Image Encoding by Independent Principal Components

Björn Arlt, Rüdiger W. Brause

FB Informatik, J.W.Goethe-Universität Frankfurt/Main, Germany
E-mail: {arlt,brause}@informatik.uni-frankfurt.de

Abstract

The encoding of images by semantic entities is still an unresolved task. This paper proposes the encoding of images by only a few important components or image primitives. Classically, this can be done by the Principal Component Analysis (PCA). Recently, the Independent Component Analysis (ICA) has found strong interest in the signal processing and neural network community. Using this as pattern primitives we aim for source patterns with the highest occurrence probability or highest information.

For the example of a synthetic image composed by characters this idea selects the salient ones. For natural images it does not lead to an acceptable reproduction error since no a-priori probabilities can be computed. Combining the traditional principal component criteria of PCA with the independence property of ICA we obtain a better encoding. It turns out that the Independent Principal Components (IPC) in contrast to the Principal Independent Components (PIC) implement the classical demand of Shannon's rate distortion theory

1 Introduction

One of the still unresolved tasks of image encoding and interpretation is the description of an image by a set of universal pattern primitives or semantic entities. By such an approach both tasks can be solved: the task of compression as well as the task of semantic content encoding.

Certainly, for compression only the most important patterns are needed. Classically, the encoding of images by only a few important components is done by the Principal Component Analysis (PCA). Here, we search for the principal directions in an input space. Since the number of pixels is treated as the number of dimensions of the input space, huge pictures can hardly be processed in reasonable time by this technique. One common solution for this problem is to cut the image into smaller patches or "subimages" which are transformed linearly by projecting them on the eigenvectors of their associated covariance matrix. It is well known that the transformed components with the highest variance (the *principal components*) yield an optimal reconstruction of the original subimages in the mean square error sense. However, for the criterion of minimal redundancy encoding, the PCA is suboptimal.

It has been pointed out by Barlow [3] that nature for encoding sensory signals in an efficient way should decrease the redundancy in the encoding. This can be done e.g. by factorial coding [4], that is, by making all components for representing a sensory event independent from another.

Recently, the Independent Component Analysis (ICA) has become subject to many research activities and several algorithms have been proposed by different authors, e.g.[1][6][7]. Here, the goal is to obtain linearly transformed components which are as independent as possible (the *independent components*). This corresponds to the minimization of the mutual information between the transformed components and therefore reduces the overall encoding amount [1][6].

Applied to image encoding, the ICA approach assumes that each observed signal vector $\mathbf{x} = (x_1,\ldots,x_n)^T$ (an image containing *n* pixels) is a linear mixture $\mathbf{x} = \mathbf{Ms}$ of *n* unknown independent source signals $\mathbf{s} = (s_1,\ldots,s_n)^T$. The unknown mixing matrix $\mathbf{M}$ must be non-singular; its columns can be viewed as "image primitives". To recover the sources signals, one has to determine a demixing matrix $\mathbf{B}$ with $\mathbf{s} = \mathbf{Bx}$.

There are several conditions involved in the demixing process [6] in general, the recovered source signals (denoted by $\mathbf{y} = (y_1,\ldots,y_n)^T$ for clarity) are scaled and permuted versions of the original sources. Furthermore, at most one of the source signals $\mathbf{s}$ should have a Gaussian probability distribution or else the separation will become ambiguous. This is why the recovered sources $\mathbf{y}$ are conventionally assumed to be non-Gaussian random variables having unit variance.

As proposed in [6][7] the determination of $\mathbf{B}$ reduces to the computation of an orthogonal matrix $\mathbf{W}_{ICA}$ if the observed signals $\mathbf{x}$ are prewhitened. This can be done by a simple PCA transform of the image vectors and scaling the obtained PCA components to unit variance. The corresponding prewhitening (or *sphering*) transform is denoted by the matrix $\mathbf{W}_{PCA}$.

Together with the convenient assumption that the recovered source signals are centered, i.e. $\langle \mathbf{y} \rangle \equiv \mathbf{0}$, we have the following ICA relation

$$\mathbf{y} = \mathbf{W}_{ICA}\mathbf{W}_{PCA}\,(\mathbf{x} - \langle \mathbf{x} \rangle) = \mathbf{B}\,(\mathbf{x} - \langle \mathbf{x} \rangle) = \mathbf{BM}\,(\mathbf{s} - \langle \mathbf{s} \rangle) = \mathbf{DP}\,(\mathbf{s} - \langle \mathbf{s} \rangle) \qquad (1)$$

where $\mathbf{D}$ is an unknown diagonal matrix and $\mathbf{P}$ an also unknown permutation matrix.

In this model the number of independent sources is assumed to be equal to the number of image pixels. Nevertheless, we expect that for a good representation covering most of the input data some of the sources are less important than others. Thus we aim for an ordering criterion which let us select the essential source signals as pattern primitives.

2 An event-oriented image model

Due to the intuitive notion of "importance" we propose that principal independent components should have a high occurrence probability. Therefore, we consider images to be composed of the superposition of many small, independent image primitives, just like a single neuron of the retina sees the world by a limited focus, which appear with a certain probability. As a further restriction, we assume that only one of two possible states is assigned to each primitive: present in the superposition or not. This leads to the formulation of *image events* ω_i (denoting the presence of primitive *i*) and $\neg\omega_i$ (denoting its absence). The task consists now of determining the most important events, i.e. those with highest probability $P(\omega_i)$.

Applied to eq. (1), the image primitives are represented by the columns of the mixing matrix **M**, and the source signals s_i encode the associated image events by

$$s_i = \begin{cases} 1 & \text{for } \omega_i \quad (\text{primitive } i \text{ is present}) \\ 0 & \text{for } \neg\omega_i \quad (\text{primitive } i \text{ is not present}) \end{cases}$$

Thus, the average $\langle s_i \rangle \equiv \bar{s}_i$ of a source signal s_i and its variance σ_{is}^2 are given by

$$\bar{s}_i \equiv \langle s_i \rangle = P(s_i{=}1)\cdot 1 + P(s_i{=}0)\cdot 0 = P(s_i{=}1) = P(\omega_i) \tag{2}$$

$$\sigma_{is}^2 = \langle s_i^2 \rangle - \bar{s}_i^{\,2} = P(s_i{=}1)\cdot 1^2 + P(s_i{=}0)\cdot 0^2 - \bar{s}_i^{\,2} = \bar{s}_i - \bar{s}_i^{\,2} = \bar{s}_i\,(1-\bar{s}_i) \tag{3}$$

Suppose that we have already computed the demixing matrix **B** in eq. (1). The recovered source signals y_i are scaled and permuted versions of the centered original sources s_i. Because the permutation **P** is unknown (and, in fact, of no interest) we assume $\mathbf{P} \equiv \mathbf{I}$ and concentrate on the non-zero scaling factors a_i satisfying

$$y_i = a_i\,(s_i - \bar{s}_i) \tag{4}$$

Since the recovered sources have zero mean and unit variance σ_{iy}^2 the following relation holds:

$$1 = \sigma_{iy}^2 = \langle y_i^2 \rangle = \langle (a_i\,(s_i - \bar{s}_i))^2 \rangle = a_i^2\,(\langle s_i^2 \rangle - \bar{s}_i^{\,2}) = a_i^2\,\sigma_{is}^2 = a_i^2\,\bar{s}_i\,(1-\bar{s}_i) \tag{5}$$

Now, if we ignore the centering terms in eq. (1), we can express the transformation of the source average $\langle \mathbf{s} \rangle$ to the observed average $\langle \mathbf{x} \rangle$ and to the recovered source average $\langle \mathbf{y} \rangle$ by

$$\langle \mathbf{x} \rangle = \mathbf{M}\langle \mathbf{s} \rangle \quad \text{and} \quad \langle \mathbf{y} \rangle = \mathbf{B}\langle \mathbf{x} \rangle = \mathbf{BM}\langle \mathbf{s} \rangle \tag{6}$$

Note that here $\langle \mathbf{y} \rangle$ is obviously non-zero unless for all *i* the probabilities $P(\omega_i)$ are zero. With eqs. (4), (6) we have

$$\langle y_i \rangle = a_i\,\bar{s}_i \tag{7}$$

Combining eqs. (5), (7) gives the desired relation for the occurrence probabilities

$$1 = (\langle y_i \rangle / \bar{s}_i)^2\,\bar{s}_i\,(1-\bar{s}_i) \quad \text{or} \quad P(\omega_i) = \bar{s}_i = \langle y_i \rangle^2 / (1+\langle y_i \rangle^2) \tag{8}$$

By this we obtained a measure to order the observed ICA components according to their decreasing occurrence probabilities, i.e. $i \geq j \Leftrightarrow P(\omega_i) \geq P(\omega_j)$.

Furthermore, if $P(\omega_i) \leq 0.5$ holds for all *i*, the components y_i are ordered by their decreasing marginal entropy $H(y_i)$, because $H(y_i)$ is a convex function of the probability $P(\omega_i)$ and monotonically increasing up to its local maximum (located at $P(\omega_j) = 0.5$), see [2].

3 Recovering the occurrence probabilities of events

To validate the theoretical results of the previous section, we computed a synthetic image according to the model in eq. (1). As image primitives we chose 16 pictures of 8×8 pixels visualising the letters ‘A’…‘P’. From these, 4096 different random linear mixtures were calculated and used as training samples. After prewhitening with the transform $\mathbf{W}_{PCA}$ we presented the samples to a hierarchical ICA network similar to the one proposed in [7] with *tanh* non-linearities. The image primitives along with the eigenimages and the recovered primitives are shown in Figure 1a-c.

For the whitened PCA components we observed near-Gaussian distributions (Figure 1d) while the distributions of the ICA components are slightly “blurred” versions of the original occurrence probabilities, see Figure 1e.

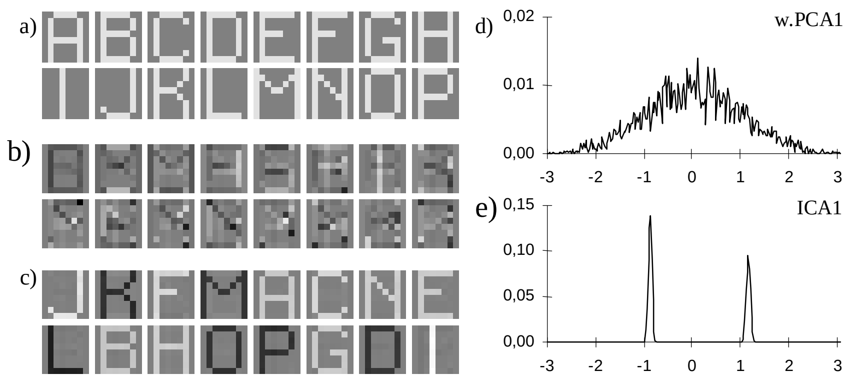

Figure 1: a) The image primitives, b) the eigenimages, and c) the recovered image primitives of the synthetic image. The probability distributions of the first whitened PCA component and of the first ICA component are shown in d) and e) respectively. To obtain the histograms the 4096 samples were quantified into 256 intervals on the horizontal axis.

As we can see, the original patterns could be recovered, although the sign and order of the components are arbitrary. So far, our approach of image decomposition by independent pattern primitives seem to work.

Now, what are the most important components? Due to our event model of section 2 this can be decided by their occurrence probability. The initial and the estimated occurrence probabilities of the first four sources are listed in Table 1 (the error is due to the imperfectly learned demixing matrix **B**). Also shown are their observed and their original marginal entropy (computed on 8 bit coefficients) compared to the marginal entropy of the first four whitened PCA components. Obviously, the single source information is reduced dramatically. Because of the “blurred” probability distributions, the marginal entropy of the recovered sources is still higher than the original entropy. However, by applying a rigorous quantization strategy we should be able to achieve further reduction, see [2].

source	probability		error		compo-nent	observed entropy	compo-nent	observed entropy	original entropy
	initial	estim.							
'J'	0.444	0.463	-0.019		w.PCA1	7.398	ICA1 'J'	3.800	0.991
'K'	0.415	0.322	0.092		w.PCA2	7.408	ICA2 'K'	4.555	0.980
'F'	0.696	0.732	-0.036		w.PCA3	7.322	ICA3 'F'	4.745	0.886
'M'	0.624	0.618	0.006		w.PCA4	7.405	ICA4 'M'	4.164	0.955

Table 1: Four of the source letters, their associated initial and estimated occurrence probabilities. Also shown are the observed and original marginal entropy of the four recovered sources and the first four whitened PCA components (in *bits*).

We can see that the initial probabilities could be approximately recovered. However, due to the approximation error the list of "important" components are not well ordered. Is this different in natural images?
Let us investigate these question more deeply for natural images instead of synthetic ones.

4 Independent components of natural images

The decomposition of natural images in order to find independent parts has also been done by other authors, see for instance Bartlett et al. [3] for parts of face images. In contrast to this the initial goal of our examinations is the efficient encoding of images with only a few important components. So, let us search for the most important components of natural images.

In our simulations a picture called *Cactus* was divided into 4543 subimages (size: 8×8=64 pixels) which were randomly chosen as training samples [2]. After centering and prewhitening of the samples we determined the matrix **B**. The corresponding image primitives were very similar to those already known in the literature, see e.g. [8].

Here, the measured probability distributions of the sources were not bimodal. This excluded the event model of section 2 for calculating the occurrence probabilities and therefore prevented an order of the sources by most probable image events. Instead, we calculated the marginal entropy of the recovered sources as the ordering criterion which is closely related to the probability ordering (see section 3).

To our deception, we found that especially all the ICA components (in contrast to the PCA components) had nearly the same information; there were no components which differed much from the others. Furthermore, the marginal entropy of the ICA components was just slightly smaller than the one of the whitened PCA components.

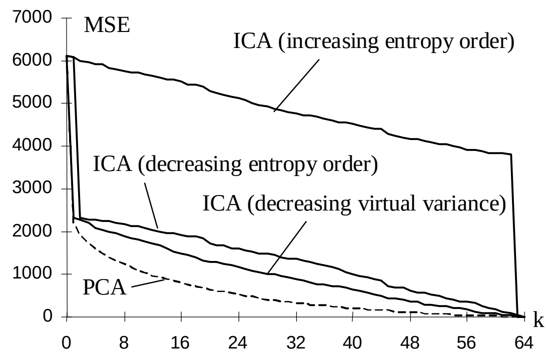

Figure 2: Decreasing the MSE by adding components.

Another measure for "importance" is the quality of the image restoration. Reconstructing the image by its first k components and comparing it with the original one gives the average error for neglecting the n–k components. Therefore we compared the optimal MSE (mean square error) contribution of the PCA components (ordered by decreasing variance) to those of the ICA components (ordered by increasing and decreasing entropy). For the latter we defined a third ordering criterion called the *virtual variance*

$$\mathrm{var}^*(y_i) \equiv \mathrm{var}\left(\frac{\mathbf{b}_i}{\|\mathbf{b}_i\|}\left(\mathbf{x}-\langle\mathbf{x}\rangle\right)\right) = \frac{\mathrm{var}(y_i)}{\|\mathbf{b}_i\|^2} = \frac{1}{\|\mathbf{b}_i\|^2} \tag{9}$$

which considers the fact that the norm of a row $\mathbf{b}_i$ of the matrix $\mathbf{B}$ is in general not equal to unity. Consequently, an ICA component with higher virtual variance is assumed to be more important. Figure 2 shows the obtained error functions. In case of the ICA, ordering the components by their decreasing virtual variance gives the best results. However, our simulations showed that the subjective quality of image restoration by a few ICA components is not acceptable.

5 Independent Components and rate distortion theory

When the number of components in the transform approach for encoding images is reduced, the full space of image components (dimensions) is reduced to a subspace. The subspace of the ICA components is characterised by its information content whereas the subspace of the PCA components is characterised by its low MSE reconstruction error.

In the previous section we found that the reconstruction criterion MSE is important also for the independent components. Here, we have two possibilities:

- We might compute the principal components first which minimise the MSE. This is done for instance by computing the PCA. Then, for the subspace of the main PCA components, we compute the ICA components. This gives us the *Independent Principal Components (IPC).*
- As most important components we might also choose the ICA components with maximal MSE as attribute. The selected components can be termed *Principal Independent Components (PIC).*

Please note that the resulting IPC and PIC components are different, not only by their number of dimensions but also because they are obtained by different statistics (subspace vs. full space).

For the IPC, the encoding yielding the MSE is reduced by the ICA. This process can be performed in two ways:

1. *Minimise the information at constant error*
   Get the first k PCA components with an acceptable MSE. Then, by an ICA transform of the k-dim. subspace, we will get the same number of encoding coefficients but with less information, i.e. less encoding bits.
2. *Minimise the error at constant information*
   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 $k+p$ base vectors of the ICA transform span the same space as the $k+p$ PCA components, the resulting image quality will be enhanced as if p more PCA components were added.

The approach starting with the search for the most important image components led us to the error-bounded maximal information for each channel. This is classically known as the *rate distortion theory* [6] and has a broad range of applications in the telecommunication area.

The IPCA feature processing procedure in this paper was done by two consecutive stages. In principal, this can also be done in one network layer by a proper learning rule. However, this is not easy to implement and is a up to future research.

6 References

[1] Amari S, Cichocki A, Yang HH. A new learning algorithm for blind signal separation. In: Touretzky D, Mozer M, Hasselmo M (ed) Advances in Neural Information Processing Systems 8, MIT Press, Cambridge MA, 1996, pp 757–763

[2] Arlt B, Brause R. The principal independent components of images. Internal Report 1/98, J.W.Goethe-Universität Frankfurt, Germany, 1998, http://www.informatik.uni-frankfurt.de/fbreports/01_.98.ps.gz

[3] Barlow H.B.: Possible principles underlying the transformation of sensory messages, in : Sensory Communication, Rosenbluth W.A.(ed), MIT Press 1961

[4] Barlow H.B.: Unsupervised Learning; Neural Computation 1, pp.295-311, 1989

[5] Bartlett M.S., Lades H.M., Sejnowski: Independent Component Representations for Face Recognition, Proc. of the SPI Symp. On Electronic Imaging (in Press) 1998. ftp://ftp.cnl.salk.edu/pub/marni/EI98.ps

[6] Comon P. Independent component analysis – a new concept?. Signal Processing 1994; 36: 287–314

[7] Hyvärinen A, Oja E. Independent component analysis by general non-linear Hebbian-like learning rules. Technical Report A41, Helsinki University of Technology, Laboratory of Computer and Information Science, 1996

[8] Olshausen BA, Field DJ. Natural image statistics and efficient coding. Network: Computation in Neural Systems 1996; 7: 333-339

[9] Shannon CE, Weaver W. The mathematical theory of information. University of Illinois Press, Urbana, 1949