just 4 independent visual objects on a picture of 256×256 = 65536 pixels. Certainly, we want to obtain a significant smaller number of outputs to describe the picture than 65536. But if we use less neurons for data compression, this becomes in conflict with the demand of the same number for sources and mixtures, the first condition for ICA cited above. What can we do ?

One common solution, taken in [BES96] and [OLS96a,b] is to cut the images into smaller patches, say 12×12=144 pixels, present many patches of many images (preferably natural scenes) and then make an ICA of the 144 channels. This gives us 144 independent "base pictures".

Nevertheless, not all ICA components are equally important. Some of them are just spurious patterns with a low occurrence probability. Since we want to obtain a stable code which covers most of the input data, we aim for the *m* ICA components with the highest occurrence probability. Here, we encounter a serious problem: how can we order the components, e.g. by an occurrence probability, which the ICA model so far did not provide? In standard ICA applications, all (time series) channels are always present, i.e. equally probable.

However, this is not the case for real world objects. In order to cover this aspect also, we have to develop a new image model which is composed by signals and events.

2 An event-oriented image model

Let us model the images as a superposition of many small, independent image patches, just like a single neuron of the retina sees the world by a very restricted focus. Our task consists now of finding the most probable ones.

2.1 Image event primitives, signals, and ICA

As an introductory example, let us consider as input events several pictures composed of four pixels. The four sample pictures are shown in Fig.2. The black pixels are coded as –1, the white ones as +1 and the gray ones as zero.

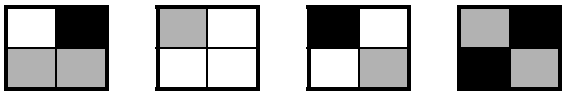

Fig.2 *The four sample pictures*

In the following state-time diagram (Fig.3) four events are presented independently. Here, each event is denoted by two states, *present* (on) or *not present* (off). The time order of the independent events is assumed to be random.

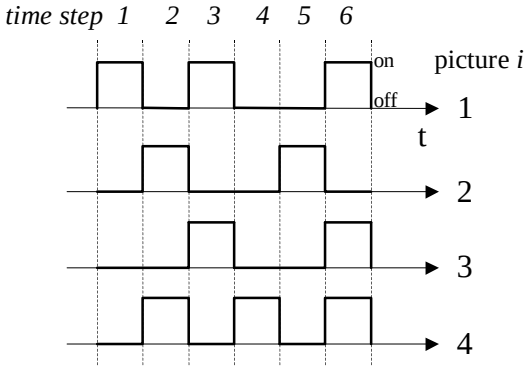

Fig.3 *The state-time diagram of input events*

Each event ω_i manifests itself on all four pixels or four channels. Assigning a signal vector $\mathbf{s}_{\omega i}$ to the event ω_i ="picture *i* appears" we note the events by the vectors

$$\mathbf{s}_{\omega 1} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{s}_{\omega 2} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \mathbf{s}_{\omega 3} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \mathbf{s}_{\omega 4} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

The picture itself can be described by the influence of the event on the pixels. Formally, we can write this as a linear mixture performed according to eq.(4) by the mixing matrix

$$\mathbf{M} = (\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4) = \begin{pmatrix} 1 & 0 & -1 & 0 \\ -1 & 1 & 1 & -1 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The superposition of the influences can be observed at each pixel as the time series of superposed signals. In Fig.4 the intensity of all four pixels is shown for the introductory example.

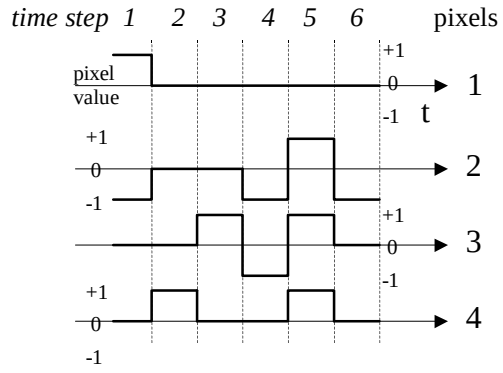

Fig.4 *The time series of the pixel channels*

In Fig.5 the corresponding images are shown.

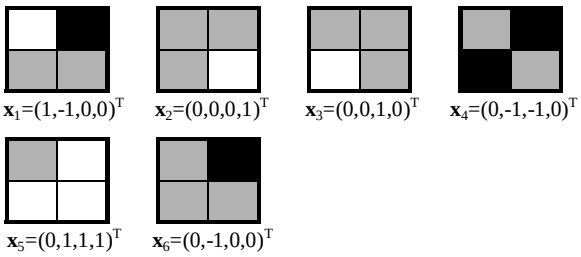

Fig.5 *The six sample pictures*

Since we assume the four events to be independent, we can see our task as not only separating the four channels of the source signal **s** from the linear mixture **x** without any knowledge about the mixture matrix **M**, but also to deduce the occurrence probabilities $P(\omega_i)$ for the independent events ω_i.

2.2 Ordering the Independent Components

To introduce the main idea for computing the probabilities for the principle independent components, we notice that the source signals are defined as

$$s_i = \begin{cases} 1 & \text{for } \omega_i \\ 0 & \text{for } \neg\omega_i \end{cases}$$

Thus, we have as the average source signal

$$\bar{s}_i \equiv \langle s_i \rangle = P(s_i=1)\cdot 1 + P(s_i=0)\cdot 0 = P(\omega_i) \qquad (10)$$

The variance σ_{is}^2 of the source signal s_i is

$$\begin{aligned} \sigma_{is}^2 &= \langle (s_i - \bar{s}_i)^2 \rangle = \langle s_i^2 - 2 s_i \bar{s}_i + \bar{s}_i^2 \rangle \\ &= \langle s_i^2 \rangle - \bar{s}_i^2 \\ &= P(s_i=1)\cdot 1 + P(s_i=0)\cdot 0 - \bar{s}_i^2 \\ &= \bar{s}_i - \bar{s}_i^2 = \bar{s}_i (1 - \bar{s}_i) \end{aligned} \qquad (11)$$

Suppose that we have already computed the demixing matrix **B** satisfying eq.(8). The recovered source signals y_i are derived from the centered source signals s_i by scaling and permutation with a matrix $\mathbf{A} \equiv \mathbf{BM} = \mathbf{DP}$. As stated in section Independent components and ICA it is impossible to determine the permutation matrix **P** so we assume $\mathbf{P} \equiv \mathbf{I}$ and $\mathbf{A} \equiv \mathbf{D}$. For one component y_i we get

$$y_i = a_{ii} (s_i - \bar{s}_i) \qquad (12)$$

where a_{ii} denotes the corresponding diagonal, non-zero, coefficient of **A**. Since y_i is centered and has unit variance σ_{iy}^2 the following relation holds:

$$\begin{aligned} 1 = \sigma_{iy}^2 &= \langle (y_i)^2 \rangle = \langle (a_{ii} (s_i - \bar{s}_i))^2 \rangle \\ &= a_{ii}^2 \sigma_{is}^2 = a_{ii}^2 \bar{s}_i (1 - \bar{s}_i) \end{aligned} \qquad (13)$$

The average $\langle \mathbf{s} \rangle$ of the source signals is transformed by the mixing matrix to the observed average signal

$$\langle \mathbf{x} \rangle = \mathbf{M} \langle \mathbf{s} \rangle \qquad (14)$$

and by the demixing matrix **B** to the average transform output

$$\langle \mathbf{y} \rangle = \mathbf{B} \langle \mathbf{x} \rangle = \mathbf{BM} \langle \mathbf{s} \rangle = \mathbf{A} \langle \mathbf{s} \rangle \qquad (15)$$

Note that here $\langle \mathbf{y} \rangle$ is obviously non-zero since we omitted the centering stage. Therefore we have

$$\langle y_i \rangle = a_{ii} \bar{s}_i \qquad (16)$$

Combining eqs.(13) and (16) gives us the relation between the observed, non-centered output and the needed occurrence probabilities

$$1 = (\langle y_i \rangle / \bar{s}_i)^2 \; \bar{s}_i (1 - \bar{s}_i)$$

or

$$P(\omega_i) = \bar{s}_i = \langle y_i \rangle^2 / (1 + \langle y_i \rangle^2) \qquad (17)$$

By this we have a measure to order the obtained ICA components according to their associated occurrence probabilities $P(\omega_1) \geq P(\omega_i) \geq P(\omega_m)$. Since the most probable events should not be neglected at all they are the most important ones.

There is also an correspondence to the average information of each component. With the definition of the average Shannon information

$$H(y) = -\sum_{\alpha \in \Omega} P(\alpha) \log(P(\alpha)) \qquad (18)$$

and setting the state space to $\Omega \equiv \{\omega_i, \neg\omega_i\}$ we obtain the marginal information for one recovered source y_i

$$\begin{aligned} H(y_i) = & -P(\omega_i) \log(P(\omega_i)) \\ & -(1-P(\omega_i)) \log(1-P(\omega_i)) \end{aligned} \qquad (19)$$

By assigning an order to the components according to their information we define with $H_1 \geq H_2 \geq \ldots \geq H_m$ another order.

How is this order related to the previous criterion of maximal occurrence probability? In Fig.6 the information of one component is shown as a function of its probability.

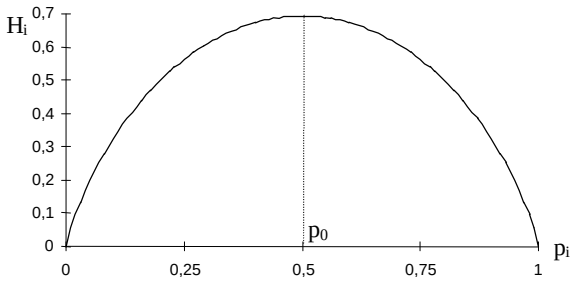

Fig.6 *The information of one component as function of its occurrence probability*

Since information is a convex function, probability and information are both monotonically increasing up to the local maximum which is located at

$$\partial H_i(p)/\partial p = -\log(p_0) + \log(1-p_0) = 0$$

or

$$p_0 = 0.5$$

Thus if we order the components in this range according to

$$i \leq j \Leftrightarrow |P(\omega_i) - p_0| \leq |P(\omega_j) - p_0| \Leftrightarrow H_i \geq H_j \quad (20)$$

we get the desired decreasing entropy order stated above.

3 Simulations and results

In this section we want to visualize our theoretical results of the previous section and show the validity of our image model.

3.1 Recovering the occurrence probabilities of events

For the start we want to show that it is possible to obtain the occurrence probabilities of independent events. For this purpose we use very basic image events. We chose 16 letters 'A'…'P', represented by a very coarse matrix of 8x8 pixels, see Fig.7.

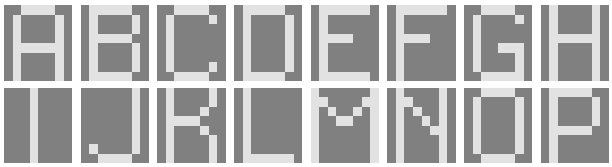

Fig.7 *The image encoding of the events*

For each one of 4096 training patterns, a random linear combination of the letters was computed and presented to a network of 16 neurons. In Fig.8 fifteen input sample pictures out of the 4096 are shown.

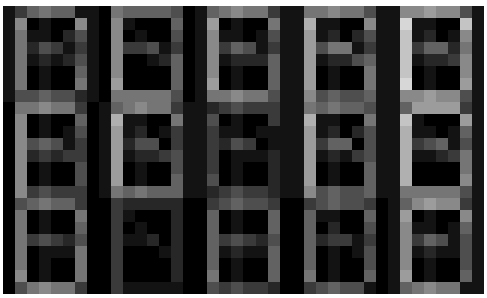

Fig.8 *Sample input pictures of mixed events*

The input events are transformed to decorrelated components by a PCA stage. Initially, we used the full alphabet, but after the PCA stage some components with zero eigenvalues were observed. This means that some letters of the alphabet can be decomposed by a linear combination of others. To obtain really independent sources we chose the subset of 16 letters shown in Fig.7.

The eigenimages formed in the PCA stage, i.e. the rows of matrix $\mathbf{W}_{PCA}$, correspond to the decorrelated components found by the PCA stage and are shown in Fig.9.

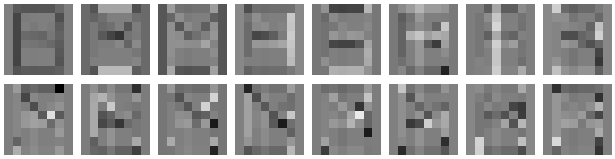

Fig.9 The eigenimages of the input pictures

Here, we observed a near-Gaussian probability distribution of the signal values, see Fig.10.

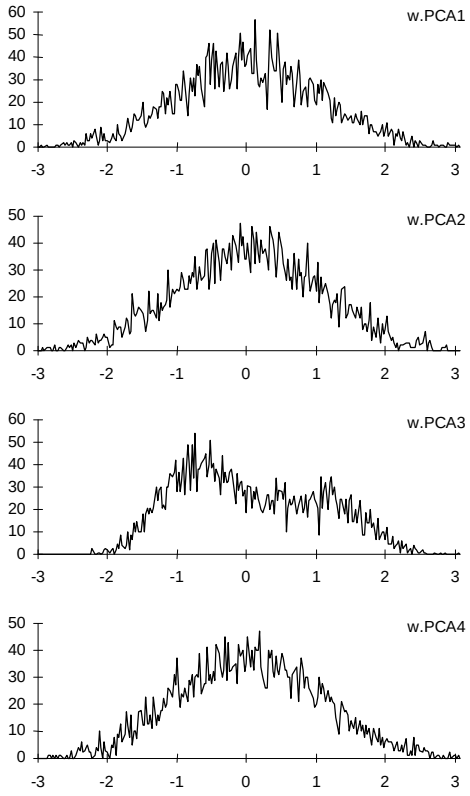

Fig.10 *The probability distribution of four image signals obtained after the PCA and whitening stages*

To obtain the histograms the 4096 samples were quantified into 256 intervals on the horizontal axis.

After the ICA stage, we recovered the source signals. Since we want to concentrate on the topic of principle components we do not describe the algorithms used to obtain the PCA and ICA in detail. Nevertheless, it should be mentioned that the statistical nature of the source signals presented a severe problem for some algorithms. For the concrete events of this section, we have exclusively bimodal source distributions with negative kurtosis, see Fig.13. Our simulations showed that some of the algorithms had problems with bimodal images, i.e. negative kurtosis [BES96], and some with the natural images of positive kurtosis [ACY96]; they did not converge for these mixtures. In order to obtain the desired results, we used versions of the algorithms described in [HYO96].

The inverse of the resulting matrix **B** is the mixing matrix **M**, containing the letters. The images corresponding to the **B** matrix are shown in Fig.11, the inverted **B** matrix gives us the reconstructed source images in Fig.12.

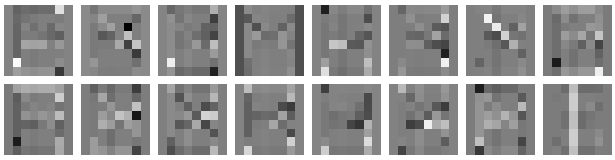

Fig.11 *The inverse source images obtained after the ICA stage*

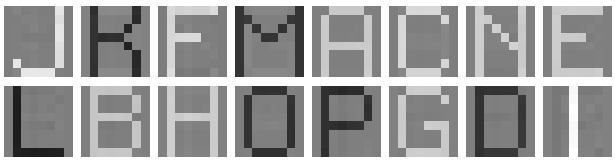

Fig.12 *The source images obtained after the ICA stage*

We see that neither the initial order nor the sign of the sources were preserved. The occurrence probability distribution of four components is shown in Fig.13.

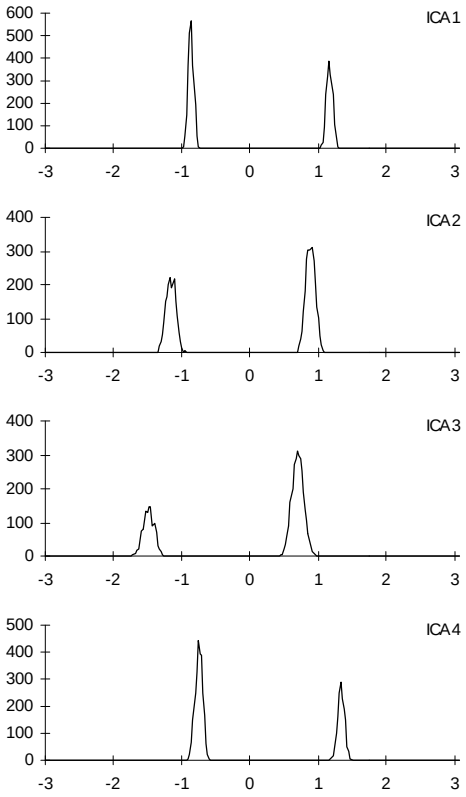

Fig.13 *The probability distribution of four image signals obtained after the ICA stage*

Ideally, the peaks seen in Fig.13 are just spikes with zero variance. Thus, the function values in a small interval around each local average can be summed up in the center of the associated interval, and set to zero afterwards. This kind of quantization should give us a better estimate of the original probability distribution.

The initial and estimated occurrence probabilities of the source letters are listed in table 1. The error is due to the imperfectly learned ICA stage.

source letter	probability used	probability observed	error
D	0,715	0,732	-0,017
F	0,696	0,732	-0,036
I	0,743	0,695	0,049
B	0,692	0,673	0,019
G	0,577	0,628	-0,051
M	0,624	0,618	0,006
L	0,520	0,534	-0,014
O	0,538	0,532	0,006
C	0,423	0,484	-0,061
A	0,492	0,466	0,027
J	0,444	0,463	-0,019
H	0,275	0,396	-0,121
E	0,454	0,362	0,092
N	0,408	0,342	0,066
K	0,415	0,322	0,092
P	0,341	0,310	0,031

Table 1 *The source letters, their associated and their recovered occurrence probabilities*

Now, our initial goal is still the efficient encoding of the image signals. This is obtained by reducing the marginal entropy of the channels. Table 2 shows the approximated average information, the entropy, of the first four channels before and after the ICA stage (calculated from the probability distribution in Fig.10 and Fig.13).

component	observed entropy	component	observed entropy	original entropy
w.PCA1	7.398	ICA1 ('J')	3.800	0.991
w.PCA2	7.408	ICA2 ('K')	4.555	0.980
w.PCA3	7.322	ICA3 ('F')	4.745	0.886
w.PCA4	7.405	ICA4 ('M')	4.164	0.955

Table 2 *The marginal entropy of four channels (in bits)*

Obviously, minimizing the mutual information dramatically reduces the single channel information. Since the probability distributions of the ICA components are slightly "blurred" their marginal entropy is still higher than the original entropy according to eq.(19). However, by applying a rigorous quantization strategy we should be able to achieve further reduction as stated above.

In linear image coding and restoration, we know that by definition the principal decorrelated components obtained after the PCA stage yield the minimal mean squared error (MSE). Thus, we cannot expect that the principal independent components will give us a smaller MSE. Nevertheless, what we can attend is that they can be encoded with a smaller number of bits. Now, for further considerations, let us change to natural images.

3.2 Reconstructing natural images

Image encoding by very few number of coefficients is still a demanding task and has a lot of applications. Perhaps, by using the ICA approach, we might obtain an encoding with a fewer number of components. For this purpose, let us regard the independent components of natural images.

The method to obtain these components is similar to the one in conventional transform coding: the whole image is split into subimages containing n pixels, and each subimage is used as one training sample.

In our simulations the picture called *Cactus* (Fig.14) was divided into 4543 subimages (size: 8×8=64 pixels) which were randomly chosen as training samples.

Fig.14 *The training picture Cactus*

First, we centered and decorrelated the 64 components of the subimage ensemble. The obtained PCA eigenimages are shown in Fig.15 (page 9).

After this, the components are transformed linearly. The transform coefficients are updated by an iterative ICA learning algorithm giving us the matrix $\mathbf{W}_{ICA}$ used in eq.(9). The columns of matrix **B** are shown as images in Fig.16.

The inverse of **B** is the mixing matrix **M.** The columns of this matrix are the source images, shown in Fig.17. The source images obtained are very similar to those already known in the literature, see e.g. [BES96], [OLS96b].

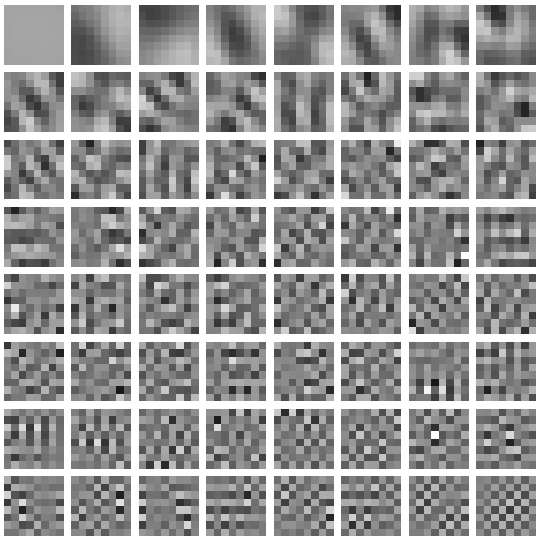

Fig.15 *The PCA eigenimages of Cactus*

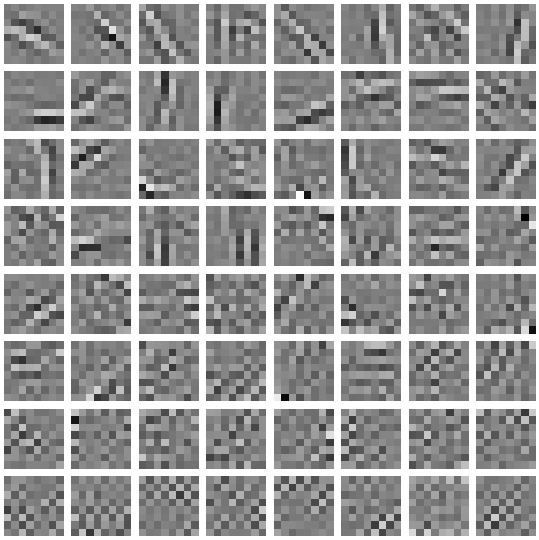

Fig.16 *The base ICA images of Cactus*

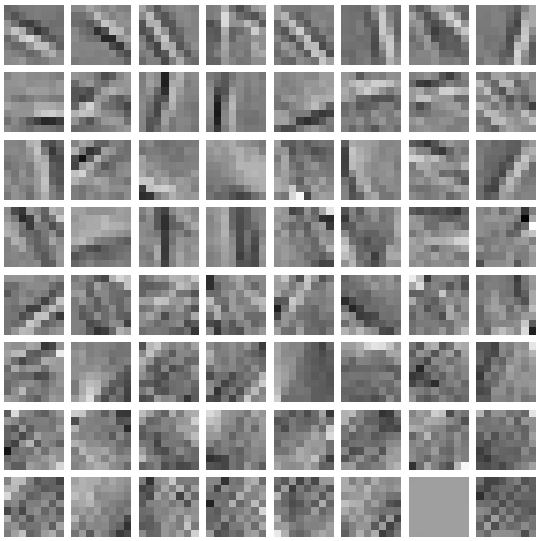

Fig.17 *The source images of Cactus*

Now, what are the most important events? Here, the measured probability distributions of the sources were not bimodal. This excluded our event model of section Recovering the occurrence probabilities of events for calculating the occurrence probabilities and therefore prevents an order of importance of the sources for analyzing the observed situations by important events. Nevertheless, we still can use the marginal information to compute the order of the components instead.

Interestingly, the initial order given by the ICA algorithm is characterized by increasing entropy. This is due to the goal of our (sequential) ICA algorithm which tries to minimize the marginal entropy for the first component by choosing the ICA component which differs the most from a Gaussian distribution, i.e. which has the smallest available entropy.

To answer the basic question if there are principal independent components which contain considerably more or less average information than others we calculated the marginal entropy of all components the same way as in section Recovering the occurrence probabilities of events. The cumulated marginal entropy of the first *k* whitened PCA components (in order of decreasing eigenvalues) and ICA components (in order of increasing entropy) is shown in Fig.18.

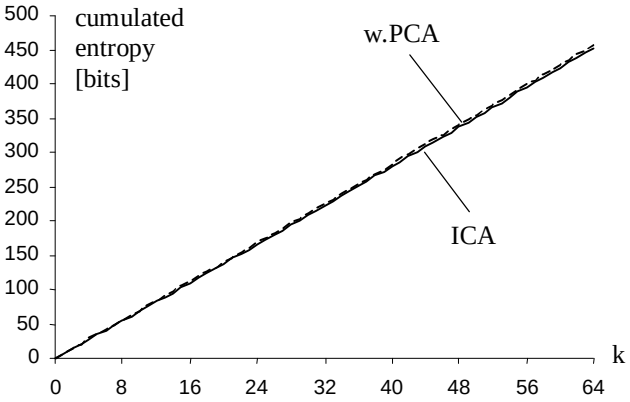

Fig.18 *The cumulated marginal entropy of the first k whitened PCA components (dotted line) and ICA components*

The difference between the two cumulation functions can hardly be seen: the marginal entropy of the ICA components is just slightly smaller than the one of the whitened PCA components. Furthermore, the cumulated entropy of both the PCA and the ICA grows approximately proportional. This means that especially all the ICA components of the image have nearly the same information; there are no components which differ much from the others.

If not in occurrence probability or average information, are there ICA components which differ in „importance“? Are there some which are more important than the others so we have to concentrate on them?

3.3 Component ordering by information

One criterion for „importance“ is the quality of the image reconstructed by the remaining components. In Fig.19 a cutout of the original image *Cactus* is shown.

Fig.19 *The cutout of the image Cactus*

The cutout, reconstructed by the 16 ICA components with the smallest average information, and by the 16 ICA components with the biggest average information, can be seen in Fig.20 and Fig.21.

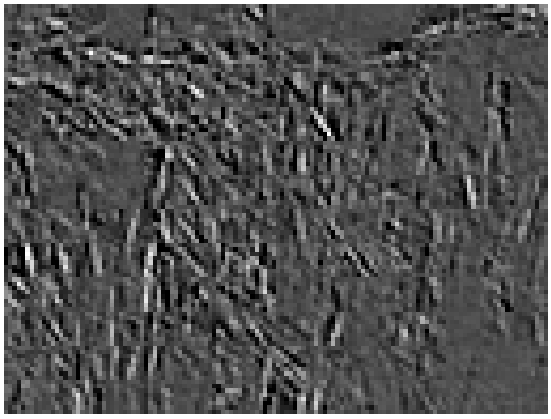

Fig.20 *The reconstruction by the first 16 ICA components*

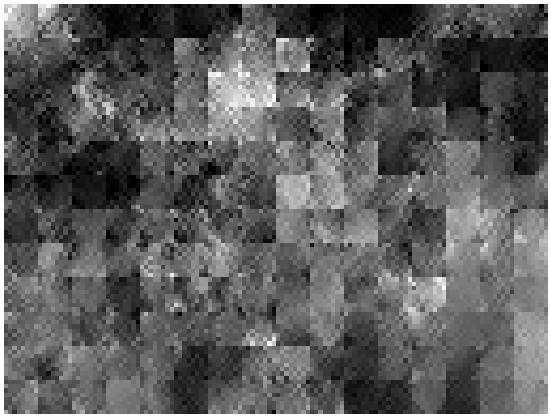

Fig.21 *The reconstruction by the last 16 ICA components*

In both cases, the reconstruction quality is not acceptable, especially, when compared with the reconstruction result of the first 16 PCA components, shown in Fig.22.

Fig.22 *The reconstruction by the first 16 PCA components*

It seems that the pure information criterion is not appropriate for image reconstruction. In contrast to this, the PCA transform seems to give better results.

Reconstructing the image by its first k components and comparing it with the original one gives us the average error for neglecting the $n–k$ components. Certainly, by using the k eigenimages of the PCA stage with the biggest eigenvalues, the mean squared error MSE is minimized because the PCA operation is defined to obtain the smallest possible MSE.

Are there principal ICA components which also minimize the error? Let us compare the MSE contribution by the PCA components by those by the ICA. In Fig.23, this is shown for the image *Cactus*. Obviously, using the components with the biggest entropy does decrease the MSE significantly faster than using the ones with the smallest entropy. Certainly, the smallest MSE is produced using the PCA components (dotted line).

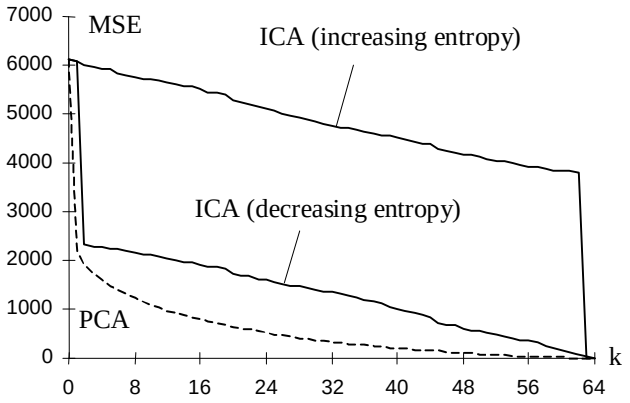

Fig.23 *Decreasing the MSE by adding components*

3.4 Component ordering by virtual variance

How can we further improve the performance of the selected ICA components? The PCA sorting criterion is the decreasing value of the eigenvalues. Since we know that the eigenvalue λ_i is

$$\lambda_i = \sigma_i^2 = \mathrm{var}(y_i)$$

equal to the variance of the component, we might order the ICA components also appropriate to their variance. Here, we encounter a problem: the ICA transform is such that all variances of the components are made equal. How to select the ones with the biggest variance?

Inspecting the transform closer we notice that the output variances are equal, but not the length of the corresponding basis vectors $\mathbf{w}_i$ of the ICA transform (rows of matrix **W**). To compare it to the PCA transform which has unit length basis vectors, we have to normalize the ICA basis vectors. Thus, we might define a *virtual variance* of a component by

$$\mathrm{var}^*(y_i) \equiv \mathrm{var}\left(\frac{\mathbf{w}_i}{\|\mathbf{w}_i\|}\mathbf{x}\right) = \frac{\mathrm{var}(y_i)}{\|\mathbf{w}_i\|^2} = \frac{1}{\|\mathbf{w}_i\|^2} \quad (21)$$

Ordering the ICA components by this criterion, we obtain a better MSE-adapted reconstruction while preserving the performance of the cumulated entropy. In Fig.24 the best ICA ordering of Fig.23 is compared to the *virtual variance* ordering.

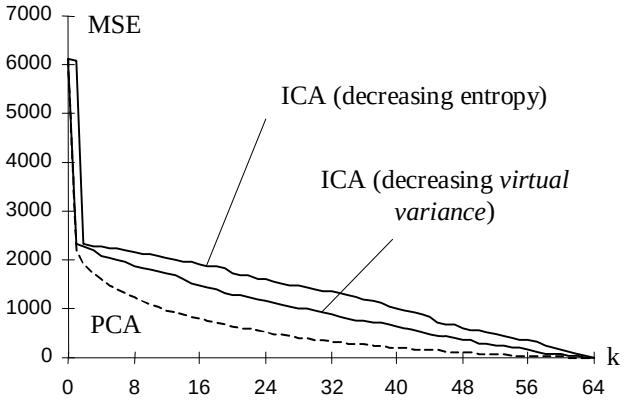

Fig.24 *The MSE of the ICA ordering by virtual variance*

To obtain an impression of the reconstruction quality, we present the reconstructed image cutout of *Cactus* by using the ICA components with the biggest virtual variance in Fig.25. Clearly, this ordering performs better than the two previous ones, but it is still inferior to the classical PCA approach.

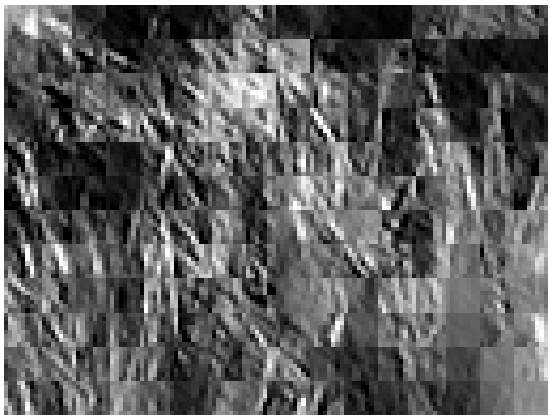

Fig.25 *The reconstructed image cutout*

Now, without using any other image reconstruction quality measure (like, for instance, the psychophysiological approach, see e.g. [CHR90]), we ask: what can the ICA approach do for encoding and reconstructing images when the minimal MSE of the reconstruction is given by the number of PCA components?

3.5 Principal independent components and rate distortion theory

When we reduce the number of components in the transform approach for encoding images we reduce the full space of image components (dimensions) to a subspace. The subspace of the ICA components is characterized by its information content whereas the subspace of the PCA components is characterized by its low MSE reconstruction error. Now, if we cannot replace the principal components of PCA for obtaining a small MSE, what about reducing their encoding information by ICA? This idea can be performed in two ways:

- Get the first *k* PCA components with an acceptable MSE. Then, by an ICA transform, we will get the same number of encoding coefficients but with less information, i.e. less encoding bits.
- For the same amount of encoding information as the *k* PCA components take, we can also get *p* more ICA transformed PCA components. Since these *p+k* base vectors of the ICA transform span the same space as the *p+k* PCA components, the resulting image quality will be enhanced like adding *p* more PCA components.

Thus our approach, starting with the search for independent image primitives, leads us to the error-bounded maximal information for each channel. This is not new: the approach of maximizing the information for a time step in a channel when an upper bound for the error (more general: for a distortion measure) exists or, vice versa, to minimize the error for a channel with a constant information per time step is classically known as the *rate distortion theory* [SHA49] and has a broad range of applications in the classical telecommunication area.

The first one of the ideas above can be expanded if we order the k ICA components according to their decreasing virtual variance and encode only the first $k' < k$ components with low additional reconstruction error. This results in a further reduction of the number of encoding bits.

To validate the latter idea we computed the ICA components of the first k PCA components (Fig.15) for $k = 16,\ldots,21$. In Fig.26a,b the ICA base vectors and images can be seen for $k = 17$. Note that they are different to those obtained in Fig.16 because the data space is also different.

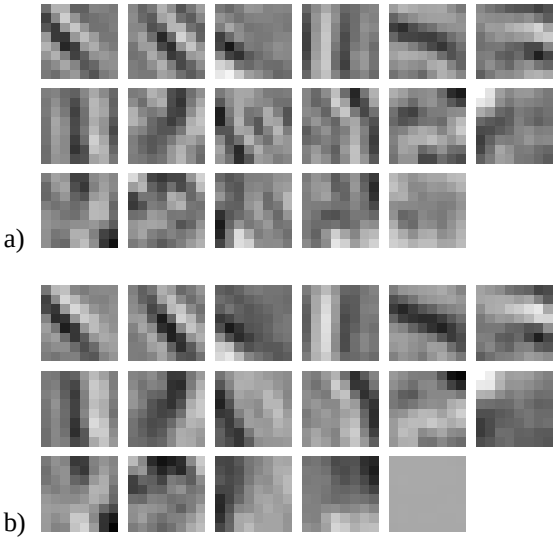

Fig.26 *The 17 ICA base vectors and 17 images*

Then the cumulated entropy was calculated and compared to the cumulated entropy of the first k whitened PCA components. We found that for the same information rate at most one additional ICA component can be encoded with an error reduction of 5%. An example for 17 ICA components is shown in Fig.27: the reconstructed image is slightly better than the one of Fig.22.

Fig.27 *The reconstrucion by 17 ICA components*

Until now we estimated the overall encoding amount by calculating the marginal entropy of the components without considering efficient quantization techniques. In the next section we shall take a closer look at this task.

3.6 Robust encoding of natural images with principal independent components

Suppose we have an image decomposed into subimages which we want to encode as efficient as possible (see section Independent components). Since we are dealing with digitized images the n components (pixels) x_i of an arbitrary subimage $\mathbf{x} = (x_1,\ldots,x_n)^T$ are discrete, i.e. each x_i stores one of N different values. Thus there is a number N^n of different image patches or "image states" that can be assigned to $\mathbf{x}$.

Obviously, a lot of these image patches are unlikely to occur in natural image data (e.g. very noisy structures) while others are quite similar (differing in only a few pixels): we assume that we have to encode only a small number $L_\varepsilon \ll N^n$ of "necessary" states of $\mathbf{x}$ which are sufficient to describe natural images at an acceptable error ε. L_ε is called the *error-bounded descriptional complexity* of the subimages [BRA93b].

The main idea of transform coding is to derive an optimized error-bounded representation $\mathbf{y} = (y_1,\ldots,y_n)^T$ of $\mathbf{x}$ according to the image statistics, i.e. $\mathbf{y}$ has to encode the L_ε necessary states of $\mathbf{x}$ as efficient as possible. Consequently, we demand the relation

$$L_\varepsilon \le \prod_i Q_i < N^n \qquad (22)$$

where Q_i denotes the number of different values that can be assigned to a component y_i [1]. The determination of the Q_i at a given error ε is a non-trivial task which will not be addressed in this paper. Instead, from an opposite point of view, we ask for the reconstruction error ε at given numbers Q_i, i.e. at a given *quantization* of the y_i.

In the previous section we used the (virtual) variance of a component y_i to decide whether its quantization number was set to Q_i=256 or to Q_i=1. But variance can tell us even more about "importance": in case of the PCA or the DCT (Discrete Cosine Transform) it is well-known that decreasing the quantization number Q_i (i.e. the resolution) of a component y_i with low variance reduces the overall encoding amount without affecting the reconstruction error perceived by the human visual system. This is why PCA or DCT components with lower variance are encoded at coarser resolution, and the same should hold for ICA.

To prove the idea we used the $k = 16,\ldots,21$ ICA and PCA components of the previous section. The ICA com-

[1] Note that the marginal entropy of a component y_i will not increase if Q_i is decreased; furthermore, y_i will be set to a constant value (e.g. *zero*) if Q_i=1.

ponents were scaled with the reciprocal norm of the associated base vectors to set their former unit variance to the virtual variance in order to be comparable to the PCA components.

Since the coefficients of both the PCA and the scaled ICA lay within an interval $\Im = [\Im_{min},\Im_{max}] \subset \Re$ we uniformly divided $\Im$ into Q subintervals $\Im_q$ of same length; the quantization was done by assigning each (PCA or ICA) coefficient $c \in \Im_q$ the arithmetical mean of $\Im_q$. After this procedure me made the following observations:

- The boundaries $\Im_{min}$ and $\Im_{max}$ of $\Im$ were given by the smallest and the biggest coefficient of the PCA component with highest variance.
- The components y_i with low variance were encoded with lower relative resolution than the components with high variance because the length of the quantization intervals were not adapted to the range of the y_i.

We computed both the MSE and the cumulated entropy for different k and Q. Fig.28 shows the resulting MSE as a function of the entropy.

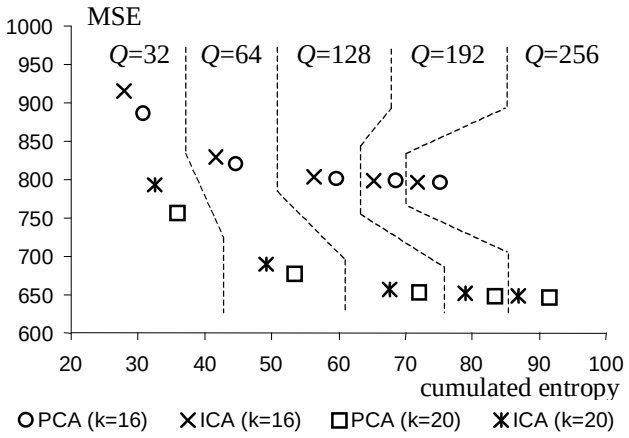

Fig.28 *The MSE as a function of the cumulated entropy at different quantization levels Q*

For both the PCA and the ICA the functional dependency of reconstruction error and cumulated entropy is approximately the same if k is equal. As in section Principal independent components and rate distortion theory, for the same amount of cumulated entropy it is possible to encode about one ICA component more than PCA components since the marginal entropy of the ICA components is lower.

Note that for k=16 components and quantization level Q=256 the MSE *and* the entropy are lower if we use *more* components (k=20) at lower resolution (Q=64). According to this observation we may state that "variety" is more important than "accuracy", i.e. to reduce the reconstruction error we should encode more components instead of increasing the quantization resolution. The systematic investigation of this behavior is subject to future research.

4 Discussion

In this paper we showed that the concept of independent components known in Principal Component Analysis (PCA) can be enlarged to cover also the occurrence probabilities and the information content of an Independent Component Analysis (ICA). Whereas the ICA approach assumes continuous source channels mixed up to the same number of channels by a mixing matrix, we applied the ICA to images assuming that they are composed by only a few image primitives.

Certainly, the components with the highest probability are also the ones which should not be neglected. As shown in section 3.2, this corresponds roughly to the mean squared error induced by neglecting the components, but is not identical to it. Theses components can be termed the *„Principal Independent Components PIC“*. For distinctive images, e.g. characters this idea gives us the most important ones.

Nevertheless, for natural images we have no a-priori probabilities. Using the ICA components with most of the information did not lead to an acceptable reproduction error. The situation changed when we applied the ICA transform to the first principal PCA components which resulted in a compact and robust encoding. This approach combines the traditional principal component criteria of PCA with the independence property of ICA. It turned out that this definition of PIC implements the classical demand of the rate distortion theory of Shannon.

5 References

[ACY96] S.Amari, A.Cichocki, H.Yang: *A New Learning Algorithm for Blind Signal Separation*; Advances in Neural Information Processing Systems 8, Touretzky, Mozer, Hasselmo (Eds.), pp.757–763, MIT Press (1996) *and available by* http://www.bip.riken.go.jp/irl/hhy/hhy/acyNIPS95.ps.Z

[BES96] A.J.Bell, T.J.Sejnowski: *Edges are the `independent components' of natural scenes*; Int. Conf. Advances in Neural Information Processing Systems NIPS 96, MIT press (1996).

[BRA93a] R.Brause: *A Symmetrical Lateral Inhibited Network for PCA and Feature Decorrelation*; Proc. Int. Conf. Art. Neural Networks ICANN-93, pp.486–489, Springer Verlag (1993)

[BRA93b] R.Brause: *The Error-Bounded Descriptional Complexity of Approximation Networks*; Neural Networks, Vol.6, pp.177–187 (1993)

[BRA94] R.Brause, *A VLSI-Design of the Minimum Entropy Neuron*; J. Delgado-Frias, W. Moore (Eds.): VLSI for Artificial Intelligence and Neural Networks, pp.53–60, Plenum Press (1994)

[BRA96] R.Brause: *Sensor Encoding Using Lateral Inhibited, Self-organized Cellular Neural Networks*; Neural Networks, Vol.9, No.1, pp.99–120, (1996)

[BRA98] R.Brause, M.Rippl: *Noise Suppressing Sensor Encoding and Neural Signal Orthonormalization*; accepted by IEEE Trans. on Neural Networks

[BUR92] G.Burel: *Blind Separation of Sources: A Nonlinear Neural Algorithm*; Neural Networks, Vol. 5, pp.937–947 (1992)

[COM94] P.Comon: *Independent Component Analysis – a new concept?*; Signal Processing, Vol.36, pp.287–314 (1994)

[CHR90] B.Chitprasert, K.Rao*: Human Visual Weighted Progressive Image Transmission*; IEEE Trans.Comm., Vol.38, No.7, pp.1040-1044 (1990)

[DEO96] G.Deco, D.Obradovic: *An Information-Theoretic Approach to Neural Computing*; Springer Verlag (1996)

[HYO96] A.Hyvärinen, E.Oja: *Independent Component Analysis by General Non-linear Hebbian-like Rules*; Helsinki University of Technology, Dep. of Comp. Sc., Report A41 (1996) *also available by* http://nucleus.hut.fi/~aapo/ps/TR_A41_genhebb.ps

[OJA92] E.Oja: *Principal components, minor components, and linear neural networks*; Neural Networks, Vol.5, pp.927–935 (1992)

[OLS96a] B.A.Olshausen, D.J.Field: *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*; Nature 381, pp.607–609 (1996)

[OLS96b] B.A.Olshausen, D.J.Field: *Natural Image Statistics and Efficient Coding*; Network: Computation in Neural Systems, No. 7, pp.333–339 (1996)

[PLUM93] M.Plumbley: *Efficient Information Transfer and Anti-Hebbian Neural Networks*; Neural Networks, Vol.6, pp.823–833 (1993)

[SIL91] F.Silva, L.Almeida: *A distributed solution for data orthonormalization*; T.Kohonen et. al. (Eds.): Artificial Neural Networks, Elsevier Sc. Publ. (1991)

[SHA49] C.E.Shannon, W.Weaver: *The Mathematical Theory of Information*; University of Illinois Press, Urbana (1949)